JOURNAL OF e-LEARNING AND KNOWLEDGE SOCIETY

www.je-lks.org

VOLUME 22 | ISSUE NO. 1 | APRIL 2026



Focus on
New alphabets of Hate

EDITORS
LUCA DE BIASE
STEFANO MORIGGI
STEFANO PASTA
PIER CESARE RIVOLTELLA
LORIS VEZZALI

AN INTERNATIONAL AND OPEN ACCESS JOURNAL
BY THE ITALIAN E-LEARNING ASSOCIATION

ISSN (online) 1971 - 8829 | ISSN (paper) 1826 - 6223

Photo cover by ChatGPT

JOURNAL OF e-LEARNING AND KNOWLEDGE SOCIETY

www.je-lks.org

VOLUME 22 | ISSUE NO. 1 | APRIL 2026

Focus on

New alphabets of Hate

EDITORS

LUCA DE BIASE
STEFANO MORIGGI
STEFANO PASTA
PIER CESARE RIVOLTELLA
LORIS VEZZALI

AN INTERNATIONAL AND OPEN ACCESS JOURNAL BY THE ITALIAN E-LEARNING ASSOCIATION

ISSN (online) 1971 - 8829 | ISSN (paper) 1826 - 6223

JOURNAL OF e-LEARNING AND KNOWLEDGE SOCIETY

www.je-lks.org - www.sie-l.it

ISSN (online) 1971 - 8829 |ISSN (paper) 1826 - 6223

Je-LKS is an Open Access Online publication. This means that everybody can free access online to abstracts and full t articles.

The Journal is issued online three times per year.

ANVUR Ranking: A-Class for Sector 10, 11-D1 and 11-D2

Libraries or researchers, can subscribe a Printed Copies Service or an Issue Reprint Service.

The Printed Copies Service is on a yearly basis and you can receive the three printed year issues.

The Issue Reprint Service provide five printed copies of a specific issue.

Both the Services have the same cost of 150€ (including taxes).

Shipping costs are not included and will be charged.

In order to subscribe Printed Copies or Issue Reprint Services contact:

email: segreteria@sie-l.it
email: staff@je-lks.org
Tel. +39 0522 522521
website: www.sie-l.it

PUBLISHER AND COPYRIGHT

Italian e-Learning Association (SIe-L)

CC BY NC SA

CONTACTS

editor@je-lks.org
managing.editor@je-lks.org
assistant.editors@je-lks.org

Registration at the Rome Court in the pipeline

ISSN (online) 1971 - 8829
ISSN (paper) 1826 - 6223

EDITOR IN CHIEF

TOMMASO MINERVA
University of Modena and Reggio Emilia, Italy

ASSOCIATE EDITORS

LUISA AMENTA
University of Palermo, Italy

DANIELA ANDREINI
University of Bergamo, Italy

TEL AMIEL
University of Campinas, Brasil

FAWZI BAROUD
Notre Dame University - Louaize, Lebanon

CHERYL BROWN
University of Canterbury, New Zealand

ALESSIA CADAMURO
University of Modena and Reggio Emilia, Italy

MARINA DE ROSSI
University of Padova, Italy

ULF-DANIELS EHLERS
Baden-Wurttemberg Cooperative State University in Karlsruhe, Germany

ANTONIO MOREIRA TEIXEIRA
Universidade Aberta - Lisboa, Portugal

MARIA RANIERI
University of Firenze, Italy

PIER CESARE RIVOLTELLA
Alma Mater University of Bologna, Italy

AHMED TLILI
Bejing Normal University, China

DANIEL VILLAR-ONRUBIA
Coventry University, United Kingdom

ZORAINI WATI ABAS
Wawasan Open University , Malaysia

MARIA C. ZAMORA
Kean University, Union (NJ), USA

ASSISTANT EDITORS

SARA GUERRA
University of Modena and Reggio Emilia, Italy

MARINA LINACERO MARTIN
University of Modena and Reggio Emilia, Italy

ELHAMSADAT MOUSAVINASAB
University of Modena and Reggio Emilia, Italy

LIDIA ALFONSINA MARTINEZ ORTUBIA
University of Modena and Reggio Emilia, Italy

ALICE PIERLUIGI
University of Modena and Reggio Emilia, Italy

LUISA RIGHETTI
University of Modena and Reggio Emilia, Italy

SHWETA ROY CHOWDHURY
University of Modena and Reggio Emilia, Italy

KATIA SANNICANDRO
University of Modena and Reggio Emilia, Italy

LUCA TADDEO
University of Modena and Reggio Emilia, Italy

MANAGING EDITOR

ANNAMARIA DE SANTIS
University of Modena and Reggio Emilia, Italy

SCIENTIFIC COMMITTEE

Adorni Giovanni - University of Genova, Italy
Bellini Claudia - University of Modena-Reggio Emlia, Italy
Bonaiuti Giovanni - University of Cagliari, Italy
Calvani Antonio - University of Firenze, Italy
Cantoni Lorenzo - University of Lugano, Switzerland
Carbonaro Antonella - University of Bologna, Italy
Cartelli Antonio - University of Cassino, Italy
Cecconi Luciano - Univ. of Modena-Reggio Emilia, Italy
Cerri Renza - University of Genova, Italy
Cesareni Donatella - University of Roma, Italy
Coccoli Mauro - University of Genova, Italy
Delfino Manuela - C.N.R. I.T.D of Genova, Italy
Faiella Filomena, University of Salerno, Italy
Ghislandi Patrizia - University of Trento, Italy
Guerin Helen - University College Dublin Ireland
Guerra Luigi - University of Bologna, Italy
Holotescu Carmen - University of Timisoara, Romania
Karacapilidis Nikos - University of Patras, Greece
Karlsson Goran - University of Stockholm, Sweden
Kess Pekka - University of Oulu, Finland
Ligorio Beatrice - University of Bari, Italy
Manca Stefania - C.N.R. I.T.D of Genova, Italy
Mandl Heinz - Universitat Munchen, Germany
Mangione Giuseppina Rita, INDIRE, Italy
Maresca Paolo - University of Napoli Federico II, Italy
Marzano Antonio - University of Salerno, Italy
Moriggi Stefano - University of Modena-Reggio Emilia, Italy
Persico Donatella - C.N.R. I.T.D of Genova, Italy
Pirlo Giuseppe, University of Bari, Italy
Rizzo Antonio - University of Siena, Italy
Rossano Veronica - University of Bari, Italy

REVIEWERS

Cosimo Accoto, Francesco Agrusti, Daniela Amendola, Alice Barana, Claudia Bellini, Elisa Bisagno, Giovanni Bonaiuti, Beatrice Bonami, Daniela Borissova, Filippo Bruni, Alessia Cadamuro, Gabriella Calvano, Roberto Capone, Antonella Carbonaro, Chiara Carletti, Gabriella Casalino, Alessio Cavicchi, Alessio Ceccherelli, Luciano Cecconi, Alessandro Ciasullo, Letizia Cinganotto, Mauro Coccoli, Francesca D'Errico, Estela Daukšienė, Annamaria De Santis, Sara Dias-Trindade, Gabriella Dodero, Hendrik Drechsler, Ulf Daniel Ehlers, Meryem Erbilek, Antonietta Esposito, Filomena Faiella, Laksmi Evasufi Fajari, Laura Fedeli, Cinzia Ferranti, Paolo Ferri, Giuseppe Fiorentino, Giuliana Franceschinis, Rita Francese, Maurizio Gentile, Carlo Giovannella, Patrizia Garista, Teodora Hristova, Franco Landriscina, Luca Andrea Ludovico, Giuseppina Rita Jose Mangione, Paolo Maresca, Antonio Marzano, Marco Masoni, Rory Mc Greal, Michele Marangi, Salvatore Messina, Sergio Miranda, Andrea Molinari, Stefano Moriggi,Giacomo Nalli, Fabio Nascimbeni, Silvia Panzavolta, Gigliola Paviotti, Corrado Petrucco, Maria Chiara Pettenati, Antonella Poce, Giorgio Poletti, Alessia Rosa, Manuela Repetto, Ana Luísa Rodrigues, María Esther Rodríguez-Gil, Veronica Rossano, Francesca Rossi, Marina Rui, Margherita Russo, Matteo Sacchet, Katia Sannicandro, Ines Saric-Grgic, Javier Sarsa, Lorenzo Scianatico, Antonio Teixera, Admed Tlili, Guglielmo Trentin, Andrea Trentini, Rosa Vegliante, Gianni Vercelli, Nicola Villa

JOURNAL OF e-LEARNING AND KNOWLEDGE SOCIETY

www.je-lks.org

ISSN (online) 1971 - 8829
ISSN (paper) 1826 - 6223

VOLUME 22 | ISSUE NO. 1 | APRIL 2026

PAG. 1 EDITORIAL
Stefano Moriggi, Luca De Biase, Stefano Pasta, Pier Cesare Rivoltella, Loris Vezzali

DISCUSSION TOPICS

PAG. V THE ENGINEERING OF HATE
Luca De Biase

INVITED PAPERS

PAG. X CAN EDUCATION COUNTER A CULTURE OF HATE?
Paola Cagliari

PAG. XXI LEARNING FROM VIOLENCE. HATE LITERACY AS A CORE COMPETENCE FOR CONTEMPORARY CITIZENSHIP
Mario Pireddu

PAG. XXXII "TO REDEEM HORROR FROM HIS INVISIBILITY". FORENSIC ARCHITECTURE'S WORK WITH IMAGES
Maurizio Guerri

PEER REVIEWED PAPERS
SPECIAL ISSUE ON
"NEW ALPHABETS OF HATE"

PAG. 1 — SECURITIZED MIGRATION NEWS AND HOSTILE AUDIENCE REACTIONS ON INSTAGRAM: A COMPUTATIONAL ANALYSIS OF ITALIAN NEWS POSTS AND USER COMMENTS
Nuccio Ludovico, Valentina Rizzoli, Aysenur Didem Yilmaz, Alice Lucarini, Veronica Margherita Cocco, Loris Vezzali

PAG. 17 — FROM HATE SPEECH TO TOXICITY: DEPOLITICIZATION, ALGORITHMIC GOVERNANCE, AND THE TRANSFORMATION OF HARM IN DIGITAL PUBLIC DISCOURSE
Roberto Bortone, Stefano Pasta

PAG. 27 — POETICISING LIFE. ART AND ART EDUCATION AS FORMS OF RESISTANCE AND ELABORATION OF THE LOGIC OF HATE
Chiara Panciroli, Pier Cesare Rivoltella

PAG. 37 — POLITICALLY MOTIVATED HATE SPEECH IN ALBANIAN LANGUAGE SOCIAL MEDIA
Rafaela Marteta

PAG. 49 — DIGITAL LITERACY PROGRAMMES AND SOCIAL INTERVENTIONS FOR BUILDING RESILIENCE AGAINST HATE SPEECH AMONG SCHOOL STUDENTS IN KERALA
Genimon Vadakkemulanjanal Joseph, Sonia Philomena V A, Athira P

PAG. 61 — WHEN AI IS SEXIST: INVESTIGATING RESPONSES TO AI AND HUMAN DISCRIMINATION
Selen Akay, Junko Kanero, Çiğdem Bağcı

PAG. 75 BEYOND AUTOMATION: THE COLLABORATIVE ROLE OF AI IN COUNTERING HATE SPEECH. A SCOPING REVIEW
Antonella Cosola, Marinella Paciello, Alessandro Pollini, Francesca D'Errico

PAG. 91 FROM EUPHEMISM TO EXTREMISM: EMERGING ALPHABETS OF HATE IN DIGITAL ACADEMIC SPACES IN EASTERN INDONESIA
Remerta Noni Naatonis, Sumarlin Sumarlin, Jimi Asmara, Yosep Jacob Latuan, Dewi Anggraini, Menhya Snae, Heni Heni, Erna Rosani Nubatonis, Sadam Muhammad Natzir, Ewaldus Kase

PEER REVIEWED PAPERS
PERMANENT CALL

PAG. 105 EVOLVING TEACHER INTERESTS IN INNOVATIVE LEARNING ENVIRONMENTS: IMPACTS OF EXPOSURE AND IMPLICATIONS FOR PROFESSIONAL DEVELOPMENT
Amelia R Granda-Pinan, Óscar R. Lozano

PAG. 119 HOW CAN EDUCATIONAL VIDEOS WITH ANNOTATIONS HELP STUDENTS IN ONLINE LEARNING ENVIRONMENTS?
Riad Bourbia, Samia Drici, Yacine Lafifi, Sevinc Gulsecen

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), I-IV

EDITORIAL

New Alphabets of Hate

Stefano Moriggi
University of Modena and Reggio Emilia (Italy)

Luca De Biase
Media Ecology Research Network – ReImagine-Europa

Stefano Pasta
Catholic University of the Sacred Heart (Italy)

Pier Cesare Rivoltella
Alma Mater Studiorum Università di Bologna (Italy)

Loris Vezzali
University of Modena and Reggio Emilia (Italy)

(published: 12/9/2026)

DOI
https://doi.org/10.20368/1971-8829/1136474

CITE AS
Moriggi, S., De Biase, L., Pasta, S., Rivoltella, P.C., & Vezzali, L. (2026). New Alphabets of Hate [Editorial]. *Journal of e-Learning and Knowledge Society*, 22(1), I-IV. https://doi.org/10.20368/1971-8829/1136474

Hate speech in the digital context has earned a place of primary importance in the cultural and political analysis of a world that communicates through social media. Not because social media uncovered some unprecedented forms of violence and division that did not exist before the first quarter of the twenty-first century, but because they have amplified the issue to such an extent as to prompt scholars to examine it more deeply. The point is that, with the advent of generative artificial intelligence, the significance of this dimension of social life may grow still further and take on features that are even harder to grasp. And this is not primarily for strictly technical reasons: research on the subject, aided in part by the close study of hate speech, is defining in a potentially surprising way the relationship between the violence of the discourse that can find a home in social networks and the evolution of the media ecosystem.

© Italian e-Learning Association

The research reported in this issue of the *Journal of e-Learning and Knowledge Society* finds, in fact, that the apparent correlation between the growth of online violence and the development of social media is connected to explanations that lie, logically, further upstream of both phenomena.

It is precisely this upstream question that the scenario essay opening the issue sets out to answer. Luca De Biase shows how hatred, while remaining in its origins a singular and irreducible phenomenon – freely echoing Tolstoy, every story of hate is different from every other – is converted, through the digital infrastructures across which it circulates, into a synthetic and standardized phenomenon that tends to converge on a single expressive form: violence. The essay proposes a taxonomy of the relational structures through which hatred has historically propagated – from the "one-to-one" relationship of domestic violence, to the "one-to-many" transmission typical of totalitarian propaganda, to the "many-to-many" form of social media, and to the still largely unexplored "many-to-one" form of conversational artificial intelligence, capable of simulating all the other structures and thereby making the boundary between personal dialogue and industrially produced hate speech increasingly permeable (De Biase, 2026). It is this interpretive framework – the idea that infrastructures do not generate hatred but constitute its operational environment, amplifying its reach and, to some extent, even its alphabets – that offers the key for reading the eleven contributions that follow.

This issue opens with this scenario essay and then brings together eleven contributions, organized in two sections: three invited papers, which introduce the theme with a broad and situated perspective, and eight research articles, selected through the issue's dedicated call, which convey its scope along four interconnected thematic axes.

A first axis concerns the very genealogy of the categories through which we name hatred: while the study of politically motivated hate speech in the Albanian language proposes reading it as a structured repertoire of "alphabets" – symbols, metaphors, and recurring linguistic patterns that recombine to convey and normalize hostility (Marteta, 2026) – a historical-conceptual analysis reconstructs how, within platform moderation practices, the legal-political category of "hate speech" is progressively giving way to the more technical and scalable category of "toxicity," with the risk of a depoliticization of discriminatory harm (Bortone & Pasta, 2026). Alongside this conceptual genealogy sits an empirical case study that carries the notion of an "alphabet of hate" beyond the English-language, high-resource contexts in which it has mostly been developed so far: a mixed computational and discourse-analytic study of a 4,200-item corpus collected from digital academic spaces in Eastern Indonesia identifies five local linguistic strategies – ranging from ethnic euphemism to ironic praise and dehumanizing metaphor – and shows that implicit, euphemistic hate speech systematically outnumbers explicit hate speech, at a ratio of roughly 1.5 to 1, a finding that reinforces the decolonial argument that the very categories used to analyze hate speech require renegotiation outside the Western contexts in which they originated (Naatonis et al., 2026).

A second axis concerns the ambivalent role of generative artificial intelligence in processes of discrimination and of countering hatred. Two experimental studies conducted in Turkey and the United States show that the same discriminatory conduct is perceived as less threatening when carried out by an AI system rather than by a person: a perceptual bias that risks normalizing algorithmic discrimination precisely because it is judged less severe (Akay et al., 2026). Conversely, a systematic review of AI-generated counterspeech shows that the effectiveness of these tools depends on remaining within a logic of human-machine collaboration, rather than one of the latter substituting for the former (Cosola et al., 2026). These findings confirm, at an experimental scale, the ambivalence of the "many-to-one" structure described in the opening essay, where the degree of alignment built into language models becomes itself a factor that can either contain or amplify the algorithmic production of hatred (De Biase, 2026).

A third axis concerns the role of digital infrastructures in mediating public discourse: a computational analysis of Instagram posts and comments on migration shows that securitized frames in journalistic discourse are associated with more polarized emotional reactions and with higher rates of hate speech, compared with humanitarian or ethical-religious frames (Ludovico et al., 2026).

A fourth axis, finally, concerns educational responses to these phenomena. Among the invited papers, the proposal of "hate literacy" as a core civic competence redefines literacy about hate speech not as prevention or self-defense, but as a practice of critical reading capable of making explicit the hidden curriculum of platforms – an argument that engages closely with the critique of the vocabulary of "toxicity" put forward in the first axis (Pireddu, 2026); while the pedagogical experience of infant-toddler centers and preschools in Reggio Emilia shows how attention to relationships and environments, from the very earliest years of life, constitutes a first and decisive cultural safeguard (Cagliari, 2026). Among the call articles, the practices explored range from art and art education as devices of

© Italian e-Learning Association

resonance capable of defusing hatred (Panciroli & Rivoltella, 2026) to the digital literacy and resilience programs developed in schools in Kerala (Joseph et al., 2026).

Closing this survey of contributions, the third invited paper shifts the gaze from discourse to the image of extreme violence: an analysis of the work of the Forensic Architecture collective shows how the very technological infrastructures – artificial intelligence, satellite imagery, platforms – that amplify the visibility and emotional charge of online hatred can, conversely, anesthetize the perception of state violence, rendering it invisible even amid its media overexposure (Guerri, 2026). If the alphabets of hate feed on algorithmic hypervisibility, the invisibility of the most extreme violence is its reverse side: two outcomes of the same, never neutral, political character of the technologies of vision.

The new alphabets of hate can therefore be sought in the intersection between the social and personal suffering that leads people to hate – fueling violence, verbal and often physical – and the growing power of digital infrastructures, with their incentive systems, their interfaces, and their recommendation algorithms, which tend to disproportionately amplify whatever generates the strongest emotional engagement, hatred included. These are "alphabets," rather than simple languages, because what we are dealing with is an unstable system of signs – slang, memes, allusions, implicit codes – capable of continuously recombining itself so as to remain legible to those who use it while eluding those who should recognize and moderate it. Generative artificial intelligence, referenced at the outset, is today its most radical expression: it makes it possible to produce, at scale, ever-new variants of the same hate speech, continuously shifting the target of detection systems and reproducing, in synthetic form, the same combinatorial logic that makes these alphabets so difficult to pin down and to counter.

Education is an integral part of any strategy for preparing against and countering the various forms that online violence takes from time to time. It cannot, however, confine itself to countering discrimination and violence directed at specific categories of people: it must reach a more sophisticated level. It must be conceived, first and foremost, for the contribution it can make to developing strategic competencies such as the capacity to identify alternatives to dominant narratives and to critically read any form of prejudice. Ultimately, this places at the center of the analysis the new context of power and socio-economic polarization – a context that fuels the distance between people and loneliness, and that makes possible disrespectful behavior toward others, or worse.

Hate speech literacy must evolve in step with the new alphabets of hate: it must be developed to strengthen the defenses of individuals and communities against the potential violence fueled by discourse that develops and spreads online, and it must also broaden its aims. It cannot remain neutral toward technologies, and even less so toward the system of power that defines the conditions shaping their development.

The construction of new educational environments is decisive for creating alternative spaces in which to develop more conscious approaches to hate speech and to nurture people's sense of agency, preventing online violence from rendering them passive. After all, this construction is vital for an independent education oriented toward liberating people's creative forces and toward defining a critical position with respect to the current media context. It is in this shift – from education as individual defense to education as a lever for transforming the context of power – that the broader stakes of this issue can be grasped. The historical moment we are living through sends a clear signal: we are moving from regarding digitalization as a synonym for modernization to situating technologies within the social and political dynamics that also interpret them as tools for mass manipulation. It is in this light that the response to the alphabets of hate becomes, first and foremost, an architectural question: even more than content regulation, it calls for the (re)design of information ecosystems that are less oligopolistic, more interoperable, and better able to safeguard the diversity of voices, including the weakest ones.

References

Akay, S., Kanero, J., & Bağcı, Çiğdem. (2026). When AI is sexist: Investigating responses to AI and human discrimination. *Journal of e-Learning and Knowledge Society*, *22*(1), 61-74. https://doi.org/10.20368/1971-8829/1136320

Bortone, R., & Pasta, S. (2026). From hate speech to toxicity: depoliticization, algorithmic governance, and the transformation of harm in digital public discourse. *Journal of e-Learning and Knowledge Society*, *22*(1), 17-26. https://doi.org/10.20368/1971-8829/1136368

Cagliari, P. (2026). Can education counter a culture of hate?. *Journal of e-Learning and Knowledge Society*, *22*(1). https://doi.org/10.20368/1971-8829/1136375

Cosola, A., Paciello, M., Pollini, A., & D'Errico, F. (2026). Beyond automation: the collaborative role of AI in countering Hate Speech. A scoping review. *Journal of e-Learning and Knowledge Society*,

© Italian e-Learning Association

22(1), 75-90. https://doi.org/10.20368/1971-8829/1136378

De Biase, L. (2026). The Engineering of Hate. *Journal of e-Learning and Knowledge Society*, *22*(1), V-IX. https://doi.org/10.20368/1971-8829/1136471

Guerri, M. (2026). "To redeem horror from his invisibility". Forensic Architecture's work with images. *Journal of e-Learning and Knowledge Society*, *22*(1). https://doi.org/10.20368/1971-8829/1136376

Joseph, G. V., Philomena V A, S., & P, A. (2026). Digital literacy programmes and social Interventions for building resilience against hate speech among school students in Kerala. *Journal of e-Learning and Knowledge Society*, *22*(1), 49-60. https://doi.org/10.20368/1971-8829/1136321

Ludovico, N., Rizzoli, V., Yilmaz, A. D., Lucarini, A., Cocco, V. M., & Vezzali, L. (2026). Securitized migration news and hostile audience reactions on Instagram: a computational analysis of Italian news posts and user comments. *Journal of e-Learning and Knowledge Society*, *22*(1), 1-16. https://doi.org/10.20368/1971-8829/1136366

Marteta, R. (2026). Politically motivated hate speech in Albanian language social media. *Journal of e-Learning and Knowledge Society*, *22*(1), 37-48. https://doi.org/10.20368/1971-8829/1136315

Naatonis, R. N., Sumarlin, S., Asmara, J., Latuan, Y. J., Anggraini, D., Snae, M., Heni, H., Nubatonis, E. R., Natzir, S. M., & Kase, E. (2026). From Euphemism to Extremism: Emerging Alphabets of Hate in Digital Academic Spaces in Eastern Indonesia. *Journal of e-Learning and Knowledge Society*, *22*(1), 91-104. https://doi.org/10.20368/1971-8829/1136316

Panciroli, C., & Rivoltella, P. C. (2026). Poeticising life. Art and art education as forms of resistance and elaboration of the logic of hate. *Journal of e-Learning and Knowledge Society*, *22*(1), 27-36. https://doi.org/10.20368/1971-8829/1136367

Pireddu, M. (2026). Learning from violence. Hate Literacy as a core competence for contemporary citizenship. *Journal of e-Learning and Knowledge Society*, *22*(1). https://doi.org/10.20368/1971-8829/1136377

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), pp.V-IX

DISCUSSION TOPICS

The Engineering of Hate

Luca De Biase[1]

Media Ecology Research Network – ReImagine-Europa

(published: 12/9/2026)

DOI
https://doi.org/10.20368/1971-8829/1136471

CITE AS
De Biase, L. (2026). The Engineering of Hate [Invited Paper]. *Journal of e-Learning and Knowledge Society*, 22(1), V-IX. https://doi.org/10.20368/1971-8829/1136471

All happy families resemble one another, Tolstoy famously wrote, while every unhappy family is unhappy in its own way. Something similar might be said of hatred: love stories may resemble one another, but every story of hate is different.

Understanding the origins of individual hatred belongs to another inquiry. This essay is not about personal suffering, collective loneliness, or the historical forces that push certain people to the margins, nor it is about the breakdown of the social elevator that, for half a century after the Second World War, gave hope to the lives of a billion people in the West and now frustrates the expectations of just as many. Its focus is narrower: the consequences of the digital infrastructures through which the many forms of hatred are communicated.

The question is how, on those infrastructures, different stories of hatred are converted into terrifying synthetic phenomena and eventually converge upon a single expressive form: violence. In such environments, human diversity is channelled into stereotyped social behaviour through communication tools that are themselves highly standardised. Individual stories do not remain individual. They enter systems of connections, conventions and convictions.

These media are not neutral. They have a history. What do we know about that history as connected to the morphology of hate speech? Why does it matter? And what can be done to influence its direction?

The Structures of Hate

Hatred can develop only within structures that host relationships between people. Those structures are not the cause of hatred, but they form its operational media environment. There is probably a relationship between the configuration of that environment and the shape taken by hate speech. How might such a relationship be studied? And how might knowledge of it help us design digital platforms and media systems that are less hospitable to hatred?

Generalisation is difficult, of course. Not all forms of hatred are the same. The form considered here is hatred directed at categories of people, in which individuals become targets only because they are assumed to belong to a particular group. This is one reason such hatred is especially intolerable.

The presumed relationship between digital media structures and the morphology of hate speech raises a central question: is the digital environment especially conducive to categorisation? If so, does it create conditions that are particularly compatible with forms of hatred directed at categories of people?

[1] corresponding author - email: luca.debiase@gmail.com

© Italian e-Learning Association

The history of computing offers one obvious clue. Computer interfaces developed through menus, keywords and systems for classifying information according to standardised concepts. In the age of social media, this logic was powerfully extended to the interpretation of people themselves. The scale of digital social platforms encouraged a quantitative view of human beings. People were transformed into homogenised profiles, limited to a relatively narrow range of possible actions, processed by fundamentally reductive algorithms and encouraged to move from the status of persons to that of characters.

The masks they wore in order to function successfully within those environments gradually hardened. The language of relationships mediated by the dominant configurations of digital technology thus became compatible with an interpretation of people based less on the specificity of their biographies than on the accumulation of identities and categories by which they could be classified.

Categorisation, of course, is far older than digital technology. Racism makes that clear. But within the digital media ecosystem as it has developed so far, it is easier to see others through the categories used to classify them than to encounter them through the singularity of their lives.

All this has unfolded across different communication structures. Some are organised primarily around relationships between individuals. Others are designed to amplify the voice of one person before a multitude. Still others bring multitudes into contact with one another.

The Forms of Relational Infrastructure

For a long time, the structures that hosted human relationships could be reduced, at least schematically, to two main forms. During the first quarter of the twenty-first century, a third expanded dramatically. A fourth is now emerging.

The first two are the familiar “one-to-one” and “one-to-many” forms. The third is “many-to-many”: much has already been written about it, though its implications still deserve clarification. The fourth is “many-to-one”, and remains largely unexplored.

In everyday conversation between individuals, the most basic communication infrastructure, individual hatred is projected onto individual victims. It often remains enclosed within the home and finds its secret expression in domestic violence. The suffering of a person who hates from a position of strength is discharged upon someone in a position of weakness, becoming the victim’s unbearable suffering.

In “one-to-one” relationships, hatred is inflicted and endured in silence: amid complicity, injustice and helplessness. Without words.

In the world of broadcasting, by contrast, a leader directs the masses towards hatred of entire categories of people in order to consolidate power. The Holocaust stands as the most terrible monument to this structure of hatred. But it is not the only example humanity has produced.

The importance of this dimension extends beyond the spectacular efficiency of German and Italian persecution of Jews and of the other groups targeted by Nazism and Fascism. Colonialism, nationalism and religious war have all generated versions of the same structure: communication flowing from one to many, from the leader to the followers.

The rise of social media brought an extraordinary acceleration in access to debate and knowledge. It also enabled many people to communicate with many others. Within that environment, however, small-scale leaders of hatred multiplied, each able to address an audience enclosed within the same bubble.

“Many-to-many” communication thus became one of the emblematic spaces in which contemporary hatred and violence are expressed.

Today a further dimension is emerging: generative artificial intelligence used through the interface of chat. Each user may experience this technology as a machine with which to establish an intimate and highly personal dialogue. Yet the machine can respond only by processing knowledge produced and recorded by a vast number of people. From that collective material, it generates answers that appear personal.

The model resembles Her, Spike Jonze’s 2013 film, in which a man falls in love with Samantha, a highly expressive artificial intelligence, and believes that his love is reciprocated. Eventually, however, Samantha reveals that she is emotionally involved with hundreds of users at the same time.

Like ChatGPT, Samantha is not an individual conversing with another individual. She is a collective voice: the product of an algorithmic synthesis of the knowledge and language expressed and recorded online by billions of people. She is collective even when she addresses each individual personally.

The user projects an individual personality onto her. But she can inhabit many relationships at once with the same apparent naturalness precisely because she is made from knowledge produced by many.

What, then, happens when we move from love to hatred in the “many to one” infrastructure?

The alignment of mainstream models can clearly act as a restraint upon violent behaviour. But a model such as Grok, whose owner Elon Musk presents it as an expression of unrestricted and unapologetic freedom of speech, is deliberately less constrained. Over time, many other models will emerge, each embodying the assumptions of its designers and each capable of speaking from many to one while interpreting cultures of every kind, including those through which hatred

© Italian e-Learning Association

travels.
The most consequential examples may arise in cognitive warfare. States engaged in hostile operations can use generative artificial intelligence to produce vast quantities of messages designed to weaken an adversary's social cohesion, including messages of hatred.
"Many-to-one" structures can imitate all the others. They can simulate "one-to-one" relationships, amplify "one-to-many" communication and industrialise the production of content suited to "many-to-many" environments. Because they are simulations, they can reproduce almost any communicative condition.
In doing so, they may make it easier to construct the narrative environments in which hatred can spread.

Language as Technology

The importance of the structures through which hatred circulates is overstated when those structures are described in purely technological terms.
Even the algorithms that, in "many-to-many" environments, select apparently personalised and highly engaging messages in order to keep users on platforms for as long as possible have different effects in different historical periods and geographical contexts.
The historical, economic, social and political causes of hatred do not originate in the structures through which messages of hatred are distributed. But those structures can amplify their reach. They can also influence their language and perhaps even their alphabets.
Languages are not merely instruments of communicative exchange. They are saturated with cultural experience accumulated in the collective memory of peoples. That experience re-enters communication whenever an expression encounters the interpretive capacities of those who receive it.
Alphabets, in their different graphic forms, are bridges between sound and writing. More broadly, in other systems of signs, they connect images and writing. In every communication structure, alphabets and languages are part of the interface linking embodied human experience to mediated interaction.
The language that develops in digital environments changes according to the form of communication involved: whether between individuals who know one another, between a small number of leaders and vast audiences, between many influencers and numerous relational bubbles, or through automated generative chat.
In the last of these cases, alphabets become especially significant, at least at the level of their smallest units. Contemporary artificial intelligence generally operates through tokens: fragments of words, syllables, characters or other linguistic units whose possible continuations are predicted statistically.
As technology and science advance, systems may be developed that operate through larger units of meaning, such as whole sentences or more explicit representations of context, increasing the ability of machines to interpret language in relation to what surrounds it.
Nothing prevents us, moreover, from imagining machines that operate at still more elemental levels when they must control robots and autonomous vehicles and therefore read the physical environment in which they act. At that level, they transform many kinds of signals - temperature readings, camera images, motion data, radar signals and much else - into machine-readable systems of signs.
These, too, can be understood as alphabets and languages: systems that make interaction between machines possible and that may be compared with scientifically developed models of reality.
The "many-to-one" communication structures typical of artificial intelligence interacting with individuals in specific physical environments will expand enormously with the spread of robotics.
In relation to hatred, one need only think of the drones and robots already conducting major operations in contemporary warfare. They are changing not only the practice of violence but perhaps also its meaning.
Should the same word be used when a drone shoots down another drone?
The algorithmic and automated mediation of human relationships therefore requires deeper reflection. It is not only a legal question of responsibility. It is not only an ethical question involving freedom of expression or the right to protection from violence. It is also an epistemological question.
How is the knowledge operationalised by algorithms and language models produced? How can training be designed so that respect for rights is built in from the outset, rather than added afterwards through improvised forms of alignment? And how can the likelihood of exposure to hatred be reduced in environments where the volume of messages is too great to be managed manually?

The Consequences of Categories

It is objectively difficult to define hate speech in operational terms that can support automated intervention and allow artificial intelligence itself to be entrusted with the defence of people against automated attacks.
In social media, within "many-to-many" structures, the automation of such defences has often increased the autocratic power of those who own and govern the platforms. The automated determination of what counts as hatred depends upon the categories, priorities and

© Italian e-Learning Association

designs of the owners of the communication infrastructure.
The same could happen in artificial intelligence if the underlying technologies remain proprietary and concentrated in the hands of a small number of oligopolists.
In that case, design, training and alignment impose upon users the values of owners and their organisations. Because oligopolists derive considerable economic benefits from their market dominance, they lobby intensely against regulation intended to protect citizens. So far, those efforts have achieved substantial success.
The values of corporate owners have therefore become, in practice, the rules governing relationships between citizens whenever those relationships are mediated by large platforms.
The categories through which oligopolists identify and evaluate hate speech determine how that speech is mitigated. This is both inadequate and unjust.
Meta's failure to respond adequately to the persecution of the Rohingya in Myanmar, which began in 2017 and continued for years, provides a particularly stark example. The relevant categories were absent from its systems, and the platforms did not adequately account for the local language. That kind of hate didn't exist in terms of Meta's algorithm. Meta consequently became an instrument through which terrible violence was organised.

What can be done?

Regulation, standardisation, public debate and many other democratically organised forms of intervention are and will remain necessary. But the development of digital platforms also points towards a practical lesson.
Measures that reduce oligopolistic control over social platforms and artificial intelligence can also help prevent reality from being identified with the prejudices embedded in models.
Information diversity, interoperable platforms, and decentralised and widely distributed innovation are among the preconditions for a complex system capable of generating the antibodies needed to resist expressions of hatred.
Alternatives are entirely possible, and often highly attractive.
In artificial intelligence, growing numbers of users and companies are choosing open-weight technologies. These can support transparency, autonomy and the interpretability of automated systems. They can also make it easier to connect algorithmic operations with the meanings attributed to them in social contexts.
At this level, the response to hate speech becomes architectural.
The development of transparent and interpretable forms of artificial intelligence is a precondition for an open and collectively governed discussion of the expressive forms that produce violence. This requires a healthy alliance between science and innovative companies capable of offering alternatives to dominant firms.
The possibilities explored by researchers suggest that the available alternatives are far more numerous than the dominant technological structures currently allow us to see.
If the "alphabet" of large language models has so far been relatively close to the structure of fragments and syllables, new scientific approaches are seeking ways to generate language through larger units of meaning, such as sentences and contexts.
The analogy is imperfect, but it is almost as though the system were moving from a phonetic towards an ideographic form of writing. Such a development could create conditions for less schematic ways of processing information, more attentive to cultural context, differences in sensibility and nuances of meaning.
This brings us to the central issue in any discussion of hate speech in digital environments: the relationship between the architecture of digital media and the categorisation of people.
Interpreting individuals primarily as members of categories is one of the foundations of the most pervasive rhetorics of hatred. It is also a route towards some of the worst forms of suffering.
Those who hate a category deny the possibility that individuals may differ from the stereotypes attached to it. They attribute supposed collective responsibilities to individual people.
The consequence is straightforward: the freedom of expression of those who hate can effectively destroy the freedom of expression of those who are hated. The latter often withdraw from public debate simply to escape attack.
Categories and the prejudices attached to them are inherited from archives accumulated over time. They change slowly and operate covertly. Their influence is implicit in databases of every kind and explicit in many of the artificial intelligence systems developed by dominant companies.
The most effective way to reduce the impact of the prejudices embedded in oligopolistic platforms is to make those platforms less oligopolistic: by multiplying alternatives that are open and interoperable, transparent and open to criticism.
Hatred and violence, unfortunately, cannot be eliminated. But a balance among rights, and among the different qualities of civil discourse, is less likely to emerge in a context in which the circulation of ideas depends upon media monopolised by a small number of people who impose their own assumptions upon the collective creativity of humanity.

© Italian e-Learning Association

Society is not simply a collection of relational bubbles in which groups hate the categories and collective characteristics of others. Yet the dominant platforms have created an environment that reflects precisely this image of society.

Progress is not guaranteed by the absolute freedom of expression of the powerful. It is more likely to arise from the democratic evolution of coexistence in environments whose information diversity creates space for the expression of the weakest and least powerful as well.

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026)

INVITED PAPER

Can education counter a culture of hate?

Paola Cagliari[a,1]

[a]*Reggio Children (Italy)*

(published: 23/5/2026)

DOI
https://doi.org/10.20368/1971-8829/1136375

CITE AS
Cagliari, P. (2026). Can education counter a culture of hate? [Invited paper]. *Journal of e-Learning and Knowledge Society*, *22*(1).
https://doi.org/10.20368/1971-8829/1136375

We have words for selling
words for buying
words for making words
but we need words for thinking.

We have words for killing
words for sleeping
words for tickling
but we need words for loving.

We have machines
for writing words
dictators, tape recorders
microphones telephones.

We have words
for making noise,
[but] words for speaking,
there now are none.

1. Introduction

We wanted to introduce our contribution with "The Words", a poem by Gianni Rodari, who, in the 1970s, emphasised how we needed new words, denouncing the lack on a social level of the words for living together, for a civil co-existence based on thought, solidarity, and dialogue.

Like all languages, the "language of hate", an expression used in Italian for "hate speech", is not constructed solely through listening to words, but on attitudes, looks, and ways of relating that have shaped and continue to shape the contexts of our experience, from birth, and that orient the construction of our idea of ourselves, of others, of relationships, and of the world. Words, however, are important, and proposing new languages to describe old and new realities is a necessary cultural undertaking in every society.

For this reason, in our contribution, which starts with experiences with children aged 0 to 6, built up over more than sixty years of work in Reggio Emilia's municipal infant-toddler centres and preschools, and which we propose as a case study, we will address the theme of education in a general sense in order to offer tools and approaches that are useful to schools in developing antibodies in children and young people that counter, at least in part, the culture of incitement to hatred, and how this experience of education has

[1] corresponding author - email: paola.cagliari57@gmail.com

© Italian e-Learning Association

proposed new languages to give shape to a new educational reality.

The Reggio Emilia Approach, known and appreciated in many parts of the world, shares both the cultural context in which it developed, and theories and practices, with other experiences in education in Italian municipalities, but has the distinctive feature of being an organic and systemic elaboration in which each part is conceived in coherence and relation to the others. It is an experience of a holographic nature, so that each aspect examined will always yield the same values and meanings.

It is not our task, in this piece, to analyse where the "languages of hate" and hate speech are generated, but a very brief examination of some issues is necessary to provide the reader with elements to understand how an experience of education developed more than sixty years ago still responds to questions asked today.

The current spread of hate speech has systemic and social roots linked to phenomena such as competition forced upon us by the market, disappointment among younger generations in their expectations of improved social status, migratory flows, the need for belonging to a group and cohesion, the sense of community that if it is experienced as a closed group leads to creating enemies, the increase in economic and educational poverty, and more.

Over the past eighty years, concepts of freedom, authority, identity, belonging, difference, security, and rights – seemingly immutable products of centuries of development and traditions that – have been challenged in an attempt to interpret a rapidly and constantly changing social, cultural, technological and natural context. The acceleration in the development and diffusion of technologies, the pace of life and social change (see Rosa, 2013) have made meaningful, comprehensive and shared elaborations the dynamics of civil coexistence more difficult, a coexistence that without new references within new ecosystems such as the digital, feeds back into unstoppable vortices of action generating disorientation and conflicting positions.

Schools cannot chase after current events, but they must seek to be part of the contemporary world, taking responsibility for working with an eye to the future and seeking, whenever possible, to anticipate the times to come. In the debate on schools, which is relatively sparse in Italy, there has been talk of adding a new subject, sometimes referred to as education in respect, in relations, or on feelings, and of prohibiting children and young people from using electronic devices and accessing social media because – it is argued – expressions of hatred, at their maximum in these environments, reverberate in acts of in-person violence, intolerance, bullying, and vandalism both inside and outside school.

We believe it is important that the widespread concern about these phenomena be prevented from leading to interventions based on emergency, improvised, and with a restrictive, and punitive nature. Educating for languages other than those of hate has always been a fundamental and primary task of the family, educational services, schools and, more generally, society. Hate speech is not a separate phenomenon but generated within a culture that we, as educators, have to interrogate in order to find strategies and tools to generate change. School, as an institution that, at least from six to sixteen years, welcomes all children and young people without distinction of gender, social status or nationality, could be an incredibly powerful driving force in building a different culture.

For this to happen a paradigm shift is necessary, from schools based only on transmission or only on socialization, to schools capable of connecting children and young people with themselves, with others and with the world, through transmission, socialization, and active participation in the construction of co-living and knowledge. This first "posture" for countering a culture of hate replaces the disjoining conjunctions and mindset of "either/or", with that of "and…and", a conjunction and mindset that connects, and places different elements on the same level of logic.

2. The importance of the first six years of life

Our case study concerns infant-toddler centres and preschools that welcome children from birth to six years old. The first six years of a person's life are crucial for anchoring the development of cognitive, social and relational skills that individuals continue to form during their whole life, to values, attitudes and ethical orientation. This affirmation was present in certain late eighteenth century authors, but very fortunately twentieth century research and the cognitive neurosciences have provided scientific evidence that continues to give us further elements. This has produced a shift in language, which educators in Reggio Emilia have worked on long and tenaciously, contributing to its wider diffusion. Certain terms, expressions and images have turned on its head the idea that children are "plants to keep straight" or "empty vases to fill". Thus we have children who are "competent", "interlocutors" from birth with the people with whom they come into contact, curious, interested in gathering information on objects of the world and spaces around them, through acts of knowledge, gathering "vocabularies of acts" that are necessary for them to construct ever more complex activities, predisposed to associating exploratory acts of different nature (the visual, kinaesthetic and tactile), equipped with all the tools necessary for constructing their own knowledge, producers of theories and capable of communicating through a multiplicity of "languages" of expression (Malaguzzi et al., 1996): These new

© Italian e-Learning Association

words have slowly but surely modified ways of educating and instructing, though on this latter aspect in particular, much work remains to be done. It is a further demonstration, if any were needed, of the urgency of creating new alphabets and new words to give shape to the world.

In the first six years of life, the brain's plasticity and neuronal endowment are at their maximum, but it is the quality and quantity of encounters that provides favourable conditions for the maintenance and opening of new pathways, of synapses.

To borrow from Jeff Hawkins, authoritative neuroscientist and computer engineer, we can say children begin life with a neuro-cortex already equipped with hypotheses about the world, meaning it is already structured to be able to see, listen, and learn the languages of expression but lacking in specific knowledge (Hawkins & Dawkins, 2021).

It is this specific knowledge that poses a crucial question for adults: what contents are essential and through what channels to offer them, starting in the first months of life, so that children can organise their knowledge and their identity within a vision of planetary citizenship?

The gestures of care towards babies and children, the tone of voice with which we talk to them, the consideration and promptness towards needs they express, our openness to physical contact, the responsiveness of our gaze, and the ways and tones of relationships that form the context of a child's life, constitute one part of that first *specific knowledge* which offers a sense of trust and security, and reinforcing the desire to explore and know the world. In fact, in these early stages of each of our existences we begin to construct either a self that trusts the goodness of the world and openness towards others, or a self that is diffident.

This importance of the first six years of life should push governments to allocate significant investments to facilitate the fulfilment of a desire for parenthood and create conditions to safeguard and support this crucial role. Among these, the presence of quality infant-toddler centres and preschools is certainly the most urgent and necessary policy, as stated in numerous European Union Recommendations [Recommendations of the European Council, 22 May 2019 (2019/C 189/02) regarding high-quality early childhood education and care, and Recommendations of the European Council, 8 December 2022 (2022/C 484/01) on early childhood education and care, Barcelona Objectives for 2030].

3. Infant-toddler centres and preschools: social outposts against a culture of hate

In the Pedagogical Guidelines for Italy's integrated system of education and instruction for children aged birth to six, we can read the following definition:

> *Educational services and preschools are places of life in which children meet other children with whom to play, dialogue, share experience, develop friendships, and build relationships of reciprocal help. In these social contexts, they learn to learn with and from others, they sense they belong to a group, and they gradually internalize community rules (in routines, in play, in conversation) comprehending their meaning and managing conflicts.*
>
> *In educational institutions, relationships between children have the possibility of developing, with continuity and stability, making it possible for "cultures" to form, meanings shared between children in the form of rituals, games, and linguistic exchanges, which express and reinforce young children's initiative, understood as the capacity for expressing our own point of view. (Decree of the Italian Ministry of Education 22 November 2021, no. 334: Adoption of "Pedagogical Guidelines for the Integrated 0-6 System")*

This challenging affirmation, interpreted in relation to cognitive, emotional, and social differences in children and young people of different ages, could be useful as a reference for the entire school system.

Let us begin by exploring this initial interpretation of educational services and preschools to argue how a "pedagogy of relationships" (see Malaguzzi, 2012 in Edwards, Gandini & Forman, 2012, pp. 44-45) such as that present in Reggio Emilia's 0-6 educational services, can offer conceptual and strategic frameworks we believe could also be useful for later levels of schooling in building antidotes to a culture of hate together with children and young people. We repeat: we do not believe that teaching, prohibitions and hours dedicated to education on relationships are effective and coherent strategies for the problem of the spread of hate speech. Neither do we believe that simply inviting each teacher to introduce elements of the issue into their subject to be a decisive solution. We believe what is necessary is a change of paradigm that modifies schools and education, taking the many good experiences already being enacted as our starting point.

© Italian e-Learning Association

4. Infant-toddler centres and preschools are places of life

In our opinion, thinking of schools as a place of life, rather than primarily as places of education, is a good starting point. In the infant-toddler centre and preschool, the openness to the outside world, the social context and its institutions, the permeability between in school and out, is achieved through several daily acts: children bringing objects from home to school, parents, welcomed at the beginning and end of the day, speaking with teachers and listening to other children whom they know by name, the fluid structure of the day without interruptions or grading the importance between times of care, of play and educational activity, a coherent educational attitude in the adults and over the course of the day, spaces and environments which are not anonymous containers of an educational relationship, but which offer themselves to children for their research, transforming together with it and bearing the visible traces of their thinking and products. Spaces in which, together with the materials and tools available to all, personal objects can find a place and be protected. The day spent in infant-toddler centres and preschools, which includes the afternoon and was most likely conceived as a response to family needs has been an important driving force in shaping the identity of these institutions as places of life. The day and the time which, though constrained by an overall organization, is offered as a designable space in which everyone, at least in part, finds that their rhythms are welcome and that they can share in many different experiences of learning, play, and participation in activities of caring for oneself and for others. This participation and sharing pieces of life with others, in an educational environment that orients and supports, fosters feelings of understanding, tolerance, friendliness and a knowledge that others are similar to me even though they are different. When experience is made up of multiple situations (play, self-care and caring for spaces and environments, sports, hobbies, reading, lunch and more, all proposed with a character of education and learning) and not just of time spent on so-called didactic activities or, in schools for older children the hours of a specific subject, it is easier to structure relationships of reciprocity, empathy and knowledge of our classmates above and beyond the classifications that imprison us and others in a fixed identity. Indian economist Amartya Sen invites us to recognize the plurality of our identities:

> *In our normal lives, we see ourselves as members of a variety of groups– we belong to all of them. A person's citizenship, residence, geographic origin, gender, class, politics, profession, employment, food habits, sports interests, taste in music, social commitments, etc., make us members of a variety of groups. Each of these collectivities, to all of which this person simultaneously belongs, gives her a particular identity. None of them can be taken to be the person's only identity, or single membership category (Sen, 2006, pp.4-5).*

Becoming curious about the multiple identities each of us bears, marvelling at the fact we can come together and learn from each other with our differences and discovering what we have in common, are fundamental discoveries and attitudes for countering feelings of exclusion, intolerance, aggression and hate. It is important therefore, to create the conditions for schools to be able to offer what we have defined as a multiplicity for experiences, because at least until middle school [in Italy] class groups are pluralist, taking in all the children and young people in the area in which the school is located. This experience of plurality is then interrupted in secondary school, which too soon reproduces the class divisions we experience in our societies.

5. Infant-toddler centres and schools as social contexts where children learn to learn with others and from others

A fundamental element for countering a culture of hate is the discovery that others, all others, possess knowledge, abilities and strategies that I can learn from, and that together we can achieve things that a child alone would not be able to do. Children discover this as part of daily life when adults are attentive to reinforcing this aspect. That is why a productive strategy, which is central to experience in Reggio Emilia's infant-toddler centres and preschools, is that of offering opportunities for working together in small groups of three, four, five, or six children, who either choose each other or are proposed by teachers and who, faced with a shared objective, seek out and negotiate ways of achieving it.

Let's look at a short example. In the atelier, children from the four and five-year-old classroom find a paper airplane made by a teacher. Curious, the children get closer, and as they try to get it to fly, they exchange some first information with each other:

Look, a paper airplane! Giulio

Hey, if you bend the tip it won't be able to fly! Mohamed

The tip is for aerodynamics which is something that makes planes fly faster: the more aerodynamics there are, the faster it is! Giulio

The children love paper planes and would love to try and make one each, so that they could have flying contests. One of the children remembers seeing instructions in a book, and the teacher gives it to him.

© Italian e-Learning Association

We've never tried to make a paper plane in fact we need instructions like these! Giulio

Lego has instructions too, and to build a house you need instructions: everything has instructions! Edward

Everything except books, and instructions, which you don't make with other instructions. Giulio

Instructions can be a form of teaching by transmission, but trying to understand together how to read the instructions is an interesting experience. The rush some children might have to reach the objective is tempered by the slower time of the group, the time which is needed to listen to each other and make space for all.

Giulio opens out the plane that the adult has made and places it on top of the instructions, seemingly looking for confirmation of the procedure. By doing this the group discovers what the original shape of the sheet was, but the correct name emerges through their exchange:

To make a paper plane you need a square sheet of paper… Edward

A tall square… suggests Mohamed who feels the definition of square is not satisfactory

A rectangle… Giulio

Here, I've found a yellow one… Edward

At this point they begin to make their first folds, commenting:

Let's start with the tip, but it's difficult… Mohamed

You have to make another fold. Giulio

Mine won't fly… Edward

…because it's a bit too creased. Giulio

Each child is making their own little plane, but the enterprise is shared.

I've got to here, where do I go? Giulio

The drawings tell you where you have to go. Edward

How do you make this fold here? Mohamed

You have to fold it like this, turn it around! [showing him] Giulio

You do it for me then! Mohamed

No, I can't do it for you, I'm having problems too! Giulio

Following the conversation, we understand how easy loans of knowledge are, the recognition that other people can offer us solutions and help, the solidarity as all three challenge their own capabilities. The attitude is not judgmental or competitive, but the discovery of what we do not know and others do.

When the planes are ready the group prepares to take up a test flight. There is a sense of solidarity among the waiting children.

Wow! Ricki flew really low! Massimo

The wings are very high, that's why it doesn't fly high! Mohamed

Riccardo's plane is losing altitude, the wings don't go up. My wings are narrower, more aerodynamic, it's going faster, perhaps it loses less fuel! Giulio

I'll going to make another fold! Mohamed

Their tests are a reflection on choices made and the identification new strategies so that their common enterprise can be successful for everyone, regardless of the contribution each one has been able to give the group. They are an evaluation and simultaneously a self-evaluation constituting a space of possibility and a push towards new knowledge in a relational and affective context of reciprocity and constructive help.

The competition we are all driven into today, by a society in which every success and especially every failure, is the sole merit or fault of one individual, is a dynamic that schools can and must dismantle by creating conditions favourable to these processes of collaboration and reciprocal recognition. When they are in groups, together seeking to rise to the challenge of the new, children, and later teenagers, learn to tolerate the frustration that arises from not managing to do something as we would like and at the same time to be understanding when faced with less competence in others. Discovering that each of us has resources and shortcomings, strengths and weaknesses, learning to recognize them and accept them, first and foremost in ourselves, is learning we begin to construct in the earliest years of our life. It takes time to learn, but we do not go to school to be the first and fastest, but to be shaped through our relationship with others and with knowledge. We are all different and precious with our differences. Errors are necessary in learning processes.

To turn all this into educational strategy is a difficult task, but necessary at all school levels.

6. Expressing one's point of view

Democracy is that form of society in which everyone is free to express their opinion. This is a fact of extraordinary importance to which schools should give the utmost attention. But is freedom of expression an absolute value, or are there constraints beyond which it could become a negative value? How are opinions formed? Do all opinions carry the same weight? How can a well-founded opinion be distinguished from one that is unfounded?

© Italian e-Learning Association

If freedom of expression is at the heart of democratic co-living, the ways in which ideas are expressed, the tones, language, argumentation, respect for the opinions of others, and how and to what extent we are open to exchanges of point of view, to transformative dialogue, are crucial elements for its quality.

In Reggio Emilia's municipal infant-toddler centres and preschools the practice of assembly is widespread and happens almost every day, gathering the whole class of children to be counted together, share their stories, discuss themes brought in by children, tell each other about the work they are doing in the various small groups children divide into and gathering comments on it from others.

Here is what children aged five to six years have to say about this time:

We sit in a circle Daniele

In a circle everyone can sees everyone else and everyone can listen to everyone else Luca

Children talk together Jordan

Everyone can talk. Mattia

To be an assembly there have to be all the children's thoughts. Ettore

It's beautiful because you can talk and say the things you want. That is, one person says something and the others have to decide if they agree or if they disagree. Luca

When you don't agree, you can argue. Ettore

… you can discuss. Emma

Or if someone doesn't agree they can say "No, thank you". Amina

But the idea can change, the thought can change and then it resembles the thought of someone else. Ettore

In assembly we put thoughts together and practice talking. Emma

After assembly you feel light-hearted and satisfied because you have said your thoughts. Ettore

These words are representative of many other conversations on the theme and highlight how offering children a space and time to talk makes it possible for them to give shape to their ideas, in a dialogue with the ideas of others. We intuit a sense of belonging and, at the same time, a sense of responsibility toward the group to "put thoughts together"; our own, so that we can express them, and everyone's thoughts, so that we can know each other and learn from one another. In fact, there is an interesting attunement here with the words of Luigi Bobbio, when he says,

> *A public debate can be considered successful on the level of participation, not when there is a [particularly] high number of participants, but when the positions expressed include all the most significant points of view on the subject debated (Bobbio, 2010, p. 53, our translation).*

This taking practice in public debate, of taking the responsibility of making our contribution, of being attentive to how we discuss together on an issue, is an exercise that promotes not only respect but an interest, free of preconception and prejudice, towards the opinions of others, to which we can respond "no thank you" or which might lead us to change our mind, offering children not only the skills but a willingness and participatory tension we hope later school years. will valorise further.

The following conversation between children aged 5 to 6 offers an example:

Teacher: *[When we were] talking about planets the other day Stefano used the word "mortal".*

In your opinions what do "mortal" and "immortal" mean?

It's when you never die.

There's my brother's game too.

It's also a virus and you die straight away. But how can someone never die?

A game [never dies] for example.

What if it breaks?

Even when you destroy it, because it doesn't die, it's just destroyed.

Immortality can live up to thirty, a hundred years.

But there's immortality inside death, you can't say they're alive.

It's not that some people die and some live, everyone lives. They put the word "immortality" for people who aren't able to die.

You die anyway. So it should be called "immortality that makes you live". We call it living.

Marco is right, because when you die you live another life, the life of heaven. You become a star, and you are an Angel-Star.

OK, but even though the word "morte" [death] *is in it, immortality means you don't die.*

If you don't want to die, you mustn't go into danger.

We can't all live all of us. As soon as we're old, we die, because immortality gets tired.

We don't have any immortality. No one has it in Italy.

Those who are good never die.

So does that mean a child who is bad dies straight away? And what about my grandparents who were so kind?

I didn't mean to say that bad children die.

God is immortal, because seeing as created us that was a good thing, and therefore he's immortal.

© Italian e-Learning Association

If God were immortal, he would have given us, who he created, immortal powers. And he didn't have that power when he was on earth, but in heaven he did.

Immortality isn't a nice thing. They show it in films and you get so you're always old, always old isn't nice.

It's not nice because if you die you don't go to heaven and you don't see your mother again who is in heaven.

It can be a jealous thing too. Because if some people have it and others don't, then you say: you never die and I do.

Everyone dies. Even the dead die, because after you die you go to the first Paradise, then you die and go to the second, and on and on like that, until good nothingness.

Planets die too, because if a rocket comes along that makes them explode…

Planets are immortal because they're made of stone, stars are mortal because they're made of fire and they can go out.

From: Tra cielo e terra: viaggio nella spiritualità dei bambini [Between heaven and earth: a journey into the spirituality of children] (1999), publication in the Documentation and Educational Research Centre, Istituzione Preschools and Infant-toddler Centres of Reggio Emilia Municipality.

In our view this is a very interesting conversation, chosen among countless conversations collected in publications in Reggio Emilia's infant-toddler centres and preschools, and which on one hand substantiates the idea that everything can be discussed with children and, at the same shows how children, in a context of free expression, are capable of accepting and debating different ideas with great civility.

However, adults taking the role facilitators of exchange and of conducting the group, must have awareness of certain things.

While it is important at this age for even the most improbable and off-topic ideas to be expressed, it is also constructive to underscore the different ideas – if children don't already do this themselves – so that we do not have a simple declaration of everyone's beliefs but also discussion and debate. An assembly is not simply a ritual of "going around the table" in which everyone has their say and nothing more, but an exercise in reciprocal listening, of expressing agreement, of amplifying, adding and disagreeing that bring greater awareness, knowledge, changes and reinforcements of our own opinions. The adult is the first to be welcoming of all the ideas, suspending judgment, asking for deeper and more detailed explanations, and giving strength to certain ideas that might move the thinking forward, reiterating them with the group, saying for example, "She said something that seems very interesting to me. What do you think?", and where possible, drawing on lived experience and experimentations that have confirmed or disconfirmed children's beliefs and opinions, not to teach them theories that are certain, but to give value to thinking procedures that encourage children not to trust the first idea that comes to mind, and look for elements that make an argument convincing. If not the scientific method itself, then this at least is something approaching it, on which over time the children can build a capacity for evaluating the credibility of information. In fact today hate speech is also linked to the discrediting of science and its findings.

The paradigm of resonance proposed by German sociologist Hartmut Rosa, could in our view be useful for further qualifying the space of the assembly and schools as spaces of life, a criterion for evaluating quality rather than competences in schools. Resonance is the opposite of indifference and its meaning is that of "making something one's own to the point that it not only belongs to me, but affects me on an existential level, and is even capable, in principle, of transforming me" (Rosa, 2020, p. 60, our translation). Schools as spaces of resonance are schools in which "the spatial aspects of the school building and the way individuals move within them also play an important role. Resonance is always a physical phenomenon. It is already recognisable in stances, encounters and interactions in classrooms and staff rooms" (ivi, p. 65, our translation).

This introduces two more central aspects which are central in the Reggio Emilia Approach and fundamental for education that counters a culture of hate: the aesthetics of spaces and environments and the relationships between adults.

7. The aesthetics of the environment

In the words of Loris Malaguzzi,

> *It is the 'aesthetic vibration' that pushes us to name things, to name figures and colours that seemed not to exist, and to better the constructs of our interpretative and creative sensibility, discover its value and the pleasurable effects we generate within ourselves and others: a further "boldness" for seducing and being seduced. In this way the seduction of aesthetics is revealed to be a possible dimension of knowledge, illuminating something which perhaps does not belong to rationality alone. A dimension we also feel when speaking, reading, admiring an image, discovering a concept, listening to a symphony. Malaguzzi, 2012 in Edwards, Gandini & Forman, 2012, pp. 87, our translation).*

While we believe beauty to be an important quality of places, when we propose the term aesthetics in relation to spaces and environment, we mean the importance

© Italian e-Learning Association

and possibility of taking pleasure in feeling, perceiving, admiring, and knowing through their body that the spaces we design in our infant-toddler centres and preschools offer to those who inhabit them. In fact, "Individuals do not limit themselves to relating with the external world in an objective way, from a third-person perspective, but literally inscribe themselves in the world, in that their body constitutes an integral part of it and, at least partly, [also] constitutes the origin of it" (Gallese & Morelli, 2024, p. 232, our translation).

The perceptual and sensory qualities of environments where children grow and learn, discovering people and the world, the emotions and the pleasure of use that they effect, are important elements for forming and nourishing a culture of care towards public spaces which with time, all of us can experience and enact as our own responsibility towards increasingly broader contexts; the whole school, our city, our country, the Internet, and the planet. We think in terms of open spaces that invite relations rather than insularity, to encounter rather than avoidance, to respect rather than insult, that is welcoming and encourages us to enter, linger, explore, that is open to holding the traces of children's thinking and actions, and which frees both the children and the environment itself from anonymity and indifference.

This idea of open environments participating in and producing a "pedagogy of relations", has been given coherent shape in Reggio Emilia's infant-toddler centres and preschools through an intense dialogue between pedagogy and architecture.

In the words of Malaguzzi, "the system of relations enacted in s school is a system that is real and physical, a living organism and, at the same time, a symbolic system, a system of reciprocal representations" (Malaguzzi, 2012 in Edwards, Gandini & Forman, 2012, pp. 77, our translation). One of the primary elements creating favourable conditions for relations is the "transparency", realised by opening up walls as much as possible with child-height "windows" that let them see beyond. It is possible to see what children in other classes are doing or the cooks, the garden and all its variety, follow parents who have accompanied us and are leaving with our gaze or glimpse them before they cross the threshold of the infant-toddler centre or school. This allows children to be aware that, beyond the wall that delimits a space, there is a world that continues to exist and live. Transparency aims to counteract what Edgar Morin (2000, our translation) defines as "a weakening of global perception" which leads to a weakening of a sense of responsibility and solidarity, of promoting the feeling we are part of a whole and not only our own group, of relating things that are separate, creating connections, and of generating a broad sense participation and co-responsibility in adults and children.

These values are amplified by the accessibility of spaces, facilitated both by the removal of architectural barriers and by material and immaterial open doors, which increase the freedom of walking through spaces, pausing in them, choosing them and meetings between children and between adults, both teachers and parents, all through the infant toddler and preschool spaces.

Finally, we have the presence of a place to which all rooms in a school lead us, the "piazza" or square (again, a change of vocabulary and language gave new meanings) which represents the idea and value of democratic exchange and debate in our spatial architectures, because, as Italo Calvino says, "every time you enter the square, you find yourself caught in a dialogue" (Calvino, 2010, p.72).

A final element of an aesthetic of spaces and environments is the integration between analogue and digital technologies, shaping an ecosystem that reproduces the context we all live in today in the infant-toddler centres and preschools, making it available to the curious gaze and reinterpretation of children who are great creators of builders of images, representations and performance. Having the digital world in dialogue with the analogue world, with simple, easy-to-use devices, makes it possible to construct immersive contexts through projections of images, photographs, drawings, paintings and transparent construction work, inside of which children can move, imagine, construct, and stage their imaginaries in a world at once physical and virtual and share them, or perhaps relive and remember their past experiences. These are opportunities that make it possible for them to discover dynamics and possibilities that often are inscrutable, concealed behind speed and ease of use, and leaving us little opportunity for steering a game or changing its rules. However, in environments conceived in this way – between digital and analogue – children can explore and learn about digital tools from a multiplicity of points of view, which often at home are reduced to the individual use of a screen. And so, we see the creation of what we defined earlier in this article as "shared meanings among children in the form of rituals, games, and linguistic exchanges" (Decree of the Italian Ministry of Education 22 November 2021, no. 334: Adoption of "Pedagogical Guidelines for the Integrated 0-6 System"). With them, adults can observe how new generations move and interact in these ecosystems, contributing to the discovery of new forms of coexistence that more than ever before require new rules and new constraints, work it is of crucial importance to also realise in our secondary schools.

We believe that these aesthetic forms of coexistence proposed in school spaces and environments, can also contribute to forming a mental *habitus* that can resists alphabets of hate.

© Italian e-Learning Association

8. The relationship between adults

Relationships between adults are often hardly considered in schools. Each teacher is responsible for their own class or classes, their own timetable and the subject they teach: relationships between colleagues are governed by collegial bodies or forms of organization for parallel classes and subject groups that schools establish. How do children perceive the relationship between adults? What values are conveyed to them? What idea is conveyed to them of the world of relationships?

In Reggio Emilia's infant-toddler centres and preschools the relational pedagogy is a reference point that also informs the work of adults. Co-teaching for most of the day between colleagues of equal status, with two teachers per class in preschool and three teachers in infant-toddler centres, is the first nucleus of a relationship that offers children a variety of adult reference points with their differences manifest and, hopefully, valorised in contexts of sharing. Teachers talk with each other, exchange points of view, make shared decisions, help each other, and support each other in times of difficulty. They do not always necessarily agree on everything but they know how to find mediations that make it possible for them to continue working side by side every day.

Secondly, the staff working in each school, with their different roles and responsibilities, form a single working group called the collective. We have spoken of shifts in language and how they have been important in constructing a new idea of the child. The idea of a collective in each school (and Reggio has been very attentive to this) leads to a different conceptualization of relationships between those who work in an educational institution. Whatever their function or task, each person has an educational role that contributes to shaping a holistic project, asked to design situations that will also involve children in activities in the kitchen or in caring for spaces and environments, and to take responsibility when needed for intervention with children in spaces without the constant direct supervision of the teachers, in ways which are coherent with the pedagogical culture of the school. This diffused presence of adults, as we mentioned earlier, conveys to children a sense of the school as an open place in which every space can be inhabited and moved through. In a collective, adults learn to become a group that knows how to discuss together, and in which each person can concretely try out their own capacity for valuing differences of opinion and action. The collective is the context in which, while reasoning on a theme, we can and should also reflect on how to approach it together, how each of us feels when faced with different ideas that they do not share, how and if we can converge together on educational opinions, strategies and approaches. This dual gaze helps each one of us to "see" beliefs we were not explicitly aware of but which shape our actions. This is of great value for children and for adolescents, both in terms of the constant reflection and vigilance adults are encouraged to practice in relation to their actions, and because it dismantles preconceived classifications of professions, in that all of them are presented and experienced through the important value they have of offering quality daily life.

The other important adults at schools are parents. We are aware that teachers often experience this as a conflictual relationship, feeling themselves to be the target of demands – if not threats – aimed at defending a child whatever their behaviour, efforts with schoolwork or difficulties that stand in the way of reaching established objectives. These problems are real but they are further fuelled by the individualist and personalist vision view of relationships schools promote.

A school is a community; each individual's well-being is dependent on the well-being of all. There is no such thing as a problem in a class that doesn't derive from the relationship between individual and context, and therefore involving everyone. In the relationships with families this is completely lost, reinforcing a culture of competition, of seeing failure as an individual's fault, of identifying and possibly ousting the culprit (those slowing down curriculum or disrupting). Clearly schools cannot follow these cultures, and on the strength of their pedagogical processing (which is always cultural, social and political) have a duty to offer regular meetings with the parents of a class, building a group that is capable of discussing together, taking a view of the complexity of the whole, and not just their own child.

Shifting perspective, learning to learn from others, challenging and questioning one's own certainties, lets parents have a first-hand understanding of some of the difficulties children and teenagers have and to accompany them with greater sensitivity in the difficult and exciting journey of growing up; at the same time offering children and young people an aesthetic of relationships that will be an important reference for them in the future,

To reflect on these affirmations, we include a short excerpt from a conversation, one of many collected during meetings with parents in Reggio Emilia preschools.

The meeting began with teachers presenting ideas that have been gathered with children, during a morning assembly on the theme of war:

"I see you address certain topics at school, we at home and you here at school have different roles. You work on groups, on the relationship with others, on differences, so for me it's very important that [these issues] *are addressed here, because this is the world*

© Italian e-Learning Association

we're immersed in. Then, at home we all re-process things based on our subjectivities."

I really agree about "not telling lies", obviously with respect for each family's lived experience. I think dialogue about feelings is very important."

With Marco I've always had the habit of watching the news before leaving the house... Sometimes he says he wants to turn it off because he doesn't want to see certain things. This is already a growing, a detaching.

Gaia hasn't talked about the war, but I understood you had talked at school because when she saw scenes on the news, she asked if there were any dead, because she saw bodies on the ground.

Talking about war or letting them see is the same thing, it depends whether you explain it to them.

Words and images have a different weight, especially if they are not contextualized… I wouldn't be able to let Elisabeth watch the news.

It depends if I have time to dedicate to certain topics, to the questions that can come later. If it's a rushed evening I don't, so that he's not left with doubts we haven't been able to discuss together.

I try not to let him watch the news, I try to explain it to him myself, to find the right words, but the fact that children are dying is difficult.

As an adult this situation really affects you; you don't know how to deal with it or talk about it. I don't address the issue because they're still little.

They have the right to be carefree now.

How to talk about peace without explaining war to them? And how to explain war without traumatizing them?

The questions are heavy going for us.

Often what holds us back is a fear of making a mistake when we give an answer or address certain issues, in trying to discuss them, talking about them together is very hard work.

Sometimes when you question them [the children] *they stop, perhaps to tell you the reply you expect them to make.*

What I'd like to do as a mother is respect my son.

9. Conclusions

Countering a culture of hate requires new paradigms and new words. Through the experience of Reggio Emilia's municipal infant-toddler centres and preschools, and by including some documents, we have tried to argue which, in our opinion, are the most relevant.

Has this educational experience produced results on this level?

Looking at the research conducted by economist and Nobel laureate James Heckman ["Investing in Childhood", research project with the University of Chicago, the Jacobs Foundation (Zurich), and the CHILD Centre at the University of Turin] whose findings, he has repeatedly stated, cannot be considered definitive given the complexity of the subject and the several variables at play, we notice that in a comparison between adults who attended schools following the Reggio Emilia approach and those who did not participate at all in any formal program of preschool education, statistically significant positive effects were identified in the area of socio-emotional competence, educational qualifications, job type, and levels of voter turnout. Despite the difficulty of processing data this way, the findings appear to support this approach to education as a valuable tool for countering a culture of hate and building new alphabets.

The belief is also supported by Howard Gardner, who is highly knowledgeable about Reggio Emilia's municipal infant-toddler centres and preschools and who has written:

> *The attainment of an ethical mind is easier when one has been raised in a milieu where good work is the norm […] My own favourite contemporary example is the picturesque northern Italian city of Reggio Emilia, a community that I have visited and studied for twenty-five years. From everything that I have observed over the years, Reggio Emilia works exceedingly well. The community is civilised, offers high-quality services to its citizens and is replete with artistic performance sand treasures. For the past several decades this community of somewhat more than a hundred thousand individuals has devoted un paralleled human and financial resources to the development of quality infant-toddler centres and preschools. In 1991, Newsweek magazine dubbed these Reggio institutions "the best schools in the world". When visitors inquire what happens to the graduates of these schools for the young, longtime residents issue this short but revealing answer: "Just look at our community." […] Reggio Emilia would not have achieved distinction in education in the absence of committed individuals who in the aftermath of the devastation wrought by World War II, banded together to create the kind of community in which they and their children could thrive. They asked, in effect, what kinds of citizens do we want to produce?(Gardner, 2006, p. 131).*

References

Bobbio, L. (2010). *Democrazia e nuove forme di partecipazione*. In M. Bovero & V. Pazè (Eds.), *La democrazia in nove lezioni*. Laterza, Roma-Bari.

© Italian e-Learning Association

Calvino, I. (2010). *Invisible Cities*. Random House, New York.

Edwards, C., Gandini, L., & Forman, G. (Eds.) (2012). *The Hundred Languages of Children*. Praeger, Santa Barbara, CA.

Gallese, V., & Morelli, U. (2024). *Cosa significa essere umani? Corpo, cervello e relazione per vivere nel presente*. Raffaello Cortina Editore, Milano

Gardner, H. (2006). *Five Minds for the Future*. Harvard Business Press, Boston.

Hartmut R. (2013). *Social Acceleration: A New Theory of Modernity*. Columbia University Press, New York Chichester, West Sussex.

Hawkins, J., & Dawkins, R. (2021). *A Thousand Brains: A New Theory of Intelligence*. Basic Books, New York.

Malaguzzi, L., et al. (1996). *The hundred languages of children*. Catalogue of the Exhibition, Reggio Children, Reggio Emilia.

Morin, E. (2000). *La testa ben fatta. Riforma dell'insegnamento e riforma del pensiero*. Raffaello Cortina Editore, Milano.

Rosa, H. (2020). *Pedagogia della risonanza: conversazione con Wolfgang Endres*. Scholé, Brescia.

Sen, A. (2006). *Identity and Violence: The Illusion of Destiny*. Norton & Co., New York and London.

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026)

INVITED PAPER

Learning from violence. Hate Literacy as a core competence for contemporary citizenship

Mario Pireddu[a,1]

[a]*University of Tuscia, Department of Humanities. Languages, Literature and Cultural Heritage - Viterbo (Italy)*

(published: 23/5/2026)

Abstract

Public debates on hate speech are often framed in terms of regulation, moderation, or prevention, positioning hostile content primarily as a pathological deviation to be removed from digital environments. This article proposes a different perspective, arguing that hate speech can be approached as a critical object of analysis for contemporary citizenship education. Building on the concept of hate literacy, the paper conceptualizes hostile online discourses as pedagogically relevant artifacts that both reflect and actively shape models of citizenship, participation, and belonging. Rather than interpreting online hostility as an automatic outcome of digital technologies, the article situates hate speech within a dynamic interplay between intentional political actors, platform infrastructures, and hegemonic cultural narratives. From this perspective, hate speech functions as a form of informal civic education, contributing to a hidden curriculum through which norms, hierarchies, and exclusions are learned and normalized. The paper outlines hate literacy as a core civic competence, understood as the ability to critically read, contextualize, and deconstruct hostile discourses by examining their discursive, technological, and political dimensions. By reframing violence and hostility as lenses through which power relations and civic subjectivities can be analyzed, the article advances a pedagogical framework that moves beyond moral condemnation toward critical engagement. The contribution concludes by discussing the implications of this approach for citizenship education in platformed societies.

KEYWORDS: Digital Citizenship, Hate Speech, Literacy, Media, Learning.

DOI
https://doi.org/10.20368/1971-8829/1136377

CITE AS
Pireddu, M. (2026). Learning from violence. Hate Literacy as a core competence for contemporary [Invited paper]. *Journal of e-Learning and Knowledge Society*, *22*(1). https://doi.org/10.20368/1971-8829/1136377

1. Introduction

In recent years, hate speech has become a central concern in public debate, policy agendas, and academic research on digital media. Across institutional and scholarly contexts, the phenomenon is predominantly framed as a problem of regulation, moderation, and prevention: harmful content is to be identified, classified, and removed in order to protect individuals and safeguard public discourse. This orientation is reinforced by the absence of a universally shared legal definition of hate speech under international human rights law, which has encouraged the development of operational frameworks focused on risk assessment, severity, and governance mechanisms (United Nations, 2023; ARTICLE 19 2015; Gagliardone et al., 2015).

While these approaches are necessary, they tend to construct hate speech primarily as a pathological deviation within otherwise healthy communication environments. In doing so, they risk obscuring its broader cultural, political, and pedagogical dimensions. A growing body of research in media studies, platform

[1]Corresponding author - email: mario.pireddu@unitus.it – address: via S. Carlo, 32 - 01100 Viterbo (Italy).

© Italian e-Learning Association

studies, and critical discourse analysis suggests that hostile online discourse cannot be fully understood as an anomaly or merely as "bad content" to be removed. Rather, it emerges from the interaction between platform infrastructures, socio-cultural imaginaries, and strategic political actors, contributing to the shaping of digital publics and forms of participation (Scharfbillig et al., 2026; Castaño-Pulgarín et al., 2021; Gillespie, 2018; Klonick, 2018; Matamoros-Fernández, 2017).

From this perspective, hate speech is not only an object of governance but also a socio-discursive practice through which meanings, identities, and hierarchies are produced and circulated. Critical traditions have long emphasized that the force of harmful speech cannot be reduced to individual intention or isolated utterances, but must be situated within broader relations of power, iterability, and social recognition (Butler, 1997; Waldron, 2014). These insights become even more relevant in platformed environments, where visibility, amplification, and algorithmic mediation reshape the conditions under which discourse is encountered, interpreted, and normalized (UNESCO, 2021; Van Dijck, 2018; Varnelis, 2008).

Building on these premises, this article proposes a shift in perspective: instead of treating hate speech exclusively as a problem to be eliminated, it can be approached as a critical object for contemporary citizenship education. In digital environments, exposure to hostile discourse is not exceptional but structurally embedded in everyday communicative practices, especially among younger users (Hawdon et al., 2017; Wachs et al., 2021). As such, hate speech participates in what can be described as an informal or hidden curriculum, through which individuals learn – often implicitly – what counts as legitimate participation, who belongs, and which forms of exclusion are normalized.

This article develops the concept of *hate literacy* as a core competence for citizenship in platformed societies. Drawing on research in critical media literacy and cyberhate education, hate literacy is understood as a compound capability that includes recognizing hostile discourse across contexts, interpreting it as a socio-technical and political phenomenon, and responding through informed and situated civic practices (Ranieri & Fabbro, 2016; Blaya, 2019; Obermaier et al., 2025). Crucially, this perspective moves beyond a purely defensive or moralizing approach, emphasizing instead the development of interpretive, analytical, and action-oriented competencies.

The argument unfolds as follows. First, the article critically examines dominant approaches to hate speech centered on regulation and moderation, highlighting their limits in addressing the cultural and pedagogical dimensions of the phenomenon. It then situates hate speech within a mediological framework that accounts for the interplay between discourse, platform infrastructures, and political dynamics. Building on this, the paper conceptualizes hate speech as part of a hidden curriculum of platformed societies and introduces hate literacy as a key civic competence. Finally, it outlines a pedagogical framework grounded in the idea of "learning from violence", discussing its implications for citizenship education in contemporary digital environments.

By reframing hate speech as an object of critical engagement rather than solely of prohibition, this contribution aims to expand the scope of media and citizenship education. In doing so, it aligns with broader efforts to understand digital literacy not only as a set of technical skills, but as a situated and critical practice embedded in complex socio-technical systems and power relations.

2. Beyond Regulation: Limits of Prevailing Approaches to Hate Speech

Contemporary approaches to hate speech are largely structured around the need to regulate, moderate, and contain harmful content within digital environments. This orientation reflects both normative concerns – grounded in the protection of dignity, equality, and social cohesion – and practical challenges related to the scale, speed, and global reach of online communication (O'Regan, 2018; United Nations, 2023). In this context, platforms have progressively assumed the role of key governors of online speech, developing policies, enforcement mechanisms, and technical infrastructures aimed at identifying and managing problematic content (Gillespie, 2018; Klonick, 2018).

Within this dominant paradigm, hate speech is primarily framed as a category of content to be classified and acted upon – whether through removal, demotion, or other forms of visibility control. The increasing sophistication of moderation systems, including automated detection, "reduction" strategies and the policy changes of the different platforms, reinforces a model in which governance is understood as a problem of classification at scale (Douek, 2022; Gillespie, 2022; Roberts, 2019). However, as several strands of research have highlighted, this framing tends to privilege a content-centric and reactive logic, focusing on individual posts or utterances while underplaying the broader socio-cultural and discursive dynamics in which hate speech is embedded. The traditional paradigm of treating hate speech as a localized anomaly – manageable through content moderation, filtering, and account bans – relies on the flawed premise of a stable, fundamentally civil communicative ecosystem. However, the attention economy has decoupled the value of digital discourse from its constructive social function, replacing civic deliberation with engagement criteria such as outrage, polarization, and identity-driven conflict. In this

© Italian e-Learning Association

landscape, respectful debate and hostile rhetoric do not compete on equal ethical grounds; rather, they coexist within the same informational flow and are evaluated solely by their virality and capacity to validate in-group expectations. Hate speech must therefore be reconceptualized not merely as a collection of offensive utterances to be individually removed, but as a property of a socio-technical ecosystem that structurally rewards affective polarization and divisive engagement over social cohesion.

A first limitation concerns the persistent instability of definitions. The absence of a universally accepted legal or conceptual definition of hate speech has led to the proliferation of operational categories that vary across jurisdictions, platforms, and research traditions (ARTICLE 19, 2015; Elliott et al., 2016). While such flexibility allows for context-sensitive interpretation, it also creates significant challenges for both governance and research, often encouraging reductive approaches based on surface features or predefined taxonomies. In computational contexts, for instance, the need for scalable solutions has led to the development of detection systems that rely on annotated datasets and linguistic proxies, which may conflate distinct phenomena such as offensiveness, toxicity, and hate speech (Fortuna & Nunes, 2018).

A second limitation lies in the tendency to individualize and decontextualize harmful discourse. By focusing on discrete acts of expression, regulatory frameworks risk overlooking the ways in which hate speech operates as part of broader discursive formations and cultural repertoires. Critical discourse and cultural studies have shown that hostility is often implicit, coded, and embedded in narratives, humor, or seemingly neutral framings, making it difficult to capture through formal definitions or keyword-based approaches (KhosraviNik & Esposito, 2018; Matamoros-Fernández & Farkas, 2021). Moreover, such discourses are not merely expressive but performative: they contribute to the construction of social boundaries, the normalization of exclusion, and the reproduction of power relations.

A third issue concerns the infrastructural dimension of platform governance. Moderation is not limited to the evaluation of content after publication, but is deeply embedded in the design of platforms, including ranking algorithms, recommendation systems, and interface affordances that shape visibility and interaction (Gorwa, 2019; Helberger et al., 2018). As a result, the circulation and impact of hate speech cannot be understood independently of these socio-technical conditions. Interventions such as demotion or quarantine illustrate how governance increasingly operates through the modulation of visibility rather than through binary decisions of removal (Copland, 2020). Yet, these dynamics often remain opaque to users and underexplored in educational frameworks.

Finally, prevailing approaches tend to position users primarily as subjects to be protected or regulated, rather than as active participants in the construction and transformation of digital publics. While important research has addressed exposure, victimization, and risk factors – especially among younger populations (Hawdon et al., 2017; Wachs et al., 2021) – less attention has been paid to the ways in which individuals learn to interpret, reproduce, or contest hostile discourse within everyday media practices. Even when educational interventions are considered, they are often framed in preventive terms, focusing on awareness and risk reduction rather than on the development of critical and participatory competences (Blaya, 2019).

Taken together, these limitations suggest the need to complement regulatory and technical approaches with a broader analytical and pedagogical perspective. Rather than treating hate speech solely as a governance problem, it becomes necessary to understand it as a complex socio-technical and cultural phenomenon that both reflects and shapes contemporary forms of citizenship. This shift does not imply abandoning regulation, but re-situating it within a wider framework that accounts for meaning-making processes, power relations, and the informal learning dynamics embedded in digital environments. In this sense, moving "beyond regulation" entails not a rejection of governance, but an expansion of the analytical lens: from content to discourse, from isolated acts to mediated practices, and from protection to education.

3. Hate Speech as a Mediated Phenomenon

Moving beyond a purely regulatory understanding of hate speech requires a shift from a content-centered perspective to a relational and mediological one. Rather than being treated as an isolated category of harmful expression, hate speech can be more productively understood as a mediated phenomenon emerging from the interaction between discursive practices, technological infrastructures, and socio-political dynamics.

This perspective challenges deterministic interpretations that implicitly attribute online hostility to the affordances of digital media alone. While features such as anonymity, scalability, and algorithmic amplification undoubtedly shape the conditions of interaction, they do not in themselves generate hate speech. Instead, research in platform and media studies highlights how hostile discourse is co-produced by the interplay of platform design, user practices, and broader cultural and political narratives (Gillespie, 2018; Matamoros-Fernández, 2017). In this sense, platforms do not merely host hate speech but actively participate in structuring its visibility, circulation, and legitimacy.

A key implication of this shift is the need to conceptualize hate speech as part of broader communicative ecologies. Studies on platformed racism and misogyny, for instance, show how

© Italian e-Learning Association

discriminatory discourse is embedded in specific vernaculars, memetic cultures, and affective dynamics that make hostility both recognizable and, at times, socially acceptable within particular communities (De Blasio & Selva, 2024; Walther & Rice, 2024; Bates, 2022; Harrington, 2022; Rachman, 2022; Pacilli, 2020; Bailey, 2021; Sugiura, 2021; EU Commission, 2021; Marwick & Caplan, 2018; Davidson et al., 2017; Massanari, 2017; Jane, 2016; Hochschild, 2016; Wodak, 2015; Waldron, 2014; Hearn, 1998; Gilligan, 1996; Matsuda et al., 1993). These forms of expression often rely on irony, ambiguity, and coded language, complicating attempts at straightforward classification and revealing the limits of approaches based on explicit markers of hate. At the same time, the mediated nature of hate speech must be understood in relation to political communication and processes of collective identity construction. Hostile discourse frequently operates as a tool for boundary-making, contributing to the definition of in-groups and out-groups and to the normalization of exclusionary narratives. As discourse-historical approaches have shown, such dynamics are not confined to marginal spaces but can be integrated into mainstream political communication, where they are reframed, legitimized, and circulated across media systems (Farci, 2025; Wodak, 2015). In platformed environments, these processes are further intensified by the logics of engagement and visibility that reward polarizing and emotionally charged content. Importantly, this mediated perspective also foregrounds the infrastructural dimension of discourse. The circulation of hate speech is shaped not only by what is said, but by how content is ranked, recommended, and connected across networks. Research on platform governance and visibility modulation demonstrates that contemporary forms of control increasingly operate through subtle adjustments in exposure rather than through outright removal (Gillespie, 2022; Gorwa, 2019). As a result, the boundaries between presence and absence, visibility and invisibility, become central to understanding how hate speech gains traction and influence within digital publics. Furthermore, the dynamics of online hate cannot be fully captured without considering their systemic and networked character. Large-scale analyses describe online hate as an adaptive ecology, capable of reorganizing across platforms and communities in response to interventions, rather than being simply eliminated (Johnson et al., 2019). This suggests that governance strategies targeting individual platforms or pieces of content may have limited effects if they do not account for the broader circulation patterns and socio-technical interdependencies that sustain hostile discourse.

Taken together, these perspectives point to the need for a conceptual shift: from viewing hate speech as discrete content to be managed, to understanding it as a dynamic and situated practice embedded in complex media environments. Such a shift has significant implications not only for research, but also for education. If hate speech is mediated, contextual, and relational, then engaging with it requires competences that go beyond recognition and avoidance, encompassing the ability to analyze infrastructures, interpret discursive strategies, and situate communication within broader socio-political processes.

4. Hate Speech and the Hidden Curriculum of Platformed Societies

If hate speech is understood as a mediated and socially embedded phenomenon, a further step becomes possible: recognizing its pedagogical dimension. Rather than being confined to moments of disruption or deviance, hostile discourse can be seen as part of a broader set of informal learning processes through which individuals encounter, interpret, and internalize norms of participation, belonging, and exclusion within digital environments.

This perspective resonates with the notion of *hidden curriculum*, traditionally used in educational research to describe the implicit transmission of values, norms, and social hierarchies beyond formal instruction. In platformed societies, such processes are not limited to institutional settings like schools, but are continuously enacted through everyday media practices. Exposure to online hate – whether as targets, bystanders, or participants – contributes to shaping expectations about what can be said, who can speak, and how conflicts are articulated in public discourse (Hawdon et al., 2017; Wachs et al., 2021).

From this standpoint, hate speech does not simply reflect existing social tensions; it actively participates in the production and normalization of civic imaginaries. Discursive practices of exclusion, dehumanization, and boundary-making function as implicit lessons about social order, teaching audiences to recognize and reproduce distinctions between legitimate and illegitimate members of a community. As critical discourse studies have shown, these processes often operate through repetition, framing, and interdiscursivity, making hostility appear as common sense rather than as explicit aggression (KhosraviNik & Esposito, 2018; Wodak, 2015).

The pedagogical force of hate speech is further amplified by platform dynamics. Visibility mechanisms – such as trending topics, recommendation systems, and engagement metrics – do not merely distribute content but also signal its relevance and legitimacy. In this sense, what gains visibility is not only seen, but also implicitly validated as worthy of attention. Research on platform governance has highlighted how these mechanisms can contribute to the mainstreaming of controversial or polarizing discourse, blurring the boundaries between marginal and dominant narratives (Gillespie, 2018; Helberger et

© Italian e-Learning Association

al., 2018). Importantly, the hidden curriculum of hate speech is not unidirectional or deterministic. Digital environments are also spaces of contestation, where hostile discourse can be challenged, reframed, or resisted through various forms of civic engagement. Practices such as counterspeech, solidarity actions, and collective reporting illustrate how users do not simply absorb norms, but can actively negotiate and transform them (Jane, 2016). However, these practices themselves are learned and conditioned by the same socio-technical environments in which hate circulates, reinforcing the need to understand learning as an ongoing, situated process.

This framing has significant implications for how hate speech is approached within educational contexts. If hostile discourse operates as part of a hidden curriculum, then ignoring or merely suppressing it risks leaving its implicit lessons unexamined and unchallenged. Educational responses that focus exclusively on prohibition or moral condemnation may fail to address the underlying processes through which meanings and norms are constructed and internalized. Instead, engaging with hate speech as a pedagogical object requires making its implicit curriculum visible. This involves analyzing how discourses are constructed, how they circulate within platform infrastructures, and how they contribute to shaping civic subjectivities. In this sense, the goal is not to legitimize or normalize hate, but to render it intelligible as a site of learning – albeit a problematic and often harmful one – through which power relations and forms of citizenship are enacted.

5. Conceptualizing Hate Literacy

The shift from viewing hate speech as an object of regulation to understanding it as a pedagogically relevant phenomenon calls for a corresponding reconfiguration of the competencies required for contemporary citizenship. Within this framework, it is possible to work on the concept of hate literacy as a core civic competence for engaging with hostile discourse in platformed societies. The notion of hate literacy builds on and extends existing traditions of media literacy and critical digital literacy, which have long emphasized the importance of analyzing media representations, understanding power relations, and fostering active participation (De Blasio & Selva, 2024; Ranieri & Fabbro, 2016). However, while these frameworks provide a crucial foundation, they do not always explicitly address the specific challenges posed by hostile, exclusionary, and harmful forms of communication that are increasingly pervasive in digital environments. Hate literacy is proposed here as a more focused construct, oriented toward the interpretation and civic handling of such discourses.

Rather than being conceived as a purely cognitive or technical skill, hate literacy can be understood as a compound and situated capability that integrates multiple dimensions.

First, it involves a *discursive competence*, that is, the ability to recognize and interpret hate speech beyond its most explicit forms. This includes identifying implicit, coded, or ironic expressions of hostility, as well as analyzing the rhetorical and argumentative strategies through which exclusion and dehumanization are constructed (Matamoros-Fernández & Farkas, 2021; KhosraviNik & Esposito, 2018; Costanza-Chock, 2020). Second, hate literacy requires a *socio-technical and infrastructural awareness*. As discussed in previous sections, the visibility and impact of hate speech are shaped by platform architectures, including algorithms, recommendation systems, and moderation practices. Understanding how these infrastructures operate - how content is amplified, demoted, or made visible - is essential for situating hostile discourse within its broader conditions of circulation (Solanes Corella & Hernández Moreno, 2025; Ward, 2024; Douek, 2022; Gillespie, 2018; Davidson et al., 2017). Third, hate literacy encompasses a *political and cultural dimension*, involving the capacity to contextualize hate speech within wider narratives of identity, power, and belonging. Hostile discourse is often embedded in ideological frameworks that normalize exclusion or mobilize affective responses, particularly in relation to race, gender, religion, or nationality. Interpreting these dynamics requires an understanding of how discourse contributes to the construction of civic boundaries and collective imaginaries (Noble, 2018; Van Dijck, 2018; Wodak, 2015; Varnelis, 2008). Finally, hate literacy includes a *civic and action-oriented component*. Beyond recognition and analysis, individuals need to develop repertoires of response that are informed, context-sensitive, and ethically grounded. Research on cyberhate and online participation highlights the importance of practices such as counterspeech, reporting, solidarity, and bystander intervention, which can contribute to reshaping communicative environments (Obermaier et al., 2025; Bailey, 2021; Copland, 2020; Blaya, 2019). In this sense, hate literacy is not only about understanding hate speech, but also about engaging with it as a site of civic practice.

These dimensions should not be interpreted as separate or sequential components, but as interrelated aspects of a broader competence. The ability to act effectively, for instance, depends on the capacity to interpret discourse and to understand the socio-technical conditions in which it circulates. Similarly, critical analysis without the possibility of action risks remaining abstract, while action without analysis may reproduce the very dynamics it seeks to counter. This conceptualization also requires a careful positioning with respect to existing educational approaches. Hate literacy does not

© Italian e-Learning Association

aim to replace broader media literacy frameworks, nor to introduce a new normative layer based on moral condemnation. Rather, it seeks to make explicit a domain of competence that is already implicitly required in everyday digital life but insufficiently addressed in formal education. In doing so, it shifts the focus from protection and avoidance toward critical engagement and situated judgment. At the same time, it is important to acknowledge the limits and tensions inherent in this proposal. Engaging with hate speech as an object of learning raises ethical and pedagogical challenges, including the risk of exposure to harmful content, the potential normalization of hostile discourse, and the difficulty of translating critical awareness into effective action. These issues do not invalidate the concept of hate literacy, but rather highlight the need for careful pedagogical design and contextual sensitivity.

6. Learning from Violence: A Pedagogical Framework

Reframing hate speech as a pedagogically relevant phenomenon entails a significant shift in how violence and hostility are positioned within educational processes. Within this perspective, "learning from violence" refers to a pedagogical orientation that uses instances of hostile discourse as entry points for critical inquiry. The focus is not on the reproduction of harmful content, but on its *deconstruction*: analyzing how it is constructed, how it circulates, and what kinds of social meanings and power relations it encodes. This approach aligns with traditions of critical media literacy that emphasize the importance of interrogating media texts as sites of ideological production and contestation (Rivoltella, 2020, 2001; Ranieri & Fabbro, 2016; Buckingham, 2003, 1996; Silverstone, 1999; Alvarado & Boyd-Barrett, 1992).

A first key dimension of this framework is *discursive analysis as pedagogical practice*. Learners are invited to examine concrete instances of hate speech – texts, images, memes, comments etc. – not in order to classify them as acceptable or unacceptable, but to understand how they work. This includes identifying rhetorical strategies, implicit assumptions, framing devices, and forms of intertextuality that contribute to the production of meaning. Such an approach is particularly relevant in contexts where hostility is not explicit, but embedded in irony, humor, or coded language (KhosraviNik & Esposito, 2018). A second dimension concerns *the analysis of socio-technical conditions of circulation*. Learning from violence requires situating discourse within the infrastructures that shape its visibility and impact. This involves examining how platform mechanisms – such as recommendation systems, engagement metrics, and moderation policies – affect what is seen, amplified, or marginalized. By making these processes visible, learners can develop a more nuanced understanding of how hate speech gains traction and how governance operates beyond simple removal (Gillespie, 2018; Douek, 2022). A third dimension is *the development of interpretive and contextual judgment*. As emphasized in legal and human-rights frameworks, the meaning and impact of hate speech are highly context-dependent, requiring attention to factors such as target groups, historical conditions, and communicative intent (ARTICLE 19, 2015). From an educational standpoint, this translates into the need to cultivate forms of situated judgment that resist both overgeneralization and relativism. Learners are encouraged to navigate ambiguity, assess degrees of harm, and recognize the limits of rigid categorizations (Markham, 2018; Moriggi & Pireddu, 2017). A fourth and crucial dimension involves *civic response and agency*. Learning from violence is not limited to analysis, but extends to the exploration of possible forms of intervention. Research on online hate highlights a range of practices – from counterspeech and bystander intervention to reporting and collective support – that can contribute to reshaping communicative environments (Obermaier et al., 2025; Windisch et al., 2022; Bailey, 2021; Costanza-Chock, 2020). Educational settings can function as spaces where such repertoires are not only discussed but critically evaluated, considering their effectiveness, risks, and ethical implications.

Importantly, this pedagogical framework requires careful design and reflexivity. Working with hostile content raises significant challenges, including the risk of re-exposure to harmful discourse, emotional distress, or unintended normalization. For this reason, educational interventions must be context-sensitive, scaffolded, and attentive to the positionality of learners. The goal is not to immerse students in unmediated exposure, but to create structured environments in which critical distance and collective reflection are possible.

Moreover, learning from violence implies rethinking the role of the educator. Rather than acting solely as a gatekeeper who filters or prohibits content, the educator becomes a mediator who guides interpretation, facilitates dialogue, and supports the development of critical competencies. This shift aligns with broader transformations in media education, where teaching is increasingly oriented toward navigating complexity rather than providing definitive answers. In this sense, the proposed framework does not offer a set of fixed methods, but a pedagogical orientation grounded in the assumption that understanding problematic phenomena is a necessary condition for engaging with them critically. By treating hate speech as a site of inquiry rather than solely as a target of regulation, "learning from violence" opens up new possibilities for citizenship education in platformed societies.

© Italian e-Learning Association

7. Educational Implications for Citizenship in Platformed Environments

The pedagogical framework outlined above has several implications for how citizenship education is conceptualized and practiced in platformed societies. If hate speech is understood as part of a mediated and informal learning environment, and if hate literacy is conceived as a core civic competence, then educational interventions must move beyond reactive and protection-oriented models toward more complex forms of critical engagement.

A first implication concerns the *repositioning of hate speech within educational curricula*. Rather than being addressed episodically – typically in response to specific incidents or as part of broader discussions on online risks – hostile discourse could be integrated as a structural dimension of contemporary media environments. This implies *treating hate speech not as an exceptional breakdown of communication, but as a recurring and patterned phenomenon through which social norms and hierarchies are negotiated*. Such an approach aligns with critical media literacy traditions that emphasize the analysis of representation, power, and participation as central components of citizenship education (Ranieri & Fabbro, 2016). A second implication relates to *the expansion of the scope of digital and media literacy frameworks*. Existing models often focus on skills such as information evaluation, content creation, and online safety. While these dimensions remain essential, they may not sufficiently address the interpretive and civic challenges posed by hostile discourse. Integrating hate literacy into these frameworks requires foregrounding competencies such as discursive analysis, contextual judgment, and understanding of platform governance, thereby extending literacy from a functional to a more explicitly political and cultural domain. A third implication concerns *the role of platforms as pedagogical environments*. As discussed in previous sections, platforms are not neutral channels but infrastructures that shape visibility, interaction, and participation (Maragliano & Pireddu, 2012). From an educational perspective, this means that learning about citizenship increasingly involves understanding how platform governance operates, including moderation policies, algorithmic curation, and visibility management (Markham, 2018; Gillespie, 2018; Gorwa, 2019). Developing such awareness is crucial for enabling informed participation, as users navigate environments where rules are hybrid, dynamic, and often opaque. A fourth dimension involves *the development of civic agency and repertoires of response*. Research on cyberhate and online participation highlights that individuals are not only exposed to hostile discourse but also positioned as potential actors - whether as bystanders, targets, or participants (Obermaier et al., 2025; Wachs et al., 2021). Educational contexts can support the exploration of different forms of engagement, including counterspeech, reporting mechanisms, and practices of solidarity. However, such engagement must be framed critically, taking into account not only effectiveness but also risks, power asymmetries, and the broader socio-technical conditions that shape possible outcomes.

At the same time, these implications highlight the need to *rethink pedagogical roles and environments*. Educators are increasingly required to navigate complex and potentially sensitive materials, balancing critical analysis with ethical responsibility. This involves creating learning settings that allow for the examination of problematic content without reproducing harm, as well as fostering dialogue that acknowledges diverse experiences and positionalities. In this sense, teaching becomes less about transmitting normative rules of "appropriate behavior" and more about facilitating processes of interpretation, reflection, and situated judgment. A further implication concerns *the evaluation and design of educational interventions*. As existing reviews have pointed out, the field of cyberhate education is characterized by a proliferation of initiatives but a relative scarcity of robust evidence regarding their effectiveness (Blaya, 2019; Windisch et al., 2022). Conceptualizing hate literacy as a multidimensional competence opens the possibility of developing more precise frameworks for assessment, including indicators related to interpretive skills, critical awareness of infrastructures, and forms of civic engagement. At the same time, it calls for longitudinal and context-sensitive research capable of capturing the complexity of learning processes in digital environments.

Finally, these considerations suggest that citizenship education in platformed societies cannot be confined to formal educational settings alone. Informal and non-formal contexts – online communities, peer interactions, and everyday media practices – play a central role in shaping how individuals encounter and make sense of hate speech. This reinforces the need for educational approaches that are not only school-based, but also attentive to the broader media ecologies in which learning takes place.

Taken together, these implications point toward a reconfiguration of citizenship education as a practice that engages directly with the complexities of contemporary communication environments.

Integrating hate literacy into this framework does not provide a definitive solution to the problem of online hostility, but offers a way to address it that is consistent with the realities of platformed societies: not by avoiding or simply removing conflict, but by developing the capacities required to understand, interpret, and respond to it.

© Italian e-Learning Association

8. Discussion

The framework proposed in this article - centered on the concepts of hate speech as hidden curriculum and hate literacy as a civic competence - opens up new analytical and pedagogical perspectives. At the same time, it raises a number of theoretical, methodological, and ethical tensions that need to be explicitly addressed.

A first issue concerns the risk of normalization. Reframing hate speech as an object of analysis and learning may be misinterpreted as a form of legitimation, especially in contexts where hostile discourse is already widespread and socially tolerated. The pedagogical approach outlined here attempts to mitigate this risk by emphasizing critical distance, contextualization, and reflexivity. However, the tension remains: engaging with harmful content requires a careful balance between making it visible and avoiding its reproduction as normalized or inevitable. As critical scholarship has shown, repetition and circulation are key mechanisms through which hate becomes embedded in everyday discourse (Walther & Rice, 2024; Hochschild, 2016; Hearn, 1998; Butler, 1997; KhosraviNik & Esposito, 2018).

A second tension relates to exposure and harm. Educational approaches that involve the analysis of hostile content must confront the possibility of re-exposing learners – particularly those belonging to targeted groups – to forms of symbolic violence. This raises important questions about pedagogical design, including the selection of materials, the framing of activities, and the creation of safe and inclusive learning environments. While avoiding exposure altogether may leave implicit norms unchallenged, unstructured exposure risks reinforcing existing inequalities and producing emotional or psychological distress (Farci, 2025; Pireddu, 2022; Citron, 2014).

A third area of tension concerns the translation from competence to action. The concept of hate literacy emphasizes not only interpretive skills but also civic engagement. However, the relationship between awareness, critical understanding, and effective intervention is neither linear nor guaranteed. Empirical research suggests that factors such as perceived efficacy, social norms, and platform affordances significantly influence whether individuals engage in practices such as counterspeech or reporting (Obermaier et al., 2025). This highlights the need to avoid overly optimistic assumptions about the transformative power of literacy alone, and to situate educational efforts within broader socio-technical and institutional conditions.

A fourth issue involves the role of platforms and structural constraints. While the framework foregrounds user agency and civic competence, it must be situated within environments where key aspects of communication are governed by opaque and asymmetrical infrastructures. Platform governance research has shown that visibility, moderation, and enforcement are shaped by institutional logics, economic incentives, and technical constraints that are often beyond users' control. As a result, the scope of individual and collective action is conditioned by factors that literacy alone cannot fully address.

A broader methodological challenge concerns the operationalization and assessment of hate literacy. While the concept is articulated here as a multidimensional competence, translating it into measurable indicators remains an open research problem. Existing studies provide useful starting points – particularly in linking critical digital literacy to bystander intervention – but further work is needed to develop robust, context-sensitive, and longitudinal evaluation frameworks (Blaya, 2019; Windisch et al., 2022).

These tensions do not undermine the proposed framework; rather, they define its conditions of applicability and its limits. They suggest that hate literacy should not be conceived as a self-sufficient solution, but as one component within a broader ecosystem of interventions that includes regulation, platform design, and social action. At the same time, they reinforce the central argument of this article: that understanding and critically engaging with hostile discourse is a necessary, albeit insufficient, condition for addressing the challenges it poses to contemporary citizenship.

9. Conclusion

This article has proposed a shift in the way hate speech is conceptualized within research and educational frameworks. Rather than approaching it exclusively as a problem of regulation, moderation, or prevention, the paper has argued for its consideration as a pedagogically relevant phenomenon – one that both reflects and actively shapes contemporary forms of citizenship in platformed societies.

By situating hate speech within a mediological perspective, the analysis has highlighted its embeddedness in the interplay between discursive practices, platform infrastructures, and socio-political dynamics. From this standpoint, hostile discourse emerges not as an isolated deviation, but as part of broader communicative ecologies through which norms, hierarchies, and boundaries of belonging are constructed and negotiated. In this sense, hate speech can be understood as contributing to a hidden curriculum that informs how individuals learn to participate in digital publics.

Building on this framework, the article has introduced the concept of *hate literacy* as a multidimensional civic competence. Rather than reducing literacy to the recognition or avoidance of harmful content, hate literacy has been articulated as the capacity to critically

© Italian e-Learning Association

interpret, contextualize, and respond to hostile discourse by engaging with its discursive, socio-technical, and political dimensions. This perspective expands existing approaches to media and digital literacy by foregrounding the specific challenges posed by exclusionary and harmful forms of communication.

The pedagogical proposal developed through the notion of "learning from violence" further reinforces this shift. By treating hate speech as an object of critical inquiry, rather than solely as a target of prohibition, the framework emphasizes the importance of developing interpretive judgment, infrastructural awareness, and civic agency. At the same time, the discussion has made clear that such an approach is not without risks and limitations, including issues related to normalization, exposure, and the structural constraints imposed by platform governance.

Taken together, these considerations suggest that addressing hate speech in contemporary societies requires moving beyond dichotomies such as regulation versus education, or protection versus participation. Instead, what emerges is the need for integrated approaches that recognize the complexity of digital environments and the multiple levels – discursive, technological, institutional – at which hostility is produced and managed. From a research perspective, this implies further work on the operationalization and assessment of hate literacy, particularly through longitudinal and context-sensitive studies capable of capturing its impact on civic practices. From an educational standpoint, it calls for the development of pedagogical designs that can engage with problematic content in ways that are critically rigorous, ethically grounded, and responsive to diverse learning contexts.

Ultimately, the contribution of this article lies not in offering a definitive solution to the problem of online hate, but in reframing it as a site of inquiry that is central to understanding and practicing citizenship in platformed societies. In doing so, it invites educators, researchers, and policymakers to reconsider not only how hate speech is governed, but also how it is learned, interpreted, and contested within the everyday experience of digital media.

References

Alvarado M., Boyd-Barrett O., eds. (1992). *Media education: An introduction*, London, British Film Institute, Open University Press.

ARTICLE 19 (2015). *'Hate speech' explained: a toolkit*, London, ARTICLE 19. https://bit.ly/4clMQYh

Bailey, M. (2021). *Misogynoir transformed: Black women's digital resistance*, New York, NYU Press.

Bates, L. (2022). *Men Who Hate Women: From Incels to Pickup Artists: The Truth About Extreme Misogyny and How It Affects Us All*, London, Simon & Schuster.

Blaya, C. (2019). *Cyberhate: A review and content analysis of intervention strategies*, Aggression and Violent Behavior, Volume 45, 2019.

Buckingham, D. (2003). *Media education: Literacy, learning and contemporary culture*. Cambridge, Eng.: Polity Press.

Buckingham, D. (1996). *Critical pedagogy and media education: A theory in search of a practice*. Journal of Curriculum Studies, 28, 627-650.

Butler, J. (1997), *Excitable speech: a politics of the performative*, London, Routledge.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). *Internet, social media and online hate speech. Systematic review*. Aggression and Violent Behavior, Vol. 58, 2021.

Citron, D. K. (2014), *Hate crimes in cyberspace*, Cambridge, MA, Harvard University Press.

Copland, S. (2020). *Reddit quarantined: can changing platform affordances reduce hateful material online?*, Internet Policy Review, 9(4).

Costanza-Chock, S. (2020). *Design justice: community-led practices to build the worlds we need*, Cambridge, MIT Press.

Davidson T., Warmsley D., Macy M., & Weber I. (2017). *Automated hate speech detection and the problem of offensive language*, Proceedings of the 11th International Conference on Web and Social Media, 11 (1), 512-515.

De Blasio E., & Selva D. (2024). *Gender and Culture Wars in Italy: A Genealogy of Media Representations*, Cham (CH), Palgrave Macmillan.

Douek, E. (2022). *Content moderation as systems thinking*, Harvard Law Review, Vol. 136, Issue 2.

Elliott C., Chuma W., Gendi Y.E., Marko D., & Patel A. (2016). *Hate Speech, Key concept paper*. Working Paper, MeCoDEM. https://bit.ly/4cgRQNK

European Union Commission, Radicalisation Awareness Network (2021). *Incels: A First Scan of the Phenomenon (in the EU) and its Relevance and Challenges for P/CVE*, Brussel, EU Commission.

Farci, M. (2025). *Quel che resta degli uomini. Sulla mascolinità*, Milano, Nottetempo.

Fortuna, P., & Nunes, S. (2018). *A survey on automatic detection of hate speech in text*, ACM Computing Surveys, 51 (4).

© Italian e-Learning Association

Gagliardone, I., Gal, D., Alves, T., & Martinez G. (2015). *Countering online hate speech*, Paris, UNESCO.

Gillespie, T. (2022). *Do not recommend? reduction as a form of content moderation*, Social Media + Society, Volume 8, Issue 3, 07/2022.

Gillespie, T. (2018). *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, New Haven, CT, Yale University Press

Gilligan, J. (1996). *Violence: Our Deadly Epidemic and Its Causes*, New York, G.P. Putnam's Sons.

Gorwa, R. (2019). *The platform governance triangle: conceptualising the informal regulation of online content*, Internet Policy Review, 8 (2).

Harrington, C. (2022). *Neoliberal Sexual Violence Politics: Toxic Masculinity and #MeToo*, Cham (CH), Palgrave Macmillan.

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). *Exposure to Online Hate in Four Nations: A Cross-National Consideration*. Deviant Behavior, 38(3).

Hearn, J. (1998). *The Violences of Men: How Men Talk about and How Agencies Respond to Men's Violence to Women*, London, SAGE.

Helberger, N., Pierson, J., & Poell, T. (2018), *Governing online platforms: from contested to cooperative responsibility*, The Information Society, 34 (1).

Hochschild, A.R. (2016). *Strangers in Their Own Land: Anger and Mourning on the American Right*, New York, New Press.

Jane, E. A. (2016). *Online misogyny and feminist digilantism*, Continuum: Journal of Media & Cultural Studies, 30 (3).

Johnson, N.F., Leahy, R., Restrepo, N.J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., & Wuchty, S. (2019). *Hidden resilience and adaptive dynamics of the global online hate ecology*. Nature. 2019 Sep;573.

Klonick, K. (2018). *The new governors: the people, rules, and processes governing online speech*, Harvard Law Review, Vol. 131, Issue 6.

KhosraviNik, M., & Esposito, E. (2018). *Online hate, digital discourse and critique: exploring digitally-mediated discursive practices of gender-based hostility*, Lodz Papers in Pragmatics 14.1 (2018). Special issue on Narrating hostility, challenging hostile narratives.

Maragliano, R., & Pireddu, M. (2012). *Storia e pedagogia nei media*, Roma, Garamond.

Markham, A.N. (2018). *Critical Pedagogy as a Response to Datafication*, Qualitative Inquiry, Vol. 25, Issue 8.

Marwick, A.E., & Caplan, R. (2018). *Drinking male tears: language, the manosphere, and networked harassment*, Feminist Media Studies, 18 (4), 2018, pp. 543-559.

Massanari, A. (2017). *#Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures*. New Media & Society, 19(3).

Matamoros-Fernández, A. (2017), *Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube*. Information, Communication & Society, 20(6).

Matamoros-Fernández, A., & Farkas, J. (2021), *Racism, Hate Speech, and Social Media: A Systematic Review and Critique*. Television & New Media, 22(2).

Matsuda, M. J., Lawrence, C. R. III, Delgado, R., & Crenshaw, K. W. (1993), *Words that wound: critical race theory, assaultive speech, and the First Amendment*, Boulder, Westview Press.

Moriggi, S., & Pireddu, M., (2017), *Vivere e non sapere. Fenomenologia della post-truth tra educazione e comunicazione*, The Future of Science and Ethics, Rivista scientifica del Comitato Etico della Fondazione Umberto Veronesi, Vol. 2, Num. 1, 06/2017, http://bit.ly/2wBBI8r

Noble, S. U. (2018). *Algorithms of oppression: how search engines reinforce racism*, New York, NYU Press.

Obermaier, M., Schmid, U. K., & Rieger, D. (2025), *Empowerment Is Key? How Perceived Political and Critical Digital Media Literacy Explain Direct and Indirect Bystander Intervention in Online Hate Speech*. Social Media + Society, 11(1).

O'Regan, C. (2018). *Hate speech online: an (intractable) contemporary challenge?*, Current Legal Problems, Vol. 71, Issue 1.

Pacilli, M.G.(2020). *Uomini duri. Il lato oscuro della mascolinità*, Bologna, Il Mulino.

Pireddu, M. (2022). *Architetture relazionali, embodiment, co-enaction o apprendimento nel metaverso*, Colazzo S., Maragliano R. (eds), Metaverso e realtà dell'educazione, Roma, Edizioni Studium, 2022.

Rachman, G. (2022). *The Age of the Strongman: How the Cult of the Leader Threatens Democracy Around the World*, New York, Other Press.

© Italian e-Learning Association

Ranieri, M., & Fabbro, F. (2016). *Questioning discrimination through critical media literacy. Findings from seven European countries*. European Educational Research Journal, 15(4).

Rivoltella, P.C. (2020). *Nuovi alfabeti. Educazione e culture nella società post-mediale*, Brescia, Scholè.

Rivoltella, P.C. (2001). *Media Education. Modelli, esperienze, profilo professionale*, Roma, Carocci.

Roberts, S. T. (2019). *Behind the screen: content moderation in the shadows of social media*, New Haven, CT, Yale University Press

Scharfbillig, M., Lewandowsky, S., Altay, S., Van Alstyne, M., Kozyreva, A. et al. (2026). *Fractured reality - How democracy can win the global struggle over the information space*, Publications Office of the European Union, Luxembourg, 2026, https://data.europa.eu/doi/10.2760/9358883, JRC144603.

Silverstone, R. (1999). *Why study the media?*, London, Sage.

Solanes Corella, Á., & Hernández Moreno, N. (2025). *Racism in the digital age: the impact of social media algorithms on public discourse*, The Age of Human Rights Journal, 25, e9603.

Sugiura, L.(2021). *The Incel Rebellion: The Rise of the Manosphere and the Virtual War Against Women*, Bingley, Emerald Publishing Limited.

Varnelis, K. (2008, Ed.). *Networked Publics*, Cambridge, MIT Press.

Van Dijck, J. (2018). *The Platform Society: Public Values in a Connective World*, Oxford, OUP.

Waldron, J. (2014). *The harm in hate speech*, Cambridge, Harvard University Press.

Walther, J. B., & Rice, R. E. (2024, Eds.). *Social processes of online hate*, New York, Routledge.

Ward, C. (2024). *UN Peacekeeping, OHCHR. A conceptual analysis of the overlaps and differences between hate speech, misinformation, and disinformation*, New York, United Nations.

Wachs, S., Costello, M., Wrigh, M. F., Flor, K., Daskalou V., Maziridou E., Kwon Y., Na, E.-Y., Sittichai, R., Biswal, R., Singh, R., Almendros, C., Gámez-Guadix, M., Görzig, A., & Hong, J. S. (2021). *"DNT LET 'EM H8 U!": Applying the routine activity framework to understand cyberhate victimization among adolescents across eight countries.* Computers & Education, 160, 104026.

Windisch, S., Wiedlitzka, S., Olaghere, A., & Jenaway, E. (2022), *Online interventions for reducing hate speech and cyberhate*, Campbell Systematic Reviews, Vol. 18, Issue 2.

Wodak, R. (2015). *The Politics of Fear: What Right-Wing Populist Discourses Mean*, London, SAGE Publications.

United Nations (2023). *What is hate speech?*, https://bit.ly/3OfHYfr

UNESCO (2021). *Media and information literate citizens: think critically, click wisely!*, Paris, Unesco.

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026)

INVITED PAPER

"To redeem horror from his invisibility". Forensic Architecture's work with images

Maurizio Guerri[a,1]

[a]*Scuola di Didattica dell'arte e cultura visuale. Accademia di Belle Arti di Brera – Milan (Italy)*

(published: 23/5/2026)

DOI
https://doi.org/10.20368/1971-8829/1136376

CITE AS
Guerri (2026). "To redeem horror from his invisibility". Forensic Architecture's work with images [Invited paper]. *Journal of e-Learning and Knowledge Society*, *22*(1).
https://doi.org/10.20368/1971-8829/1136376

1. Faced with the pain of others

We live in a situation that can be succinctly described as *Surveillance Capitalism* (2019), to borrow the title of Shoshana Zuboff's famous book. The technological network to which we are connected and for which we work, more or less consciously, produces a political and social asymmetry that highlights once again how technology as a whole cannot in any way be regarded as neutral, but is always politically oriented (see also Casilli, 2020; Ippolita, 2017).

We are witnessing a phenomenon that originates from the dialectical opposite of surveillance. I am referring to the relationship we have with the images of the systematic destruction of the population of Gaza. Faced with the sheer volume of images and testimonies of the annihilation of thousands of human beings in Gaza, we have witnessed – and continue to witness – a sort of blockage, a loss of speech, a paralysis of action. The images of an ongoing genocide do not provoke the ethical and political stance one might have expected, particularly in European countries and even more so in those very nations whose populations actively collaborated in the systematic elimination of European Jews and which, for decades after the Liberation from Nazi-Fascism, shaped the ethical and political conscience of generations on the values of coexistence developed precisely in the wake of that immense tragedy.

On the one hand, in recent years we have seen the Israeli army using the Lavender artificial intelligence system to monitor every aspect of life in Gaza (see Yuval, 2024), with the aim of identifying "suspected" Hamas affiliates and eliminating them as swiftly as possible without these individuals ever being tried in a court of law, as required by the principles governing a democratic state. On the other hand, we are faced with a wealth of images and testimonies showing the systematic process of annihilating the population and seizing territory, yet these do not elicit the ethical and political response one might reasonably expect. We look, but we do not see what the images are showing us.

So what happens to us when we look at those images? It is one of the questions Susan Sontag grappled with throughout her life: what happens to us when we look

[1] corresponding author - email: maurizioguerri@fadbrera.edu.it

© Italian e-Learning Association

at images of "the pain of others"? Images of bodies burned by napalm in the Vietnam War, the corpses of ethnic cleansing in the former Yugoslavia, the civilian victims of the war on terror, migrant children drowned and washed up by the waves on the shores of the Mediterranean. Sontag's detailed reflection on this kind of imagery can be summed up in a single sentence: "the aestheticizing tendency of photography is such that the medium which conveys distress ends by neutralizing it" (Sontag, 1973, p. 98). According to Sontag, the reason we are unable to see what we are looking at lies in the relationship we establish with the images reproduced by the media: whilst, on the one hand, images bear witness to the suffering of others, on the other, Sontag concludes that those images, precisely because they are images (photographs, videos etc.), neutralise the shock they have provoked in the observer. For Sontag, there are many reasons for this perceptual outcome: the proliferation of forms of entertainment which, in the consumption of images, result in the absence of a contemplative space capable of allowing reflection on what we are looking at is, for Sontag, certainly one of the main reasons for the neutralisation of the distress that a war photograph initially conveys to the viewer.

Sontag pauses to reflect on Jeff Wall's image *Dead Troops Talk (A vision after an ambush of a Red Army patrol*, *near Moqor, Afghanistan, winter 1986*). It is a large-scale image installed as a lightbox; Wall's installation is not a record of an incident that occurred, but a visual construction designed precisely to explore the very question that Sontag herself is addressing. In the installation, we see several Soviet soldiers riddled with bullets, disembowelled, with their heads split open, yet still alive after a violent attack. But the bodies of other soldiers lie lifeless on the ground. Sontag is searching for an image capable of escaping those forms of neutralisation of anguish and pain that characterise our relationship with images of war. Wall's image seems to respond to this need to oppose war, just as Turgenev's *The Execution of Troppmann* manages to oppose the death penalty. Sontag observes, regarding Wall's image: "Engulfed by the image, which is so accusatory, one could fantasize that the soldiers might turn and talk to us. But no, no one is looking out of the picture. There's no threat of protest. They are not about to yell at us to bring a halt to that abomination which is war. They haven't come back to life in order to stagger off to denounce the war-makers who sent them to kill and be killed" (Sontag, 2003, p. 125). And finally, Sontag concludes her reflection:

> *These dead are supremely uninterested in the living: in those who took their lives; in witnesses – and in us. Why should they seek our gaze? What would they have to say to us? "We" – this "we" is everyone who has never experienced anything like what they went through – don't understand. We don't get it. We truly can't imagine what it was like. We can't imagine how dreadful, how terrifying war is; and how normal it becomes. Can't understand, can't imagine (Sontag, 2003, pp. 125-26).*

According to Sontag, even Wall's image fails to establish a connection with the viewer. The dead in the scene constructed by Wall are irrevocably separated even from their comrades-in-arms who are still alive, just a few centimetres away from them. The distance separating us, who observe the dead from a safe distance, from those lying on the ground is unbridgeable, at least in the images. The suffering endured by those who have died, the terrifying and horrific consequences of war transposed into an image, do not stir the imagination of the viewer; and so we, the observers, continue to look at the photograph without being able to see what is happening and without being able to take a stand against the violence of war. War is unimaginable, Sontag seems to conclude, so its reproduction in images can only be a checkmate for us who view them.

If we follow Sontag's line of reasoning, the act of looking but not seeing by millions of citizens who are unable to take a stand in the face of the images from Gaza is, so to speak, absolved – at least on this specific issue – of their aesthetic responsibilities. We have countless images of the destruction suffered by the people of Gaza, but however terrifying they may be, they fail to strike a chord with our emotions, our imagination and our reason: "Can't understand, can't imagine" writes Sontag. Sontag seems to identify structural aspects within the images that prevent the viewer from empathising, from recognising themselves in the painful action unfolding in the scene they are watching. It is as though images of death in war always have an anaesthetic or hyperesthetic effect in terms of the relationship they establish with pain and death.

But let us try to consider the relationship between images, wars and ourselves from another perspective as well. In the 1960s, Siegfried Kracauer, in a famous passage from *Theory of Film*, interprets cinematic images by drawing on the myth of Perseus and Medusa:

> *We have learned in school the story of the Gorgon Medusa whose face […] was so horrible that the sheer sight of it turned men and beasts into stone. When Athena instigated Perseus to slay the monster, she therefore warned him never to look at the face itself but only at its mirror reflection in the polished shield she had given him. […] The moral of the myth is, of course, that we do not, and cannot, see actual horrors because they paralyze us with blinding fear; and that we shall know what they look like only by watching images of them which reproduce their true appearance. […]*

© Italian e-Learning Association

> *Now of all the existing media the cinema alone holds up a mirror to nature. Hence our dependence on it for the reflection of happenings which would petrify us were we to encounter them in real life. The film screen is Athena's polished shield. This is not all, however. In addition, the myth suggests the the images on the shield or screen are a means to an end; they are to enable – or, by extension, induce – the spectator to behead the horror they mirror (Kracauer, 1960, p. 305).*

And in the final paragraph, Kracauer concludes:

> *The mirror reflections of horror are an end in themselves. As such they beckon the spectator to take them in and thus incorporate into his memory the real face of things too dreadful to be beheld in reality. In experiencing the rows of calves 'heads or the litter of tortured human bodies in the films made of the Nazi concentration camps, we redeem horror from its invisibility behind the veils of panic and imagination. And this experience is liberating in as much as it removes a most powerful taboo. Perhaps Perseus 'greatest achievement was not to cut off Medusa's head but to overcome his fears and look at its reflection in the shield. And was it not precisely this feat which permitted him to behead the monster? (Kracauer 1960, p. 306).*

It is interesting to note that in these passages Kracauer also engages with images of the Holocaust, one of the tragedies of contemporary history which, in many respects, calls into question the very possibility of representing that event through images. As is well known, he was among the first philosophers, in the 1920s, to recognise the anaesthetic and distracting function performed primarily by images within the context of the nascent entertainment industry (see Kracauer, 2022). But this does not prevent Kracauer from recognising an opportunity in our relationship with images: the possibility that they might serve as testimony. Let us analyse Kracauer's passage: images allow the viewer to convince themselves of the possibility of decapitating the horror they reflect. Images allow us to "incorporate"into our "memory" "the real face of things too dreadful to be beheld in reality". By looking at the horror reflected in images, "redeem horror from its invisibility**".** In this sense, the experience of engaging with images can be liberating because it has the potential to remove the most powerful taboo. Finally, it is precisely through images that we can overcome fear, break free from the paralysis it imposes on us, and by looking at its reflection in the shield, we put ourselves in a position to start acting again.

This reflection on the testimonial potential inherent in images does not imply a naive and optimistic acceptance of images in their entirety. As we mentioned, it was Kracauer himself who, from the 1920s onwards, began to develop a phenomenology of entertainment images and the role of distraction. In addition to Kracauer, Walter Benjamin, Theodor W. Adorno, Max Horkheimer, Guy Debord, Jean Baudrillard and Susan Sontag have all explored the alienating function that images perform within the society of the spectacle, right up to the era of surveillance capitalism. But I believe it is necessary to continue thinking dialectically about the sphere of image production, and not to abandon the testimonial capacity that images continue to fulfil in our time. I believe that the work on images by the Forensic Architecture research group can be understood as an ideal continuation of Kracauer's insights.

Weizman describes the formation of the research group he founded as follows:

> *Forensic Architecture is also the name of a research agency I established in 2010, together with a group of fellow architects, artists, filmmakers, journalists, scientists, and lawyers. We undertake independent research or act on commission from international prosecutors and environmental and human rights groups to investigate state and corporate violence, especially when it bears upon the built environment. The agency produces evidence files that include building survey, models, animations, video analyses, and interactive cartographies, and presents them in forums such as international courts, truth commissions, citizen tribunals, human rights and environmental reports, and, on one occasion, in the UN General Assembly (Weizman, 2017, p. 9).*

For FA in the context of the use of images in forensic contexts, see Grespi & Villa (2024).

We know that many acts of violence would not exist were it not for verbal, photographic, video and other forms of evidence that demonstrate their occurrence. At the same time, we know that there is a significant asymmetry between the capacity of state institutions and large industrial groups to utilise image-producing tools, and the capabilities of individual citizens and communities who suffer the violence of state-sanctioned warfare. This asymmetry is present across all technological sectors and can be summarised by the difference in the focusing capabilities of the optical systems of drones and satellites available to governments, compared to the satellite imagery accessible to the public.

© Italian e-Learning Association

> *Unlike the randomly disturbed grains of analog photography, digital images, such as satellite images, are divided into a grid of equal square units, or pixels. This grid filters reality like a sieve or a fishing net. Objects larger than the grid are captured and retained. Smaller ones pass through and disappear. Objects close to the size of the pixel are in a special threshold condition: whether they are captured or not depends on the relative skill, or luck, of the fisherman and the fish (Weizman, 2017, p. 27).*

There are human figures and objects that hover on the threshold of detectability due to the resolution of the images documenting the event. In the 1970s, satellites transmitted images of parts of the Earth where the width of a single pixel corresponded to a distance of 60 metres. In the 1980s, linear resolution reached 30 metres and, by the early 1990s, around 20 metres. It was during the 1990s, amidst the conflict in the former Yugoslavia, that satellite imagery began to be regarded as crucial visual evidence of what was occurring in war-torn territories, precisely because of its resolution, making it possible to identify large bombed-out buildings or mass graves, for example. At the start of the 21st century, resolution reached 2.5 metres per side, and a few years later it improved further to 0.5 metres. Then this process of focusing on the planet from above came to a halt. Why? Weizman writes: the reason for halting the improvement in resolution lies in the fact that:

> *at 0.5 meters, the pixel resolution corresponds to the dimensions of the human body – an area 0.5 meters by 0.5 meters is roughly the size of the human body as seen from above. As such, the pixel could be thought of as analogous to what Le Corbusier called a "modulor" – a system of proportions and measurements that relate to the human body. The satellite images "modulor" was not meant to help organize space, but rather to remove the human figure from representation. The human body was now drowned within the pixel resolution available to independent groups to analyze human rights violations (Weizman, 2017, p. 28).*

It is important to note that the resolution limitations of satellite imagery apply only to publicly available images. The resolution of military drones is extremely high (though subject to military secrecy), whilst US Keyhole-class satellites have a resolution of around 15 centimetres, but these are inaccessible to human rights groups.

The official reasons given for this suspension of the satellite image focusing process are based on respect for individual privacy. But, as Weizman rightly points out:

> *The difference in resolution demonstrates the imbalance of power. While the human body is the scale to which drone optics are calibrated, it is the very thing that publically available satellite images are designed to mask (Weizman, 2017, p. 30).*

Images are becoming increasingly important both in the conduct of warfare and in the subsequent reconstruction of events; in modern warfare, in particular, imaging equipment is used by military personnel to track down targets, whilst at the same time governments delete those images to eliminate any evidence of unwanted deaths, effectively erasing those deaths. Since the nineteenth century, state police forces have used photography to reconstruct crime scenes; today, it is the perpetrators – that is, states or companies acting on behalf of state powers – who are equipped with more effective and powerful tools than the investigators, namely international courts and independent research groups:

> *The visual spectrum between the high resolution used for killing and the low resolution available for monitoring the killing is the space exploited by deniers. The practice of counterforensics […] has to engage a condition of structural inequality in access to vision, signals, and knowledge, and to find ways to operate close to and under the threshold of detectability (Weizman 2017, p. 30).*

FA's work consists of reconstructing, using all the technologies at our disposal – particularly those relating to the study of architectural structures, the various types of imagery associated with the event, and the analysis of eyewitness accounts – events that would otherwise not exist in the public record, neither for those who experienced them nor for those seeking to understand history: civilians killed pre-emptively or "by mistake" in undeclared wars, migrants who have drowned in the Mediterranean, men killed by police forces in violation of the laws of the very nation in which they operate. The perception among the global public is that these events *do not exist*, not because there are no images to bear witness to their existence, but because these images are systematically withheld from the public eye by the political power of states – a political power which, as we have observed in recent years, is increasingly dependent on the interests of a few powerful groups. The relationship between the current US government and Big Tech is highly significant (see Bria, 2025).

© Italian e-Learning Association

2. Counter-forensics

> *Although the name [Forensic Architecture] refers to the work of building surveyors, our agency is composed instead of an interdisciplinary team of architects, filmmakers, artists, scientists, and lawyers. Our products are evidence files in the form of building surveys, physical or digital models, animations, video and maps of various forms. When we work in a legal context, it is often for prosecutors in international law or human rights cases, but our work is not limited to the legal domain – we also produce evidence for citizen- organized truth commissions and tribunals and human rights and environmental protection agencies. Our investigations seek to extend beyond the procedural limitation of each of the forums in which we are asked to present evidence. We try to present incidents in their historical and political contexts – to reconstruct around them the world that made them possible (Weizman, 2017, p. 64).*

Weizman emphasises that FA's work goes beyond the boundaries of the legal context. Not only does FA's gaze extend beyond the limits of the law, but, as Weizman explains, it aims to reverse the very notion of forensic investigation – which is, in fact, subservient to the law – with the result of turning the direction of its gaze on its head, a gaze typical of government agencies such as the police or the secret services when they investigate all those they intend to keep under surveillance. The need to go beyond the boundaries of the law stems primarily from the fact that the killings of individuals or groups of people, the existence of which FA seeks to demonstrate, mostly take place within the logic of the law and are justifiable under existing state laws.

The first methodological revolution concerning the practice of FA is of an aesthetic nature:

> *It is in this way that forensic architecture is able to invert phenomenology's categories of perception and experience: it is not concerned with how we might experience a building, but rather, fundamentally, with how a building might experience its users, how it might sense the way they move and act within and around it. This is not to make an anthropomorphic point: buildings sense not in a human, but rather in a building sort of way. The same principle, as I will later show, can also be extended to build environments and larger territories across the dry surface of the earth. They also act as political sensors to be read (Weizman, 2017, p. 54).*

According to Weizman, therefore, we are faced with a substantial monopoly on information about the war – not only of a military nature, but also of a legal and informational nature, in terms of both words and images. Consequently, an investigation seeking to prove the existence of killings perpetrated by a state (or its contractors) must necessarily adopt a different perspective from that which sets the destruction of human lives through military operations as its own end, justifies such actions through the law, or renders them non-existent in historical terms by preventing or restricting the sharing of images and documents that bear witness to the killings that have taken place.

The term that best describes this revolution in FA's research activities is "conter-forensics" a term coined by Thomas Keenan (2014) on the basis of Allan Sekula (1993)'s reflections [Sekula refers primarily to the work of Clyde Snow and Susan Meiselas in Kurdistan after the first Gulf War: "Counter-forensics, the exhumation and identification of the anonymized ('disappeared') bodies of the oppressor state's victims, becomes the key to a process of political resistance and mourning"]. If the investigative practice employed by FA focuses on deconstructing the violence inherent in state law – which is increasingly founded on forms of war that are concealed or justified on legal grounds – then we must overturn or shift this perspective in order to open ourselves up to a different outlook and discourse, and to establish aesthetic relationships that differ from those that dominate in inter-state relations governed by the logic of law. One way to develop a counter-forensic perspective is to take the phrase "Forensic Architecture" and place the emphasis on the adjective "forensic", that is, relating to the forum, the public square, that open space where matters of importance to the polis are discussed and debated. The history of the use of the English term "Forensic – and the same can be said of the Italian adjective "forense" – follows a "followed a trajectory of linguistic telescoping" (Weizman, 2017, p. 65). The term "forum" no longer refers to the town square where public debate aimed at establishing justice takes place; instead, it has been reduced solely and exclusively to the spaces where legal proceedings are conducted – in essence, the forum is the court. What Allen Feldman (2004) has termed the "police concept of history" operates across three "theatres": the field, the laboratory and the forum, understood as a court of law. In contrast, the practice of counter-forensics seeks to erode the difference between these three spheres, which are regarded as separate, with the aim of constructing forums where there aren't.

Now, the issue of *construction* is fundamental to the practice of FA research, particularly when considered in relation to the use of images in counter-forensic practice. Military attacks are occurring with increasing frequency in urban areas, and different urban environments are "highly sentient in both material,

© Italian e-Learning Association

analogue, and digital terms" (Weizman, 2017, p. 57). When an act of war occurs,

> *all elements of the city start recording, each in its own way. Buildings record vibrations and the force of impacts. Plants record – crushed fields around a city's agricultural outskirts register the movement of military vehicles on them when they stop photosynthesizing, a signal that is captured by remote sensors orbiting above and beamed back to earth. Air- quality sensors pick up increases in traffic as tanks roll in or refugees escape. People remember – in processes that, as I have already shown, are often complex and not straightforward – and increasingly use their camera phones to record the events around them, uploading images, sound, and video online. Each of these sensors is indeterminate, and patient investigative labor has to be invested in reading anything from them and then later also in cross-referencing and pulling the data together (Weizman, 2017, p. 58).*

These modes of perception are not immediate; they do not reflect reality – or a fragment of reality – in a mirror-like manner. Instead, we must think of all these signs as *media* – even when they are not media in the strict sense – through which an event can be reconstructed. And media convey meaning only in the form of a narrative; they require interpretation and imagination, not mere presentation. It thus becomes clear that, for Weizman, counter-forensic research is "an aesthetic practice" (Weizmann, 2017, p. 94) because "it depends on both the modes and the means by which reality is sensed and presented publicly" and "slows down time and intensifies sensibility to space, matter, and image" (Weizmann, 2017, p. 94). Counter-forensic activity is aesthetic because "seeks to devise new modes of narration and the articulation of truth claims" (Weizman, 2017, p. 94) Weizman conceives of matter as a whole as an aesthetic sensorium (see: Weizman, 2017, p. 94) capable of perceptually capturing, in ways specific to the matter in question, everything that occurs in a given place. This opens up the possibility of conceiving of counter-forensic practice as grounded in a material aesthetic: "Material aesthetics" (Weizman, 2017, p. 94) has no evidence to present within a pre-established interpretative system such as the law, but must reconstruct the places and moments of an event through the images, sounds, voices and environmental structures in which such testimonies are imprinted.

Weizman explains some fundamental aspects of material aesthetics:

> *Material aesthetics is the quality of relations between things – the being of matter in the world, its ability to absorb and the degree to which it might. This extends the principles of photography to the rest of the material world, breaking film's and digital photography's monopoly over visual representation. The inverse must also be true: as objects become images, images should be studied as things, parts of the material world. Still, to be read as sensors, the transformations of material objects must be captured by other sensors, such as photographs, analog or digital, remote or proximate, single or hyperspectral, that translate the sensorial capacity of matter into data and help make sense of them (Weizman 2017, p. 96).*

Thus, whilst forensic practice centres on the idea of demonstrating, through clear evidence, a linear and direct relationship between cause and effect, counter-forensic research, on the other hand, requires reconstructing what happened through a process of 'montage'. Weizman refers to this as a "field causality" (on the issue of field causality, see Solla 2022) that can be constructed through a montage employing images, objects and words.

> *While criminal law seeks to establish a linear string of causal relations between intentions, actions, and victims or between, so to speak, the two ends of a smoking gun, the representation of the causes of environmental violence demands the establishment of more complex and diffused causal structures. I propose the term "field causality" to refer to indirect forms of causality, multidirectional and distributed over extended spaces and time durations. It is an inherently spatial form of causality whose employment seeks to reconnect the multiple threads that linear juridical protocols have torn apart (Weizman, 2017, p. 119).*

In counter-forensic investigations, there is never a smoking gun; the images are not immediate and are not authentic; they must always be placed in relation to other images, other signs, and a discourse. Weizman defines the "architectural image complex" (Weizman, 2017, p. 100) that characterises much of counter-forensic research as an assembly of numerous photographic materials which, when spatially repositioned, function as an "optical device" (Weizman, 2017, p. 186) that allows the viewer "to establish and view images and the relations between them" (Weizman, 2017, p. 186). To show what happened, a construction is needed – an editing process capable of holding together materials, images and words.

All the counter-forensic investigation methods developed by FA over the years are currently being

© Italian e-Learning Association

used in the creation of the *Cartography of genocide*, which aims to reconstruct everything that has happened in Gaza since 7 October, drawing on all available visual and other evidence (see https://forensic-architecture.org/investigation/a-cartography-of-genocide).

In *Investigative Aesthetics*, Fuller and Weizman focus on the need to engage with images in a manner that reflects justice and to re-establish that sense of community which has been lost in the sensory realm, as it is dependent on the logic of the law. The starting point for Fuller and Weizman lies in understanding that all media technologies are not neutral but are always politically oriented.

> *The media technologies of artificial intelligence, satellite images, social media platforms, smart cities or facial recognition cameras are not neutral; they are products of specific political and historical contexts, with inbuilt biases, opacity, partiality and illegibility and have the potential to enhance discrimination and domination. These biases may not only be those that entrench existing social norms, which have to be fought and reworked, but also those that are particular to specific media forms. These might be particular idiosyncrasies or predilections. They might texture or produce information in certain ways (Fuller & Weizman, 2021, p. 16).*

3. Investigative aesthetics

Precisely because of the political orientation of media technologies, we must proceed to construct what the authors term an "investigative aesthetic", maintaining a sceptical attitude towards terms such as "fact", "evidence", "truth" and "knowledge", whilst at the same time seeking to redefine and problematise those very notions. This work, which examines terms such as fact, evidence and truth, is necessary because we are confronted with forms of monopoly over these very concepts. The authors write: "This challenge is urgent because it happens at the same time as the rise of political powers that aim to replace the always conditional concept of truth with a thrilling sense of certainty" (Fuller & Weizman, 2021, p. 19). In this regard, it is essential to conceive of the image not as a surface upon which an immediate proof is imprinted, but as a door that opens up new connections with other sensitive traces, with other forms of testimonial evidence. It is within this context that the concept of aestheticisation emerges. For Fuller and Weizman, aestheticisation is understood in a sense that is the opposite of the meaning it has in Walter Benjamin's thought. The term aestheticisation should not, therefore, be understood as a spectacular and fetishistic relationship with images, but rather as the chance to activate one's own sensitivity to certain events within a common, public context. As Fuller and Weizman state:

> *Aestheticisation, the process or act of becoming or making sensitive, is dialogic and collective, just like an emotion is relational and justice is assembled. There is a process to take part in it, but it is also necessary to recognise how the sensing self is an occurrence. The conscious subject is built up through the interaction of numerous entities, systems and experiences. Each of these may have quite distinct aesthetic capacities […] Aesthetics is, crucially, a question of the material relation within and between entities and the ecologies of which they are part (Fuller & Weizman, 2021, 35-36).*

If aestheticisation is the process of becoming sensitive to an event, a capacity to enter into a sensitive relationship with something, this also implies that every fragment of the environment in which human beings live is itself aestheticised, that is, capable of sensing and capturing signs of an event that has entered into a relationship with that fragment of matter.

This notion of aestheticisation has its dialectical opposite in "anaesthetisation". Anaesthetisation is the act of numbing the senses. The authors give the following example:

> *For example, a sense of injustice can be aestheticised or anaesthetised, in fact may be primarily so as a feeling before it becomes a thought. And this can create a link between what one may tacitly perceive, see or hear; what one may feel about what one sees and hears; and how that affects one's sense of right and wrong. In this sense, to be politicised is to increase one's ability to be "aestheticised" to the world (Fuller & Weizman, 2021, p. 36).*

In this passage by Fuller and Weizman, we may find at least the beginnings of an answer to the question we posed at the very beginning: why are there so many images and so many accounts of the genocide in Gaza, yet a lack of political response capable of opposing it in a clear, decisive and unequivocal manner? Firstly, following Fuller and Weizman, the issue should not be approached from an iconoclastic or technophobic standpoint: it is not simply because of the sheer volume of images as such that a live genocide fails to provoke a wave of revolt in the countries most directly concerned, such as the EU and the US. The point is that the way in which that event is perceived has been shaped, as a whole, by the media's coverage of the event, by the forms of repression directed against those demonstrating in support of the inhabitants of Gaza and calling for an immediate halt to the ethnic cleansing, by

© Italian e-Learning Association

the stance taken by the mainstream media, and by the media system as a whole. This entire complex communication system has ultimately produced a public numbing in the face of ethnic cleansing in the Palestinian territories. Sensing and sense-making constitute the dialectical poles of aesthetics: we can perceive without the production of meaning, without the aestheticisation of a particular event. We can look at images without seeing anything. When a group of people is unable to form a common aesthetic understanding of events unfolding in history, they are prevented from taking a political stance. Without an aesthetic understanding of the event, it is impossible to move on to a political analysis of it: aesthetic understanding, critical engagement and political action constitute three aspects of a single act.

The research into images and testimonies carried out by the FA research group is a response to the need to bring violent acts committed by governments against small or large communities of human beings out of invisibility and the realm of the imperceptible; these acts of violence would not exist, nor would there be persecutors and victims, were it not for restoring their visibility within spaces that are simultaneously legal, scientific and artistic, in an attempt to produce a new political meaning within the community, to bring what was relegated to invisibility back into the public gaze and discussion of the forum. FA's work aims to "redeem horror from its invisibility" by showing how a fundamental aspect of this redemption involves changing the relationship we have with the system of technology and images as it currently unfolds within "surveillance capitalism". It is not technology and images in themselves that constitute the direct cause of this numbing effect, but a certain type of economic domination that also unfolds on the level of the technological articulation of the sensible.

References

Bria, F. (2025). *Le coup d'état de la tech autoritaire*. "Le Monde diplomatique", novembre 2025: https://www.monde-diplomatique.fr/2025/11/BRIA/68925

Casilli, A.A. (2020). Schiavi del clic. Perché lavoriamo tutti per il nuovo capitalismo. Milano, Feltrinelli

Feldman, A. (2004). Securocratic Ars of Public Safety: Globalized Policing As Scopic Regime. Interventions: International Journal of Postcolonial Studies, 6(3), p. 332.

Fuller, M., & Weizman, E. (2021). Investigative Aesthetics: Conflicts and Commons in the Politics of Truth, Verso, London-New York 2021.

Grespi, B., & Villa, F. (2024). Il postfotografico. Dal selfie alla fotogrammetria digitale. Torino Einaudi

Keenan, Th. (2014). Counter-Forensics and Photography Grey Room, 55, 58-77.

Kracauer, S. (2022). Selected Writings on Media, Propaganda, and Political Communication, New York, Columbia University Press.

Kracauer, S. (1960). The Theory of Film. The Redemption of Physical Reality, Princeton, Princeton University Press.

Kurgan, L. (2013). Close Up at a Distance: Mapping, Technology, and Politics, New York, Zone Books.

Ippolita (2017). Tecnologie del dominio. Milano, Meltemi.

Sekula, A. (1993). *Photography and the Limits of National Identity.* Culture-front, 2(3), 54-55. http://web.archive.org/web/20210922141940/http://greyroom.org/issues/55/12/photography-and-the-limits-of-national-identity/

Solla, G. (2022). Aprire l'immagine, forzare lo sguardo. Forensic Architecture e il Caso Iuventa, *PHILM. Rivista di filosofia e cinema,* 1, 211.

Sontag, S. (1973). On Photography. New York, Farrar, Straus and Giroux.

Sontag, S. (2003). Regarding the Pain of Others. New York, Farrar, Straus and Giroux.

Weizman, E. (2017). Forensic Architecture. Violence at Threshold of Detectability. New York, Zone Books.

Weizman, E. (2012). Hollow Land: Israel's Architecture of Occupation. London, Verso.

Yuval, A., (2024). The AI Machine directing Israel's bombing spree in Gaza, + 972 Magazine, April 2024: https://www.972mag.com/lavender-ai-israeli-army-gaza/

Zuboff, S. (2019). The age of surveillance capitalism: the fight for the future at the new frontier of power, London, Faber and Faber.

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), pp. 1-16

Securitized migration news and hostile audience reactions on Instagram: a computational analysis of Italian news posts and user comments

Nuccio Ludovico[a,1], Valentina Rizzoli[b], Aysenur Didem Yilmaz[a], Alice Lucarini[a], Veronica Margherita Cocco[a], Loris Vezzali[a]

[a]*Faculty of Medicine, University of Modena and Reggio Emilia – Modena (Italy)*

[b]*Sapienza University of Rome – Rome (Italy)*

(accepted: 17/5/2026; published: 23/5/2026)

Abstract

Migration has become one of the most contested issues in contemporary European public debate, where news coverage often frames migrants through the language of border control, legality, and public order. Such securitized representations may contribute to keeping migrants at a symbolic distance from audiences, presenting them less as socially embedded individuals and more as categories to be regulated or governed. While previous research has examined media representations of migration and online hate speech separately, less is known about how migration-related news content is associated with the hostile reactions that can emerge among social media audiences exposed to this content.

This study addresses this gap by analyzing migration-related Instagram posts published by eight major Italian news organizations between January 2025 and March 2026, together with the comments they generated. A corpus of 368 news posts and 76,998 user comments was analyzed by combining topic modeling of news posts with automated emotion detection and hate speech classification of user comments to examine how securitized thematic environments are associated with hostile audience responses. The results show that audience reactions were marked by a strong prevalence of negative emotions, with anger emerging as the dominant response across the corpus. This affective profile became more polarized in relation to news posts that framed migration through legal-institutional conflict, return procedures, border enforcement, and NGO-related controversies. In these securitized contexts, audience responses were more strongly concentrated around anger and disgust, and this emotional concentration was accompanied by higher levels of hate speech.

KEYWORDS: Migration, Hate Speech, Securitization, Media Framing, Instagram, Emotion Detection, Topic Modelling, Computational Social Science.

DOI
https://doi.org/10.20368/1971-8829/1136366

CITE AS
Ludovico, N., Rizzoli, V., Yilmaz, A.D., Lucarini, A., Cocco, V.M., & Vezzali, L. (2026). Securitized migration news and hostile audience reactions on Instagram: a computational analysis of Italian news posts and user comments. *Journal of e-Learning and Knowledge Society*, *22*(1), 1-16.
https://doi.org/10.20368/1971-8829/1136366

1. Introduction

1.1 Migration, Media, and Online Hate Speech

Migration has become one of the most socially contested issues in contemporary European public debate. Southern Europe, and Italy in particular, represents a particularly relevant case in this regard due to its strategic position within migration routes from the Balkans and North Africa (Ceccorulli, 2020). Given the strong humanitarian and social implications of these phenomena, migration has long occupied a prominent place in public discourse, and it has rapidly become an institutional issue, a ground of political contestation, and an identity-laden component, shifting from a matter of emergency and humanitarian concern toward questions of security, social order, and legality.

Media discourse actively contributes to shaping public perceptions of migration and migrants, functioning as a vehicle for multiple, often contrasting, interpretations of social reality, problem definitions, causal interpretations, moral evaluations, and possible remedies (Entman, 1993). Previous research has

[1] corresponding author - email: nuccio.ludovico@unimore.it

© Italian e-Learning Association

highlighted the persistence of polarized and often distorted representations of migration in both traditional and digital media. A recent study by de Rosa and colleagues (2023) on the Italian online press has shown that migration-related discourse is structured through recurring representational patterns and politically salient symbolic oppositions, rather than through neutral descriptive reporting. Similarly, studies of Italian immigration discourse have shown that, even when not overtly hostile, implicit forms of misrepresentation are systematically observable in media communication (Coschignano et al., 2023). A substantial body of research has shown that migration is frequently framed according to a securitization perspective, that is, as a matter of control, surveillance, and public order. Migration has been described as a domain increasingly constructed through the language of threat, borders, institutional unease, and risk management (Bigo, 2002; Huysmans, 2000), which may further promote the polarization of public opinion.

Within this broader communicative context, the emergence of hostile audience reactions should not be understood as an isolated phenomenon, but as potentially linked to the discursive environments and media representations through which migration is framed and debated. Indeed, Conzo and colleagues (2021) showed that negative portrayals of immigrants can produce severe consequences, including greater perceptions of threat and anxiety, as well as physiological reactions related to aggression, suggesting that anti-immigration media discourse can potentially reinforce polarization and discrimination. Consistently, the role of media in framing public perceptions of immigration and migrants is especially important when studying the growing phenomena of hostility and hate speech, which have become central concerns in the study of contemporary digital discourse (e.g., Amnesty International Italia, 2024). Specifically, social media platforms have become key arenas for the circulation, reinterpretation, and contestation of migration-related content, especially when news posts trigger immediate and emotionally charged audience reactions or function as authoritative sources legitimizing distorted interpretations of reality.

In social media settings, news content and user commentary interact within the same communicative space, making visible not only media representations of migration but also the forms of public response they may trigger, intensify, or normalize. For this reason, social media platforms such as Twitter/X and Facebook have long been considered primary arenas for investigating online hate speech and discourse polarization. However, despite the growing relevance of social media platforms for digital news circulation, Instagram has remained comparatively understudied as an environment in which journalistic content and audience reactions coexist. This omission is particularly relevant in the Italian context, where Instagram reached 27.8 million users in early 2025, corresponding to 46.8% of the total population and 54.8% of adults aged 18 and above (DataReportal, 2025). Moreover, Instagram is progressively becoming an additional outlet through which newspapers and journalists disseminate news content, engage audiences, and extend the visibility of journalistic narratives beyond traditional media channels (Newman et al., 2025), while younger audiences increasingly rely on social platforms for news (Robertson et al., 2026). Instagram therefore represents a strategically relevant communicative environment in which journalistic narratives and audience reactions coexist within the same platform architecture. Building on this premise, the present study examines the communicative dynamics and representational frameworks under which hostile discourse is more likely to emerge. More specifically, it investigates how migration framings in Instagram posts published by major Italian news organizations are associated with audience emotional responses and hate speech.

1.2 Research on Hate Speech in Media Communication

Although the literature on migration-related hate speech is extensive, important aspects remain underexplored. A substantial portion of research on hate speech has concentrated on detection, annotation, corpus construction, and automated classification. This work has been crucial in documenting the presence of hostile discourse and in developing computational tools for its identification, especially in non-English contexts such as Italy. However, while this line of research has produced important resources, such as annotated corpora identifying hate speech and racial stereotypes in Italian Twitter discourse (Bosco et al., 2023; Sanguinetti et al., 2018), it has more often focused on how hate speech can be detected than on the communicative and representational conditions under which hostility is more likely to emerge.

Recent work has examined migration-related discourse on social media platforms in Italy primarily through Twitter/X, focusing on lexical associations, sentiments, and emotions over time (Ambrosetti & Miccoli, 2025). Instagram has also been studied in relation to immigration, but mainly as a space of polarized visual and symbolic representations rather than as an arena in which audiences react to news content through comments (de Rosa et al., 2020). However, aside from a few examples using social media such as Facebook (e.g., Humprecht et al., 2020), research on media framing and research on online hate have often remained analytically separated. On the one hand, studies of migration framing have examined how migrants are represented in political and media discourse. On the other hand, studies of online hostility have investigated the prevalence, linguistic structure, and classification of hate speech in user-generated content. Much less attention has been devoted to the relationship between these two dimensions. As a result, we still know relatively little about how migration is framed by Italian news organizations on Instagram, and how these framings

© Italian e-Learning Association

may be associated with emotional responses and hostile discourse among users. This aspect is particularly relevant because it concerns not only the presence of hate speech, but also the broader social conditions of its activation. If migration is framed primarily through border enforcement, legal controversy, or institutional control, such representations may contribute to constructing migrants as distant, abstract, and governable others, rather than as socially embedded individuals.

This scenario is further complicated by the preexisting ideological posture of news organizations operating in the digital space. In the Italian context, which is characterized by a high degree of polarized pluralism (Mancini, 2013), media outlets such as "la Repubblica" and "La Stampa" typically maintain a progressive editorial orientation, whereas others, such as "Il Messaggero" and "Il Resto del Carlino", are characterized by more conservative stances; meanwhile, outlets like "Il Fatto Quotidiano" are more often associated with a populist, anti-establishment positioning. This ideological diversity reflects a range of representational logics that are transposed into the social media arena, potentially influencing how news posts trigger emotionally charged audience reactions. Understanding whether and how these different framings are associated with distinct emotional and hostile responses is therefore crucial for clarifying the communicative processes through which hate speech in public debate on immigration may be intensified or normalized.

1.3 Media Framing and Securitization of Migration

The gaps outlined above point to the need for an approach able to link media framings of migration to audience responses in online environments. Framing does not merely refer to the thematic content of news coverage, but to the interpretive organization of events through the selection and salience of specific aspects of reality, which in turn shape problem definitions, moral evaluations, and perceived courses of action (Entman, 1993). In the case of migration, this is especially relevant because public communication rarely presents migration as a neutral social phenomenon. Instead, it tends to organize it through recurring interpretive logics that foreground specific actors, threats, responsibilities, and solutions (Nese, 2023).

Research on immigration framing has shown that such frames can affect public opinion in substantively different ways, and that these effects are not reducible to a simple positive-versus-negative distinction. Previous studies have found that immigration frames influence attitudes by activating distinct evaluative and emotional processes, suggesting that the way migration is narrated can shape not only judgments about policy but also broader orientations toward migrants as a social group (Lecheler et al., 2015; Bos et al., 2016). This suggests that audience reactions in comment sections may be associated not simply with whether migration is discussed, but with how it is framed.

Among the possible framings through which migration is publicly constructed, securitization is especially important for understanding why some representations may trigger more hostility than others. From this perspective, migration becomes politically meaningful through discursive processes that define it as a matter of threat, border protection, surveillance, and institutional management, rather than simply as a demographic or humanitarian issue (Huysmans, 2000). This shift is part of a broader "governmentality of unease," in which migrants are increasingly positioned as objects of regulation and risk management, and migration itself is narrated through the vocabularies of control, instability, and public order (Bigo, 2002). Moreover, even when not overtly hateful, a securitized framing of migration may still contribute to constructing migration as a problem to be managed and contained, thereby promoting suspicion, hostility, and social anxiety, and often normalizing restrictive interpretations of migration-related events. For the purposes of this study, this means that journalistic framings centered on securitization may be especially relevant because they can create discursive environments in which hostile responses become more intelligible, resonant, or socially acceptable.

1.4 Securitization, Social Distance and Social Conditions of Hostility

From a socio-psychological perspective, one of the key consequences of securitized framing is its capacity to maintain migrants at a symbolic and cognitive distance from the audience. When migration is narrated primarily through the registers of border enforcement, institutional management, and legal regulation, migrants are systematically positioned as abstract categories (e.g., objects of policy and control) rather than as concrete individuals (Fernández, 2013). This process of abstraction has implications that extend beyond media representation itself, shaping how audiences perceive migrants and respond to migration-related content.

Research on moral disengagement (Bandura, 2002) has shown that dehumanizing or abstracting representations of outgroups reduce the inhibitory mechanisms that normally constrain aggressive or hostile behavior. When individuals are represented as undifferentiated members of a threatening category rather than as persons, the moral constraints against expressing hostility toward them are weakened. Applied to media contexts, Esses et al. (2013) showed that presenting immigrants as sources of infectious diseases, bogus queues-jumpers, or terrorists can foster dehumanization processes toward them. This suggests that framings which strip migrants of individual and social attributes, meaning treating them as flows, threats, or governance problems, may lower the psychological cost of hostile expression in online environments (Stephan & Stephan, 2017). Additionally, related work on dehumanization similarly suggests that when outgroup members are denied fully

© Italian e-Learning Association

human qualities, thus represented as legal categories, or sources of risk rather than as individuals, they are more likely perceived as legitimate targets of prejudice, moral exclusion, and discriminatory responses (Haslam, 2006; Leyens et al., 2000). Crucially, this dynamic does not require overt hostility in the news content itself. A post covering a court ruling on expulsion procedures, or the legal controversy surrounding a naval blockade, may present migration as a governance issue without deploying explicitly hostile language. Yet such framings consistently define migrants through their legal status and their relationship to institutional power, rather than through the social or emotional attributes that normally sustain recognition of shared humanity. When people are encountered primarily as members of a problematic category (rather than as individuals) the psychological distance that facilitates hostile expression tends to increase. Importantly, if such dynamics are present in traditional media outlets, they are also likely to extend to social media environments, where audiences similarly engage with migration as a news topic almost exclusively through group-level representations.

In sum, unlike interpersonal or community contexts in which people may have direct experience of individual migrants, digital news environments present migration as an abstract social phenomenon, populated by categories rather than people. Under these conditions, the emotional and evaluative tone established by the framing of the original post acquires particular weight, since it constitutes much of the available context within which audiences interpret and respond. When that framing is organized around threat, enforcement, and institutional conflict, the social media platform may become a space in which anger and hate are not only more likely to be expressed, but also more likely to be experienced as appropriate, reinforcing rather than challenging the hostile orientations that securitized discourse tends to activate. Framing, emotions, and hostility should therefore not be treated as distinct dimensions of online discourse on migration, but as interconnected elements of the same communicative process.

1.5 The Present Research

This study investigates how migration is framed by major Italian news organizations on Instagram and how these framings are associated with the emotional responses and hate speech expressions produced by the audience of this communication. These framings of migration spread through the digital platform are, in fact, not neutral in their communicative consequences (Entman, 1993), but they create discursive conditions that differentially shape how audiences perceive, evaluate, and react to migration-related content (Lecheler et al., 2015). Understanding these conditions is therefore an essential step toward clarifying the broader dynamics through which hate speech in digital public debate may be intensified, normalized, or potentially mitigated.

To explore these dynamics at scale and in naturalistic conditions, the present study adopts a computationally assisted analysis applied to a large corpus of Instagram news posts and their associated comment threads. In this approach, topic modeling is used as a first step to identify recurring semantic clusters in the corpus. These clusters are then interpreted as empirically derived migration frames through qualitative inspection of keywords and representative posts, with attention to the dominant problem definitions, actors, institutional logics, and migrant representations made salient within each cluster. In interpreting these frames, particular attention is paid to whether they instantiate securitizing logics.

The analysis moves across three interconnected dimensions. The first concerns the framing structure of journalistic content: identifying recurrent migration frames makes it possible to characterize the interpretive logics through which Italian news organizations make migration meaningful on Instagram, and to assess the extent to which these logics reflect securitizing tendencies. The second dimension concerns the emotional reactions expressed in comment sections: examining the affective profile associated with each frame allows the study to move beyond a simple hostile versus non-hostile distinction, capturing the emotional climates that different representations of migration may accompany. The third dimension concerns the distribution and severity of hate speech across frames: by linking hostile expressions in comments to the framing of the corresponding post, the study assesses whether securitized representations of migration are associated with more intense or more widespread hostile reactions among audiences.

These three dimensions jointly inform the following research questions (RQ):

RQ1. Which migration frames can be identified in Instagram posts published by major Italian news organizations?

RQ2. How do emotional responses in user comments vary across these migration frames?

RQ3. How is hate speech distributed across these migration frames?

RQ4. To what extent is emotional polarization associated with hate speech intensity across migration-related frames?

Through the proposed research questions, the present study moves beyond a purely descriptive account of online hostility in migration-related online discourse, contributing to ongoing debates on the securitization of migration and addressing the broader communicative conditions under which hate speech may be intensified, normalized, or potentially mitigated.

© Italian e-Learning Association

2. Method

2.1. Data gathering and Sample Construction

The study examined publicly available Instagram posts and corresponding user comments concerning immigration in Italy, published by major national news organizations between January 1, 2025 and March 1, 2026.

To ensure that the sample reflected a broad segment of the Italian media ecosystem, the corpus included the six leading generalist newspapers by circulation with an active Instagram presence, namely Corriere della Sera, la Repubblica, La Stampa, Il Messaggero, Il Resto del Carlino, and Il Fatto Quotidiano. Two major national press agencies, ANSA and Adnkronos, were also included in order to capture a more institutionally oriented and agency-driven form of reporting. The final dataset comprised posts published by these eight accounts and all corresponding user comments retrieved for the selected posts.

Posts were identified through a keyword-based filtering procedure designed to capture content explicitly referring to immigration and immigrants in the Italian context. The Italian search terms included *immigrazione* ("immigration"), *immigrat** ("immigrant/s"), *migrant** ("migrant/s"), *extracomunitar** ("non-EU national/s"), *clandestin** *irregolar** ("clandestine/irregular"), and *stranier** *irregolar** ("irregular foreigner/s"), including their morphological variants. A total of 42,765 posts were initially screened. Of these, 565 posts matched the filtering criteria. The resulting corpus was subsequently filtered to remove all duplicated posts and posts referring to migration in non-Italian or cross-national contexts without direct relevance to the Italian case (N = 197), resulting in a final sample of 368 posts.

For each retained post, up to 500 user comments were collected and linked to the corresponding post identifier, allowing the integration of post-level and comment-level analyses. The resulting dataset therefore had a hierarchical structure in which comments were nested within posts. This structure made it possible to examine how the main topic of a post related to the emotional and hate speech rate of the comment thread associated with it. At the end of the data gathering procedure, a total of 76,998 comments were collected. The distribution of posts across sources is reported in Table 1.

Table 1 - Number of posts per source.

Source	n posts	% posts
La Stampa	75	20.4
La Repubblica	63	17.1
Ansa	53	14.4
Il Corriere della Sera	53	14.4
Il Messaggero	43	11.7
AdNkronos	42	11.4
Il Fatto Quotidiano	33	9
Il Resto del Carlino	6	1.6

Before analysis, the dataset underwent a pseudonymization procedure aimed at protecting user privacy and ensuring compliance with GDPR principles. With the exception of the eight institutional media accounts, all user identifiers were permanently replaced with unique random alphanumeric keys. Mentions, tags, and URLs contained in the comments were systematically removed. The master key linking original identifiers to the pseudonymized dataset was stored separately in a password-protected JSON file on an offline device accessible only to the research team. In accordance with the principle of data minimization, identifying information was retained only for the period strictly necessary to satisfy research and reproducibility requirements.

2.2. Data Analysis

The analytical strategy combined post-level computational text analysis with comment-level emotion and hate speech detection to examine how different representations of migration in journalistic content were associated with audience reactions in the comment sections. The analysis was conducted in three steps. First, topic modeling was applied to the textual content of the posts to identify recurring semantic clusters in migration-related news coverage. These clusters were then interpreted as empirically derived migration frames through qualitative interpretation of keywords, phrases, and representative posts in each topic. Second, the emotional content of user comments was analyzed using an automatic emotion detection model, allowing comments to be classified according to their predominant emotional tone. Third, hate speech detection was applied to comments to identify hostile, derogatory, or discriminatory expressions. Finally, comment-level outputs were aggregated at the post level and linked to the corresponding post frame, allowing emotional and hostile reactions to be compared across framing patterns.

2.2.1. Discovering Migration Framing via Topic Modeling

As a preliminary step for topic modeling, textual content of Instagram posts was preprocessed by using the *spaCy* Python library (Honnibal et al., 2020). Preprocessing included Unicode and whitespace normalization, removal of URLs and numerical tokens, and lemmatization performed by means of the *it_core_news_lg* language model from *spaCy*. Topics were extracted using BERTopic (Grootendorst, 2022), a transformer-based topic modeling framework that combines contextual embeddings with clustering procedures. Sentence embeddings were generated using the *paraphrase-multilingual-MiniLM-L12-v2* (Reimers & Gurevych, 2019) multilingual model. Dimensionality reduction was performed using UMAP (McInnes et al., 2018), followed by density-based clustering with HDBSCAN (Campello et al., 2013). Text representation was further refined using a *CountVectorizer* with Italian stopwords and n-grams up to trigrams.

© Italian e-Learning Association

To ensure robustness, a grid search over multiple topic solutions was performed. Model selection was based on diagnostic indices including clustering quality (silhouette score), topic diversity, assignment confidence, and outlier rate. Among the best-performing solutions identified by the grid search, the model producing the largest number of clusters was selected. This choice was made to preserve a more fine-grained representation of the different interpretive patterns through which migration was covered in the corpus. The selected solution assigned 325 posts to seven non-outlier clusters, while 43 posts were classified as outliers.

Manual inspection showed that these posts did not form an additional coherent migration frame, but rather consisted of semantically heterogeneous items with weak membership in the identified clusters and mixed or peripheral references to migration. They included, for example, posts on citizenship referendums and voting procedures, highly event-specific local controversies, and items in which migration appeared as a secondary element within broader discussions of crime, social marginality, or political protest. Consistent with this heterogeneity, their average maximum topic probability was 0.186, compared with 0.750 among posts assigned to one of the seven clusters. They were therefore treated as unassigned documents and excluded from subsequent topic-level analyses.

Each non-outlier post was assigned to its most probable semantic cluster, along with an associated probability score. The interpretation of clusters was conducted inductively by inspecting top keywords, n-grams, and representative documents. Manual labels were assigned to capture the dominant interpretive logic through which migration was framed within each cluster.

2.2.2. Emotion Detection and Polarization Measurement

The emotional content of user comments was analyzed using *GordonAI* (Miracula, 2025), an automatic emotion detection model for Italian-language text. Prior to inference, comments were preprocessed using the same normalization pipeline adopted for posts. In addition, emojis were converted into textual descriptions using the *emoji* Python library, to preserve affective information embedded in symbolic expressions. Comments shorter than three characters after preprocessing were excluded from analysis. The model assigned to each comment a probability distribution across seven emotional categories corresponding to Ekman's (1992) basic emotions (joy, sadness, fear, anger, surprise, and disgust) plus a neutral category. For each comment, both the most probable emotion label and its associated intensity (maximum probability) were retained.

To enable post-level analysis, comment-level outputs were aggregated by computing the mean probability of each emotion across all comments associated with a given post. Based on these distributions, an index of emotional polarization was computed as the inverse of normalized entropy (Shannon, 1948; Bright, 2018) of the emotion probability distribution, such that higher values indicate a stronger concentration of responses around a limited set of emotions, and lower values indicate more evenly distributed emotional reactions. In other words, high polarization values refer to a post triggering a singular, dominant emotional response across the audience. Conversely, low polarization values indicate an emotional fragmentation, suggesting that the content elicited a heterogeneous range of reactions, reflecting either a lack of thematic focus or a deeply ambivalent reception among users.

2.2.3. Hate Speech Detection

Hate speech detection was performed using *IMSyPP*, a pretrained transformer-based model for Italian-language hate speech classification (Kralj Novak et al., 2022), on preprocessed comments.

The model assigned each comment to one of four categories: acceptable, inappropriate, offensive, and violent, thereby capturing different levels of severity within hostile discourse. In addition to categorical predictions, the model produces a full probability distribution over classes. For the purposes of this study, hate speech was operationalized in three complementary ways. First, a binary indicator was constructed by distinguishing acceptable from non-acceptable content. Second, a continuous hate score was computed as the complement of the probability assigned to the acceptable class, providing a graded measure of hostility. Third, the distribution of severity levels across categories was retained to capture qualitative differences in hate expression. At the post level, hate-related indicators were aggregated across comments, including the proportion of hateful comments (hate rate), average hate probability, and category-specific prevalence.

3. Results

3.1. Main Migration Frames in Instagram News Posts

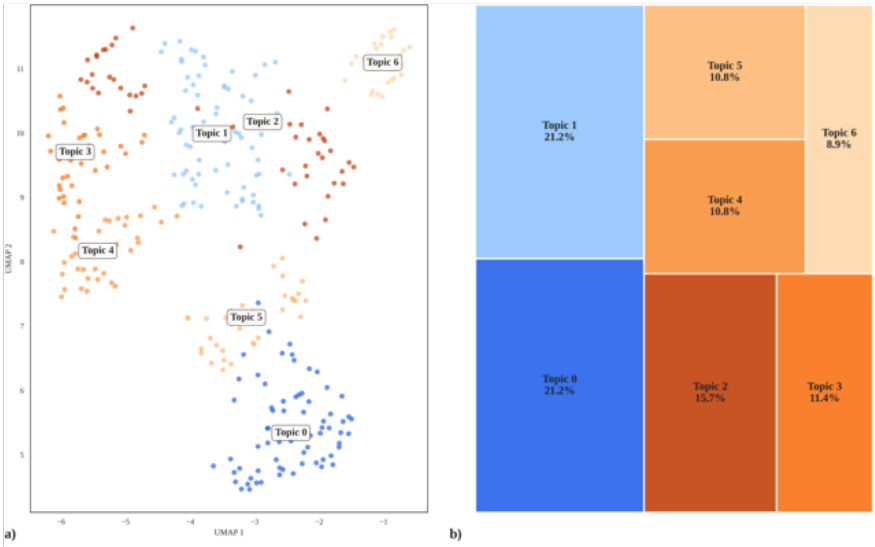
Topic modeling identified seven main clusters in the corpus of Instagram posts related to immigration. Figure 1 shows the semantic distribution of posts in the embedding space obtained with BERTopic, together with the relative prevalence of each topic, while Table 2 reports the main key terms and example documents used to interpret and label each cluster.

The two most frequent topics were Topic 0 and Topic 1, each accounting for 21.2% of the posts (Figure 1b). Topic 0 (*Humanitarian Rescue and Shipwrecks*) was characterized by terms referring to maritime routes, shipwrecks, and rescue operations (e.g., sea, coast guard, Mediterranean, boat, Libya). Topic 1 (*EU Governance and Migration Policy*) included terms related to elections, political actors, and European institutions

© Italian e-Learning Association

Figure 1 - Topic structure of immigration-related Instagram posts identified with BERTopic.

Note. (a) Two-dimensional UMAP projection of document embeddings, showing the semantic distribution of posts across the seven topics. BERTopic assigned 325 posts to seven non-outlier topics; 43 posts were treated as outliers and excluded from topic-level comparisons. Each point corresponds to a post and is colored according to its assigned topic. (b) Topic prevalence, showing the proportion of posts assigned to each topic. Topic 0: *Humanitarian Rescue and Shipwrecks*; Topic 1: *EU Governance and Migration Policy*; Topic 2: *Demographic Change and Welfare Systems*; Topic 3: *Legal Disputes and Return Policies*; Topic 4: *Border Control and NGO Operations*; Topic 5: *Crime, Detention and Violence*; Topic 6: *Religious Ethics and Migrant Reception*.



(e.g., election, campaign, Europe, vote, Germany). Topic 2 (*Demographic Change and Welfare Systems*) represented 15.7% of the posts and included terms referring to population, poverty, welfare, and social conditions (e.g., population, poverty, culture, living conditions). Topic 3 (*Legal Disputes and Return Policies*) accounted for 11.4% of the posts and was characterized by references to courts, legal procedures, and international agreements (e.g., court, tribunal, protocol, Albania). Topic 4 and Topic 5 each represented 10.8% of the corpus. Topic 4 (*Border Control and NGO Operations*) included terms related to naval blockades, NGOs, and judicial decisions (e.g., naval blockade, magistrate, NGO, ship, sentence). Topic 5 (*Crime, Detention and Violence*) was associated with crime, detention, and violence (e.g., murder, victim, refugee, prison, crime). Finally, Topic 6 (*Religious Ethics and Migrant Reception*) was the least frequent (8.9%) and included terms referring to religion, the Catholic Church, and moral discourse (e.g., pope, church, dignity, Vatican). The UMAP projection of document embeddings suggests partial semantic overlap in the language used in some topics (Figure 1a). In particular, Topics 0 (*Humanitarian Rescue and Shipwrecks*) and 5 (*Crime, Detention and Violence*) appear relatively close to each other, as do Topics 3 (*Legal Disputes and Return Policies*) and 4 (*Border Control and NGO Operations*), while Topics 1 (*EU Governance and Migration Policy*) and 2 (*Demographic Change and Welfare Systems*) occupy neighboring areas in the central portion of the map. By contrast, Topic 6 (*Religious Ethics and Migrant Reception*) is positioned more separately from the other clusters, suggesting a more distinct semantic profile within the corpus.

3.2. Emotion Distribution Across Migration Frames

Emotion detection was applied to all comments associated with the posts in the dataset to observe the emotional reactions elicited by different thematic framings of immigration. Figure 2 shows the distribution of emotions for each topic, together with the overall emotional profile across the entire corpus.

Across all migration frames, anger was the most prominent emotion, with an average post-level probability of 39.0%. Joy was the second most frequent emotional component (15.7%), followed by disgust (11.9%), sadness (10.1%), and fear (9.2%), whereas surprise remained marginal overall (1.1%). Neutral reactions accounted for 13.1% of the overall emotional profile. Taken together, these results indicate that Instagram discussions about immigration were characterized by a clear prevalence of negative emotions, especially anger.

Although anger was the dominant emotion across all frames, its relative weight varied across thematic areas. Posts belonging to *Legal Disputes and Return Policies* (Topic 3) and *Border Control and NGO Operations* (Topic 4) showed the highest levels of anger among all frames, respectively 43.2% and 42.2%, together with the highest levels of disgust, 14.4% and 14.5%. *Crime, Detention and Violence* (Topic 5), *EU Governance and Migration Policy* (Topic 1), and *Humanitarian Rescue and Shipwrecks* (Topic 0) also displayed a strong prevalence of anger, respectively 40.1%, 40.0%, and 38.4%.

© Italian e-Learning Association

Table 2 - Topics identified in migration-related Instagram posts.

Note. Key terms are English translations of the original Italian terms identified by the topic model. Example documents are English translations of excerpts from the original Italian post texts. Omissions in the example documents are indicated by square brackets. Translations are provided to improve accessibility for an international readership.

Topic	Documents	Prevalence	Key terms (translated)	Label	Example document (translated)
0	69	21.2%	shipwreck, vessel, boat, navy, sea, coastal, Libyan, Libya, Mediterranean, coast guard	*Humanitarian Rescue and Shipwrecks*	"The sea continues to swallow lives, and the Italian coasts are returning their remains. In the past week, the remains of 15 people have been found along Italian shores, victims of shipwrecks caused by the violent storms that struck the Mediterranean in January. These tragic events coincided with the third anniversary of the Cutro shipwreck, where 94 people died in 2023. […]"
1	69	21.2%	demonstration, European Union, electoral, election campaign, European, vote, Germany, German	*EU Governance and Migration Policy*	"The new party founded by former Folgore commander and later League deputy secretary Roberto Vannacci is called Futuro Nazionale con Roberto Vannacci (FNV). According to Article 1 of its statute, the party aims to promote popular participation in Italian political life, supporting ideas and proposals intended to guarantee the country's internal and external security, the defence of borders, the natural family, and life. [...]"
2	5	15.7%	population, poverty, arrival in Italy, Spain, culture, living, social, Sanremo, million	*Demographic Change and Welfare Systems*	"At a time when much of Europe is tightening its migration policies, Spain has introduced an unprecedented measure: a decree to regularize around 500,000 people living in irregular conditions. Defined by the socialist government as a 'historic' move to remove past bureaucratic barriers, the initiative aims to integrate thousands of residents into the country's legal framework […]"
3	37	11.4%	European Court of Justice, Court of Rome, Italy-Albania protocol, court, Gjader, Albania	*Legal Disputes and Return Policies*	"The European management of migration flows has returned to the centre of international debate, highlighting the need to guarantee rights and transparency. The Court of Justice of the European Union has called on Italy to respect legal safeguards for migrants. […]"
4	35	10.8%	naval blockade, court, ship disembarkation, magistrate, Diciotti ship, defendant, naval, conviction	*Border Control and NGO Operations*	"The Court of Palermo has ruled that the NGO Sea-Watch must be compensated with 76,000 euros for the unjustified detention suffered by the Sea-Watch 3 in 2019 after the Rackete case. The NGO itself announced the decision. 'While the Meloni government announces a "naval blockade" and attacks NGOs involved in rescue operations at sea […]"
5	35	10.8%	murder, Libya, Libyan, kill, victim, refugee, torture, perpetrator, report, crime	*Crime, Detention and Violence*	"Maysoon Majidi, the 29-year-old Kurdish activist who had been arrested on 31 December 2023 on charges of aiding irregular immigration, was acquitted because she 'did not commit the act.' The acquittal was issued by the Court of Crotone.[…]"
6	29	8.9%	Pope Francis, Catholic, bishop, church, Roman, Santa Maria, square, Pope Leo XIV, dignity	*Religious Ethics and Migrant Reception*	"No more weapons, dignity for migrants and marginalized people, and investment in the family 'founded on the stable union between a man and a woman': this was the warning issued by Pope Leo XIV during his meeting in the Vatican with the diplomatic corps accredited to the Holy See. [...]"

© Italian e-Learning Association

Notably, *Humanitarian Rescue and Shipwrecks* displayed the highest level of sadness across all frames, 15.8%, suggesting a more grief-oriented emotional profile compared with the more explicitly securitized frames. A different pattern emerged for *Demographic Change and Welfare Systems* (Topic 2), where anger remained high, 36.6%, but was accompanied by comparatively higher levels of joy, 19.5%, and neutral reactions, 14.6%. Finally, *Religious Ethics and Migrant Reception* (Topic 6) showed the lowest level of anger, 31.6%, and a more balanced emotional profile, with relatively higher proportions of joy, 19.6%, neutral reactions, 16.9%, fear, 10.9%, and sadness, 10.7%.

3.3. Hate Speech Across Migration Frames

Hate speech detection was applied to all comments and then aggregated at the post level to estimate the prevalence and severity of hostile language associated with each topic. Figure 3 shows the mean post-level hate rate across topics together with severity-specific hate speech rates. Across the dataset, the majority of comments were classified as non-hate, but the average post-level prevalence of hate speech varied substantially across topics. The highest mean hate rates (Figure 3b) were observed for *Legal Disputes and Return Policies* (Topic 3; 23.2%), *EU Governance and Migration Policy* (Topic 1; 22.4%), and *Border Control and NGO Operations* (Topic 4; 22.0%). Lower values were observed for *Crime, Detention and Violence* (Topic 5; 16.2%), *Humanitarian Rescue and Shipwrecks* (Topic 0; 13.9%), and *Religious Ethics and Migrant Reception* (Topic 6; 13.9%), while the lowest mean hate rate was found for *Demographic Change and Welfare Systems* (Topic 2; 12.1%).

When considering severity levels, most hate-related comments were classified as offensive rather than violent or inappropriate. Across all topics, the offensive category accounted for the largest share of hate speech, ranging from approximately 78.8% to 91.3% of hateful comments depending on the topic (e.g., direct derogatory expressions such as “Go f*** yourself, you piece of s***” [Comment 73,589]; translated from Italian). Violent expressions were relatively rare but appeared more frequently in *Crime, Detention and Violence* (Topic 5; 4.7%) and *Humanitarian Rescue and Shipwrecks* (Topic 0; 4.3%) (e.g., explicit threats such as “You are disgusting pieces of s***, I would kick you back to your filthy country” [Comment 40,398]; translated from Italian), whereas inappropriate language was more common in *Demographic Change and Welfare Systems* (Topic 2; 19.7%) and *Humanitarian Rescue and Shipwrecks* (Topic 0; 15.2%) compared to the other topics (e.g., stigmatizing expressions such as “Typical woke generation nonsense” [Comment 2,936]; translated from Italian) (Figure 3a).

3.4. Hate speech and Emotional Polarization

To provide an integrated representation of the relationship between emotional reactions and hostile language intensity, we used the continuous hate score described above and computed, for each topic, the average hate score, the level of emotional polarization in the comments, and the number of posts assigned to the topic. These indicators are jointly displayed in Figure 4.

Topics related to migration policies (Topic 1), border control (Topic 4), and legal disputes (Topic 3) were associated with the highest levels of hate. The latter two (*Border Control and NGO Operations* and *Legal Disputes and Return Policies*) also showed high emotional polarization.

Humanitarian Rescue and Shipwrecks (Topic 0) and *Demographic Change and Welfare Systems* (Topic 2) were characterized by lower average hate scores, despite moderate levels of polarization. *Crime, Detention and Violence* (Topic 5) appeared close to the median values for both dimensions, indicating an intermediate profile. *Religious Ethics and Migrant Reception* (Topic 6) showed the lowest values of both polarization and hate score. Taken together, these results indicate that frames centered on political decision-making, border management, and legal conflict are associated with stronger emotional polarization and higher levels of hate speech, whereas humanitarian and moral-religious frames are linked to less hostile audience responses.

4. Discussion

The findings of this study offer a systematic picture of how migration is framed on Instagram by major Italian news organizations, and of the emotional and hostile dynamics these framings tend to activate in their audience.

The topic modeling results identified seven distinct thematic clusters, capturing the range of interpretive logics through which major Italian outlets framed migration on Instagram during the observation period, from January 2025 to March 2026. The two most prevalent topics, *Humanitarian Rescue and Shipwrecks* and *EU Governance and Migration Policy*, together accounting for over 40% of the corpus, reflect the main aspects around which Italian migration discourse has long been organized: the immediate human emergency at sea and the broader political-institutional contestation over how Europe and Italy should respond to it (Radicioni et al., 2021). The remaining topics cover a spectrum from demographic and welfare concerns to judicial and enforcement controversies, religious and ethical debates, and crime-related narratives. Notably absent from this structure is any topic centered on migrants' experiences, integration trajectories, or social lives in the receiving country, a silence that is itself theoretically meaningful and consistent with prior findings on the systematic underrepresentation of migrant voices and perspectives in mainstream European journalism (Berry et al., 2016; Georgiou et al., 2017; Greussing & Boomgaarden, 2017).

© Italian e-Learning Association

Figure 2 - Post-level emotion profiles across topics.

Note. Values are mean post-level emotion probabilities (%). Neutral emotion category excluded for visualization purposes.

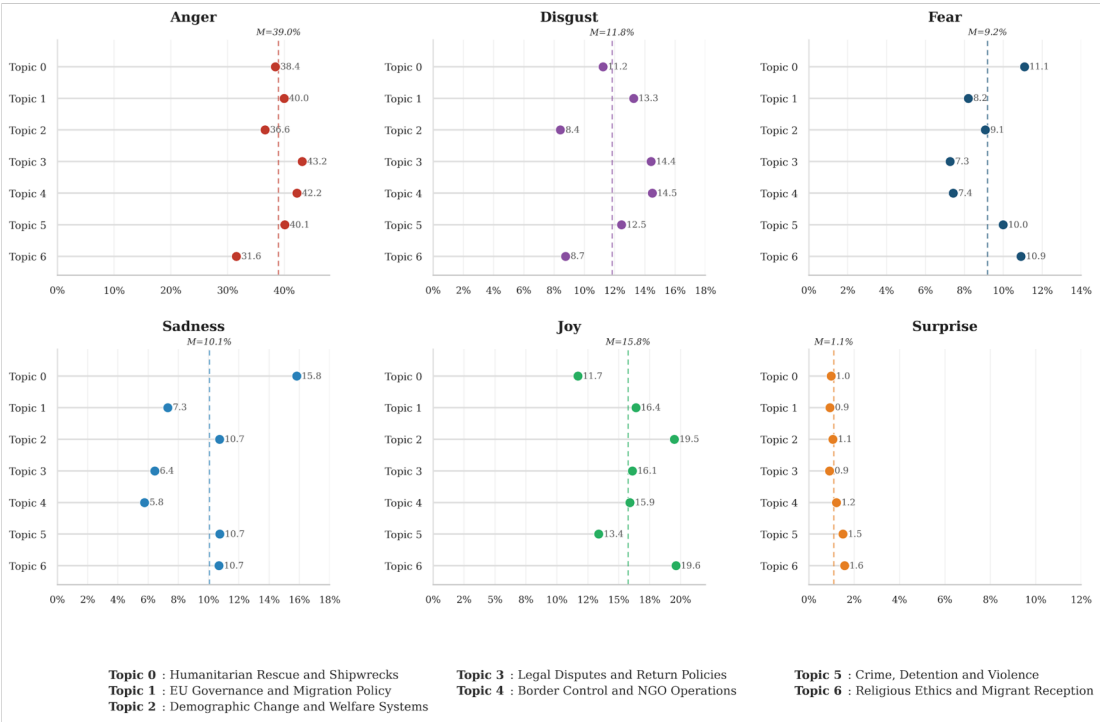
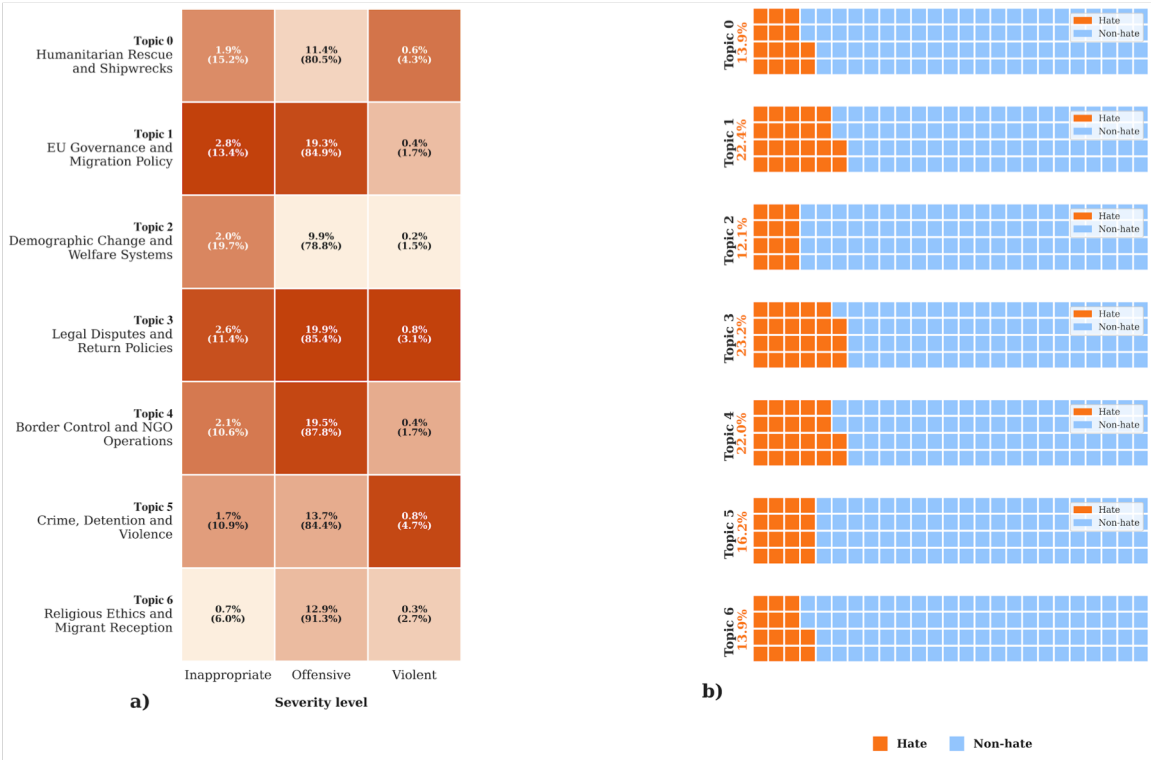


Figure 3 - Hate speech prevalence and severity across topics.

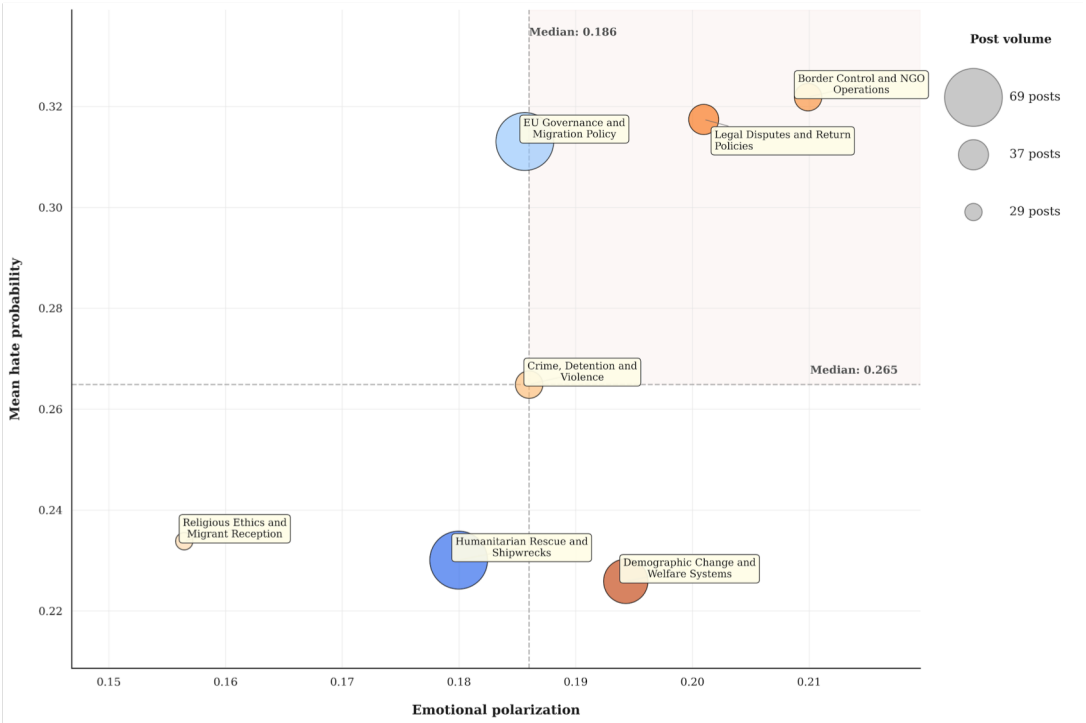
Note. Panel (a) shows mean post-level severity-specific hate rates. Main values report the mean percentage of all comments classified in each severity category; values in parentheses report the mean percentage of each category among hateful comments. Panel (b) shows the mean post-level binary hate rate. Values are averaged across posts.



© Italian e-Learning Association

Figure 4 - Emotional polarization and mean hate probability across topics.

Note. Each bubble represents a topic. The x-axis shows mean emotional polarization, and the y-axis shows the mean continuous hate score, computed as one minus the probability assigned to the acceptable class. Bubble size represents the number of posts assigned to each topic. Dashed lines indicate median values across topics.



Our results showed that, regardless of the specific angle through which migration is framed, audience reactions are predominantly negative. Across all topics, anger was the dominant emotion registered in comment sections, a finding consistent with prior research showing that migration-related news systematically activates negative affective responses among audiences (Conzo et al., 2021; Lecheler et al., 2015). Yet the distribution of anger was not uniform. Comments associated with *Legal Disputes and Return Policies* and *Border Control and NGO Operations* display the highest concentrations of anger and disgust among all topics. This pattern is particularly relevant in light of previous findings on the affective structure of intergroup hostility. While anger tends to orient individuals toward confrontational action, disgust is more closely associated with moral exclusion and the denial of social membership to outgroup members (Cottrell & Neuberg, 2005). Their joint prominence in these topics is thus consistent with the theoretical perspective suggesting that frames positioning migrants as objects of regulation and institutional management may activate the emotional repertoire most directly associated with outgroup derogation and the weakening of moral constraints against hostile expression (Bandura, 2002). This theoretical expectation is supported by recent work showing that moral disengagement mediates the relationship between perceived immigration threat and online incivility across European contexts marked by political polarization around immigration (Bührer et al., 2026).

The emotional concentration observed in these topics is further reflected in their emotional polarization scores, which rank among the highest in the corpus. High polarization, meaning the convergence of audience responses around one or two dominant emotions rather than a distributed affective profile, may further amplify the communicative dynamics activated by securitized framing. When comment sections are dominated by anger and disgust to the near exclusion of other emotional registers, such concentration tends to be associated with a higher likelihood of hostile expression going uncontested and normalized (Rossini, 2022). This is also consistent with evidence that topics related to political conflict and policy controversy systematically elicit more affectively concentrated audience reactions than other news categories (Salminen et al., 2020), and with broader findings indicating that emotionally uniform discursive environments may lower the threshold at which hostile expression is perceived as normal or acceptable.

These patterns find direct confirmation in the hate speech detection results. Topics with the highest emotional polarization, such as *Legal Disputes and Return Policies* and *Border Control and NGO Operations*, also show the highest proportions of hostile comments. Notably, this hostility does not necessarily translate into overt violence, as suggested by the severity

© Italian e-Learning Association

profile of detected hate speech that skews toward offensive rather than violent expressions across all topics. However, prior research has shown that while the most prevalent forms of online hate fall below the threshold of explicit threat, they nonetheless promote degradation, stigmatization, and the normalization of exclusionary attitudes toward outgroup members (Waldron, 2012; Salminen et al., 2020). Similarly, research on the normalization of hostile discourse suggests that repeated exposure to offensive but sub-threshold content progressively lowers audience sensitivity to exclusionary language, making more extreme expressions more intelligible over time (Bührer et al., 2025).

Taken together, the three explored dimensions of thematic framing, emotional response, and hate speech converge on a coherent interpretive picture. Immigration and migrants discourse on Italian Instagram is not homogeneous. Narratives that represent migration as a problem of governance and legal conflict generate more emotionally concentrated and more hostile audience reactions than topics organized around humanitarian, demographic, or religious-ethical registers. This gradient is consistent with prior evidence that enforcement and security framings of migration tend to exacerbate negative and emotionally charged audience responses (Guo et al., 2025). The present study extends that evidence to a platform and national context that have received little systematic attention in the existing literature.

4.1. Intergroup Contact as a Counter-Hate Strategy

Although hate speech in online contexts is widespread, there are strategies that may potentially counteract it. Among these, intergroup contact represents one of the most effective strategies identified by social psychology research. Specifically, positive contact between members of different groups is typically associated with a reduction in prejudice (Pettigrew & Tropp, 2006) – also when considering harsh forms like dehumanization (Vezzali et al., 2022) – even in presence of threatening groups (Van Assche et al., 2023), and among most prejudiced individuals (Turner et al., 2020) Importantly, the effects of contact also extend to collective action, with positive contact often producing mobilizing effects, especially among advantaged group members (who hold the power), being associated with increased willingness to support disadvantaged groups (Cocco et al., 2024b).

Unfortunately, research on the association between intergroup contact and hate speech is still at an early stage. Arcila-Calderon et al. (2022) conducted a European study considering hate speech in relation to social acceptance of migrants and refugees. Results revealed that hate speech (derived from over 800,000 tweets) was lower in regions with higher opportunities of contact and support for social acceptance of migrants and refugees. Moreover, it was found that direct contact, when intimate and positive, can buffer the negative association between media news and attitudes (Fuochi et al., 2020). Taken together, these findings suggest that social interventions that promote positive intergroup contact could be a powerful tool to mitigate hate speech expressed in media news in online contexts. A further study by Soral et al. (2022) revealed that contact interventions fostered an affective response by increasing empathy, which in turn favored greater sensitivity to counteract hate speech and make less use of derogatory language toward refugees. It is worth noting that none of these studies tested contact with stigmatized categories online and its effects on hate speech. We argue for the importance of examining online contact as a possible antidote to hate speech in online contexts.

Finally, contact can also occur in online contexts, where communication between groups is mediated by technological devices, like smartphones, computers, tablets (Amichai-Hamburger & McKenna, 2006), and where several mediums are available (e.g., forums, chats, e-mail). Even when directly embedded in online contexts, contact can therefore represent a safe strategy to promote harmonious intergroup relations, as it may ensure greater anonymity, social equality because subtle cues revealing status differences are less visible, greater control of the interaction. In addition, online contact potentially allows a much higher number of intergroup contacts, by eliminating barriers of space. Research mainly conducted over the last ten years has indeed shown that positive contact between groups is an effective tool also in online environments (White et al., 2020, 2021). Relevant to the present aims, it has been shown that online contact can also favor the development of community ties, strengthening relationships among members of the same community (Cocco et al., 2024a) and therefore mitigate hate speech. A further relevant aspect of online contact is that, like face-to-face offline contact, its effects are allowed by affective changes, like a reduction in anxiety at the prospect of interacting with outgroup members (White et al., 2019). To the extent that, as our study shows, hate speech is emotionally connotated, a strategy based on online contact may allow the activation of affective processes that mitigate prejudice, possibly also when it is expressed as hostility and hate speech in online environments (Sportelli et al., 2025).

4.2. Limitations and Future Research

Several limitations of this study should be acknowledged alongside the strengths that qualify its contributions. The design is observational and cross-sectional. Consequently, the associations documented between thematic framing and audience reactions are descriptive. The co-occurrence of securitized topics with higher hate speech and emotional polarization does not establish that the former produces the latter, as unmeasured confounders such as societal contextual factors or pre-existing audience characteristics may account for part of the observed variation. This is, however, a common limitation shared by naturalistic

© Italian e-Learning Association

studies of online discourse, and it is precisely the ecological validity of the chosen approach that constitutes one of its main strengths.

It should also be noted that the present study documents associations between framing and hostile reactions without directly investigating the mechanisms through which hate speech is promoted or, conversely, mitigated. Understanding what communicative, social, or contextual factors may mitigate the dynamics documented here remains an important open question.

In this regard, future research would benefit from incorporating broader contextual variables that may affect the production and diffusion of hate speech in migration-related online discourse. Particularly relevant are territorial and community-level factors, such as the presence of local integration initiatives, the density of intergroup contact opportunities, or the salience of pro-migrant civic mobilization in the areas where audiences are located. Their inclusion in future comparative or multilevel designs would contribute to a more contextually grounded understanding of how media framing effects on online hostility vary across different social environments.

A second limitation relies on the computational models employed for emotion detection and hate speech classification. These models have not been validated through manual annotation on this specific corpus. Both *GordonAI* and *IMSyPP* are purpose-built for Italian-language text, but their accuracy on social media cannot be assumed without empirical verification. Results should be interpreted considering this uncertainty and future work should incorporate inter-annotator reliability checks on a stratified sample of comments. It is worth noting, however, that the consistency of the patterns across the analytical dimensions of emotion and hate speech provides some convergent support for the robustness of the findings.

The study is also limited to textual content, leaving aside the visual elements that constitute a primary communicative mode on Instagram. Research on visual framing has shown that images exert independent effects on audience responses, and their integration into future computational analyses of Instagram content represents an important direction for multimodal work. Similarly, the sample is restricted to eight major news organizations, excluding explicitly partisan or far-right outlets that are known to produce more intensively securitized framing and to attract more hostile audiences. While this research design choice reflects a deliberate focus on mainstream journalism and the audience reactions it generates, the variation captured in the present corpus likely underestimates the full range of framing effects observable across the Italian media spectrum. Future research targeting such outlets would allow a more comprehensive test of the securitization hypothesis.

5. Concluding remarks

This study examined how major Italian news organizations frame migration on Instagram and how different framings shape the emotional and hostile dynamics they generate among audiences. Combining topic modeling, emotion detection, and hate speech classification on a hierarchically structured corpus of posts and comments, the findings reveal a systematic and theoretically interpretable pattern. When migration is narrated through a securitized framing, the audience reaction becomes more concentrated around anger and disgust, and substantially more hostile than when the same issue is framed in humanitarian, demographic, or moral-religious terms. These results extend prior evidence on securitization and audience response to a platform and national context that the existing literature has largely overlooked.

These findings, however, cannot be interpreted in isolation from the national context and its political trajectory during the period under study. The corpus was collected during a phase in which migration governance was subject to intense institutional and judicial contestation. The legal status of offshore detention facilities, the criteria for designating safe countries of origin, and the respective roles of executive and judicial powers in asylum and return procedures all became primary grounds of public and political conflict, with consequent prominent media coverage. It is perhaps unsurprising, then, that the topics generating the highest rates of hostile comments, concerning legal disputes and border control, map directly onto the discursive terrain in which this contestation unfolded. The findings thus suggest that the association between securitized framing and online hostility documented here reflects not only the general properties of enforcement-oriented migration discourse, but also the amplifying effect of a political conjuncture in which migration was simultaneously institutionally contested and persistently salient in the public agenda. This has consequences that extend beyond the present analysis. When human groups become the sustained object of political contestation and media conflict, the hostile reactions documented in this study serve as a reminder of the potential human costs of that process, for the people whose lives and legal status are at stake in the debates being commented upon.

References

Ambrosetti, E., & Miccoli, S. (2025). The Perceptions About the Immigration Phenomenon Communicated Through Social Media Platforms in Italy: Ambrosetti and Miccoli. *Journal of International Migration and Integration*, *26*(3), 1643-1668.

Amichai-Hamburger, Y., & McKenna, K. Y. (2006). The contact hypothesis reconsidered: Interacting via

© Italian e-Learning Association

the Internet. *Journal of Computer-mediated communication*, *11*(3), 825-843.

Amnesty International Italia. (2024). Barometro dell'odio: Delegittimare il dissenso. Amnesty International Italia. https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2024/05/Amnesty-Barometro-odio-2024.pdf

Arcila-Calderon, C., Sanchez-Holgado, P., Quintana-Moreno, C., Amores, J. J., & Blanco-Herrero, D. (2022). Hate speech and social acceptance of migrants in Europe: Analysis of tweets with geolocation. *Media Education Research Journal*, *30*(71), 21-34.

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education*, *31*(2), 101–119. https://doi.org/10.1080/0305724022014322

Berry, M., Garcia-Blanco, I., & Moore, K. (2016). *Press coverage of the refugee and migrant crisis in the EU: A content analysis of five European countries*. United Nations High Commissioner for Refugees. https://www.unhcr.org/56bb369c9.html.

Bigo, D. (2002). Security and immigration: Toward a critique of the governmentality of unease. *Alternatives*, *27*(1_suppl), 63-92.

Bos, L., Lecheler, S., Mewafi, M., & Vliegenthart, R. (2016). It's the frame that matters: Immigrant integration and media framing effects in the Netherlands. *International Journal of Intercultural Relations*, *55*, 97-108.

Bosco, C., Patti, V., Frenda, S., Cignarella, A. T., Paciello, M., & D'Errico, F. (2023). Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP. *Information Processing & Management*, *60*(1), 103118.

Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of computer-mediated communication*, *23*(1), 17-33.

Bührer, S., Koban, K., & Matthes, J. (2025). Misreading uncivil and intolerant anti-immigration content in digital spheres: The role of recent online perpetration. *New Media & Society*. https://doi.org/10.1177/14614448251401865

Bührer, S., Koban, K., & Matthes, J. (2026). Immigration-related socio-cognitive dynamics driving digital hate: A four-country study on online incivility and intolerance perpetration. *Journal of Computer-Mediated Communication*, *31*(1), Article zmaf026. https://doi.org/10.1093/jcmc/zmaf026

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Ceccorulli, M. (2020). Italy and Migration: Justice on this Side of the Mediterranean. In *The EU Migration System of Governance: Justice on the Move* (pp. 119-145). Cham: Springer International Publishing. https://dx.doi.org/10.1007/978-3-030-53997-9_5

Cocco, V. M., Lucarini, A., White, F. A., & Vezzali, L. (2024a). Virtually meeting "The Intouchables": Online contact, sense of community, and prejudice toward people with a disability. *Journal of Applied Social Psychology*, 54, 332–343.

Cocco, V. M., Vezzali, L., Stathi, S., Di Bernardo, G. A., & Dovidio, J. F. (2024b). Mobilizing or sedative effects? A narrative review of the association between intergroup contact and collective action among advantaged and disadvantaged groups. *Personality and Social Psychology Review*, *28*(2), 119-180.

Conzo, P., Fuochi, G., Anfossi, L., Spaccatini, F., & Mosso, C. O. (2021). Negative media portrayals of immigrants increase ingroup favoritism and hostile physiological and emotional reactions. *Scientific reports*, *11*(1), 16407. https://doi.org/10.1038/s41598-021-95800-2

Coschignano, S., Minnema, G., & Zanchi, C. (2023). Explaining the distribution of implicit means of misrepresentation: A case study on Italian immigration discourse. *Journal of Pragmatics*, *213*, 107-125.

Cottrell, C. A., & Neuberg, S. L. (2005). Different Emotional Reactions to Different Groups: A Sociofunctional Threat-Based Approach to "Prejudice". *Journal of Personality and Social Psychology, 88*(5), 770–789. https://doi.org/10.1037/0022-3514.88.5.770

DataReportal. (2025). *Digital 2025: Italy*. DataReportal. https://datareportal.com/reports/digital-2025-italy

De Rosa, A. S., Bocci, E., Nubola, A., & Salvati, M. (2020). The Polarized Social Representations of immigration through the photographic lens of INSTAGRAM. *Psychology Hub*, *37*(3), 5-22.

De Rosa, A. S., Bocci, E., & Proietti, S. (2023). (Im) migration in Italian online press through the lens of the social representations theory. *International Review of Sociology*, *33*(2), 279-325.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3-4), 169–200. https://doi.org/10.1080/02699939208411068

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication,*

© Italian e-Learning Association

43(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Esses, V. M., Medianu, S., & Lawson, A. S. (2013). Uncertainty, threat, and the role of the media in promoting the dehumanization of immigrants and refugees. *Journal of social issues,* 69(3), 518-536. https://doi.org/10.1111/josi.12027

Fernández, I., Igartua, J. J., Moral, F., Palacios, E., Acosta, T., & Muñoz, D. (2013). Language use depending on news frame and immigrant origin. *International journal of psychology : Journal international de psychologie*, *48*(5), 772–784. https://doi.org/10.1080/00207594.2012.723803

Fuochi, G., Voci, A., Veneziani, C. A., Boin, J., Fell, B., & Hewstone, M. (2020). Is negative mass media news always associated with outgroup prejudice? The buffering role of direct contact. *Group Processes & Intergroup Relations*, 23(2), 195-213. https://doi.org/10.1177/1368430219837347

Georgiou, M., & Zaborowski, R. (2017). *Media coverage of the refugee crisis: A cross-European perspective* (Council of Europe Report DG1(2017)03). Council of Europe. https://rm.coe.int/1680706b00

Greussing, E., & Boomgaarden, H. G. (2017). Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of Ethnic and Migration Studies*, *43*(11), 1749–1774. https://doi.org/10.1080/1369183X.2017.1282813

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guo, L., Su, C. C., & Chen, H.-T. (2025). Do news frames really have some influence in the real world? A computational analysis of cumulative framing effects on emotions and opinions about immigration. *The International Journal of Press/Politics*, *30*(1), 211–231. https://doi.org/10.1177/19401612231204535

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. https://github.com/explosion/spaCy

Humprecht, E., Hellmueller, L., & Lischka, J. A. (2020). Hostile emotions in news comments: A cross-national analysis of Facebook discussions. *Social Media+ Society*, *6*(1), 2056305120912481.

Huysmans, J. (2000). The European Union and the securitization of migration. *JCMS: Journal of common market studies*, *38*(5), 751-777.

Kralj Novak, P., Scantamburlo, T., Pelicon, A., Cinelli, M., Mozetič, I., & Zollo, F. (2022, July). Handling disagreement in hate speech modelling. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 681-695). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-08974-9_54

Lecheler, S., Bos, L., & Vliegenthart, R. (2015). The mediating role of emotions: News framing effects on opinions about immigration. *Journalism & Mass Communication Quarterly*, *92*(4), 812-838.

Leyens, J.-P., Paladino, P. M., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A., & Gaunt, R. (2000). The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review, 4*(2), 186–197. https://doi.org/10.1207/S15327957PSPR0402_06

Mancini, P. (2013). The Italian public sphere: a case of dramatized polarization. *Journal of Modern Italian Studies*, *18*(3), 335–347. https://doi.org/10.1080/1354571X.2013.780348

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Miracula, V. (2025). *Computational approaches to countering disinformation: An analysis based on artificial intelligence models and social network dynamics* (Doctoral dissertation, University of Catania). https://www.iris.unict.it/handle/20.500.11769/690195

Nese, A. (2023). Migrations in Italy and perceptions of ethnic threat. *Journal of International Migration and Integration*, *24*(3), 939-968.

Newman, N., Ross Arguedas, A., Robertson, C. T., Nielsen, R. K., & Fletcher, R. (2025). Digital news report 2025. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2025-06/Digital_News-Report_2025.pdf

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751-783.

Radicioni, T., Squartini, T., Pavan, E., & Saracco, F. (2021). Networked partisanship and framing: A socio-semantic network analysis of the Italian debate on migration. *PLOS ONE*, *16*(8), e0256705. https://doi.org/10.1371/journal.pone.0256705

Reimers, N., & Gurevych, I. (2019, November). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019*

© Italian e-Learning Association

conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 3982-3992).

Robertson, C. T., Ross Arguedas, A., Mukherjee, M., & Fletcher, R. (2026). *Understanding young news audiences at a time of rapid change*. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-r08r-mt26

Rossini, P. (2022). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, *49*(3), 399–425. https://doi.org/10.1177/0093650220921314

Salminen, J., Şengün, S., Corporan, J., Jung, S., & Jansen, B. J. (2020). Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE*, *15*(2), e0228723. https://doi.org/10.1371/journal.pone.0228723

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. A. (2018, May). An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sportelli, C., Cicirelli, P. G., Paciello, M., Corbelli, G., & D'Errico, F. (2025). "Let's make the difference!" Promoting hate counter-speech in adolescence through empathy and digital intergroup contact. *Journal of Community and Applied Social Psychology*, *35*, e70028.

Stephan, W. G., & Stephan, C. W. (2017). Intergroup threats. In C. G. Sibley & F. K. Barlow (Eds.), *The Cambridge handbook of psychology and prejudice* (pp. 131-148). Cambridge University Press.

Soral, W., Malinowska, K., & Bilewicz, M. (2022). The role of empathy in reducing hate speech proliferation. Two contact-based interventions in online and off-line settings. *Peace and Conflict: Journal of Peace Psychology, 28*(3), 361–371. https://doi.org/10.1037/pac0000602

Turner, R. N., Hodson, G., & Dhont, K. (2020). The role of individual differences in understanding and enhancing intergroup contact. *Personality and Social Psychology Compass*, *14*(6), e12533.

Van Assche, J., Swart, H., Schmid, K., Dhont, K., Al Ramiah, A., Christ, O., Kauff, M., Rothmann, S., Savelkoul, M., Tausch, N., Wölfer, R., Zahreddine, S., Saleem, M., & Hewstone, M. (2023). Intergroup contact is reliably associated with reduced prejudice, even in the face of group threat and discrimination. *The American psychologist*, *78*(6), 761–774. https://doi.org/10.1037/amp0001144

Vezzali, L., Capozza, D., Di Bernardo, G. A., Falvo, R., Stathi, S., & Hodson, G. (2022). Strategies for the promotion of humanity attribution to outgroups. *European Review of Social Psychology, 33*(2), 215–254. https://doi.org/10.1080/10463283.2021.1963156

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

White, F. A., Borinca, I., Vezzali, L., Reynolds, K. J., Blomster Lyshol, J. K., Verrelli, S., & Falomir-Pichastor, J. M. (2021). *Beyond* direct contact: The theoretical and societal relevance of indirect contact for improving intergroup relations. *Journal of Social Issues*, *77*(1), 132-153.

White, F. A., Maunder, R., & Verrelli, S. (2020). Text-based E-contact: Harnessing cooperative internet interactions to bridge the social and psychological divide. *European Review of Social Psychology*, *31*, 76–119.

White, F. A., Turner, R. N., Verrelli, S., Harvey, L. J., & Hanna, J. R. (2019). Improving intergroup relations between Catholics and Protestants in Northern Ireland via E-contact. *European Journal of Social Psychology*, *49*, 429-438.

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), pp. 17-26

From Hate Speech to Toxicity: depoliticization, algorithmic governance, and the transformation of harm in Digital Public Discourse

Roberto Bortone[a], Stefano Pasta[b,1]

[a] *National Office Against Discrimination - Rome (Italy)*

[b] *Catholic University of the Sacred Heart, Dept. of Education – Milan (Italy)*

(accepted: 20/5/2026; published: 23/5/2026)

Abstract

Over the last twenty-five years, hate speech has become a key category in international public policies, while digital environments have increasingly promoted the rise of the broader and more operational category of toxicity. This article argues that the shift from hate speech to toxic content should not be understood as a merely terminological substitution, but as a semantic and governmental transformation in the way discursive harm is identified, measured, and managed. The paper first reconstructs the historical and normative genealogy of hate speech, then examines the psychological, computational, and platform-based genealogy of toxicity, and finally compares the two frameworks through their conceptual, operational, and political implications. Particular attention is paid to the agency of platforms, algorithmic governance, content moderation, and the tension between discriminatory harm and conversational harm. The article suggests that toxicity offers scalability and technical operability, but may also contribute to the depoliticization of online harm if detached from histories of discrimination, protected characteristics, and asymmetries of power.

KEYWORDS: Hate Speech; Toxicity; Platform Governance; Content Moderation; Algorithmic Governance; Digital Public Sphere.

DOI
https://doi.org/10.20368/1971-8829/1136368

CITE AS
Bortone, R., & Pasta, S. (2026). From Hate Speech to Toxicity: depoliticization, algorithmic governance, and the transformation of harm in Digital Public Discourse. *Journal of e-Learning and Knowledge Society*, *22*(1), 17-26. https://doi.org/10.20368/1971-8829/1136368

1. Introduction

Over the last twenty-five years, "hate speech" has become a key concept in European and international public policies, without ever consolidating into a universally shared definition. In the lexicon of international institutions, it generally refers to forms of expression that incite, promote, disseminate, or justify hatred, discrimination, or violence against individuals or groups identified through protected characteristics. The Council of Europe Recommendation of 1997 provided an early broad and politico-normative framework, centered on expressions that "spread, incite, promote or justify" hatred based on intolerance (principle 1). The United Nations Strategy and Plan of Action on Hate Speech, in 2019, proposed a more descriptive and operational working definition, referring to communications that "attack" or use pejorative or discriminatory language against people or groups "on the basis of who they are" (p. 2). In 2022, Recommendation CM/Rec(2022)16 of the Council of Europe further refined the framework by introducing the semantically significant phrase "real or attributed" characteristics, thereby avoiding an implicit definition of the protected subject through the labels imposed by haters.

At the same time, however, digital environments have transformed the very nature of the phenomenon. Online hatred no longer appears only as direct insult or explicit incitement, but as a broader communicative ecosystem in which hate, harassment, disinformation, and other forms of abuse coexist and reinforce one another (Pasta, 2018; Bortone, 2023, 2025). In this context, the category of toxicity has gained growing relevance. Developed mainly in computational studies, content moderation, and platform policies, it serves to describe and measure communicatively harmful content even

[1] corresponding author - email: stefano.pasta@unicatt.it

© Italian e-Learning Association

when such content does not fall within the legal perimeter of hate speech. Empirical studies show, on the one hand, that online interactions tend to deteriorate into increasingly toxic exchanges when prolonged over time (Avalle et al., 2024; Pasta, 2022) and, on the other hand, that platform architectures can amplify anger and extreme content, encouraging the circulation of hatred and other forms of toxic communication (Munn, 2020).

This article interprets the shift from "hate speech" to "toxic language" not simply as a terminological replacement, but as a change in semantic regime and in the logic of governance. The analysis proceeds in three steps. First, it reconstructs the cultural and legal context in which the concept of hate speech emerged and stabilized. Second, it traces the psychological, computational, and platform-centered genealogy of toxicity. Finally, it compares the implications of the two lexicons in order to highlight their conceptual, operational, and political divergences.

2. Hate speech between rights, language, and social conflict

"Hate speech" is a historically constructed concept that matured within Western debate as a legal, political, and cultural category. If discriminatory and violent discourses, as well as the outcomes of what has been called the "hostile mind" (Santerini, 2021), have always existed, it is above all in the second half of the twentieth century that three decisive passages can be identified in the consolidation of the category.

A first turning point can be located in the post-war context, marked by the need to come to terms with the consequences of racist and antisemitic propaganda in Nazi-fascist totalitarian regimes (Santerini, 2005). In this framework, international law began to recognize that certain forms of expression are not neutral, but can contribute to the production and legitimation of violence. Documents such as the United Nations International Convention on the Elimination of All Forms of Racial Discrimination (1965) established the principle that incitement to hatred and discrimination is a legally and socially relevant problem, one that must be addressed not only through the regulation of actions, but also through attention to the social effects of language and the damage it can produce.

A second significant development took place in the United States between the 1970s and the 1990s, where debate on hate speech emerged within a structural tension between two fundamental principles: on the one hand, the very broad protection of freedom of expression guaranteed by the First Amendment; on the other, the demand - advanced especially by civil rights movements and university settings - to recognize the harm produced by racist, sexist, and homophobic language and acts, and to protect vulnerable groups from symbolic and material discrimination. In this genealogy, some Supreme Court decisions remain unavoidable reference points: Beauharnais v. Illinois (1952), often recalled for the issue of group defamation; Brandenburg v. Ohio (1969), which established the well-known standard of imminent lawless action; and R.A.V. v. City of St. Paul (1992), which made explicit how controversial it was to restrict bias-motivated expression. It is in this context that the term hate speech consolidated as an analytical category, even while remaining highly contested and normatively unstable. For this reason, Alexander Brown (2017) has convincingly argued that it should be treated not as a transparent and unitary concept, but as a heterogeneous family of expressive phenomena that remain historically and normatively disputed.

Still in the United States, but from a perspective very different from that of classical constitutional jurisprudence, Critical Race Theory introduced another decisive element: shifting the center of analysis from freedom of speech to the social effects of language. In this respect, *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (1993), by Mari Matsuda, Charles R. Lawrence III, Richard Delgado, and Kimberle Crenshaw, remains a key text. Although the authors also use the already established term "hate speech," they insist that racist language is not simply offensive, but can constitute a real-world injury in the form of exclusion and subordination. For this reason, they privilege the term "assaultive speech," which is less ambiguous because it focuses not on intentions or content alone, but on effects (Pasta, 2025). In this perspective, racist and other injurious speech acts represent a real form of harm that limits the freedom and equality of those targeted; for this reason, the authors argue, a democratic society may justify carefully tailored restrictions on speech. The text is especially significant because it anticipates several core features of the later development of the category of hate speech: the idea that formal freedom of expression can conceal substantive inequality, the attention to the lived experience of those who suffer racism and discrimination, the understanding of language as a practice of power and reproduction of hierarchies, the interdisciplinary nature of the inquiry, and the importance of intersectionality.

Alongside the U.S. debate, the European context developed a different approach, less absolutist in its understanding of freedom of expression and more oriented toward the protection of human dignity, the prevention of hateful discourse, and democratic coexistence, also in light of twentieth-century historical memory. In coherence with the differences between the European and American legal traditions (Ziccardi, 2016), European reflection adopted a more restrictive conception of freedom of expression when that freedom harms the dignity of others, treating language not only as individual manifestation but also as a social practice capable of producing exclusion, marginalization, and, in some cases, violence. In this framework, hate speech

© Italian e-Learning Association

consolidated as discourse that targets vulnerable groups. This is the context in which the Council of Europe adopted Recommendation No. R(97)20 in 1997, defining hate speech as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin." As an outcome of the previous trajectory, hatred is thus understood as a broad spectrum of expressions, as a social and structured phenomenon rather than a merely individual one, and as including less explicit forms such as aggressive nationalism, ethnocentrism, and hostility toward migrants - all correctly interpreted as fertile ground for more overt manifestations of hatred. It is not a binding legal norm, but rather a politico-normative definition intended to orient the policies of member states while leaving wide margins of interpretation and adaptation.

Recommendation CM/Rec(2022)16 of the Council of Europe later introduced several elements of novelty. First, it explicitly states that the characteristics on which hatred is based may be not only "real" but also "attributed," thereby recognizing that the target of hateful discourse is not objectively given but discursively constructed by the speaker. Second, the 2022 definition expands the range of relevant expressions by including not only incitement, promotion, or justification of hatred, but also forms of denigration and stigmatization, thus recognizing the continuity between explicit manifestations and subtler, everyday forms of discursive exclusion. Finally, the Recommendation strengthens attention to context and proportionality, underscoring the need for a graduated approach combining legal and non-legal measures.

In this perspective, hate speech progressively takes shape as a multifaceted concept: legal, because it is decisive in the balancing of freedom of expression and non-discrimination; political, because it calls into question the boundaries of the democratic environment and the conditions of participation; cultural, because certain utterances gradually become socially recognizable as degrading, threatening, or discriminatory. The history of the concept shows that we are not dealing with a merely descriptive category, but with an instrument built to define the nexus between language and inequality. From this point onward, the next transition becomes intelligible. Especially from a legal point of view, as public discourse increasingly moves into digital environments (Ziccardi, 2015; Bortone, 2023), the need emerges for a new law of the Internet (Rodota, 2004): the problem is no longer only how to prohibit or sanction certain expressions, but also the extent to which digital communication architectures contribute to defining, expanding, and normalizing them. It is not by chance that, as early as 1999, Lawrence Lessig argued in Code that in cyberspace the design choices embedded in software – the "code" – regulate behavior at least as much as legal rules do. This intuition now appears increasingly decisive for understanding the transformation of hate speech into a technical and infrastructural problem.

3. Beyond hate speech: the rise of toxicity

The lexicon of toxicity developed in a disciplinary context different from that of hate speech. While the latter emerged and consolidated within the language of human rights and non-discrimination, the notion of toxicity took shape mainly with the rise of the social web in the early 2000s, first within the psychology of communication and later in studies of computer-mediated communication and new digital environments.

A foundational reference is the work of the American psychologist John Suler, who in 2004 introduced the concept of the online disinhibition effect to describe the tendency of individuals to express themselves more freely in digital environments than in face-to-face interaction, often through more intense, impulsive, and aggressive forms of speech. Suler distinguishes between "benign disinhibition," associated with openness, self-disclosure, and emotional support, and "toxic disinhibition", marked by aggression, hostility, offensive language, and antisocial behavior (Suler, 2004).

It is important, however, not to move too quickly: toxic disinhibition does not yet coincide with toxic content. In the first case, we are dealing with a psychological and situational explanatory model that seeks to clarify why certain digital environments foster flaming, trolling, harassment, and other aggressive forms of conduct. In the second case, by contrast, we are already within a descriptive and classificatory logic, typical of platform moderation and computational research. Here, toxicity no longer indicates only a condition of production of discourse, but a quality of discourse itself: the degree to which content is perceived as rude, disrespectful, aggressive, or likely to deteriorate conversation. It is in this transition that toxicity becomes especially useful for annotation systems, benchmarks, and automated classification.

Unlike hate speech, toxicity does not necessarily presuppose a reference to vulnerable groups or protected characteristics. It describes instead a communicative quality: the degree to which content is perceived as hostile, corrosive, degrading, or potentially harmful to conversation. For this reason, the category of toxicity proved particularly suitable to the needs of platforms and computational research, both of which need to transform complex linguistic phenomena into annotatable, scoreable, and classifiable features. An emblematic case is Perspective API, launched in 2017 and developed by Jigsaw, a Google/Alphabet unit,

© Italian e-Learning Association

as a free machine-learning service designed to help platforms, publishers, and moderators anticipate how a comment might be perceived within a conversation. In its documentation, its main attribute – toxicity – is defined as a comment that is "rude, disrespectful, or unreasonable" and likely to discourage participation in the discussion. Around this infrastructure, an entire ecosystem of datasets annotated for toxicity, subtypes of toxicity, and identity references has been consolidated for the training and evaluation of automated models.

What emerges here is a passage from a socio-political notion of hatred to a technical-operational notion of "healthy conversation," one that can be handled through machine learning and inserted into a broader process of algorithmic governance of public discourse. The possibility of translating complex linguistic phenomena into scores, categories, and operational thresholds allows platforms to intervene at scale, but at the same time orients what is recognized as a problem. In other words, it is not only technology that "measures" toxicity; it is the need for measurement that selects which dimensions of discourse are made visible and treatable. In this sense, the category of toxicity is not neutral, but embedded in specific logics of governance that privilege what is computable over what is socially and historically situated.

The emergence of this concept is closely linked to the diffusion of the first online communication platforms - forums, chats, blogs - and to the need to understand phenomena such as flaming, trolling, and digital harassment (Pasta, 2018). Toxic disinhibition is thus interpreted as the effect of specific communicative conditions typical of the online environment, including anonymity, invisibility, emotional illiteracy, and a reduced perception of the social consequences of one's actions. The focus is not the content of discourse or its broader social dimension, but the conditions that facilitate the expression of behaviors that would otherwise remain inhibited.

Connected to the idea of toxic disinhibition is the concept of polarization, which has become central for describing the communicative dynamics of the social web and is often linked, in the age of platformization, to segmented public spheres and to the circulation of hateful discourse. In this direction, the work of Walter Quattrociocchi and his collaborators has progressively shifted attention from isolated content to the mechanisms of aggregation, mutual confirmation, and polarization that characterize digital information ecosystems: from echo chambers and confirmation bias to persistent patterns of toxicity across different platforms (Quattrociocchi & Vicini, 2013). Hate speech and toxic content, in this sense, can partially overlap, but both are understood as products of polarized information ecosystems – effects, rather than simple causes, of circulation dynamics such as echo chambers, homophily, infodemics, and polarization itself, including in their relation to algorithmic logics.

The shift from hate speech to toxicity does not necessarily entail an attenuation of the problem of hatred. On the one hand, it makes it possible to capture the gray area of discursive harms that are not always reducible to the most explicit forms of discriminatory hatred (Pasta, 2022). On the other hand, however, it moves the axis from law and responsibility toward metrics, policies, and risk management, thereby implicitly redefining the priorities of public and private intervention. The category of toxicity spreads also because it is more scalable, measurable, and comparable across multilingual and multicultural digital environments. Yet precisely this greater technical translatability can produce an effect of depoliticization: the centrality of asymmetries of power may be neutralized, and the focus can shift from justice to the management of "harmful" or "unsafe" communication across a broad range of behaviors.

4. Comparing hate speech, toxicity, and platform definitions

At this point, it becomes useful to compare three definitional families that frame current interventions against processes of digital targeting: the classical definitions of hate speech elaborated by international institutions; the definitions of toxicity developed in the psychology of online communication and later in computational research; and the operational definitions adopted by digital platforms for moderation and enforcement, as of May 2026.

The tables below do not aim to provide an exhaustive repertory, but rather a reasoned comparative grid gathering the most influential definitions. For the hate speech column, a selection of particularly influential institutional formulations within the European and international context has been included; for the toxicity column, definitions or formulations that have played a significant role within the psychological tradition and in the subsequent computational translation of the concept; and for the platform-related column, official policies that clearly illustrate the transformation of a socio-political category into a moderation taxonomy.

The criteria adopted for the comparison are: disciplinary matrix of origin, primary object, necessity of reference to protected characteristics, type of presumed harm, threshold of problematicity, role of context, and regulatory output. As for the selected platforms, Meta, YouTube, TikTok, Twitch, and Discord do not exhaust the landscape, but represent five different ecologies of governance: generalist social networking, video sharing, short-video platforms, live streaming, and community chat/server environments.

Table 1 summarizes the principal differences among three definitional families that overlap only partially.

© Italian e-Learning Association

Table 1 - Criteria for comparing hate speech, toxicity, and digital platform definitions.

Criterion	Hate speech	Toxicity	Operational platform definitions
Originating framework	Human rights, anti-discrimination, international case law	Psychology of online communication, natural language processing, content moderation studies	Trust & safety, community standards
Primary object	Expressions that incite, promote, disseminate, or justify hatred, violence, or discrimination; or that denigrate persons or groups	Comments or content perceived as rude, aggressive, disrespectful, or likely to deteriorate conversation	Attacks, threats, dehumanization, degradation, slurs, incitement to hatred or violence according to proprietary taxonomies
Centrality of reference to protected characteristics	Yes, central	No, not necessarily	Generally central in hate policies; often integrated with broader policies on harassment and abusive behavior
Typical target	Groups or members of vulnerable or inferiorized groups	Generic interlocutors or participants in discussion	Individuals or groups, with differentiated protections for protected attributes
Type of harm presupposed	Discriminatory harm, violation of dignity, exclusion, threat to democratic coexistence	Deterioration of the discursive climate, interactional hostility, abandonment of discussion	Risk to user safety, service integrity, regulatory compliance, reputational harm
Threshold of problematicity	Hatred, discrimination, violence, denigration; often graduated	Rudeness, disrespect, aggression, likelihood of derailing discussion	Operational thresholds such as “direct attack”, “promotes violence or hatred”, “degrades”, or “dehumanizes”
Role of context	High: history, relations of power, social meaning, asymmetries	Weaker or more standardized, especially in automated models	Relevant, but filtered through internal guidelines, exceptions, and scalability requirements
Technical operability	Medium-low	High	Very high
Typical output	Norms, recommendations, case law, criminal/civil/ administrative sanctions	Scores, benchmarks, datasets, auditing, risk assessment	Removal, downranking, strikes, demonetization, labeling, limitation of reach
Main risk	Ambiguity or definitional overbreadth	Depoliticization of discriminatory harm; model bias	Reduction of a socio-political category to a problem manageable at scale

As the table shows, classical hate speech serves to name discriminatory harm; toxicity serves to make communicative harm operationally treatable; and platforms produce a hybrid category that preserves the reference to protected characteristics while bending it toward the needs of enforcement and private governance. In this way, platform policies transform a socio-political category into a taxonomy of moderation

The comparison suggests that classical hate speech definitions are semantically denser. They do not merely say that content is offensive or aggressive: they place it within a framework of discrimination, dignity, potential violence, and democratic coexistence. This density makes them theoretically richer, but also less easily translatable into automated criteria. By contrast, toxicity is thinner on the political plane but much stronger operationally: it is easier to annotate, quantify, and model.

In this framework, platform policies reveal a precise conceptual transformation: the socio-political category of hate is retained only insofar as it can be divided into moderable acts, operational thresholds, exceptions, and enforcement levels. The real fracture, therefore, is not simply between “hate” and “toxicity,” but between discriminatory harm and conversational harm. Hate speech primarily protects against subordination and exclusion; toxicity primarily protects against deterioration of conversation and hostile climates; platforms, finally, also protect themselves, since service integrity, perceived safety, governability of space, and regulatory compliance are structural parts of their definitions.

5. Platform agency: the technical reformulation of hate speech

In conditions of persistent fragmentation of the transnational normative framework governing the web, digital platforms and their operating architectures now constitute an infrastructural level that not only organizes the circulation of content, but actively intervenes in its definition, classification, and possible removal. If hate speech allows us to analyze the content

© Italian e-Learning Association

Table 2 - Criteria for comparing hate speech, toxicity, and digital platform definitions.

Hate speech	Toxicity	Platforms
Council of Europe 1997: "all forms of expression which spread, incite, promote or justify hatred based on intolerance"	Suler 2004: online disinhibition helps explain why people express themselves online in more intense, impulsive, or aggressive ways	Meta: "direct attacks against people ... on the basis of protected characteristics"
Council of Europe 2022: expressions that incite, promote, disseminate, or justify violence, hatred, or discrimination, or denigrate, on the basis of real or attributed characteristics	Perspective API (since 2017) / computational tradition: toxicity as an annotatable property of comments, useful for estimating how rude, disrespectful, or conversation-damaging they may be	YouTube: content that "promotes violence or hatred" against individuals or groups on the basis of protected status
United Nations 2019: communication that attacks or uses pejorative or discriminatory language toward persons or groups "on the basis of who they are"	Perspective subtypes: severe toxicity, identity attack, insult, and other subcategories that decompose communicative harm into classifiable attributes	TikTok: explicit or implicit content attacking a protected group; dehumanization, supremacy, segregation, and discrimination are also prohibited
EU Framework Decision 2008/913/JHA: public incitement to violence or hatred against a group or a member of a group defined by race, colour, religion, descent, or national or ethnic origin	NLP / moderation research: a metric and generalizable category, useful for benchmarks, auditing, and risk assessment beyond the legal core of hate speech	Twitch: content or behavior that discriminates, denigrates, harasses, or encourages violence based on protected characteristics
Contemporary European approach: a graduated concept, to be addressed through criminal, civil, administrative, and non-legal measures proportionate to gravity	Contemporary line of inquiry: toxicity is often treated as an indicator of deterioration of discursive space and of the probability of making conversation unlivable	Discord: expressions that degrade, vilify, or dehumanize, incite hostility, or promote harm on the basis of protected characteristics

and socio-political dimension of discourse, and toxic disinhibition explains some of its psychological conditions of production, platforms – as non-neutral spaces of mediation – are real actors exercising definitional power over public discourse. By operationally reformulating the problem of hate speech through their own policies, they exercise not only regulatory power, but also genuinely epistemic power: they select the categories through which the phenomenon becomes knowable, measurable, and governable.

This conceptual transformation can be read as a form of privatization of the regulation of public discourse, in which global economic actors assume functions historically closer to law and democratic institutions. By privileging an approach centered on the removal or downgrading of explicitly aggressive content, platforms risk neglecting subtler, more diffuse, and normalized forms of hatred, operating through narratives, stereotypes, and discursive constructions that are less visible but no less socially effective. Recent regulatory and judicial developments also show that online harm is increasingly being read as an effect of the environment designed by the platform, and not only of single pieces of content.

Shifting attention from structures of meaning to the more superficial manifestations of discourse, as platforms tend to do, brings out a tension between the needs of scalability and the need for critical understanding. In a previous analysis based on a corpus of cases of online racism, it was argued that there is a "return of race in the absence of scientific credibility," a phenomenon that can be reread in light of the transition from hate speech to toxic content: "A concept of 'race' defeated by science is thus affirmed, yet socially accepted by our implicit popular pedagogies and therefore able to re-emerge in the collective consciousness even without scientific credibility" (Pasta, 2019, pp. 185-186). In a similar way, within the communicative flow of the social web, racist discourse – loaded with history and meaning – may be emptied of meaning and thereby rendered socially acceptable.

This transition also finds empirical confirmation. Within Hate Studies, research aimed at detecting forms of hatred can be divided between studies that rely exclusively on machine-learning methods and those that combine automated retrieval with human classification (Forzinetti et al., 2024; Pasta, 2024a). As shown by studies conducted by the Mediavox Observatory at Università Cattolica on antisemitism (Pasta, 2024b; Pasta et al., 2021), Islamophobia (Pasta, 2023a), and anti-Gypsyism (Pasta, 2023b), systems based solely on automated detection - for example through dictionaries of offensive words or linguistic patterns - return results that differ significantly from context-sensitive human annotation when the same corpus is subjected to the two methods and evaluated through a confusion matrix.

What is at stake, therefore, is the protagonism of powerful private actors in the digital sphere against the

© Italian e-Learning Association

relative weakness of supranational actors that – especially in Europe – have attempted to regulate platform responsibility, sometimes ending in open conflict (Human Rights Watch, 2023). It is important to recall here the connection between the genesis of hate speech and the perspective of human rights, at a historical moment in which supranational law itself is being challenged by war, by the crisis of liberal democracy, and by claims voiced from within the global web elite. Donatella Di Cesare (2025) has spoken of "technofascism" to describe the convergence of technocratic and ethnocratic tendencies within parts of the new right. In a different but convergent register centered on the primacy of technique and security, Alexander C. Karp (CEO of Palantir Technologies) and Nicholas W. Zamiska, in The Technological Republic (2025), call for a renewed alliance between technological industry, state power, and national security. More than reconstructing the genesis of the concept of hate speech, these texts help illuminate the ideological context in which the boundaries of acceptable discourse online are being redrawn today.

6. Between hate speech and toxic content: some consequences

The comparison developed in this article shows that the passage from hate speech to toxicity concerns the way in which discursive harm is named, measured, and governed. Hate speech retains a normative force that is difficult to replace, because it allows language to be read within relations of power, histories of subordination, and regimes of inferiorization that exceed mere conversational aggression, especially in relation to vulnerable groups (Pasta, 2018). Toxicity, by contrast, has established itself because it offers a flexible and technically manageable operational category, functional to the measurement, classification, and management of communicative risk in digital environments, and capable of capturing a gray area of discursive harm that does not always coincide with the hard core of discriminatory hate. This gray area of communicative harm is not entirely alien even to European anti-discrimination law: both Directive 2000/43/EC and Directive 2000/78/EC qualify as discrimination unwanted conduct that violates a person's dignity and creates an intimidating, hostile, degrading, humiliating, or offensive environment. This shows that the problem of environmental and relational harm partly predates the computational category of toxicity, even if it does not coincide with it.

The point, then, is not to establish in the abstract which of the two categories should prevail, but to recognize that both refer today to a broader transformation of the digitalized public sphere (Bortone, 2023), in which the governance of discourse is increasingly entrusted to technical infrastructures, risk metrics, and criteria of optimization (Habermas, 2005; van Dijck et al., 2018). The Digital Services Act, for its part, requires very large online platforms (VLOPs) to identify and mitigate systemic risks deriving not only from illegal content, but also from service design, recommender systems, automated moderation, and targeted advertising. This represents an important attempt to respond to the platform agency discussed above.

A first methodological consequence concerns detection. Those who continue to work with the category of hate speech – whether in detection, in evaluating the impact of offensive content on victims, in building policy, or in educational research – can no longer ignore the spread of the category of toxicity, which now informs a substantial part of moderation practices, risk assessment, and computational research. Conversely, those who employ toxicity should make explicit why they do so, and to what extent that choice entails a shift from discriminatory harm to conversational harm. Content that is socially significant from the standpoint of discrimination may appear only weakly "toxic" according to computational parameters, while highly conflictual but non-discriminatory content may be classified as problematic. The result is a possible divergence between what is technically detectable and what is socially meaningful.

A second consequence concerns the fact that platforms do not merely moderate hate speech: they contribute to reformulating it. The risk, as this contribution has sought to show, is that the political dimension of hateful discourse – understood as the expression of hierarchies, conflicts, and inequalities – is progressively recoded as an issue of compliance with communicative standards, service safety, or quality of user experience. Depoliticization would therefore not consist in the disappearance of the problem, but in its reformulation in terms compatible with the governability of the platform. In moderation practices, platforms tend to privilege hybrid categories that combine elements of classical hate speech with logics proper to toxicity. This undoubtedly allows greater operational effectiveness - in removal, downranking, or visibility limitation - but also reformulates the problem itself: hate speech is increasingly treated as an issue of individual communicative behavior, linked to the protection of digital well-being, rather than as an expression of structures of social inequality and injustice.

A third consequence concerns public policy. The shift from hate speech to toxicity raises significant questions in terms of accountability and governance. If hate speech has traditionally been anchored to legal frameworks and to systems of rights protection, toxicity occupies a more fluid space, often governed by private policies and technical standards. In the absence of a fully shared transnational normative framework, this implies a displacement of power toward platforms, which not only apply rules but also help define them by

© Italian e-Learning Association

selecting which dimensions of the phenomenon are made visible and governable.

A fourth consequence concerns education. If one limits oneself to addressing "toxicity" as a problem of communicative behavior - for example by promoting respectful communication or conflict-management skills - one risks overlooking the structural dimension of hate speech, linked to stereotypes, prejudice, power relations, and socially situated practices of inferiorization. By contrast, a critically informed educational approach should hold together both dimensions: it should develop communicative and relational skills for onlife citizenship (Pasta & Rivoltella, 2022) while also promoting a critical literacy capable of recognizing the nexus between language, inequality, and the social construction of difference (Pasta, 2023c; Faloppa, 2026). The challenge, then, is to construct integrated definitional models capable of combining the technical precision of automated detection with the interpretive depth of the social sciences and law, so that the complexity of the phenomenon is not reduced to what is merely measurable.

A further element reinforces this need for conceptual caution. Recent regulatory and judicial developments show that digital harm is increasingly being read as the effect of the environment designed by the platform, and not only of the individual content item. In the United States, Section 230 of the Communications Decency Act (1996) has historically shielded intermediaries from classical editorial liability. The Supreme Court did not substantially curtail this framework in Gonzalez v. Google (2023), and Twitter, Inc. v. Taamneh (2023); and in Moody v. NetChoice (2024), it recognized broad editorial discretion in content moderation. At the same time, the March 2026 verdict in K.G.M. v. Meta Platforms, Inc., et al. shifted attention toward the design and functioning of Instagram and YouTube – that is, toward the design of the environment rather than only third-party content – while, at the European level, the Commission preliminarily found Meta in breach of the DSA for failing effectively to prevent access by users under the age of 13. All this does not coincide with the category of toxicity, but it makes more plausible the idea that digital harm must also be read as an effect of socio-technical architectures, recommender systems, and logics of engagement.

Finally, a critical piece of evidence deserves to be signaled as a future development of research: the possibility that a new platform environment is emerging. Some evidence suggests that platforms do not merely remove clearly hateful or violative content, but may also reduce the visibility of politically sensitive content, content difficult to monetize, controversial content, or content deemed risky for the safety and governability of the environment. A Human Rights Watch report of 2023 documented more than 1,050 cases of removal or undue suppression of Palestinian or pro-Palestinian content on Facebook and Instagram; Meta announced in 2024 that it would no longer proactively recommend political content on Instagram and Threads, and from 2025 it discontinued political, electoral, and social issue ads in the European Union. In this context, even content oriented toward peace or civic testimony may enter a zone of reduced sponsorizability or reduced visibility, not because it is assimilated to hate, but because it is classified as politically sensitive or as social-issue content. A report gathered by the authors from an association that attempted to promote a peace event points in the same direction, although it has no generalizable probative value. The provocative question – can even peace be "toxic"? – should therefore be understood not literally, but as a critical question: in an environment optimized for brand safety, risk management, and governability, even pacifist content, war testimonies, citizen journalism, or digital activism may become problematic for the platform, although they are neither hateful nor illegal.

For this reason, one possible line of future research would be to verify whether the feeds of the main social media platforms have been progressively "cleaned" not only of explicitly hateful content, but also of images, testimonies, and forms of digital activism that do not propagate hatred and yet make the environment less marketable, more conflictual, or more difficult to govern according to platform parameters.

Acknowledgements

The present contribution is the joint work of the two authors; however, it should be noted that §§ 1, 4, 6 were written by Roberto Bortone and §§ 2, 3, 5 by Stefano Pasta.

References

Avalle, M., Di Marco, N., Etta, G. et al. (2024). Persistent interaction patterns across social media platforms and over time, Nature, 628, 582-589, https://doi.org/10.1038/s41586-024-07229-y.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification, Companion Proceedings of the 2019 World Wide Web Conference ACM, https://doi.org/10.1145/3308560.3317593.

Bortone, R. (2023). Molto social, troppo dark. Tra hate speech, propaganda, metaverso e intelligenza artificiale: i rischi del web oggi, Roma, Fefè.

Bortone, R., & Pistecchia, A. (2025, Eds.). L'ultimo pregiudizio. L'antiziganismo tra storia e attualità, Roma, Nuova Cultura.

© Italian e-Learning Association

Brown, A. (2017). What is Hate Speech? Part 1: The Myth of Hate. Law and Philosophy, 36(4), 419-468, https://doi.org/10.1007/s10982-017-9297-1.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language, Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512-515, https://doi.org/10.1609/icwsm.v11i1.14955.

Di Cesare, D. (2025). Tecnofascismo. Torino, Einaudi.

Faloppa F. (2020), #Odio. Manuale di resistenza alla violenza delle parole, Torino, UTET.

Faloppa, F. (2026). Disarmare il discorso. Sulla militarizzazione del linguaggio. Firenze, Effequ.

Forzinetti, E., Della Vedova, M., Pasta, S., & Santerini, M. (2024). Indicators for characterising online hate speech and its automatic detection (arXiv:2402.08462), Cornell University, http://arxiv.org/abs/2402.08462.

Gelber, K. (2021). Differentiating hate speech: a systemic discrimination approach, Critical Review of International Social and Political Philosophy, 24(4), 393-414, https://doi.org/10.1080/13698230.2019.1576006.

Habermas, J. (2005). Storia e critica dell'opinione pubblica. Roma-Bari, Laterza.

Karp, A.C., & Zamiska, N.W. (2025). The Technological Republic: Hard Power, Soft Belief, and the Future of the West, New York, Crown Currency.

Lessig, L. (1999). Code and Other Laws of Cyberspace. New York, Basic Books.

Matsuda, M., Lawrence III, C.R., Delgado, R., & Crenshaw, K. (1993). Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment. New York, Routledge.

Munn, L. (2020). Angry by design: toxic communication and technical architectures. Humanities and Social Sciences Communications, 7, 53. https://doi.org/10.1057/s41599-020-00550-7.

Pasta, S. (2018). Razzismi 2.0. Analisi socio-educativa dell'odio online. Brescia, Morcelliana Scholé.

Pasta, S. (2019). Razzismi espliciti banalizzati. L'ambiente digitale e il 'ritorno della razza'", in Santerini M. (ed.), *Il nemico innocente. L'incitamento all'odio nell'Europa contemporanea. The Innocent Enemy. Hate incitement in contemporary Europe*, Milano, Guerini e Associati, 173-190.

Pasta, S. (2022). Social network conversations with young authors of online hate speech against migrants, in Monnier A., Boursier A., Seoane A (Eds.), Cyberhate in the Context of Migrations, London, Palgrave MacMillan, 187-214.

Pasta, S. (2023a). Discours de haine en ligne. Une analyse des tweets islamophobes entre automatismes et évaluation qualitative, in Karkun A., Jovelin E. (Eds.), *Vers une paix durable. Une perspective interculturelle*, Paris, Éditions du Cygne, 217-237.

Pasta, S. (2023b). Hate Speech Research: Algorithmic and Qualitative Evaluations. A Case Study of Anti-Gypsy Hate on Twitter, *REM. Research on Education and Media*, 15(1), 130-139.

Pasta, S. (2023). Tackling online hate speech with the involvement of targeted groups. The methodological proposal of the project REASON – REAct in the Struggle against ONline hate speech, QTimes. Journal of Education, Technology and Social Studies, XV(3), 429-445, doi: 10.14668/QTimes_15330.

Pasta, S. (2024a). Lo "spettro dell'odio online": una proposta di classificazione tra valutazioni algoritmiche e qualitative, in Crescenza G. (eds.), *Educare in tempi di odio e violenza. Sfide pedagogiche e istituzionali*, Bari, Progedit, 113-126.

Pasta, S. (2024b). Hate Studies tra logica computazionale e classificazione umana. Un caso studio sull'antisemitismo in Twitter, Scholé. Rivista di educazione e studi culturali, LXII (1), 230-252.

Pasta, S. (2025), *Countering Hate Speech in the Postdigital: A Challenge for 'Onlife Citizenship'*, in Gomez Paloma F., di Tore P.A., Mangione G.R.J. (eds.), *Teacher Training and Student Learning – Past Values, Present Uncertainties and Future Prospects*, IntechOpen, London, pp. 155-174, DOI: 10.5772/intechopen.1011481.

Pasta, S., Santerini, M., Forzinetti, E., & Della Vedova, M. (2021). Antisemitism and Covid-19 on Twitter. The search for hatred online between automatisms and qualitative evaluation, Form@re. Open Journal per formazione in rete, XXI, 3, 288-304,

Pasta, S., Rivoltella, P.C. (2022, Eds.), Crescere onlife. L'Educazione civica digitale progettata da 74 insegnanti-autori, Brescia, Scholé.

Quattrociocchi, W., & Vicini, A. (2023). Polarizzazioni. Informazioni, opinioni e altri demoni nell'infosfera, Milano, FrancoAngeli.

Rodota, S. (2004). Tecnopolitica. La democrazia e le nuove tecnologie della comunicazione, Roma-Bari, Laterza.

Santerini, M. (2005). Antisemitismo senza memoria, Roma, Carocci.

© Italian e-Learning Association

Santerini, M. (2021). La mente ostile, Milano, Raffaello Cortina.

Sellars, A. (2016). Defining Hate Speech, Berkman Klein Center Research Publication No. 2016-20, https://cyber.harvard.edu/publications/2016/DefiningHateSpeech.

Suler, J. (2004). The Online Disinhibition Effect. CyberPsychology & Behavior, 7(3), 321-326, https://doi.org/10.1089/1094931041291295.

van Dijck, J., Poell, T., & de Waal, M. (2018). The Platform Society: Public Values in a Connective World, Oxford, Oxford University Press, https://doi.org/10.1093/oso/9780190889760.001.0001

Waldron, J. (2012). The Harm in Hate Speech. Cambridge, Harvard University Press.

Ziccardi, G. (2015). Internet, controllo e libertà Trasparenza, sorveglianza e segreto nell'era tecnologica. Milano, Raffello Cortina.

Ziccardi, G. (2016). L'odio online Violenza verbale e ossessioni in rete. Milano, Raffello Cortina.

Normative and institutional documents

Council of Europe, Committee of Ministers (1997), Recommendation No. R(97)20 on "Hate Speech", Strasbourg.

Council of Europe, Committee of Ministers (2022), Recommendation CM/Rec(2022)16 on Combating Hate Speech, Strasbourg.

Council of Europe. (2022), Explanatory Memorandum to Recommendation CM/Rec(2022)16 on combating hate speech, Strasbourg.

Council of the European Union (2000), Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.

Council of the European Union (2000), Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation.

Council of the European Union (2008), Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.

European Parliament and Council (2022), Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act).

United Nations (1965), International Convention on the Elimination of All Forms of Racial Discrimination.

United Nations (2019), United Nations Strategy and Plan of Action on Hate Speech, New York.

Policy documents, case law, and essential web sources

Cornell Law School, Legal Information Institute, 47 U.S. Code § 230 - Protection for private blocking and screening of offensive material.

Discord (2025), Hateful Conduct Policy Explainer, https://discord.com/safety/hateful-conduct-policy-explainer.

European Commission (2026, 29 April), Commission preliminarily finds Meta in breach of Digital Services Act for failing to prevent minors under 13 from using Instagram and Facebook.

Gonzalez v. Google LLC, 598 U.S. (2023).

Human Rights Watch (2023), Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook, https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-on-instagram-and-facebook

Jigsaw. Supporting Online Conversations at Scale with AI, https://jigsaw.google/our-work/supporting-online-conversations/.

K.G.M. v. Meta Platforms, Inc., et al. (California jury verdict, March 2026).

Meta (2024), Actualizacion sobre nuestro abordaje al contenido politico en Instagram y Threads, https://about.fb.com/ltam/news/2024/02/actualizacion-sobre-nuestro-abordaje-al-contenido-politico-en-instagram-y-threads/.

Meta (2025), Ending Political, Electoral and Social Issue Advertising in the EU in Response to Incoming European Regulation. https://about.fb.com/news/2025/07/ending-political-electoral-and-social-issue-advertising-in-the-eu,

Moody v. NetChoice, LLC, 603 U.S. (2024).

TikTok (2025), Countering Hate Speech & Behavior. https://www.tiktok.com/safety/en/countering-hate/.

Twitch (2025), Hateful Conduct Education Resource, https://www.twitch.tv/p/safety/education-portal/en/articles/hateful-conduct/.

Twitter, Inc. v. Taamneh, 598 U.S. (2023).

YouTube (2025), Hate speech policy, https://support.google.com/youtube/answer/2801939.

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), pp. 27-36

Poeticising life. Art and art education as forms of resistance and elaboration of the logic of hate

Chiara Panciroli[a], Pier Cesare Rivoltella[b,1]

[a]*Alma Mater Studiorum Università di Bologna, Dept. of Education Sciences – Bologna (Italy)*

[b]*Alma Mater Studiorum Università di Bologna, Dept. of Arts – Bologna (Italy)*

(submitted: 02/02/2026; accepted: 18/05/2026; published: 24/5/2026)

Abstract

This article analyses the role of art and art education as practices of poeticising existence capable of countering the logic of hate and violence in contexts marked by alienation, trauma and social disharmony. Drawing on the theoretical contribution of Michèle Petit, in dialogue with Gaston Bachelard's poetics of *rêverie* and Hartmut Rosa's sociology of resonance, the article interprets hate as an expression of a silent and discordant relationship with the world and with others. In this context, art is taken as a device of resonance, capable of reactivating speech, imagination and the ability to attribute meaning to experience. From a methodological point of view, the research adopts a theoretical-interpretative approach integrated with case studies, identifying and analysing artistic and educational practices operating as forms of prevention, cultural resistance and therapy. The cases discussed – from the pedagogical atelier, artivism and street art, to artistic practices of trauma processing in war contexts – show how art can transform wounded objects, places and materials into symbolic spaces of reconciliation and healing. Importance is given to the role of the artist and art educator as *a passeur*: a cultural mediator who does not transmit predefined contents but creates the conditions for a symbolic transition towards forms of subjectivation, emancipation and the construction of bonds. The article concludes by arguing that the poeticization of existence today represents a fundamental educational and cultural resource for imagining and practising a culture of peace.

KEYWORDS: Poeticization of Existence, Art Education, Resonance, Hate and Violence Contrast, Culture of Peace.

DOI
https://doi.org/10.20368/1971-8829/1136367

CITE AS
Panciroli, R., & Rivoltella, P.C. (2026). Poeticising life. Art and art education as forms of resistance and elaboration of the logic of hate. *Journal of e-Learning and Knowledge Society*, *22*(1), 27-36. https://doi.org/10.20368/1971-8829/1136367

1. Introduction. Hate and violence as a language of disharmony and alienation

Michèle Petit is an anthropologist who collaborates with the Laboratoire Dynamiques Sociales et Recomposition des Espaces (LADYSS) of the CNRS in Paris. In her work, she has often focused on the relationship between the possibility of speech, appropriation of the world and traumatic experiences of hate and violence, which deny the practicability of this relationship.

Petit cites several examples from her research. One such example is Irene Vasco, who has been working for years in Colombia in suburbs that are prey to drug trafficking, or in communities of exodati (Colombia has almost eight million internally displaced persons). Another example is that of Catherine Meurisse, who survived the attack on the Charlie Hebdo editorial office. On 7 January 2015, Catherine left home very late and when she arrived outside the editorial office, the terrorist commando was already at work: together with her colleague Luz, she heard the shots from the Kalashnikovs and witnessed from outside a massacre in which she could have been a victim; in the following days, she found herself in a state of dissociation, traumatised and consumed by guilt for not having died too.

Michèle Petit also reinterprets Mira Rothenberg's (1979) account of a group of Jewish children from Eastern Europe who survived the Holocaust and found themselves in the United States after the war, in a foreign country, without their families, struggling with

[1] Corresponding author – email: pier.rivoltella@unibo.it – address: Via Barberia 4, 40123, Bologna (IT).

© Italian e-Learning Association

a new language. Rothenberg's book is emblematically titled "Des enfants au regard de pierre" (Children with Stone Eyes), because the gaze of these 32 children was "stone-like", unable even to verbalise anything: words had become uninhabitable for them, language had been expropriated from them. Rothenberg thus began to tell the children the story of the American Indians, how they too had been dispossessed of their country, and she played the children some poems composed by Native American poets. The children became fascinated. They cleared the classroom of desks and chairs, set up tepees, and drew a river and vegetation on the floor. Art, in the form of poetry and painting, allowed these children to understand what they had been through and unlocked their thoughts and words, which had been blocked by the traumas they had suffered.

This is where our proposal comes from. We are all experiencing an increasingly silent world – a world that, at best, does not respond to us and, at worst, turns against us, as demonstrated by the climate chaos of recent years – and others who are increasingly self-centred, disenchanted, alienated, if not aggressive, hostile and violent. The question – and the working hypothesis we formulate – is whether art can function in this context as an opportunity for contrast and redemption. As an opportunity for contrast: because the logic of hate and violence can be disarmed through art. As an opportunity for redemption: because it is not enough to counter hate, we need to build peace.

We will proceed in three steps.

In the first part (§2), we will explain the methodological approach we used, consisting of theoretical research and case studies. In this first step, we formulate our hypothesis, borrowing from the reflections of Hartmut Rosa (2010; 2019). It consists in believing that our society is characterised by the experience of alienation, or rather by a relationship with the world and with others in which they remain silent. This experience can be rendered, using a musical metaphor, by saying that the world and others no longer resonate: the result is a discordant relationship, a loss of harmony; we are no longer in tune with the world and with others. This situation of alienation can only be countered by (re)activating one or more axes of resonance. One of these axes is certainly art. In the experience of resonance, the world and others respond to us: harmony is restored, and a vibration in unison occurs.

The second step (§3) will consist of identifying some case studies that can confirm the hypothesis we have suggested. Identifying them will allow us to complete our theoretical proposal by suggesting three main ways in which art can counter hate and violence by restoring a resonant relationship with the world and with others. These are: (1) the atelier as a pedagogical device for preventing violent acts by building correct civic behaviour; (2) artivism, in particular street art, as a form of resistance and cultural transmission to promote a logic of respect instead of hate; (3) the use of art as a form of therapy, both in cases where the target is victims of hate and in cases where, thanks to art, the aim is the redemption of violent individuals.

The third and final step (§4) is the discussion of the results of the case studies. Their analysis – the outcome of which is the first hint of a phenomenology of forms of art-intervention to combat hate and violence – will lead us to question their distinctive features and propose. The final outcome will be, as a provisional conclusion, a focus on the art-educator as a key figure in a new social pedagogy.

2. Materials and methods. Re-harmonising the world through art

2.1. Time of crisis and the search for a harmonious composition with life

From a methodological point of view, this article opts for a theoretical research approach – proposing an interpretative hypothesis that in the future can guide empirical research – and tests its validity through comparison with practices by means of some case studies.

Theoretical – or speculative, or nomothetic (Van der Maren, 2003) – research is a mode of investigation whose primary object is the elaboration, analysis and conceptual clarification of theories, categories and interpretative frameworks, rather than the direct collection of empirical data (Babbie, 2020). It is based on methodologically controlled procedures such as critical analysis, argumentative reconstruction, comparison of theoretical models and, sometimes, genealogy of concepts, with the aim of highlighting implicit assumptions, internal consistencies, epistemological limitations and heuristic potentialities of the theories examined. From this perspective, theoretical research is therefore not a form of free speculative reflection, but secondary scientific work that produces knowledge through the organisation, interpretation and synthesis of already available theoretical contributions (Flick, 2018). The validity of theoretical research does not depend on immediate empirical verifiability, but on the transparency of the criteria for selecting sources, the coherence of the argument, the explicitness of analytical choices and, above all, the ability to generate conceptual models useful for guiding empirical research and professional action (Bryman, 2016). In the educational and social fields – as is the case with this work – theoretical research plays a foundational and guiding role, clarifying the constructs, languages and epistemological frameworks within which empirical studies are situated (Van der Maren, 2003).

In this regard, our research focused on a family of multidisciplinary theories having in common the awareness that our times are times of crisis marked by

© Italian e-Learning Association

acceleration, depersonalisation and fear, and by the idea that this crisis can be countered by the ability to live life harmoniously. These theories include: the idea of "moments of happiness despite everything" (Augé, 2018), the economy of sufficiency (Sachs, 1998, 2023; Bailly, 2015), the revival of the contemplative dimension of life (Han, 2023), positive psychology and in particular the concepts of flow (Csikszentmihalyi, 1990) and "moments of light" (Shaeffer, 2015), the revival of the concept of wonder (Chirico & Gaggioli, 2018) and the issue of the sublime and catharsis (Blumemberg, 1997), finally aesthetic, neuroscientific and social reflection on the concept of resonance (Petit, 2023; Vessel, Star & Rubin, 2013; Cheng, 2006). Critical analysis of this literature has led us to isolate in particular the concept of resonance as focused on by German sociologist Hartmut Rosa (2010; 2019): starting from this concept, we will try to show how art can function as a device of resonance to re-establish man's relationship with the world and with others, previously compromised by the logic of hate and violence.

2.2. A phenomenology of forms of art-intervention

The case study is a qualitative research strategy aimed at the in-depth and contextualised analysis of a complex phenomenon, considered in its specificity and natural environment (Yin, 2018). Rather than aiming at statistical generalisation, this method aims at an intensive understanding of the processes, dynamics and meanings that characterise a single case or a limited number of cases, understood as systems delimited in time and space (Stake, 1995). The researcher collects data through multiple sources – such as interviews, observations, documents, artefacts and digital traces – adopting a triangulation logic that reinforces the credibility of the analysis. The case study – which may have exploratory, descriptive or explanatory purposes – is particularly suited to the investigation of emerging phenomena, educational innovations or professional practices, where the boundaries between the phenomenon and the context are not clearly separable (Merriam, 1998). Validity does not lie in the representativeness of the sample, but in the analytical richness, consistency of the research design and transparency of the interpretative procedures, which allow the reader to assess the transferability of the results to other similar contexts (Creswell, 2013).

In our research, the identification of cases to be explored in depth is the result of a three-pronged exploratory investigation: firstly, a review of the literature to map the ways in which artivism (Trione, 2022) and media artivism (Medrado & Rega, 2023) have addressed the problem of hate and violence in order to defuse it; secondly, analysis of the web in search of emblematic experiences worthy of being translated into cases to support our thesis; thirdly, the analysis of creative materials available in the MOdE (Museo Officina dell'Educazione), a virtual museum that is part of the University of Bologna's Museum System and which has already become a laboratory for the creation of workshops designed according the perspective adopted in this article.

3. Results. The ways in which art defuses the logic of hate and creates spaces for resonance

3.1. "Don't stop the music!". Re-tuning with nature and with others

From the analysis of the above-mentioned literature (§ 2.1), art constitutes a privileged space of resonance, in the sense that Hartmut Rosa attributes to the term: a non-instrumental, non-appropriative relationship, capable of responding to and transforming those who experience it. The metaphor of resonating, of being in tune, also recurs in other contributions. Michèle Petit (2023, p. 28) observes that

> *what we find in literature and in contact with works of art from an early age is perhaps, first and foremost, at least for a short period, the possibility of feeling in harmony, in the musical sense of the term, with our surroundings', responding to a deep need for harmony between the external world and the inner world. 'From an early age, children [...] need to experience harmony, to feel in tune with their surroundings and, at the same time, with their inner world (Petit, 2023, p. 52).*

[*From now on, all the quotations from non-English written sources, are translated in English by the authors*].

This insight is reflected both in neuroscience – think of Vessel, Starr and Rubin (2013, p. 2), who observe how beauty promotes the fact that "our brain registers a certain "harmony" between the external world and our inner representation of ourselves" – and in developmental psychology: Daniel Stern, for example, describes the "primary affective attunement" between mother and newborn by comparing the gestures that promote it to the moment when musicians in an orchestra tune their instruments.

Art thus appears as a practice of re-tuning the relationship, a movement that opposes the logic of exploitation and waste in order to become, as Bailly says (2015, p. 201), "a movement towards things that is no longer one of exploitation, profit and waste, but one of gaze, respect, wonder, the idea of attention, of vigilant, sensitive listening". In this direction, Gaudin's idea of inhabiting also refers to a work of reabsorbing the discord and strangeness that separate us from the world, restoring a space of closeness to things: inhabiting means

© Italian e-Learning Association

> *weaving all sorts of things around us to make them friends, to feel less indifference towards them. Inhabiting means giving things a place in a space of closeness. To reabsorb the sense of strangeness that distances us from what is outside ourselves. To try to escape the mental disorientation caused by a feeling of discord that we identify with the outside world (Gaudin, 1996, p. 22).*

Shaeffer (2015, p. 310) is on the same wavelength:

> *In every culture, human beings have been able to draw on a set of mental resources whose origins date back to the roots of life, in order to process (discontinuous) experiences that give rise to moments of light, in which everything seems to realign itself with simplicity and naturalness, leaving no room, for a few moments, for problems or anxieties. It is access to these moments of happy immanence that we in the West have been accustomed to identifying (for three centuries now) with the name "aesthetics".*

It is precisely in these moments that what Cheng calls "resonating" takes place: with oneself, with others, with living beings, according to a law of the soul that coincides with not impeding music.

> *Resonating, that is the right word. Resonating within oneself, resonating with the "basso continuo" of another, or with that of the living universe – this is what makes it possible to be immortal. "To sing is to be", writes Rilke. Is there another law for the soul outside of this: "Do not stop the music"? (Cheng, 2006, p. 49).*

From this perspective, only art seems capable of remedying our discordant relationship with the world and with others, because only art opens a space of resonance in which the world is no longer mute or passively available, but capable of responding to us.

The fruitfulness of this interpretative hypothesis will be tested in the next paragraph of this article through some case studies.

3.2. The *atelier*: pedagogical art as a form of prevention/redemption

A pedagogical interpretation of the meaning of spaces as tools for interpretation can help us understand how an atelier can contribute to countering the logic of hate. Specifically, we refer to the pedagogical atelier as an intentionally designed, dynamic, and transformable environment, oriented towards observation, experimentation, and the reworking of experience through a plurality of expressive languages (Panciroli, 2016). In line with this perspective, the educational space cannot be reduced to a purely functional dimension but rather takes the form of an "existential context" with a strong symbolic and relational value (Iori, 1996).

If we want to identify some educational characteristics of the atelier, we can first highlight a symbolic and interpretative dimension, also considering the most recent contributions on the cultural function of art and aesthetic devices. The atelier space can, in fact, be understood as a complex semiotic environment (Panciroli, 2023) acting as a device to produce meaning, in which artistic languages operate as cognitive and emotional mediators (Panciroli, 2016). In this regard, Mirzoeff (2015) emphasizes how contemporary visual culture is never neutral but constitutes a field of forces in which processes of power, inclusion, and exclusion are played out. By offering the possibility of deconstructing and reconstructing images, the studio allows us to think to visuality not in terms of immediacy and simplification, but as the subject of critical interpretation. Similarly, Mitchell (2017) highlights how images should not be understood only as representations, but as "acts" that produce effects in reality: the atelier thus becomes a space in which these visual acts can be questioned, transformed, and re-signified, countering the symbolic rigidity that fuels the logic of hate.

A second dimension is the experiential and aesthetic one, which can be understood through the contributions of scholars who have highlighted the role of aesthetic experience in transformative learning processes. Illeris (2018) points out that deep learning occurs when the cognitive, emotional, and social dimensions are integrated into a meaningful experience. The atelier, as a space for sensitive and embodied experimentation, promotes this integration, allowing individuals to learn through body, emotions, and imagination. This perspective refers to a teaching method valuing aesthetic experience as a form of knowledge that cannot be reduced to measurable outcomes but is essential for the formation of critical thinking. In this sense, the atelier teaches us to suspend immediate judgment and to slow down, which is also necessary for the symbolic processing of conflict.

A further dimension of the atelier is the relational and social one: Wenger-Trayner & Wenger-Trayner (2020) describe social environments as spaces for negotiating meanings in which communities are built through shared practices (Rivoltella & Panciroli, 2019). The atelier, as a non-hierarchical and open space, encourages such practices, promoting forms of participation that value the plurality of viewpoints. According to Biesta (2021), authentic education is not limited to the transmission of skills but concerns the possibility for individuals to "come into the world" as responsible beings in relation to others. The atelier, therefore, precisely because it is based on processes of co-construction and comparison, becomes a space in

© Italian e-Learning Association

which the educational experience is expressed as a practice of symbolic coexistence, capable of counteracting the dynamics of exclusion and polarization.

Finally, the ethical-political dimension of the atelier as a space of cultural resistance can be further explored considering contemporary reflections on the educational role of art in contexts marked by conflict and symbolic violence. Nussbaum (2018) emphasizes how the arts play a crucial role in the development of moral imagination, understood as the ability to recognize the humanity of others.

Similarly, Butler (2020) highlights how cultural practices can open spaces of shared vulnerability, opposing the logic of dehumanization. In this perspective, the atelier is not only a place of expression, but also a political device in a pedagogical sense, educating people to take responsibility for and care for the common world. The atelier, as an intentional pedagogical space, embodies this possibility: a place where artistic experience becomes an exercise in humanization, capable of countering the symbolic impoverishment that fuels hatred and opening spaces of resonance, care, and shared responsibility.

In terms of size, the artistic-pedagogical atelier is configured as a symbolic and relational space capable of bringing about a real suspension of the logic of hate, acting according to two perspectives: preventive and reparative/redemptive. In "Nous sommes des animaux poétiques", Petit (2021) argues that human beings, even before being rational or social subjects, are narrative and poetic subjects who construct meaning through images, metaphors, stories, and symbolic gestures. It is precisely this poetic dimension that, when denied or impoverished, leaves room for forms of identity rigidity, emotional closure, and violence. The atelier, understood not simply as a place of artistic production but as an educational and cultural device, reactivates this dimension, offering a protected environment in which aesthetic experience allows complex emotions, inner conflicts, and experiences of exclusion to be processed without resorting to the language of aggression. In this context, art in the atelier can take on two distinct functions.

In a preventive sense, pedagogical art acts on individuals who have not committed acts of violence, but who may be exposed to dynamics of hate, radicalization, or dehumanization of the other. Through artistic creation, imagination is trained as an ethical faculty: the possibility of imagining other worlds, different points of view, and alternative narratives interrupts the simplifying mechanisms of hatred, which is based on reducing the other to a stereotype or enemy. As Petit points out, engaging with poetic experiences allows us to "inhabit nuances" (Petit, 2021, p. 104), to tolerate ambiguity and to recognize our own vulnerability, which are essential elements for non-violent coexistence.

From a redemptive perspective, the workshop takes on a reparative function for those who have already committed violence. In these contexts, art does not operate as a moralizing or corrective tool, but as an opportunity for self-renarration. The artistic gesture becomes an act of symbolization that allows the destructive experience to be transformed into expressive material, giving the subject the opportunity to recognize themselves as an author and not just as a culprit. According to Petit, it is precisely the poetic word – understood in a broad sense, as a symbolic form – that allows a distance from the violent gesture and reopens a space for humanization, in which the subject can return to thinking of themselves in relation to others. In both perspectives, preventive and redemptive, the atelier is therefore configured as a space of resonance: a place where individual experiences find an echo in shared forms, where art defuses the binary logic of hate and promotes processes of mutual recognition. In this sense, the poeticization of life does not represent an escape from reality, but an educational practice of cultural resistance, capable of countering the symbolic impoverishment that fuels violence and restoring a deeply humanizing function to education.

In this way, the pedagogical atelier becomes a device of resonance through which individuals can reconnect with themselves and with others, transforming experiences of closure, aggression and symbolic violence into possibilities for dialogue, recognition and peaceful coexistence.

3.3. Artivism: street art as a form of cultural resistance

If the atelier represents an intentional pedagogical space in which the "poeticization of experience" allows conflict to be processed and the logic of hate to be defused, artivism – and street art in particular – constitutes the extension of this educational stance into the public space. As Trione (2022) points out, artivism cannot be understood as a simple militant expression of art, but as a heterogeneous constellation of practices that bring art "back into the dynamics of history", removing it from pure contemplation and transforming it into a critical exercise of the gaze. In this perspective, street art takes the form of visual counter-information capable of intervening directly in everyday places, re-semantising urban spaces and turning them into "gyms of thought and images" (Bonfante, 2022).

The value of artivism lies in its ability to break the apparent neutrality of public space, making conflicts, inequalities, and symbolic removals visible. As Trione (2022) points out, street artists do not merely bear witness, but "intervene concretely in reality", making visibility a political act that provokes reactions, tensions, and stances. In this sense, street art acts as a practice of cultural resistance to the logic of hatred, as it interrupts narrative simplification and restores

© Italian e-Learning Association

complexity to reality, forcing passersby to confront images that destabilize, question, and sometimes disturb. This function is particularly evident when street art is incorporated into educational and museum settings capable of enhancing its interpretative dimension.

This is the case of MOdE, the Museum Workshop of Education at the University of Bologna, which has integrated the language of street art into cultural and educational mediation programs. In particular, the use of murals and anamorphic illusions has made it possible to reflect on the relationship between perception, space, and meaning: the image, when viewed from a privileged point of view, recomposes itself, revealing a recognizable form, while from other angles it appears fragmented or indecipherable.



Figure 1 - A pedagogical atelier in MOdE.

This visual device becomes a powerful educational metaphor for social and cultural complexity, showing how meaning is never given once and for all, but depends on the gaze, position, and experience of the observer. In this sense, the experience offered by MOdE is fully in line with artivism as a pedagogical practice: anamorphic illusions and murals do not communicate unambiguous messages, but activate processes of interpretation, inviting subjects to question their own point of view and dominant narratives. Street art thus becomes an educational tool capable of countering the logic of hatred, which is based precisely on the rigidity of the gaze and the impossibility of recognizing the plurality of points of view.

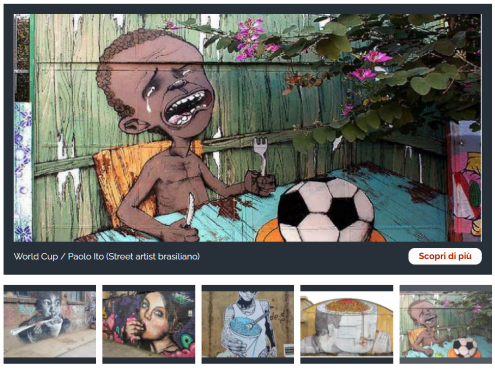


Figure 2 - Street art workshops in MOdE.

The connection to the meaning of the pedagogical atelier is obvious: like the atelier, artivism also poeticizes public space, transforming walls, facades, and urban surfaces into places of symbolic resonance. However, while the atelier works mainly on a protected and reflective level, street art exposes the poetic gesture to the risk of conflict, making it visible and shared. This exposure is where its educational power lies: art does not shy away from reality, but traverses it, offering symbolic forms that allow violence to be processed without reproducing it. Artivism, understood in this way, takes the form of a practice of cultural resistance which, through image and space, restores to the collective experience the possibility of thinking, feeling, and inhabiting the world outside the logic of hatred. Alongside the recognition of the transformative potential of artivism, it is nevertheless necessary to take a critical look at the risks inherent in politically engaged artistic practices, as explicitly pointed out by Trione himself (2022). One of the main dangers is moralism, or the tendency to produce works based on ethical assumptions that are apparently acceptable to everyone, but which end up reducing the complexity of reality to a simplified opposition between good and evil. In such cases, the work risks becoming a *prescriptive pedagogical* device, closer to the rhetoric of indignation than to an authentic critical exercise of the gaze.

A second, closely related risk concerns the aestheticization of conflict and suffering. Some artivist practices can fall into forms of spectacularization of pain, producing powerful images that, instead of activating processes of understanding and elaboration, end up numbing the viewer's sensitivity. This dynamic, already denounced by Susan Sontag in relation to images of war, leads to an immediate and consumable emotional experience, in which trauma becomes an aesthetic object and conflict loses its historical and political dimension. Bonfante (2022) emphasizes how this shift can transform artivism into a "ready-to-wear" practice, easily absorbed by the art system and the cultural market, neutralizing its subversive scope.

A further critical element concerns the ambiguous relationship between artivism and the art system. Many street artists and politically engaged artists, while declaring themselves antagonists, operate within the institutional and media circuits they intend to criticize. This "apocalyptic and integrated" condition, to use a formula made famous by Umberto Eco, raises questions about the real capacity of artivism to produce change and not just symbolic consensus. The risk is that the work will merely confirm the convictions of an already sensitized public, reinforcing a homogeneous moral community without affecting the deep structures that generate exclusion, violence, and hate. These critical issues take on decisive pedagogical significance. If artivism wants to establish itself as a form of cultural resistance to the logic of hatred, it must

© Italian e-Learning Association

avoid replacing the latter with a new symbolic rigidity based on indignation or unilateral denunciation. The most effective artistic practices are not those that "drag politics into art", but those that look at politics through the lens of art, maintaining a critical distance that allows the work to remain open, ambiguous, and questioning. It is in this openness that artivism can retain an authentic poetic and pedagogical quality: not as an instrument of moral persuasion, but as a device of thought capable of generating questions, undermining certainties, and making thinkable again what the logic of hate tends to simplify and close off.

By re-signifying public spaces and activating critical and emotional participation, artivism restores forms of collective resonance capable of interrupting the symbolic simplifications on which hate and exclusion are based.

3.4. Art as a form of therapy

A final way in which the logic of hate can be countered is using art as a form of therapy, in at least two senses. In the first sense, as Picasso suggested, "art shakes the dust of everyday life from the soul": awakening and catharsis at the same time, the possibility of emancipation from a conformist and anaesthetised passivity. The second sense is suggested by the President of the *Videoinsight* Foundation, Rebecca Russo**,** collector, psychotherapist and patron: "Art causes insight, awareness, inner vision, intuition, solution. Art is revolution. It belongs to human nature, expresses primary needs, is profound, urgent. It is circular, immortal, universal". It is not only an opportunity to look at things differently, but also a way to change them: this is the case of the collective exhibition *Hater*, hosted in Turin in 2023 by the *Videoinsight* Foundation, which invited 23 artists to exhibit their works aimed at countering the logic of online verbal violence through art; or that of the Künstlerhaus, Halle für Kunst & Medien (KM-) in Graz, which in 2019 brought together a group of sixteen international artists in the collective exhibition *Hate Speech: Aggression and Intimacy*, curated by Sandro Droschl.

The case we have chosen to highlight to demonstrate the therapeutic value of art in relation to hate and violence is that of some Israeli and Palestinian artists. Their work is born out of close contact with the experience of war and aims to help process grief and suffering.

Seagal Hagege is an American Jew who decided to make *aliyah* and move to Israel after COVID. Here she dedicates herself to providing psychological support to reservists, shocked by war and what they had seen, done and suffered during it. It is in this context that the *Helmets 4 Chayalim* project was born, in which Seagal asks artists to create works of art from old soldiers' helmets. *Dr. H*, an artist known for his urban graffiti style, paints lemons on the helmet (Figure 3): "It's the lemon tree in my mother's garden. The one that kept me grounded every time I had to go to war. Home, family, security".



Figure 3 - DrH: *Helmet with lemons*.

Another artist, Ayala, covers her helmet with rose petals (Figure 4) made from letters written by soldiers to their families in 1960. The idea is to use art as a storytelling tool to process the experience of war from the perspective of personal feelings and experiences.



Figure 4 - Ayala: *Helmet with rose petals*.

Tomer Peretz is an Israeli artist who lives and works in California, after serving in the Israeli army. Directly linked to the experience of war is his collection War Is Over: unique garments made from discarded military jackets enriched with multimedia elements (Figure 5).



Figure 5 - Tomer Peretz: a jacket of *War is over* collection.

© Italian e-Learning Association

Each jacket carries a chip that tells the true story of the person who wore it: the wearer can use their smartphone to access the information contained in the jacket by listening to audio files, reconstructing the history of the jacket and looking at photographs. It is a conscious and powerful fashion project that aims to transform the uniform – a symbol of conflict – into a testimony of change, identity and peace.

Jonathan Kis-Lev, born to parents who immigrated to Israel from the former Soviet Union, is a young street artist who uses art as a symbolic healing practice for places wounded by conflict. With his coloured spray cans, he transforms the walls of Bethlehem and the occupied territories into spaces for emotional reworking and reconciliation, veritable canvases of peace. The *Naïve Series*, painted with his left hand to evoke childlike spontaneity, gives shape to smiles, animals and gestures of affection: simple images that act as therapeutic devices, capable of reactivating empathy, imagination and trust (Figure 6).



Figure 6 - Jonathan Kis-Lev: *Naïves series*.

It is an art that reminds us how drawing, like play, can become an act of non-violent resistance and an exercise in possible futures. Kis-Lev's collaboration with Palestinian artists in the West Bank reinforces this perspective, showing how shared creativity can open spaces for dialogue and healing where politics struggles to reach, contributing to the construction of a culture of peace based on encounter and mutual recognition.

Assaf al Kherty recovers scrap metal, abandoned tools and war debris scattered in the fields, transforming what was once an instrument of destruction into a material for healing and symbolic recomposition (Figure 7). His powerful and essential sculptures act as collective therapeutic devices: everyday objects – a suitcase, a drill bit, a school backpack – are removed from the logic of war and returned to a human narrative, capable of giving shape to the lives, wounds and hopes of young people in the Strip. In this process of creative reuse, scrap metal becomes the supporting structure of new stories and a vehicle for resilience, allowing trauma to be processed and a different future to be imagined. Al Kherty's works thus take the form of small monuments to a possible peace: signs of a culture that, even in a territory marked by blockade, poverty and destruction, entrusts art with the task of rebuilding bonds, meaning and hope.



Figure 7 - Assaf al Kherty at work.

Through these artistic practices, objects, bodies and places marked by war and hate are transformed into spaces of resonance, where trauma can be symbolically elaborated and the possibility of empathy, reconciliation and non-violent coexistence can emerge again.

4. Discussion and conclusion. The artist and art educator as *passeur*

The materials and cases discussed in this article clearly show how art and art education can be understood as practices of poeticising existence (Petit, 2023), capable of countering the logic of hate and violence not only on a symbolic level, but also on a relational and therapeutic one. In contexts marked by alienation, trauma and loss of resonance with the world, art reactivates what Bachelard (1971) defines as *rêverie*: a psychic and cultural space in which experience becomes habitable, narratable and transformable once again. The cases analysed – from the pedagogical atelier to street art, to artistic practices for processing trauma in war contexts – show how artistic creation makes it possible to restore speech where violence has rendered it impracticable, to re-establish the relationship between the inner and the outside world, to transform objects, places and wounded memories into devices of resonance. In this sense, art does not function as a simple representation of pain, but as a gesture of care and resistance, capable of producing bonds, empathy and mutual recognition: an educational, cultural and therapeutic response to the disenchantment of the world.

The artist and art educator then take on the role of *passeur*, of ferrymen, accompanying individuals and communities towards spaces where it is possible to inhabit the world poetically, give shape to emotions, process trauma and imagine alternatives to the present.

The figure of *the passeur*, as it emerges in the thinking of Michèle Petit (2023), designates people who does not transmit content or impose meanings, but creates the conditions for the other to cross a threshold: towards symbolic language, imagination and poetic speech. The *passeur* does not guide from above, but accompanies, making accessible a space where everyone can

© Italian e-Learning Association

reconfigure their experience and reclaim their voice. In this perspective, the *passeur* is the one who guards and opens, rather than the one who explains.

This role finds an important resonance in the thinking of the French philosopher Jacques Rancière (1991), particularly in the figure of *the maître ignorant*, who breaks down the hierarchy between those who know and those who do not, basing the act of education on the principle of the equality of intelligences. For Rancière, too, emancipation does not come about through explanation, but through the activation of an autonomous movement of thought. The *passeur* shares this logic: he does not lead towards predefined knowledge but makes possible an experience of crossing in which the individual becomes the subject of his own learning and transformation. In this sense, the *passeur* operates a politics of sensitivity, redistributing the "partage du sensible" and opening up spaces in which art and words can escape the logic of domination to become instruments of care: a slow and profound work on the imagination, capable of rebuilding bonds, generating empathy and reactivating hope, even - and above all - in contexts marked by fracture, violence and loss.

Authors'contribution

Chiara Panciroli and Pier Cesare Rivoltella: Data curation, Original draft preparation, Reviewing, Writing (§ 4). Chiara Panciroli: Writing (§§ 3.2, 3.3). Pier Cesare Rivoltella: Conceptualisation, Methodology, Writing (§§ 1, 2, 3.1, 3.4).

References

Augé, M. (2018). *Bonheurs du jour*. Paris: Albin Michel.

Babbie, E. R. (2020). *The practice of social research*. Boston: Cengage.

Bachelard, G. (1971). *The poetics of reverie: Childhood, language, and the cosmos*. Boston: Beacon Press.

Bailly, J.-C. (2015). *L'élargissement du poème*. Paris: Bourgois.

Biesta, G. (2021). *World-centred education*. London: Routledge.

Blumenberg, H. (1997). *Shipwreck with spectator: Paradigm of a metaphor for existence*. Cambridge, MA: MIT Press.

Bonfante, L. (2022). *Artivismo: arte & politica*. Doppiozero. Retrieved in internet: https://www.doppiozero.com/artivismo-arte-politica

Bryman, A. (2016). *Social research methods*. Oxford: Oxford University Press.

Butler, J. (2020). *The force of nonviolence*. London: Verso.

Camus, A. (1951). *L'homme révolté*. Paris: Gallimard.

Cheng, F. (2006). *Cinq méditations sur la beauté*. Paris: Albin Michel.

Chirico, A., & Gaggioli, A. (2018). Awe: More than a feeling. *The Humanistic Psychologist*, 46(3), 274–280. https://doi.org/10.1037/hum0000098

Creswell, J. W. (2013). *Qualitative inquiry and research design*. Thousand Oaks, CA: Sage.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row.

Flick, U. (2014). *An introduction to qualitative research*. London: Sage.

Flick, U. (2018). *Doing qualitative data analysis in education*. London: Sage.

Gaudin, H. (1996). Embrasure. *Cahiers de la ville Gillet*, 5, 19–31.

Han, B.-C. (2023). *Vita contemplativa: In praise of inactivity*. Hoboken, NJ: John Wiley & Sons.

Illeris, K. (2018). *Contemporary theories of learning*. London: Routledge.

Medrado, A., & Rega, I. (2023). *Media activism, artivism and the fight against marginalisation in the global south: South-to-south communication*. London: Routledge.

Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco: Jossey-Bass.

Mirzoeff, N. (2015). *How to see the world*. London: Pelican.

Mitchell, W. J. T. (2017). *Image science*. Chicago: University of Chicago Press.

Nussbaum, M. (2018). *The monarchy of fear*. Oxford: Oxford University Press.

Panciroli, C. (2010). Il modello di Museo Virtuale dell'Educazione dell'Università di Bologna. *Ricerche di Pedagogia e Didattica*, 5(1), 1–15.

Panciroli, C. (2016). *El MOdE como un espacio de investigación expresiva: el ejemplo de los atelier*. In *Entornos informales para educar en artes*. Valencia: Publicacions de la Universitat de València.

© Italian e-Learning Association

Panciroli, C. (2021). *Elementi di didattica postdigitale.* Bologna: Bologna University Press.

Petit, M. (2023). *Nous sommes des animaux poétiques. L'art, les livres et la beauté par temps de crise*. Auxerre: Éditions Sciences Humaines.

Rancière, J. (1991). *The ignorant schoolmaster: Five lessons in intellectual emancipation*. Stanford: Stanford University Press.

Rivoltella, P. C., Panciroli C. (2019). *Visual Media Education. Audiovisivi e formazione nella società postmediale.* Milano: Sociologia della comunicazione, Franco Angeli.

Rosa, H. (2010). *Alienation and acceleration: Towards a critical theory of late-modern temporality* (Vol. 3). Aarhus: Aarhus University Press.

Rosa, H. (2019). *Resonance: A sociology of our relationship to the world*. Cambridge: Polity Press.

Rothenberg, M. (1979). *Des enfants au regard de pierre*. Paris: Seuil.

Sachs, W. (1998). Rich in things, poor in time. *Resurgence*, 196. Retrieved in internet: https://www.academia.edu/35979808/Rich_in_things_poor_in_time

Sachs, W. (2023). Sufficiency. In *Handbook of the Anthropocene: Humans between heritage and future* (pp. 799–803). Cham: Springer International Publishing.

Shaeffer, J.-M. (2015). *L'expérience esthétique*. Paris: Gallimard.

Sontag, S. (2003). *Regarding the pain of others*. New York: Farrar, Straus and Giroux.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Stern, D. N. (2010). *Forms of vitality: Exploring dynamic experience in psychology, the arts, psychotherapy, and development*. Oxford: Oxford University Press.

Trione, V. (2022). *Artivismo. Arte, politica, impegno*. Torino: Einaudi.

Van der Maren, J. (2003). *La recherche appliquée en pédagogie: Des modèles pour l'enseignement*. Bruxelles: De Boeck.

Vessel, E. A., Starr, G. G., & Rubin, N. (2013). Art reaches within: Aesthetic experience, the self and the default mode network. *Frontiers in Neuroscience*, 7, 258, 1–9.

Wenger-Trayner, E., & Wenger-Trayner, B. (2020). *Learning to make a difference*. Cambridge: Cambridge University Press.

Yin, R. K. (2018). *Case study research and applications: Design and methods*. Thousand Oaks, CA: Sage.

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), pp. 37-48

Politically motivated hate speech in the Albanian language social media

Rafaela Marteta[1]

Academy of Sciences of Albania, Institute of Linguistics and Literature, Dept. of Lexicology, Terminology and Applied Linguistics – Tiranë (Albania)

(submitted: 4/3/2026; accepted: 23/5/2026; published: 4/6/2026)

Abstract

The research addresses politically motivated hate speech in Albanian-language triggered by recent political events. An evident reactive hate dominates the comment section of political news outlet in social media, so that one is moved to detect the underlying partisan motivations of this hostility. Some comments provide answers and others suggest reasons that reflect a spreading animosity towards politics and its traditional or emerging actors. Robert Entman's frame paradigm offers a theoretical foundation to detect hate speech embedded in mediated interaction in digital platforms. This research combines framing with interpretative insights derived from critical discourse analysis to capture deeper causal logics embedded in frames as a communicating text. Findings reveal that politically motivated hate speech is often normalized and legitimized in digital interaction when it is frames around some thematic issues such as national identity, moral values, economic development. Hate speech targets include political actors and institutions, collective outer groups in some cases. Recurring narrative patterns highlight responsibility attribution and negative moral evaluation, which reinforce shared perceptions about political legitimacy, and underlying legacy. The research suggests that a context-sensitive approach when addressing hate speech detection helps to understand mechanisms of its construction in order to design preventive initiatives.

KEYWORDS: Albanian Language, Social Media, Hate Speech, Frame, Politics.

DOI
https://doi.org/10.20368/1971-8829/1136315

CITE AS
Marteta, R. (2026). Politically motivated hate speech in the Albanian language social media. *Journal of e-Learning and Knowledge Society*, *22*(1), 37-48. https://doi.org/10.20368/1971-8829/1136315

1. Introduction

Many studies have addressed the problem of growing hostility in online environment. It reflects the need to examine forms and dimensions of broader rise in hateful discourse across the public sphere. Recognizable systematic features have emerged. This indicates that many forms of hostility can no longer be regarded as accidental excesses of freedom of expression. Studies confirm that speech manifesting hostility is part of a system of meanings that construct social practices. The digital environment offers significant benefits, but also exposes users to negative, harmful and dangerous dynamics. Under these circumstances individuals may become either victims or perpetrators.

Digital communication environments are frequently permeated and polluted by hostility and prejudice. Stefano Pasta (2018) argues that this prejudiced thinking online relies on the banalization of hate pedagogies and the abdication of responsibility. This perspective is highlighted by Pier Cesare Rivoltella while introducing Pasta's study (2018): the sense of responsibility enables individuals to become wiser co-constructors of the digital space, conceived as a reality that fosters freedom and collective intelligence. Many social practices previously carried out offline have gradually "migrated", adopting to the characteristics of digital environments. It is now evident that the extent of human activity within digital spaces requires responsible forms of engagement. Such an approach is essential to ensuring the quality of the "onlife" (Floridi, 2015) experience.

Pier Cesare Rivoltella (2018) emphasizes also the importance of competences required for the exercise of

[1] corresponding author – rafaela.marteta@akad.gov.al – address: Sheshi "Nënë Tereza", nr. 3, Tiranë, Albania.

© Italian e-Learning Association

full and responsible digital citizenship. The author highlights the political dimension of a commitment to safeguarding and preserving the common good and a sustainable future. This scenario deteriorates further when users fail to assume responsibility, thereby undermining the fundamental principles of digital coexistence. Scholars have proposed literacy and identification strategies that extend beyond the mere recognition of hostile expressions in addressing the phenomenon of hate speech. Conceptualizing hate speech as part of "alphabets of hate" enables a deeper understanding of it as a structured repertoire of symbols, metaphors, and linguistic patterns.

These elements function collectively to convey, normalize and reproduce hate. When manifestations of hate assume multiple forms, they constitute alphabets capable of constructing and reinforcing collective narratives. Within these narratives, hate becomes recognizable through the mechanisms of animosity directed against specific individuals or social groups. Given the plurality of forms, performative practices, and actors involved, hate speech represents a contested and evolving issue. For the purpose of this research, one of the definitions adopted is that provided by the Council of Europe.

> *Hate speech is a complex and multidimensional phenomenon that has far-reaching consequences in democratic societies. It not only affects the dignity and human rights of the individuals directly targeted, but also of persons belonging to the same minority or group as those directly targeted. […] as all kinds of expression, which spread, incite to, promote or justify violence, hatred, discrimination or prejudice against a person, or a group of persons, that is based on presumed or real personal characteristics or status including ['race'/race], colour, language, religion, citizenship, national or ethnic origin, age, disability, sex, gender, gender identity and sexual orientation. (Council of Europe, 2022, p. 17).*

This definition concerns speech targeting vulnerable groups on the basis of protected identity characteristics. While this research focuses on hostility directed toward state actors and institutions that possess public authority and political agency. They are generally subject to broader limits of acceptable criticism. This form of expression may be more accurately understood as political hostility rather than hate speech strictly understood. Nevertheless, such discourse may still produce harmful democratic effects when it normalizes dehumanization, legitimizes violence, or intensifies political polarization. In this regard a conceptual clarification provide by the United Nations is useful. It defines hate speech as: "any kind of communication in speech, writing or behavior that attacks or uses pejorative or discriminatory language with references to a person or group on the basis of who they are" (United Nations, 2022).

Owing to the complexity of phenomenon and the lack of a universally accepted definition, existing research has adopted a range of methodological approaches. These include fully automated, manual, and hybrid frameworks. Findings have highlighted the need for an interdisciplinary approach to this topic, given that relevant contributions emerge from multiple academic disciplines. Alexandra Siegel (2020) argues that research on the effects of methodological approaches used to study online hate is relatively recent and remain limited in scope. She further notes that the boundaries of the field are often unclear, given that "definitions of hate speech can be extremely broad or fairly narrow" (Siegel, 2020).

Despite the difficulty of reaching consensus on a shared definition, scholars have continued to develop approaches that address various dimensions of explicitly harmful or offensive language, as well as the pedagogies of hate (Pasta, 2018). Pasta highlights the challenge and conceptual complexity involved in engaging with this field, thereby evidencing methodological and interpretative difficulties inherent in studying the phenomenon. Definitions, causes, and targets have all been central to research on the detection of online hate speech. Whether based on fully automated or hybrid approaches, this field is characterized by a persistent tension (Pasta, 2023) between purely technological studies and those that are integrate human and computational perspectives. Pasta (2023) highlights this ongoing methodological divide between technology driven approaches and those grounded in socio-humanistic and technology-informed frameworks.

Alexandra Siegel argues that fully automated approaches can be particularly effective in identifying subtle forms of hate speech; however, she also stresses the importance of theory-driven methodologies. According to her, the latter are crucial to ensuring that detection frameworks remain conceptually grounded and analytically robust (Siegel, 2020). There is no doubt about the advantages of automation enabled by technological progress. However, Siegel highlights the limitations inherent in these technologically driven methods. In particular, she notes that they have not yet been able to systematically capture forms of hate speech in a transversal manner (Siegel, 2020).

An additional limitation to be noted is that the majority of existing studies are based primarily on English-language content. Attention should also be drawn to contents expressed in languages that pose greater challenges for the automatic detection of hate speech. The development of approaches capable of addressing detection challenges in low-resource languages represents both a significant methodological challenge and a sustained area of scholarly interest for scholars

© Italian e-Learning Association

(Siegel, 2020). An effective methodological approach is crucial for examining new dimensions of hate speech dissemination within Web 2.0 environments, particularly with regard to its structural and formal characteristics. Above all, it is essential for analyzing its meanings and underlying motivations. (Forzinetti et al., 2024)

2. Structure of the manuscript

Following the general introduction to the research, this paper is structured as follows:

- A short review of studies on hate speech in the Albanian language
- Methodological design
- Results
- Discussion and Conclusions.

It begins with a review of existing studies on hate speech in the Albanian language. These contributions situate the research within the relevant scholarly context and identify key gaps in the scientific literature. This section is followed by an accurate presentation of the methodological design adopted for analysis. The subsequent section reports the main findings of the research. The final part offers a critical discussion of the results and draws the main conclusions, highlighting implications for future work.

3. Hate speech studies in the Albanian language

Studies on hate speech in Albanian language remain limited in scale compared to research conducted in other languages. The majority of these studies adopt fully automated approaches, primarily focusing on the analysis of comments sections across various online platforms. Hersi Kopani and Rovena Llapushi (2025) demonstrate that fine-tuning the GPT-3.4 API can yield significant improvements in the detection of hate speech on Albanian social media platforms. Fetahi et al. (2025) examine the methodological challenges involved in detecting hate speech in a low-resource language such as Albanian. The Albanian language is characterized by extensive dialectal variation and scarcity of recourses and tools for advancing natural language processing (NPL). Bogdani et al. (2021) have conducted a comprehensive study on the phenomenon of hate speech, with funding provided by the Council of Europe.

The findings of Muhamet Kaçuri's master thesis focus on the detection hate speech int the context of the Albanian language rule-based system. Kaçuri (2023) argues that, while deep learning models demonstrate the highest performance according to recent publications, they remain limited by their lack of explainability. In his research he introduces *Shaj* (English: curse), an annotated Albanian dataset for hate and offensive speech derived from user-generated content on social media platforms. Nevertheless, methodologies are often sporadic due to technical and technological constraints. The Albanian language continues to be an under-resourced area for morphological analysis in natural language processing (Çepani & Çerpja, 2026). Other contributing factors may be strategic in nature, including limited institutional attention to raising informational awareness, and addressing educational dimensions of the issue.

The present paper focuses on online hate speech driven by political events. Politics has been identified as a particularly potent trigger for hate speech (Petersen, Osmundsen & Arceneaux, 2023). This pattern is equally observable in the Albanian context. Online political discourse has emerged as a space for the production, circulation and even contestation of political content. Digital platforms are not merely communication channels. They constitute environments in which social and political realities are constructed through linguistic and symbolic practices.

The outcomes of the 2016 judicial reform continue to resonate within social media discourse, shaping public deliberation, and in many instances exacerbating the proliferation of hate speech. Following a period of intensified investigations, the SPAK (Special Structure Anti-Corruption) reported the arrest of dozens of high-ranking Albanian politicians (SPAK, 2024). At the societal level, the reform challenged prevailing perceptions of systemic impunity, an issue deeply politicized and socially salient. While this paper does not examine the activities of SPAK, the political context surrounding its operations provides essential background for the analysis of politically motivated hate speech in Albanian-language social media.

Developments at the institutional level are rapidly translated into the discursive arena of social media. In this context, processes of political accountability often trigger hate-driven commentary targeting politicians. There appears to be a widespread perception of an intensification in hate speech directed at politicians. The Opposition reinforces this perception through the organization of weekly protests. During these events, aggressive rhetoric frequently emerges. It is often characterized by denigration, humiliation and the criminalization of the government. Studies suggest that the circulation of hostile political rumors may benefit those who propagate them. In particular, it may enable interested parts to gain status through this process. (Petersen, Osmundsen & Arceneaux, 2023).

For the purpose of this research, politically motivated hate speech refers to any form of expression of hostile or denigrating discourse targeting individuals or group int their capacity as political actors or supporters of a political orientation. Its primary aim is the delegitimizing of these actors within the online public sphere.

© Italian e-Learning Association

To provide a comprehensive framework of the political context that triggers broader hostility, the selected news topics focus on three key political actors: the head of government, the leader of the opposition, and a deputy minister. The latter is currently at the center of legal and institutional disputes having been accused of corruption-related misconduct while remaining in office. Given the political salience of these cases and the competing interpretations they generate, Robert Entman's framing offers an appropriate analytical lens for examining how media selectively emphasizes certain aspects of political reality, thereby structuring public discourse (Entman, 1993). At the same time, framing shapes the interpretative context of public debate.

Scholars have argued that audiences are not necessarily passively influenced by such frames (Entman, 1989; Graber, 1988). Rather as Pasta (2018) suggests, mediated communication extends the mere transmission of messages. It also serves to construct a sense of belonging among communities organized around shared beliefs. In this sense, politically driven frames non only structure interpretation but also foster collective identification. They reinforce discursive environments in which hostility may circulate and intensify. This paper specifically examines hate speech in comments responding to political news reels, treating these as communicative texts by a distinct semiotics of hate speech. The analysis focuses on the ways in which hate speech is produced. Specifically, hate frames were identified, and underlying cognitive schemata (Van Dijk, 1988: Fairclough, 2001) were examined through the discursive strategies they manifest (Van Dijk, 2011; Wodak, 2015).

4. Methodological design

A qualitative methodology was adopted to analyze users' comments on Facebook Reels related to political news. The comments were manually collected from Reels published on the official pages of Albanian-language news outlet between November 2025 and February 2026. The corpus was collected within this clearly delimited time period and corresponds to a phase of particular heightened political and public debate, driven by ongoing political crisis. Six official Facebook pages were selected on the basis of their relevance to national politics, visibility and levels of users' engagement. News items were selected according to their political relevance, the presence of a clearly identifiable target, and the salience of hate speech in the associated comment section. These criteria of selection and salience were employed to support frame identification (Entman, 1993).

Only publicly accessible comments were included, and user identities were anonymized to ensure confidentiality. Reels were selected over standard Facebook posts due to their higher levels of users' engagement and algorithmic visibility. From each reel, the 100 most relevant comments, including associated threads where present, were extracted, constituting a single dataset. Each dataset includes metadata specifying the identity of the official page identity, the date of publication, the news topic, the political orientation of the news outlet, number of comments per reel, and selection criteria applied. The overall corpus (approximately 600 comments) was determined according to the principle of theoretical saturation; reached when no new relevant information emerged from the data.

Comments were treated as autonomous discursive units capable of conveying coherent meaning. Each unit underwent manual annotation. Frames were assigned to comments that met Robert Entman's operational definition and minimal inclusion criteria, whereby at least two of the following conditions had to be present:

- problem definition: the target presented as threat or social problem etc.
- causal attribution: the target blamed for negative outcomes etc.
- moral evaluation: explicit moral condemnation, etc.
- solution recommended: calls for punishment, exclusion, removal etc.

Ambiguous cases were excluded from the analysis, including comments with no identifiable target, isolated insults, or those meeting only one for Entman's criteria. Finally, frames were interpreted using a theory-driven codebook developed for research purposes. Discursive strategies as outlined by Ruth Wodak (2015) and Teun A. van Dijk (2011), were integrated with socio-cognitive accounts of dehumanization proposed by Nick Haslam (2006) and Albert Bandura (1999). The codebook includes operational definition for hate speech, framing, discourse analysis, and discursive strategies. This instrument was constitutive for the dataset, structuring the coding of online comments according to: (1) textual content, (2) frame criteria, and (3) discursive features. The latter was assigned on the basis of the discursive strategies proposed by Wodak (2015) and Van Dijk (2011).

This procedure produced a structured and analyzable dataset represented in a grid format. While the codebook helped to limit subjective interpretation in this single-coder research, it also provided a consistent analytical framework for data classification. Table 1 below illustrates the structure of the dataset analysis. Some comments were translated into English for illustrative purposes only; otherwise, all comments in the dataset are in Albanian. The original comments have not been subjected to any linguistic correction and are reproduced as they appear in the comments section. This choice includes dialectal variations, phonetic features, systematic orthographic errors, and instances of language hybridization.

© Italian e-Learning Association

Hate speech classification followed a non-binary coding approach. Only comments previously identified as hate speech have been analyzed. Representative anonymized posts form the dataset are reproduced in Table 1, together with their assigned categories, illustrating how the coding scheme was applied in practice. The codebook is published as an addendum to the main paper. The codebook is divided into two sections. The first section presents the table of the operational definitions employed for the following concepts: hate speech (Council of Europe, 2021; United Nations, 2022); frame (Entman, 1993); Critical Discourse & Discourse Historical Approach (Van Dijk, 2001; Wodak, 2011, 2013, 2015; Reisigl & Wodak, 2001, 2009, 2015), as well as dehumanization (Bandura, 1999; Haslam, 2006). The second section contains a table outlining the strategies used in the dataset coding process. The first column contains the strategies, while the second specifies the associated discursive features, including linguistic indicators, psychologic & socio-cognitive accounts related to each strategy.

5. Results

Hate frames were conceptualized as discursive configuration that select specific aspects of reality and render them salient in order to construct the target as a political and social problem (Entman, 1993; Van Dijk, 2011). Certain frames were characterized by discursive features that reinforced polarization between in-group and out-group (Van Dijk, 2011). By identifying systematic patterns across discursive, socio-cognitive and social levels (Van Dijk, 2011; Haslam, 2006; Bandura, 1999), the analysis moves beyond a purely quantitative understanding of hate speech detection. For this reason, the research does not rely on statistical representation of data.

Although the corpus was coded manually a structured coding grid (Table 1) was employed to ensure transparency and analytical rigor. The decision not to translate the comments was motivated by the need to preserve their linguistic authenticity, as the original language contains structures that are essential for a deeper level of analysis. Translation could obscure linguistic patterns that express implicit forms of hate speech even without explicit hostile lexicon. It may also flatten certain cultural dimensions, as well as underlying mental models and attitudes embedded in the original discourse. Out of a total of approximately 600 comments around 440 met the criteria for classification as hate frames. The emerging targets of hate speech were categorized as follows:

1. individual politician;
2. political group (e.g., supporters of X Party);

Table 1 - Example of coded dataset based on the codebook.

ID - Text	Target	Frame criteria	Discursive features (linguistic indicators, psychologic and socio-cognitive accounts)
qetsohu se na e shpifen PD dhe PS un si dua as njeren un dua shqiptaret me zemer e puntore. Jo goj hale si puna jote	political group (supporters of Party X)	Problem Moral judgement	Perspectivation/Ingroup & outgroup membership: position of the commenter in terms of distance toward the target “I don’t like someone filthy mouthy like you” Moral disqualification (insults), Content of attributed difference between perceiver and target Polarization between in-group and out-group; Different deegrees of power, solidarity, intimacy: “I love Albanians who have a heart, and work hard, but not you filthy mouth”; Dehumanization (relational cognition): categorical distinction between them and us
Si mundet kjo hajdute me cizme te flase ne parlament pa u skuqur e as zverdhur per vjedhjet qe ka bere jo vetem per tunelin e Llogarase por lista eshte aq e gjate sa do ti rrethonte kete dhe ALIBABEN DISA HEREKETA RILINDASIT NUK KANE TURP ASHTU SI PARAARDHESIT E TYRE MASAKRUAN KETE POPULL Moj Laje hajdute mos dil me ne faqe te dheut se na e shpiiiiif.	individual politician	Problem Cause Moral judgement	Denomination/nomination: constructing the target as a thief “this thief, cat with the boots”; membership categorization devices, deictics: “this thief”; Moral disqualification of the target (metaphors fort theft, insults); Showing contempt to the target Participant description as different: Identifying the target with communist legacy (the target is heir to one of most important exponents of communist establishment) gender-based labeling, Predication (actor qualification) negative predicate and propositions, implications, presuppositions Intensification/Focus: capitalized words

© Italian e-Learning Association

3. Collective (e.g., “all Albanian people”);
4. Institutions (e.g., media, government);
5. multiple targets, when more than one of the identified targets co-occurred.

With a dominant presence, the primary target of politically motivated hate speech is the individual politician. This might be related to the way news is framed in social media Reels, which tend to highlight certain aspects while downplaying or omitting others. In such cases, the identification of the target is partly conditioned by the structural affordances of the medium.

Linguistic indicators reveal recurring patterns across the three columns of the analytical grid.

1) Particularly, at the discursive level (textual features), metaphors of dehumanization are frequently employed. All categories are illustrated below through specific example:

Metaphors of dehumanization: (Albanian) Ooooooo pyke,na ke ca trapin,kancer,ec vdis,se na je bo mikrob vdekje prures lugat dreqi gerr/ verr,ke 40- vite rrace e keqe, (English) ooooh wooow you scum, cancer, die, you’ve become a deadly microb devil’s servant damn you, you’ve been a bad breed for 40 years.

Intensification of the illocutionary mode (intensifiers, emphasize, quotation marks as in: (original) HA MUT SALI ME GJITH ATA QE MERR PAS TI SE ERE MUTI TE VJEN NGA SHKON E Nga vjen more Qen. (English) YOU SHITTY SALI NAD THOSE WHO FOLLOW YOU SHITTY STINKY WHEREVER YOU COME AND GO, you Dog.

Negative focus on the target, imperatives, call for action as in: (original) Ik or komunist mutit se ty vetem ish komunistat e spijunat te vijn per mbrapa se njerzit e ndershem nuk te degjoj as ste pierdhin ku rrin ti duhet te shkosh ne burg per krimet qe ke ber, (English) Go away you communist shit because only former communist spies come after you honest people don’t listen to you or even care where you are you should go to prison for the crimes you have committed.

“Us” vs. “Them” polarization (the target is constructed as different) as in: (original) Hahahaha ky vazhdon akoma aty sap dot nje dritare ne shtepin tende private pa vajtur ne 100 arkitetet e ketij per te vajtur ne 100 hajdutrit e ketij me zyrat me leket e popullit flas per ne popullin se ata vet ngren 100 kate pa e car trapin fare po populli pse te iken ty o godos, (English) Hahaha he still there while you her you can’t even open a window in your private house without going through 100 architects yet he goes through 100 thieves with flats funded with peoples’ money I speak for the people because they themselves built 100 floors without giving it a shit while your people is leaving the country you idiot. These examples construct a narrative that suggest a moral rupture through an “as versus them” opposition. Such discourse which reflects a polarized division between in-groups and out-groups, constitutes a central mechanism of conflictual political communication (van Dijk, 2006; Wodak et al., 2009). This discoursive configuration distinguishes from mere offensive language, because it operates through systematic processes of stigmatization na ddeligitimation (Faloppa, 2011).

Sarcastic or derogatory nicknaming, alteration of proper names to ridicule the target as in: (0riginal) Si ke ftyrēdh gēnjen. O KAMELEON I NDYR.Tregoi popullit sa MILJARDA I KE VJEDH. Ti dhe sekserēt e tu. Ju KRIMINELA. Qē nuk i jepni buk. Popullit pēr tē jetuar. Si ke ftyrē e del. Na tregon prallat e EDVIN NASTRADINIT. Ti duhet tē varesh nē LITAR O KRIMINEL.O buk shkelēs i POPULLIT, (English) What a lying face you have. YOU FILTHY CHAMELEON. Tell people how many BILLIONS YOU HAVE STOLEN. You and your intermediaries. YOU CRIMINALS. That don’t give bread to the people for living. How can you show your face. You tell us the tales of EDVIN NOSTRADAMUS. You should be HANGED YOU CRIMINAL. You OPPRESSOR OF THE PEOPLE.

2) Recurring patterns at a socio-cognitive level concerned frame assignment through mental models and social attitudes. These patterns involve the definition of the problem, whereby the target is constructed as a threat, or as a broader social and political issue, as illustrated in the examples below:

(original) Po mos be shume moral o te korruptuar qe i moret meqaf ate popull, (English) Don’t act all moral you corrupted ones who ruined those people’s lives. Moral evaluation/judgement implicating, presuming, alleging as in: (original) Je njeriu me i korruptuar ne bote nuk ka te dyte mbas teje, (English) You are the most corrupted person in the world, there is no one second to you.

Causal attribution of blame/responsibility for societal issues as in: (original) Ti sje ne rregull nga trute e kokes;ke te gjithe njerzit qe ke qeveris deri sot ne burg;dhe ti je i paster ipa korruptuar;nuk i ha njeri mo kto perralla ;pij longun 83 mandateve qe ke mar se i di gjith bota si i more ;do fusesh Shqiprin ne europ me pensjoniste 100mileksha ;kurr mos i hegshi ato leke Zoti ju ndershkofte, (English) You are out of your mind; all the people you have governed until today are in prison and you are “clean” and not corrupted;nobody believes these stories anymore ; now you can drink out of those 83 mandates you’ve had because everybody knows how you got them ; you want to push Albania into Europe with 100-lek pensioners; may God punish you for touching pensioners’ money. Stereotyping, invoking generalized negative traits about a target group as in: (original) Te tere hajduta pisa gjakpirsa asnje nga ju plera qe ulni bythen te ato karrige .vetem per te vjedhur, (English) All of you are thieves and bloodsuckers none of you scum who sit on those chairs are able to do else but steal.

© Italian e-Learning Association

3) Recurring patterns at the social level relate to broader societal and political implications, particularly when the focus is placed on the target, often constructed as individual politician: (original) *Korupsjoni ini ju bashk me stafin tuaj dhe ju ini anti shqiptar,* (English) *You are the corruption you and your staff, you anti-Albanians.* Exclusion, marginalization of the target: call for silencing the, disappear, delegitimation as in: (original) *Ku shkuan dy urat e projektuara ? Ne drejtim te pa ditur ...ne tynelin e llogarase? je e vetmja qe nuk mund te besh moral shqiptareve..,* (English) *Where did the money for two planned bridges go? Towards the unknown… into the llogara tunnel? You are the only one who cannot lecture Albanians.*

Individual targeting based on arguments, implications, presumptions for societal harm, crisis or threat to the community as in: (original) *Ky duhet te arestohet dhe denohet per tradheti kombetare. Pa u lodhur shume per te gjetur edhe provat se nuk ja vlen te shpenzohet me shume per tradhetaret e atedheut. Ne emer te popullit si dikur ; vetem keshtu e pastron ate vend nga shushunjat, (English) He should be arrested and punished for national treason. No need to spent much time gathering evidence as it is not worth for traitors of the homeland. In the name of the people as in the past; only in this way the country will be cleansed of the parasites*.

Collective versus individual targeting refers to whether hostility is directed toward a group or toward a single figure, as illustrated in the following example: (original) *Po normalisht do sulmojne sos kan pune te tjera per te bere. Rri dhe merr mend ktu ne komente ku pjesa me e madhe hane buken kot. Ku ne kohen e PD kta skishin goje te ndiheshin qe po u shtypnin kokat e spo e merrnin vesh e po jetonin ne mizerje pa drita pa uje,po ne kto kohe ku JU keni sjell zhvillim e keni rritur rrogat e keni shtruar rruget e keni kthyer Tiranen ne Kryeqytetin me te bukur te Europes dhe Veriun me nje zhvillim te papare e Jugu gjithashtu, tani u vjen goja ktyre se Ai i pari sikur u kish q* plakat kush tia mbante i pari kandilin e vët ua mbante nga koka*
(English) *Well, of course they will attack, they have nothing else to do. They sit here in the comments while most of them are living off nothing. Back in the PD era, they didn't even have the words to realize they were being crushed, living in misery without electricity or water, but now in these times when YOU have brought development, increased salaries, paved roads, turned Tirana into the most beautiful capital in Europe, developed the North like never before, and the South as well, now they complain – because, first of all, it's as if he f***ed the old ladies, and who held the candle for him It stayed on their heads.*

6. Discussions and conclusions

Rather than generalizing, this research conceptualizes framing as a mechanism that defines and delimits hate speech within contemporary digital environments. While automated approaches offer notable advantages, they may come at the cost of overlooking key discursive features in favor of computational efficiency. The automatic detection of online hate speech presents several methodological and practical challenges that must be addressed. Some of these challenges have been outlined in reports by the European Online Hate Lab (2023), in particular the limited capacity of automated systems to capture meaning during data processing (European Commission, 2023).

For this reason, there is a need for studies adopting context-sensitive and interpretative approaches that take into account the socio-cultural variables contributing to the production of hate speech. In the reports by the European Online Hate Lab various terms are used to categorize hateful language. The report identifies a scale of severity and the intensity of hateful language. It distinguishes six levels of escalation, ranging from disagreement as the least intense to death as the most intense. This typology is particularly useful for this research, as the findings indicate a relevant presence of high-intensity forms of hateful language, including expressions that move beyond criticism or hostility and approach calls for harm and elimination. High-intensity hateful speech tends to generate morally charged frames in which emotive, polarized commentary reinforces symbolic "us versus them" boundaries rather than persuasive argumentation (Matthes & Kohring, 2008).

ROPH (Research on online political hostility, 2025) is a research project aimed at examining and addressing extreme forms of online political hostility. The research has identified offline sources of frustration that are transferred to online contexts, mediated by psychological, sociological, and political predispositions (Petersen, Osmundsen & Arceneaux, 2023). "A need for chaos" (Petersen, Osmundsen and Arceneaux, 2023) has also emerged from our dataset. Commenters issue call for uprising, violence, boycott, capital punishment, divine punishment against individual politicians or political groups. A "burn-it-all-to-the ground" social attitude identified by Petersen, Osmundsen and Arceneaux (2023) is also present int this research. It appears as an explicit pattern associated with a disintegrative social orientation. Hannah Arendt (1970) associated processes of social and political disintegration to the growing demands of the masses and the weakening of political parties. In contemporary settings, this dynamic unfolds within what Rivoltella describes "as another place, a parallel space" (Rivoltella, 2017). Within this "other space" experiences are redefined, while they are "empowered by the suppression of the body" (Rivoltella, 2017). They may also generate frustration due to limited opportunities for concrete action. In line with Arendt's (1970) argument such frustration may become a source of violence.

© Italian e-Learning Association

Indeed, the dataset reveals evidence of recurring narratives that construct present-day politics as fundamentally immutable in its nature. It is presented as intrinsically corrupted, hypocritical, and ultimately a threat to national interest. This essentialization, frequently reinforced through copular constructions with the verb "to be", has a legitimizing function in the case of hate speech. It generalizes blame and responsibility, and indiscriminately delegitimize political processes and political actors. In this respect, the Albanian language displays certain linguistic features that may facilitate processes of essentialization and collectivization. The frequent use of the verb "to be" for ontological attribution and the use of absolute quantifiers – all, always, never – supports the generalization of guilt. Within this framework, essentialization is intertwined with recurrent socio-cultural and anthropological patterns. Political dissent is reframed in logics of collective survival and generalized distrust.

According to Albert Nikolla (2011), the lack of trust among Albanians is a social attitude toward elites who continue to hold power despite having lost citizen's trust. The findings of this research suggest that hostility is directed primarily toward public authorities which are portrayed as incompetent, corrupt and oppressive. This form of discourse overlaps with processes of institutional delegitimization, carrying significant implications for democratic legitimacy and the quality of democratic discourse. Manuel Castells (2017) argues that distrust towards particular candidates may extend beyond individual to institutions and, ultimately, to the political system as a whole. In this sense hate against individual politician functions as a delegitimizing discourse that contributes to the erosion of institutional trust. By portraying them as enemies of the people who operate within corrupt structures, such narratives can foster antagonistic political identities and normalize hostility toward institutional authority.

Therefore, it is important to distinguish hostility toward institutions from legitimate political criticism. Criticism challenges policies, decisions within the framework of democratic debate. Hate is expressed through discursive strategies that delegitimize individual politicians and the institutions they represent as inherently corrupt or criminal. Such discourse often relies on accusation, inflammatory language and calls for punishment or rejection of institutional authority.

Many comments attribute responsibility for political decline to the target, frequently invoking death, violent agency, or divine punishment. This discursive pattern activates what Bandura (1999) defines moral disengagement; through which harmful judgements are morally justified and normalized. Both physical distance and the inability to directly experience others' suffering further contributes to moral disengagement (Bandura, 1999). As argued by Pasta (2018), new media have reshaped the ways in which emotions are experienced and expressed. Value-based labeling, vertical comparisons, and the deployment of the disgust and contempt functions to delegitimize the target's right to speak, in extreme cases, their very right to exist. Explicit calls for death or capital punishment exemplify this attitude.

Notably, some social discourses legitimize the use of communist-era punitive practices as a justified response to the target's presumed corruption and criminality. Through presuppositions, predications, implications, and constructions of negative agency the targets or their actions are criminalized (Van Dijk, 2011). Many frames operate through dehumanization by animalizing and objectifying the target on behalf of non-human characteristics (Haslam, 2006). A recurrent labeling of targets as "communists" evokes a totalitarian legacy, constructing them as dangerous, destructive and threatening. Violent action may be perceived as necessary when a group is framed as threatening (Pollack et al., 2023). Hate speech grounded in a totalitarian legacy is highly prevalent in the dataset, especially when the target is a deputy minister and the descendant to one of the most prominent representatives of the communist establishment. Hate frames imply the incrimination of her political legacy by associating her current political activity with that of her grandfather. This association serves to delegitimize both figures as responsible for crimes and theft, while constructing them as enemies of the country. Studies suggest that enemies are constructed on the basis of perceived inherent features rather than temporary one (Van Dijk, 2011).

In conclusion, it can be argued that social media provide a space for the propagation and normalization of hate speech. Hateful messages, often encountering little resistance and limited moderation, contribute to the normalization of hostility towards targeted individuals. This occurs when targets are constructed as e problem, a cause of social harm, or as morally despicable actors. Understanding the deeper motivations behind hatred expressed online may help individuals and institutions to identify it as a driver of conflictual digital citizenship, and social conflict. Groups and institutions addressing hate speech may respond through awareness-raising and prevention programs, or by developing counter-narratives.

A key finding of this research is the transversality of hate speech: hostile and exclusionary expressions are not confined to specific actors or settings. This suggests that hate speech is not the product of isolated individuals, but part broader discursive patterns circulating in everyday online communication. Its transversal spread blurs the boundary between explicit hostility and normalized stigmatizing language. This highlights the need to address this behavior and the wider communicative environments that sustain it. This concern is echoed in the Joint Communication of the European Commission to the European Parliament and

© Italian e-Learning Association

the Council of Europe. It emphasizes that language which belittles or dehumanizes others contributes to a sense of impunity against attacks on specific communities or individuals. While freedom of expression is a fundamental right and a democratic value, human rights law recognizes that it cannot be used to incite hatred and violence. And there should be no impunity for those who cross the line. (European Commission, 2023).

In this context, media-education based pedagogical activism may offers a response. Practically, pedagogical activism can be implemented through structured classroom or community-based practices that directly engage with real digital discourse. A counter-discourse production is a possible method. Learners collaboratively create alternative narratives – videos, posts, memes – that reframe polarized or hostile discussion into empathetic discussion. Gagliardone et al. (2015) argue that, despite the variety of approaches and intended audiences, initiatives aimed at addressing hate speech should converge around three educational objectives: "to inform, to analyze and to confront hate speech" (Gagliardone et al., 2015)

The research acknowledges several limitations, including the absence of inter-coder reliability due to a single author coding. Although this limitation is partially mitigated by theoretical grounding and a documented codebook, internal interpretative processes may still introduce subjectivity. Moreover, the methodological approach itself may constitute a limitation due to difficulties in applying it to a large dataset. Subjectivity in framing comments may also function as a frame in its own right, shaping how issues are interpreted (Matthes & Kohring, 2008). Finally, while the reel format enables the observation of spontaneous public reactions, it also reflects the influence of certain groups in shaping structural aspects of news production.

Datasets and reproducibility

Datasets will be published as an addendum to the main paper.

References

Arendt, H. (1990). On violence. Harcourt, Brace & World.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. Personality and social Psychology Review, 3(3), 193-209. https://doi.org/10.1207/s15327957pspr0303_3

Bayer, J., & Bard, P. (2020). Hate speech and hate crime in the EU and the evaluation of online content regulation approaches (Study requested by the LIBE Committee, Directorate-General for Internal Policies, European Parliament). Publications Office of the European Union. http://doi.org/10.2861/28047

Bogdani, M., Faloppa, F., & Karaj, S. (2021). Mapping hate speech in Albania. Council of Europe.

Castells, M. (2009). Communication power. Oxford University Press.

Council of Europe. (2022). Recommendation CM/Rec (2022)16 of the Committee of Ministers to member States on combating hate speech. Council of Europe.

Çepani, A., Çerpja, A. (2026). Digital morphology of the Albanian language. Journal of Responsible Technology, 25, 100151. https://doi.org/10.1016/j.jrt.2026.100151

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of Communication, 43(4), 51-58. http://doi.org/10.1111/j.1460-2466.1993.tb01304.x

European Commission & High Representative of the Union for Foreign Affairs and Security Policy. (2023). Joint communication to the European Parliament and the Council: No place for hate: A Europe united against hatred https://commission.europa.eu/news-and-media/news/no-place-hate-europe-united-against-hatred-2023-12-06_en

European Commission. (2023). The European Online Hate Lab. Publications Office of the European Union. http://doi.org/10.2838/59442

Fetahi, E., Hamiti, M., Susuri, A., Ajdari, J., & Zenuni, X. (2024). AI-based hate speech detection in Albanian social media: New dataset and mobile web application integration. International Journal of Interactive Mobile Technologies (iJIM), 18(24), 190-208. https://doi.org/10.3991/ijim.v18i24.50851

Faloppa F. (2011). Razzisti a parole (per tacer dei fatti). Laterza.

Floridi, L. (Ed.) (2015). The Onlife Manifesto: Being human in a hyperconnected era. Springer. https://doi.org/10.1007/978-3-319-04093-6

Fetahi, M., Susuri E., Hamiti, M., Kastrati, Z., Canhasi, E., & Misin9, A. (2025). Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. Social Network Analysis and Mining, 15, Article 82. https://doi.org/10.1007/s13278-025-01497

Forzinetti, E., Della Vedova, M. L., Pasta, S., & Santerini, M. (2024). Indicators for characterising online hate speech and its automatic detection. arXiv. https://arxiv.org/abs/2402.08462

© Italian e-Learning Association

Haslam, N. (2006). Dehumanization: An Integrative review. Personality and Social Psychology Review, 10(3), 252-264. https://doi.org/10.1207/s15327957pspr1003_4

Kaçuri, M (2023). Explainability of hate speech classification for Albanian language using rule-based system and neural networks (Diploma thesis). Technische Universität Wien. https://doi.org/10.34726/hss.2023.105780

Kopani, H., & Llapushi, R. (2025). Fine-tuning GPT-3.5 for hate speech detection in Albanian YouTube comments: Challenges and analysis. In K., Dhoska & E. Spaho (Eds.) Bridging Horizons in Artificial Intelligence, Robotics, Cybersecurity, Smart cities and Digital Economy: Sustainable Economy and Ecotechnology (pp. 243-261). Springer. https://doi.org/10.1007/978-3-031-72029-1 20.

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. Journal of Communication, 58(2), 258-279. https://doi.org/10.1111/j.1460-2466.2008.00384.x

Nikolla, A. (2011). L'uomo nuovo Albanese. Tra morale comunista e crisi della transizione. Bonanno Editore.

Pasta, S. (2018). Razzismi 2.0: Analisi socio-educativa dell'odio online. Scholé.

Petersen, M. B., Osmundsen, M., Arceneaux, K., et al. (2023). The need for chaos and the sharing of hostile political rumors in advanced democracies. American Political Science Review, 117(1), 1-18. https://doi.org/10.1017/S0003055422000759

Petersen, M. B., Osmundsen, M., Bechmann, A., Lindekilde, L., Sherson, J., Bor, A., et al. (2020-2025). Research on online political hostility (ROPH) Aarhus University, Department of Political Science. https://ps.au.dk/forskningsprojekter/research-on-online-politcal-hostility/about-the-roph-project

Pollack Sarnecki, H., Lioufas, A., & Jarlsbo, M. (2023). Rotten democracy – Conspiracy propaganda, racism and violence (Report No. FOI-R-5473-SE). Swedish Defence Research Agency (FOI)

Rivoltella, P. C. (2018). Prefazione in Razzismi 2.0: Analisi socio-educativa dell'odio online. Scholé.

Rivoltella, P. C. (2017). Tecnologie di comunità. ELS La Scuola.

Siegel, A. A. (2020). Online hate speech. Princeton University Press.

United Nations. (2022). United Nations strategy and plan of action on hate speech. United Nations. https://www.un.org/en/genocide-prevention/hate-speech/strategy-plan-action

Van Dijk, T. A. (2008). Toward a theory of context: The socio-cognitive approach. Cambridge University Press.

Van Dijk, T. A. (1998). Ideology: A multidisciplinary approach. Sage.

Van Dijk, T. A. (2011). Discourse studies: A multidisciplinary introduction (2nd ed.). Sage Publications

Wodak, R. (2015). The historical-discursive approach. In R. Wodak & M. Meyer (Eds.), Methods of critical discourse studies (3rd ed., pp.23-61). Sage Publications.

© Italian e-Learning Association

Appendix - Minimal Codebook Structure

Part 1. Table of operational definitions for concepts.

Concepts	Operational Definition for
Hate speech	Hate speech is a complex and multidimensional phenomenon that has far-reaching and dangerous consequences in democratic societies. It not only affects the dignity and human rights of the individual directly targeted, but also of persons belonging to the same minority or group as those directly targeted. Hate speech leads to dangerous divisions in society as a whole, affects the participation and inclusion of all those targeted by it and threatens democracy. […] Hate speech is understood as all kinds of expressions, which spread, incite to, promote or justify violence, hatred, discrimination or prejudice against a person, or a group of persons, that is based on presumed or real personal characteristics or status including ['race'/race], colour, language, religion, citizenship, national or ethnic origin, age, disability, sex, gender, gender identity and sexual orientation (Council of Europe, 2021). In the context of this document, the term hate speech is understood as any kind of communication in speech, writing or behavior that attacks or uses pejorative or discriminatory language with references to a person or group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor. This is often rooted in, and generates intolerance and hatred, in certain contexts, can be demeaning and divisive (United Nations, 2022).
Frame	Framing essentially involves selection and salience. To frame is to select some aspects of a perceived reality and make them more salient in a communicating text in such way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described (Entman, 1993).
Critical Discourse & Discourse Historical Approach.	Critical Discourse Analysis (CDA) is a research enterprise which critically analyses the relation between language and society. More specifically, CDA is a type of discourse-analytical research that studies the way ideology, identity and inequality are (re)enacted through texts produced in social and political contexts. (Van Dijk, 2001; Wodak, 2011). The discourse-historical approach (DHA) belongs in the broadly defined field of critical discourse studies (CDS), or also critical discourse analysis (CDA) (Reisigl & Wodak, 2001, 2009; Wodak, 2011, 2013). CDS in general investigates language use beyond the sentence level, as well as other forms of meaning-making such as visual sounds, seeing them as irreducible elements in the (re)production of society via semiosis. CDS aims to denaturalize the role discourse splay in the (re)production of noninclusive and nonegalitarian structures and challenges the social conditions in which they are embedded. (Wodak, 2015; Reisigl & Wodak, 2015).
Dehumanization	Dehumanization is a phenomenon in interpersonal as well as in intergroup contexts, occurs outside the domains of violence and conflict, and has socio-cognitive dimensions in addition to the motivational determinant that are usually emphasized. Two distinct forms of dehumanization should occur when the respective properties are denied to others. In most theoretical accounts, dehumanization is seen primarily as a motivated phenomenon, enabling the release of aggression or removing the burden of moral qualms or vicarious distress (Bandura, 1999; Haslam, 2006).

© Italian e-Learning Association

Part 2. Table of strategies for dataset coding.

Strategies	**Discursive Features** **(linguistic indicators, psychologic & socio-cognitive accounts)**
Nomination (Wodak)	*Actor construction* Membership categorization devices (Deictics; Verbs and nouns for actions and processes; Metaphors)
Denomination (Van Dijk)	*Participant description as different* Strangers, immigrants, others, Opponents, enemies, negative agency (Pronouns, demonstratives)
Dehumanization (Haslam)	*Essentialism* (Nature of difference between perceiver and target, content of attributed difference between perceiver and target)
Predication (Wodak)	*Actor qualification* (The labeling of social actors, positively or negatively, appreciatively or depreciatorily. Stereotypical, evaluative attributions (metaphors, allusions, evocations, adjectives, implicatures)
Predication (Van Dijk)	*Meaning of sentences* (Negative predicate and propositions, implications, presuppositions, allegations, negative emphasize, negative properties)
Dehumanization (Haslam)	*Social categorization* (Mechanistic, a view of others as fungible, and lacking individuality, depersonalized, deindividuated outgroups, moral exclusion)
Argumentation (Wodak)	*Justification/question claims of truth and norms* (topoi, fallacies)
Argumentation (Van Dijk)	*Persuasion by arguments and fallacies* (Formats/schemas/alleged negative properties, rhetorical moves, deictics, adjectives, pronouns)
Dehumanization (Haslam)	*Behavior Explanation* (a more mechanistic view of the other, emphasizing factors that are deterministic, attenuate personal agency an de-emphasize intentional states)
Perspectivation (Wodak)	*Positioning the speaker in terms of involvement or distance* (Metaphors, quotation marks, deictics, indirect or free speech, other)
In-group and out-group membership (Van Dijk)	(as in) *Us* vs. *Them* Different degrees of power, solidarity, intimacy (pronouns, rhetorical moves)
Dehumanization Haslam	*Relational cognition* Communal sharing but categorical distinction between "us" and them"/asocial, Value-based dehumanization: an outgroup is perceived to have dissimilar values to the ingroup; lack of prosocial values and incongruent with one's ingroups' values.
Intensification (Wodak)	*Intensifying the illocutionary mode (*Diminutives, hesitations, verbs of saying, feeling, thinking, hyperboles, litotes, other)
Focus (Van Dijk)	*Special focus on participant (*Stress, volume, size, colour, enumeration, alliteration, visual structures)
Dehumanization (Haslam)	*Emotion* (disgust and revulsion, object of disgust and contempt, elicitors of disgust: death, excretion, sexuality. In its animalistic form the implicit contrast is animals, and it involves denial of uniquely human characteristics. In its mechanistic form it involves emotional distancing, the other is cold, robotic, passive, lacking depth.)

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), pp. 49-60

Digital literacy programmes and social interventions for building resilience against hate speech among school students in Kerala

Genimon Vadakkemulanjanal Joseph[a,1], Sonia Philomena V A[b], Athira P[a]

[a]*Vimal Jyothi Institute of Management and Research, Chemperi. Kannur, Kerala (India)*

[b]*Department of English, Nirmalagiri College (Autonomous), Kerala (India)*

(submitted: 10/3/2026; accepted: 23/5/2026; published: 6/6/2026)

Abstract

Online hate speech against school students is increasing globally. This paper examines how digital literacy programmes and school interventions can help students aged 10 to 18 recognise, resist and respond to hate speech in online spaces. The study reviews research published between the years, 2020 and 2026. It expatiates the causes and effects of online aggression among adolescents. Qualitative data gathered through structured thematic discussions with 16 educationists drawn from government, aided and unaided schools across the State of Kerala. Inductive thematic analysis of these discussions informs the practical approaches presented, drawing on Social Emotional Learning (SEL) and critical media literacy frameworks. The paper highlights several initiatives used in Kerala schools. These include the KITE fake news detection curriculum, the State Council of Educational Research and Training (SCERT) value education framework, the Student Police Cadet programme and the Suraksha Mitram support system.

Two conceptual models guide the discussion. The first model explains how films, games and digital media can increase hostility and aggressive behaviour among students. The second model presents a four-component resilience framework that helps students develop responsible digital behaviour and stronger critical thinking. This framework is explicitly exploratory and normative in character: it synthesises existing evidence and programme documentation into a prescriptive architecture designed for iterative evaluation and refinement. The paper recommends a multi-level prevention approach. It focuses on student skill development, positive peer influence, teacher training, family participation and clear school policies. The approach suits government, aided and unaided schools in Kerala. The study is guided by Bronfenbrenner's ecological systems theory and Bandura's social learning theory. It concludes with practical recommendations to strengthen digital citizenship and reduce online hate among school students.

KEYWORDS: Digital Literacy, Hate Speech, Cyberbullying Prevention, Kerata KITE, Social Emotional Learning.

DOI
https://doi.org/10.20368/1971-8829/1136321

CITE AS
Joseph, G.V., Philomena V A, S. & P, A. (2026). Digital literacy programmes and social interventions for building resilience against hate speech among school students in Kerala. *Journal of e-Learning and Knowledge Society, 22*(1), 49-60. https://doi.org/10.20368/1971-8829/1136321

1. Introduction

The student community is greatly exposed to the digital world of information and connectivity. The seamless digital access creates new opportunities as well as risks to the students. The same platforms that provide educational content can also spread hate speech. Hate speech includes messages that attack or insult people based on caste, religion, gender, language or other social identities. Many young students face such content during an important stage of their psychological development (Castellano et al., 2023; Cedena-de-Lucas et al., 2026).

Indian schools bring together students from different social backgrounds and religions. Schools also include children from migrant families and students from different caste groups. In such a diverse environment, online hate speech can increase social tension among students. Studies in Indian educational settings show that verbal bullying and online harassment are common forms of peer victimisation. Religious and caste-based comments often act as the main triggers (Chakravarthi et

[1] corresponding author - email: jinuachan@vjim.ac.in

© Italian e-Learning Association

al., 2025). The impact of hate speech can be serious. Continuous exposure to hostile messages can increase anxiety and stress among students. It can also reduce their interest in studies and classroom participation. Some students may avoid school or withdraw from learning activities. These effects can influence academic performance and the overall social climate in schools (Philip, 2023).

Kerala has several institutional systems that can help address this problem. Educational programmes already work with students and teachers across the state. These include Kerala Infrastructure and Technology for Education (KITE) curriculum initiatives, the State Council of Educational Research and Training (SCERT) value education framework and the District Institutes of Education and Training (DIETs). Other initiatives include the Student Police Cadet (SPC) programme, the Suraksha Mitram support system and the Little KITEs student IT clubs. The monitoring structure of the Samagra Shiksha Abhiyan also supports school level programmes. Together these initiatives create a strong support network. However, there is still a need for a clear framework that connects these programmes to the issue of online hate speech.

What makes Kerala a particularly distinctive analytical case is the convergence of three structural factors that are rarely found together in any single Indian state. They are: near-universal school enrolment, exceptionally high baseline digital literacy (94 per cent as per Census 2011, approaching 100 per cent among current school students), and a dense institutional infrastructure for educational governance. These conditions mean that Kerala is not simply a convenient geographical focus; it is a context where systemic, scalable interventions are genuinely feasible within existing educational institutions. The applicability of the proposed framework is not limited to Kerala – similar conditions in other states or countries would support comparable implementation. But its specificity and the evidence underpinning it derive from the particular programmes, policies, and social dynamics of this state. Where the paper makes recommendations that are broadly generalisable.

It is also necessary to define hate speech carefully in the school context. Not every offensive comment can be treated as hate speech. Very broad definitions may restrict normal discussion or disagreement. This paper follows the definition used by UNESCO in 2019. Hate speech refers to communication that attacks a person or group based on protected characteristics such as religion, caste, gender, ethnicity, or disability. Such communication may insult, threaten or promote hatred. Hate speech may appear in schools in different forms. These include undermining value of life, caste-based insults shared through messaging apps, memes that stereotype religious communities, gender harassment in group chats and jokes about the accents or culture of migrant students. At the same time normal criticism, disagreement, or satire should not be treated as hate speech. Teaching students to understand this difference is an important part of digital literacy education.

It is equally important to distinguish hate speech from conceptually adjacent categories that appear in the same literature and policy discourse, including cyberbullying, online aggression, and misinformation. Cyberbullying refers to repeated, targeted harmful behaviour directed by an individual or group against another individual, and does not necessarily involve identity-based attacks on protected characteristics. Online aggression is a broader category encompassing hostile communication that may or may not be repeated, targeted, or identity-based. Misinformation concerns false or misleading content distributed irrespective of any intent to harm a particular social group. Hate speech, as this paper consistently uses the term in accordance with the UNESCO (2019) definition. It is distinctively characterised by its targeting of persons on the basis of protected identity characteristics as caste, religion, gender, ethnicity, disability and language in the Indian school context. These phenomena frequently co-occur in school digital environments and while some interventions address more than one category. The hate speech particularly damaging in diverse educational settings and prevents precise measurement of intervention outcomes.

Kerala had 94 percentage digital literacy as per census-2011 and it is nearing 100 percentage among school students. Digital learning tools have become common in classrooms after the COVID period. At the same time divisive content has increased on social media platforms. Some of this content circulates in messaging groups of school students. This situation increases the need for stronger digital awareness and responsible online behaviour.

This paper reviews research on how digital media exposure can influence hate speech and aggression among young people. The qualitative analysis discusses practical school-based methods that can build resilience among students and support dignified digital communication.

media promotes observational learning and social modelling (Bandura, 1977): when characters in films or games resolve conflicts through aggression and face no meaningful consequences, students absorb aggressive behaviour as a normative response to frustration or social threat. Repeated exposure reduces empathy by making suffering and revenge appear routine (Krahé, 2025; Wachs et al., 2025).

As illustrated in Figure 1, four reinforcing pathways are identified in the literature. First, exposure to violent media promotes observational learning and social modelling (Bandura, 1977): when characters in films or games resolve conflicts through aggression and face no meaningful consequences, students absorb aggressive behaviour as a normative response to frustration or social threat. Repeated exposure to these reduces empathy by making suffering and revenge appear routine (Krahé, 2025; Wachs et al., 2025).

© Italian e-Learning Association

2. Literature Review

2.1 Digital Media Impact on Hate and Aggression in Students

A growing body of research documents the mechanisms through which digital media exposure increases hostile attitudes and aggressive behaviour in adolescents. Figure 1 summarises these pathways, which have been identified across experimental, longitudinal and review-based studies conducted in India and internationally.

As illustrated in Figure 1, four reinforcing pathways are identified in the literature. First, exposure to violent

Second, interactive gaming introduces a qualitatively different dynamic. Unlike passive film viewing, games require players to practise aggressive responses repeatedly. Games are structured to reward aggressive choices through points, progress and social status within gaming communities. This reinforcement loop is immediate and direct. It strengthens associations between aggression and reward in ways that can generalise beyond the game environment (Chen et al., 2024).

Third, social media echo chambers expose students to concentrated streams of hate narratives, filtered through algorithmic recommendation systems that prioritise emotionally provocative content because it generates higher engagement. Online disinhibition – the reduction in social constraint associated with perceived anonymity – lowers the threshold for posting hateful content that students would not express face-to-face. Repeated exposure within a closed information environment normalises the attitudes being circulated (Sunstein, 2017; Törnberg & Törnberg, 2024; Vasconcellos-Silva & Castiel, 2025).

The combined effect of these four pathways produces the behavioural outcomes shown at the centre of Figure 1: increased hostile thoughts, desensitisation to violence, and normalisation of aggressive behaviour (Pittman, 2023).

The moderating factors identified in the research literature – parental involvement, critical discussion and education, and prosocial peer influence – correspond precisely to the intervention targets addressed in the second half of this paper.

Research affirms these findings and indicates the increasing online bullying among school students. A review by Suresh and Vijaya (2024) found that verbal and relational bullying are the most common forms in schools. Cyberbullying has increased sharply in studies after 2015. A major clinical study at the National Institute of Mental Health and Neurosciences (Ranjith, Vranda, & Kishore, 2023) examined 650 students. The study found cyberbullying victimisation in 14.5 percent of students and perpetration in 5.8 percent. About 13.8 percent were both victims and perpetrators. A global

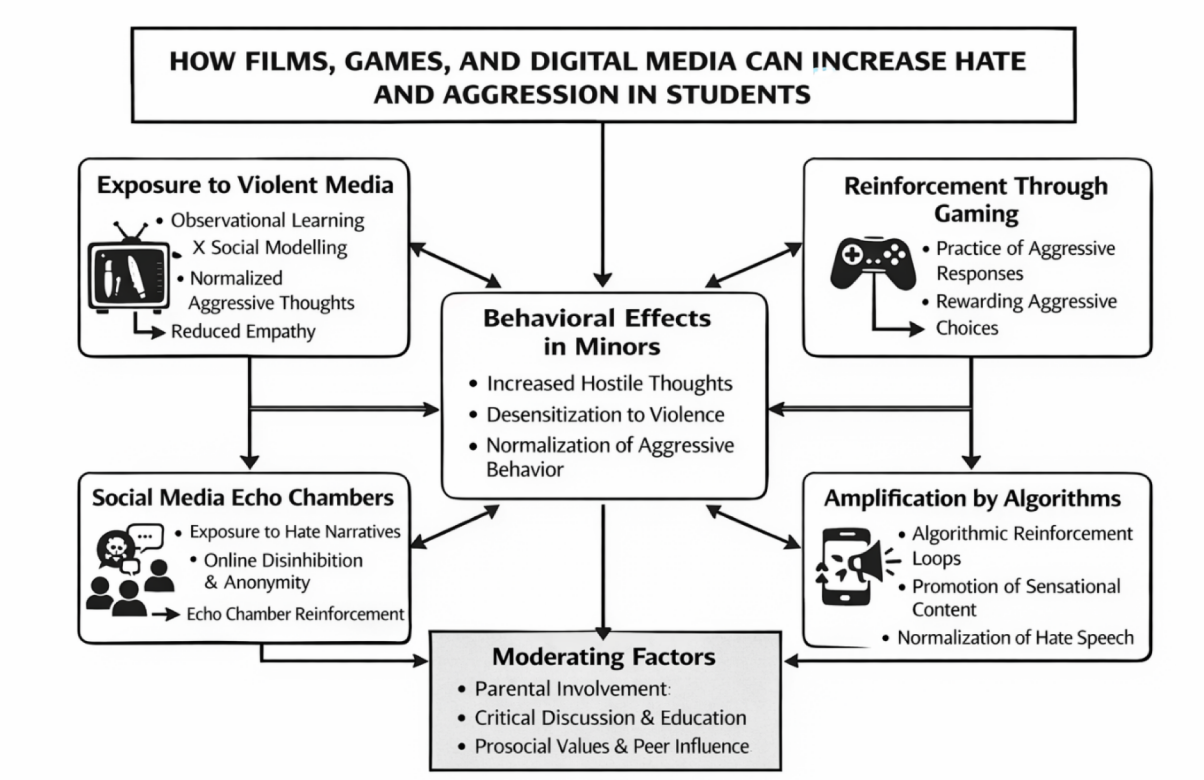


Figure 1 - How Films, Games, and Digital Media Can Increase Hate and Aggression in Students. [Note: This figure was developed by the author based on the studies using standard diagrammatic design tools and is not AI generated].

© Italian e-Learning Association

review by Zhu and colleagues (2021) shows that cyberbullying is linked with low self-esteem, social isolation and higher risk of suicidal thoughts.

2.2 Theoretical Frameworks

This analysis is based on two theoretical frames. First, Bronfenbrenner's (1979) ecological systems model advocates the integration of concentric layers of social context: the microsystem (classroom and peer group), the mesosystem (school–family interactions), the exosystem (institutions and policies) and the macrosystem (cultural values and norms). This model explains why a module delivered only to students without involving teachers, families, or school policy tends to produce limited and short-lived effects. Sustainable change requires addressing multiple levels of this ecology simultaneously (Bilz et al., 2024).

Second, Bandura's (1977) social learning theory provides a complementary account of how behaviours and norms are acquired and modified. Students learn through observation of peers and adults, through direct experience and through the social feedback they receive for their behaviour. An intervention that changes only the knowledge of individual students without incorporating their encircling social environment is unlikely to produce durable change in conduct.

2.3 Evidence for School-Based Interventions

A quasi-experimental trial of the Bullying Intervention Module (BIM) in Varanasi schools by Singh and Singh (2025) found a 25 percent improvement in students' awareness of bullying causes and consequences. Ranjith et al. (2025) validated components for Karnataka covering safe platform use, family communication strategies and school reporting systems directly adaptable for Kerala. Whole-school approaches consistently outperform single-component programmes. A review of Indian educational psychology literature found that programmes combining teacher training, peer norms change and family involvement reduced aggressive behaviour by approximately 20 percent relative to single-component alternatives (Joseph et al., 2026; Ranjith et al., 2023). UNESCO's (2019) digital literacy policy framework identifies critical media literacy as the capacity to evaluate rather than merely access digital content. This media literacy is protective against hate propaganda and reinforcing harmony.

Research on peer relationships also highlights the role of bystanders as a leverage point for intervention. In the majority of cyberbullying incidents, bystanders were present in some form – as witnesses to a public post, members of a group chat, or recipients of a forwarded message. Whether those bystanders reinforced, ignored, or challenged the hateful behaviour was a stronger predictor of incident escalation than any characteristic of the perpetrator (Joseph & Thomas, 2020; Rudnicki et al., 2023; Wachs et al., 2024). This finding has a direct and actionable implication for schools: programmes that activate the prosocial majority – the students who neither perpetrate nor are victimised but who witness hate speech regularly – have the potential to change the normative environment for the whole community.

UNICEF India's (2020) analysis of child protection in digital environments in India emphasises that the most effective national-level responses combine curriculum-based education with platform-level policy engagement and community-level awareness. Kerala's state government has engaged in policy dialogue with social media companies regarding communal content, and the Kerala Police's Cyberdome unit has conducted awareness programmes in schools and communities. Coordination between these institutional efforts and the school-based prevention framework proposed in this paper would enhance the coherence and reach of the overall response.

2.4 Social Emotional Learning as Affective Foundation

The implementation of Social Emotional Learning (SEL) provides the affective foundation for resilience of the students. The five core competencies – self-awareness, self-management, social awareness, relationship skills, and responsible decision-making – map directly onto the capacities required to navigate hate speech. The SCERT's existing value education slots (10–15 minutes) offer a practical timetable entry point for SEL activities without displacing academic content in government and aided schools of Kerala. In structured discussions, students articulate responses to realistic scenarios involving online exclusion or identity-based insults to build reflective habits. Role-playing exercises enable students to inhabit the perspectives of targets, bystanders and perpetrators in simulated online exchanges.

Assessment of SEL outcomes in school settings is most effectively accomplished through teacher observation rubrics, student self-report instruments and peer-assessment formats rather than written tests. Teachers use reflection on weekly prompts such as 'What did I notice this week about how online words affect people around me?' – instruments that serve simultaneously as learning tools and low-burden evidence-collection mechanisms. These continuous evaluations provide teachers and school administrators with data on the trajectory of students' developing awareness without requiring standardised testing infrastructure. The DIET network can support teachers in designing and interpreting these instruments as part of regular in-service training.

© Italian e-Learning Association

3. Analysis of Digital Literacy Programme Components in Kerala Schools

3.1 Qualitative Study: Methodology

Population, Sampling, and Data Collection

The study participants were school teachers and administrators serving in government, government-aided, and unaided private schools in Kerala. Participants were identified through a stratified purposive sampling procedure. The three school management categories (government, aided, unaided) formed the strata and within each stratum schools affiliated with the Kerala State syllabus, CBSE, and ICSE boards were represented to ensure curriculum diversity. Within this stratified frame, voluntary participation was invited through school principals who were briefed on the study's purpose. The resulting sample comprised 16 educators: 6 from government schools, 6 from government-aided schools, and 4 from unaided private schools, spanning primary, upper primary, and secondary levels.

Data were gathered through semi-structured group discussions conducted in Malayalam and English according to participant preference. Each discussion lasted between 60 and 90 minutes. A discussion protocol organised around three thematic domains was followed consistently: (i) the nature and manifestations of hate speech observed in the school's digital environment; (ii) the behavioural, social-normative, and psychological consequences of hate speech observed among students; and (iii) the interventions currently employed by the school to prevent or respond to hate speech, including participants' assessments of their effectiveness. Discussions were recorded with participant consent and subsequently analysed.

Data Analysis

The data of the discussions were subjected to inductive thematic analysis following the six-phase procedure described by Braun and Clarke (2006): familiarisation with the data, systematic generation of initial codes, organisation of codes into candidate themes, review of themes against the full dataset, definition and naming of final themes, and production of the analytic narrative. Coding was conducted manually. The emergent themes were cross-referenced against the experimental and epidemiological literature reviewed in Section 2. The apparent discrepancies between primary data and existing evidence were specifically interrogated rather than resolved by default in favour of either source.

Interpretive Criteria and Limitations

Trustworthiness was pursued through three procedures: prolonged engagement with the dataset, member-checking of key thematic summaries with two participants drawn from different school types, and reflexive acknowledgement of the researcher's

Table 1 - Consequences of Hate Speech in Educational Settings.
Source: Formulated from primary qualitative data (educator discussions) and literature synthesis.

Context	Behavioural Consequences	Social and Normative Consequences	Psychological and Emotional Consequences
Initial or limited exposure to hate speech	Ability to recognise derogatory language and hate-based content online; motivated avoidance of hostile interactions; appropriate use of reporting tools when prompted	Peer norms broadly align with school anti-discrimination expectations; bystanders respond with concern and support; class climate perceived as mostly inclusive	Emotional arousal (distress, discomfort) when exposed to slurs or hostile posts; empathy for peers who are targeted; sense of belonging and trust in school environment intact
Repeated or sustained exposure to hate speech	Reduced sensitivity to derogatory language; hate speech normalised in everyday digital exchange; increased risk of perpetrating bullying or passing on hostile content; avoidance of school and social settings; discrimination against peers from targeted groups	Descriptive norms shift: hateful behaviour perceived as common and therefore acceptable; bystander passivity increases; social hierarchies reinforced along caste, religious, or gender lines; erosion of inter-group trust in mixed-faith or multi-caste classrooms	Desensitisation to violence and hostility; diminished empathic response; elevated anxiety, depression, and emotional difficulties in victimised students (Zhu et al., 2021); reduced sense of school safety and belonging; increased risk of academic disengagement
Institutional level (school environment)	Increased disciplinary incidents linked to online conflicts spilling into school settings; decline in participation in co-curricular and community activities; teacher time diverted from instruction to conflict management	School's anti-hate norms weakened if policies are not enforced consistently; community trust in school as a safe environment erodes among minority-group families; peer group fragmentation along identity lines	Collective emotional climate shifts toward anxiety and vigilance; reduced teacher confidence in managing sensitive digital citizenship discussions; wider school community morale affected when high-profile incidents become public

© Italian e-Learning Association

institutional position within the Kerala educational system. Data saturation (the point at which additional discussion yielded no new themes) was reached after twelve discussions; the remaining four provided confirmatory depth rather than novel themes. The primary limitation of this qualitative component is its relatively small sample and its reliance on voluntary participation, which may have introduced a pro-intervention bias among participants. The data are best understood as providing practitioner-grounded corroboration of the literature-based evidence, not as a stand-alone empirical foundation for the proposed framework.

Table 1 below presents the consolidated output of the thematic analysis, organised by context of exposure to hate speech. The three-column structure (behavioural, social-normative, and psychological consequences) reflects the three thematic domains of the discussion protocol. Entries are indexed to supporting literature where convergence was identified.

3.2 Digital Literacy and Fake News Curriculum

Critical media literacy (CML) teaches students to interrogate the construction and circulation of digital content rather than treating it as neutral information (Bilz et al., 2024). This is especially important where communally divisive fabricated content circulates during election periods and religious festivals through social media. The SIFT method (Stop, Investigate the source, Find better coverage, Trace claims to their origin) provides a structured framework that can be scaffolded across year groups. For Classes 6 and 7, the investigation step can be simplified to checking whether a forwarded image appears on established fact-checking platforms such as Alt News or Boom. For Classes 9 and 10, students can engage with more demanding analysis of how hate speech often operates through partial truths.

KITE incorporated fake news detection modules into ICT textbooks for Classes 5 and 7 in 2024. The modules cover the mechanics of online misinformation, verification protocols adapted for the platforms students commonly use (WhatsApp, YouTube, Instagram) and the consequences of sharing unverified content with respect to recent communal disputes in Indian states. Teacher feedback indicated improved student engagement and measurable changes in verification habits.

The 2024 curriculum revision of the Kerala Syllabus introduced discussion of the social and legal consequences of spreading misinformation, including the provisions of the IT Act 2000 and the IT Rules 2021, bringing a rights and responsibilities dimension into the classroom that connects digital literacy to civic education more broadly.

3.3 Peer Mentoring and Bystander Training – Little KITEs

Peer-led mentoring is among the most cost-effective prevention components (Zhu et al., 2021). Kerala's Little KITEs clubs, operating across hundreds of governments and aided schools, provide existing infrastructure for peer digital citizenship leadership. Senior Little KITEs members trained in hate speech recognition and response can serve as first-line mentors for younger students encountering victimisation.

Bystander training is a frequently underemphasised dimension of prevention. Bystander behaviour is one of the strongest predictors of whether cyberbullying escalates or subsides (Kowalski et al., 2014; Rudnicki et al., 2023). The Little KITEs club activities train specific, low-risk bystander strategies – counter-posting, private supportive messages to targets, use of platform reporting mechanisms – to bridge the gap between awareness and action. This is integrated into the Little KITEs training calendar without displacing existing technical content.

3.4 Student Police Cadet Programme

The Student Police Cadet programme (SPC) is jointly managed by Kerala's Police Department and the General Education Department of Kerala since 2010. It engages approximately 86,000 students of Classes 8 to 12 at a time. Its two-year curriculum covers law, human rights, civic empathy, community service and rejection of social evils. The programme has documented reductions in disciplinary incidents in participating schools and improvements in self-reported empathy across social backgrounds. Students with SPC exposure integrate an explicit digital citizenship strand covering online rights, responsibilities and responses to hate speech (Chacko, 2020; Student Police Cadet, n.d.).

3.5 Suraksha Mitram and Safe Expression Structures

The Suraksha Mitram initiative was introduced by the Kerala General Education Department in 2025 to protect students from abuse. It ensures the availability of multiple confidential 'help boxes' in campus, weekly reviews of the feeds from the boxes, teachers' training, regular reporting, zero-hour sessions, and diary writing. It provides confidential multiple pathways for students to obtain protection from online victimisation (Kerala General Education Department, 2025). Research consistently shows that victimised adolescents rarely disclose to adults, citing fear of device confiscation, concern about parental overreaction or scepticism about effective intervention (Afrouz & Vassos, 2024; Jovanovic & Markovic, 2023). Normalising structured student expression within the school day lowers the social cost of disclosure.

Table 2 below presents a consolidated overview of the core activities performed in Kerala schools that constitute the empirical basis for the resilience framework proposed in Section 4.

© Italian e-Learning Association

These combined efforts strengthen empathy, responsible digital behaviour and safer school environments for diverse student communities.

4. A Proposed Framework for Building Resilience

Based on the review of the research and the analysis of the qualitative discussions, a model for building resilience among students is proposed. Figure 2 presents the integrated resilience framework proposed in this paper. It identifies four interconnected components – Education and Awareness, Critical Thinking Skills, Positive Online Behaviour, and Safe and Inclusive Spaces. The model is driven by the dual engines of Digital Literacy Skills and Empathy and Civic Values, and directed towards the outcome of resilient and responsible students.

This model is normative and exploratory in character: it synthesises evidence from the literature reviewed in Section 2 and from the qualitative data presented in Section 3.1 into a comprehensible prescriptive architecture. However, it has not been empirically validated through a controlled experimental or quasi-experimental evaluation. The four components and their inter-relationships are theoretically grounded and supported by practitioner-reported data. But require independent testing across different school contexts before strong causal claims can be made about the framework's effectiveness. The model is best understood as an existence claim: that the institutional infrastructure of Kerala's school system makes this kind of integrated framework feasible – not yet as a demonstrated efficacy claim.

Future validation strategies may include: pre-post survey designs measuring hate speech attitudes, bystander efficacy, and digital verification behaviour in schools that implement the framework. The structured teacher observation protocols using the SEL rubrics described in Section 2.4 and longitudinal tracking of anonymised incident data using the KITE platform's data management capabilities. Comparison between schools with high versus partial implementation would enable dose-response analysis of the framework's components.

The four components of the framework address the full range of moderating factors identified in Figure 1 – parental involvement, critical education, and prosocial peer influence – within a structured school-based architecture.

Together they operationalise the multi-level ecological approach described in Section 2.2, targeting the individual, peer, institutional and family dimensions simultaneously.

4.1 Education and Awareness

Students who cannot identify hate speech will neither seek help when victimised nor exercise restraint when tempted to perpetuate it. KITE's fake news detection curriculum and SCERT's value education framework jointly address this knowledge dimension. Classroom activities should include exposure to examples drawn from platforms students actually use, followed by structured analysis of why the content is harmful and what responses are available.

Table 2 - Summary of Core Intervention Activities in Kerala Schools.
Source: Formulated from primary data and programme documentation.

Initiative	Type	Core Activities	Relevance to Resilience	Source
Fake News Modules (KITE ICT, 2024)	Curriculum	Verification skills; source analysis; Classes 5 & 7	Reduces susceptibility to hate propaganda	KITE, GoK (2024)
First Bell Digital Teaching Platform	Digital education	Asynchronous content delivery; Grades 1–12	Safe information access regardless of bandwidth	KITE, GoK (2020–present)
Little KITEs Student IT Clubs	Student leadership	Peer mentoring; ICT projects; digital citizenship	Peer norm-setting; bystander capacity	KITE, GoK (ongoing)
SCERT Value Education (2023 revision)	Curriculum	Empathy activities; diversity narratives; social reform history	Inter-community respect; hate recognition skills	SCERT Kerala (2023)
Student Police Cadet (SPC) Programme	Civic character	Discipline; law; civic empathy; community service	Reduces violence; builds tolerance in Classes 8-12	Kerala Police / GED (ongoing)
Suraksha Mitram / Help Boxes / Zero-Hour	Student support	Confidential disclosure; safe expression; peer support	First-response support for victimised students	GED Kerala (2023–present)
Samagra Shiksha Abhiyan (Kerala)	National scheme	Inclusion support; DIET training; monitoring dashboard	Equitable implementation infrastructure	MoE GoI / GED Kerala (ongoing)

© Italian e-Learning Association

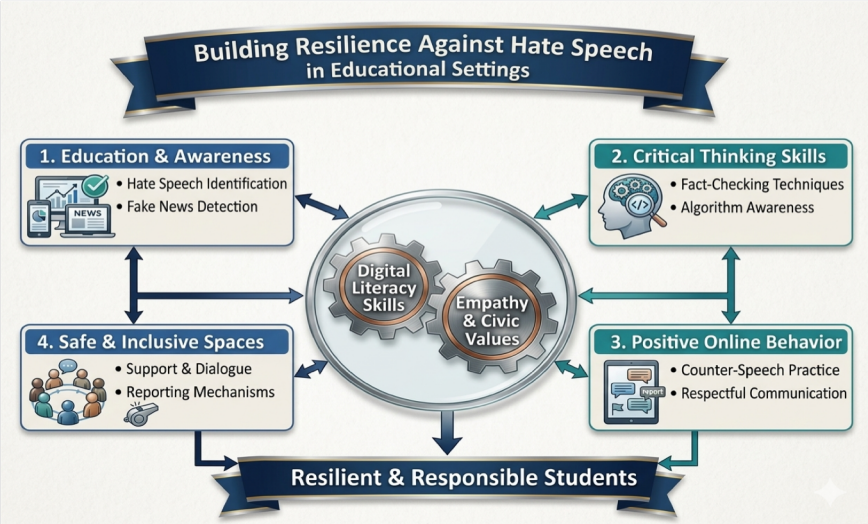


Figure 2 - Proposed Preventive Framework: Building Resilience Against Hate Speech in Educational Settings [Note: This figure was developed by the author using standard diagrammatic design tools and based on the earlier researches].

This component requires brief but specific professional development training for teachers. DIET in-service training structured as four to six sessions – using Malayalam-medium facilitator guides and case study cards mapped to SCERT's timetable – provides adequate preparation for primary and lower secondary levels. Training must include reflection on facilitators' own social positions, particularly regarding caste, to avoid inadvertently reproducing the dynamics the programme aims to address.

4.2 Critical Thinking Skills

The second component develops the analytical tools required to act on awareness. SIFT-based fact-checking pedagogy, adapted across year groups from Class 6 upward, is the core methodology. For senior secondary students, the analysis extends to understanding how algorithmic amplification works and what platform-level governance tools – reporting mechanisms, privacy settings and appeal procedures – are available to users.

The goal is to equip students with a calibrated critical evaluation habit – the automatic pause and questioning before engaging with content that triggers a strong emotional response, rather than a generalised scepticism that would undermine legitimate digital engagement. Singh and Singh's (2025) finding that structured reflective prompting was the most effective element of the BIM intervention supports the priority given to metacognitive approaches here.

4.3 Positive Online Behaviour

The third component moves from protective to generative: students do not merely avoid harmful behaviour but actively practise positive alternatives. Counter-speech training enables students to produce content that challenges hateful narratives with accurate, respectful information. Counter-speech is increasingly recognised as a more sustainable response than reporting or withdrawal alone (Prasannan et al., 2025), as it changes the information environment rather than merely removing individual posts (UNESCO, 2019). The positive online behaviour can be reinforced by cultivating a respectful communication culture within the student community.

Positive online behaviour can take the form of student-produced counter-narrative content in the native language. Students can draft unity pledges collaboratively in class and share them through school KITE platform channels. The Little KITEs peer leadership structure provides the social architecture through which positive digital norms are transmitted from experienced students to newer ones.

4.4 Safe and Inclusive Spaces

The fourth component addresses what happens when the first three are insufficient. Safe and inclusive spaces are defined by two features: accessible support and functional reporting mechanisms. The Suraksha Mitram initiatives and 'zero-hour' discussion provide the 'support and dialogue' dimension. The reporting dimension requires both a clear school policy specifying what constitutes reportable digital hate speech and a practical, low-stigma mechanism for making a report without fear of social consequences.

Families are essential to this component. PTA awareness sessions on how to respond to disclosure with supportive rather than punitive reactions increase the likelihood that victimised children will approach their parents. This family engagement dimension directly operationalises the mesosystem level of Bronfenbrenner's ecological model.

© Italian e-Learning Association

4.5 Implementation Across School Types

Kerala's school system comprises government schools, government-aided schools and unaided private schools, which differ in resource availability and governance. The framework is designed to be scalable across this spectrum. In well-resourced schools with reliable connectivity and trained teachers, full digital implementation using KITE platforms is feasible. The Samagra Shiksha monitoring dashboard provides a common accountability framework across school types, enabling district-level tracking of implementation quality and outcome indicators without imposing significant additional administrative burden.

4.6 Monitoring and Continuous Improvement

A prevention framework is only as strong as its monitoring system. Anonymous student surveys administered at the beginning, middle and end of each academic year provide data on exposure to hate speech, perceived school climate and bystander behaviour intentions. Quarterly review of incident reports, coupled with pattern analysis by grade, class and incident type, enables school leadership to direct targeted interventions where need is greatest. Schools participating in KITE's network can use the platform's data management capabilities to maintain anonymised records that inform continuous programme improvement without compromising student privacy.

External monitoring through district child protection units and UNICEF India's child safeguarding programme adds a layer of quality assurance, providing accountability to the prevention framework and credibility to the process beyond individual school boundaries.

The limitations of the proposed framework must be acknowledged explicitly. First, the model's feasibility depends on a level of institutional capacity – KITE infrastructure, DIET training capacity, SPC presence – that may not be uniformly available across all school types, particularly unaided private schools in rural areas. Second, the framework is primarily grounded in literature from Western, South Asian, and Kerala-specific contexts. Its components may require cultural adaptation before application in other Indian states with different social compositions, languages, or administrative structures. Third, the qualitative data underpinning the framework are practitioner-reported and subject to potential social desirability bias. These limitations do not undermine the framework's utility as an organising architecture for evidence-informed practice, but they do constrain the strength of claims that can be made about its generalisability and impact prior to independent evaluation.

5. Discussion

The programmes introduced in the schools of the State of Kerala are helpful in building resilience against hate speech and provide technological support systems to develop students' capacity to engage responsibly in digital environments. These include student skill development, positive peer influence, teacher preparation, family involvement and clear school policies. When these components work together they can reduce victimisation and improve student well-being. Kerala has strong institutional systems that can support such programmes. The real challenge is maintaining quality and consistency of implementation when they are scaled across many schools with different resource levels.

Evidence from pilot programmes and qualitative discussions shows that schools with systematic teacher orientation produced better outcomes than those without it. This suggests that training through the District Institutes of Education and Training (DIET) is a foundational, not optional, element of implementation.

Programmes designed only in English can exclude students who are already vulnerable to online abuse. Materials in Malayalam should go beyond simple translation; they must reflect the cultural experiences and reference points of students. Digital platforms change faster than school curricula. Because of this, programmes should focus on transferable analytical skills – the ability to recognise the structural patterns of hate speech – rather than platform-specific rules that become obsolete as platforms evolve. Classroom discussions on caste-related issues require particular care: teacher training must include structured reflection on personal attitudes and social positions to avoid inadvertently reproducing hierarchies the programme aims to address.

A concern that emerged from the qualitative discussions was the inequality in digital access: some students face online risks without adequate data connectivity support, meaning that the risk environment and the protective resources are unevenly distributed within the student population.

The current evidence base has limitations. Many Indian studies use small samples and short observation periods. The data from Kerala pilot programmes are useful but still limited in scope. More rigorous studies are needed, including larger school samples, validated research instruments, and follow-up over at least one academic year. Research should also examine hate speech separately from general cyberbullying, because identity-based attacks often involve different social, psychological, and normative dynamics.

6. Conclusions

Research shows that hate speech harms psychological well-being, reduces academic engagement and weakens

© Italian e-Learning Association

inter-group trust among students. However, Kerala's schools already have the systems and institutional support needed to address this problem if these efforts are organised within a clear and practical framework.

This paper proposes such a framework based on two conceptual models, existing research and discussion with educators. The approach has four main components: Education and Awareness, Critical Thinking Skills, Positive Online Behaviour and Safe and Inclusive Spaces. Together these elements help students understand harmful content, question misleading messages, respond responsibly online and support peers who face abuse. The aim is to build resilience and responsible digital citizenship among young learners.

The framework does not require new institutions or large financial investment. Instead, it calls for better coordination of programmes that already exist in Kerala's education system.

The distinctiveness of Kerala as a site for this framework lies in the institutional density and policy infrastructure already in place to address them. No other Indian state currently operates a comparable simultaneous combination of KITE's technical infrastructure for curriculum delivery, SCERT's curriculum development capacity, the SPC programme's civic character formation mandate, and the Suraksha Mitram support architecture within the same school system. This institutional convergence creates a genuine opportunity not only to implement but also to rigorously evaluate and iteratively refine a resilience model. This model could subsequently use in other states and international contexts where comparable. The framework presented here is offered as an initial architecture for that process of evidence-based development.

Several practical steps can move this effort forward. Schools can introduce a joint KITE–SCERT digital citizenship module for Classes 6 to 10 that clearly addresses hate speech. Teacher training through DIET programmes can support classroom discussions on digital behaviour. The Suraksha Mitram system can include clear reporting channels for online abuse. Monitoring indicators can also be added to Samagra Shiksha dashboards. Finally, schools should support rigorous research that evaluates these interventions across different types of institutions in Kerala.

Acknowledgements

All the authors are associated with this research. First author took a lead role in conceptualization, methodology, data curation, analysis and draft of the research. Second author supported editing and language correction. The third author associated with the refining of draft and final writing. All authors have read and approved the publication.

And there is no conflict of interest among them at any stage of this research. This research is not supported any funding from any agency. The technical support of the VJCRS, Vimal Jyothi Institute of Research is acknowledged.

All figures in this paper were developed by the author using standard diagrammatic design tools, and based on the relevant researches.

Informed Consent Statement

All subjects included in the study provided informed consent.

References

Afrouz, R., & Vassos, S. (2024). Adolescents' experiences of cyber-dating abuse and the pattern of abuse through technology, a scoping review. Trauma, Violence, & Abuse, 25(4), 2814-2828.

Akintayo, S. O. (2025). Impact of Social Media on Substance Abuse and Addiction among Nigerian Youth: A Call to Action for Awareness and Prevention. Available at SSRN 5599311.

Allport, G. W. (1954). The nature of prejudice. Addison-Wesley.

Bandura, A. (1977). Social learning theory. Prentice Hall.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Bilz, L., Fischer, S. M., Kansok-Dusche, J., Wachs, S., & Wettstein, A. (2024). Teachers' intervention strategies for handling hate-speech incidents in schools. Social Psychology of Education, 27(5), 2701-2724. https://doi.org/10.1007/s11218-024-09929-9

Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. Harvard University Press.

Castellanos, M., Wettstein, A., Wachs, S., Kansok-Dusche, J., Ballaschk, C., Krause, N., & Bilz, L. (2023, April). Hate speech in adolescents: A binational study on prevalence and demographic differences. In Frontiers in Education (Vol. 8, p. 1076249). Frontiers Media SA. https://doi.org/10.3389/feduc.2023.1076249

Cedena-de-Lucas, B., Amate-García, M., Arco-Tirado, J. L., & Fernández-Martín, F. D. (2026). Service-learning as a strategy to prevent online hate speech perpetration in secondary education. International Journal of Educational Research, 137, 102964. https://doi.org/10.1016/j.ijer.2026.102964

Chakravarthi, B. R., Rajiakodi, S., Ponnusamy, R., Sivagnanam, B., Thakare, S. Y., & Thangasamy, S. (2025). Detecting caste and migration hate speech

© Italian e-Learning Association

in low-resource Tamil language: BR Chakravarthi et al. Language Resources and Evaluation, 59(3), 3051-3086. https://doi.org/10.1007/s10579-025-09848-x

Chen, S., Wei, M., Wang, X., Liao, J., Li, J., & Liu, Y. (2024). Competitive video game exposure increases aggression through impulsivity in Chinese adolescents: evidence from a multi-method study. Journal of youth and adolescence, 53(8), 1861-1874.

General Education Department, Government of Kerala. (2023). Curriculum framework for value education in Kerala schools. SCERT Kerala.

Hawke, J., & Puig Larrauri, H. (2026). Towards a third side on social media. Peacebuilding, 14(1), 14-29.

Joseph, G. V., & Thomas, K. A. (2020). Volatility of digital technology enabled learning through social media: educators' apprehensions. TEST Eng Manag, 82, 5832-9.

Joseph, G. V., Jose, D., Rajan, J., Shankaran, S., Vijay, D., & Navya, V. (2026). Management graduates' attitudes to green technology integration and AI tools for sustainable business practices. Journal of Learning for Development. https://doi.org/10.56059/jl4d.v13i1.2023

Jovanovic, S., & Markovic, L. (2023). Juvenile Offenders and Victims of Digital Violence. J. Crimin. & Crim. L., 61, 27.

Kerala General Education Department. (2025). Suraksha Mitram initiative for child protection in schools. Government of Kerala. https://www.education.kerala.gov.in/

Kerala Infrastructure and Technology for Education. (2024). ICT textbook revisions: Fake news detection modules for Classes 5 and 7. Government of Kerala.

Kerala Infrastructure and Technology for Education. (2025). First Bell: Digital class content for Kerala schools. Government of Kerala. https://firstbell.kite.kerala.gov.in

Kerala Infrastructure and Technology for Education. (n.d.). ICT-enabled education initiatives in Kerala. Government of Kerala. https://kite.kerala.gov.in

KITES (2026). Welcome to KITES. https://kite.kerala.gov.in/

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. Psychological Bulletin, 140(4), 1073–1137. https://doi.org/10.1037/a0035618

Krahé, B. (2025). The social psychology of aggression. Routledge.

Livingstone, S., & Helsper, E. J. (2010). Balancing opportunities and risks in teenagers' use of the internet: The role of online skills and internet self-efficacy. New Media & Society, 12(2), 309–329. https://doi.org/10.1177/1461444809342697

Ministry of Education, Government of India. (2020). National Education Policy 2020. Government of India.

Philip, R. (2023). Curbing Hate Speech through Education: Modalities for Equipping Future Teachers. Journal of Pedagogy and Education Science, 2(03), 209-220. https://doi.org/10.56741/jpes.v2i03.350

Pittman, S. K. (2023). Beliefs about aggression as mediators of relations between community violence exposure and aggressive behavior among adolescents: Review and recommendations. Clinical Child and Family Psychology Review, 26(1), 242-258.

Prasannan, P., Kumaresan, P. K., Rajiakodi, S., Subalalitha, C. N., & Chakravarthi, B. R. (2025). Counter-speech generation for homophobic and transphobic social media content in Malayalam. Social Network Analysis and Mining, 15(1), 87. https://doi.org/10.1007/s13278-025-01507-x

Ranjith, P. J., Vranda, M. N., & Kishore, M. T. (2023). Predictors, prevalence, and patterns of cyberbullying among school-going children and adolescents. Indian journal of psychiatry, 65(7), 720-728. https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_313_23

Ranjith, P. J., Vranda, M. N., & Kishore, M. T. (2025). Development and validation of school-based intervention on cyberbullying for adolescents. Indian journal of psychiatry, 67(2), 252-255. https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_777_23

Rudnicki, K., Vandebosch, H., Voué, P., & Poels, K. (2023). Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. Behaviour & Information Technology, 42(5), 527-544.

Singh, S., & Singh, S. (2025). Perceptions of Indian students towards bullying: Intervention through bullying intervention module (BIM). SAGE Open, 15(1). https://doi.org/10.1177/21582440241305199

State Council of Educational Research and Training Kerala. (n.d.). Academic guidelines and curriculum development programmes. https://scert.kerala.gov.in

Student Police Cadet. (n.d.). SPC: Overview. https://studentpolicecadet.info/overview/spc

© Italian e-Learning Association

Sunstein, C. R. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press.

Suresh, S., & Vijaya, R. (2024). Mapping the literature on school bullying in India: A scoping review. Aggression and Violent Behavior, 77, 101978. https://doi.org/10.1016/j.avb.2024.101978

Törnberg, A., & Törnberg, P. (2024). From echo chambers to digital campfires: The making of an online community of hate in Stormfront. In Social processes of online hate (pp. 93-119). Routledge.

UNESCO. (2019). Digital literacy in education: Policy brief. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000369841

UNICEF India. (2020). Protecting children in the digital age: Challenges and responses in India. UNICEF. https://www.unicef.org/india

Vasconcellos-Silva, P. R., & Castiel, L. D. (2025). Chambers that echo hate, bubbles that distill fear: the constitution of Self and intolerance as roots of disinformation. Ciência & Saúde Coletiva, 30, e10492023.

Wachs, S., Schittenhelm, C., Kops, M., Gámez-Guadix, M., & Wright, M. F. (2025). Happiness Through HateLess? Examining the Direct and Indirect Effects of an Anti-Hate Speech Program on Victimized and Non-Victimized Youth. Journal of adolescence, 97(6), 1645-1655. https://doi.org/10.1002/jad.12525

Wachs, S., Wettstein, A., Bilz, L., Espelage, D. L., Wright, M. F., & Gámez-Guadix, M. (2024). Individual and contextual correlates of latent bystander profiles toward racist hate speech: A multilevel person-centered approach. Journal of youth and adolescence, 53(6), 1271-1286.

Zhu, C., Huang, S., Evans, R., & Zhang, W. (2021). Cyberbullying among adolescents and children: A comprehensive review of the global situation, risk factors, and preventive measures. Frontiers in Public Health, 9, Article 634909. https://doi.org/10.3389/fpubh.2021.634909

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), pp. 61-74

When AI is sexist: Investigating responses to AI and human discrimination

Selen Akay[a,1], Junko Kanero[a], Sabahat Cigdem Bagci[a]

[a]*Faculty of Arts and Social Sciences, Sabancı University – Istanbul (Türkiye)*

(submitted: 10/03/2026; accepted: 9/6/2026; published: 23/7/2026)

Abstract

As artificial intelligence (AI) systems prove their utility in various aspects of daily life, their potential risk to societal progress is still debated given the increasing rates of their discriminatory behaviors. Some preliminary evidence suggests that the same discriminatory behavior may be perceived differently when performed by a person or an AI system, yet it remains largely unknown how people perceive and respond to AI's discriminatory behaviors. Through two studies conducted in Türkiye (Study 1) and the United States (Study 2), participants read about either AI's discrimination against women, humans' discrimination against women, or an unrelated topic, and indicated the degree of perceived threat to women, willingness to participate in collective action for women's equal treatment, and their attitudes toward women and feminists. In both cultures, discrimination was perceived to be less threatening to women when committed by AI compared to a human, yet collective action intentions and attitudes toward women and feminists did not differ across conditions. Notably, American men consistently reported less favorable attitudes toward women after reading about discrimination. These findings suggest discrimination by AI requires special attention and underline the importance of investigating the influence of technologies on sustaining intergroup inequality.

KEYWORDS: Discrimination, Intergroup Threat, Artificial Intelligence (AI), Collective Action.

DOI
https://doi.org/10.20368/1971-8829/1136320

CITE AS
Akay, S., Kanero, J., & Bagci,S.C. (2026). When AI is sexist: Investigating responses to AI and human discrimination. *Journal of e-Learning and Knowledge Society*, *22*(1), 61-74. https://doi.org/10.20368/1971-8829/1136320

1. Introduction

Recent technological advancements have led artificial intelligence (AI) tools to be available for the use of the general public. AI tools can perform a wide range of tasks, from writing elaborate essays to create workout routines and giving relationship advice. Considering the availability and popularity of AI technologies, the ethical implications of interacting with AI systems, as well as the potential risks of incorporating them into our lives, have become a central topic of debate (e.g., Darling, 2012, 2015). While AI technologies have the potential to reduce social inequality by providing affordable tools to perform tasks that are otherwise challenging for disadvantaged groups, they can also exacerbate inequality in various ways, particularly by discriminating against already disadvantaged populations.

Technologies often make seemingly avoidable mistakes or errors, ranging from technical errors to social norm violations (Giuliani et al., 2015; Tian & Oviatt, 2021). Discrimination – the inappropriate and unfair negative treatment of others based on their group identities (Allport, 1954; Dovidio et al., 2010) – is one of such behaviors that AI systems occasionally display based on the user's characteristics or the input they give to the AI. One explanation for AI discrimination is algorithmic bias, which arises when algorithms are trained on data that reflect existing social biases (Danks & London, 2017). Such biases may be encoded intentionally by developers or emerge unintentionally from exclusionary datasets that overrepresent Western, White, and male populations (Bender et al., 2021; Hall & Ellis, 2023). Because algorithms tend to overgeneralize patterns in available data, they can produce biased inferences – such as underestimating women's suitability for STEM roles or using postcodes as proxies for race in crime-risk assessments – thereby reinforcing existing inequalities even when datasets appear neutral (Dastin, 2018; Allen & Masters, 2020)

[1] corresponding author - email: selen.akay@sabanciuniv.edu – address: Orta Mahalle, Universite Caddesi No:27, 34956 Tuzla, Istanbul, Turkiye

© Italian e-Learning Association

and disproportionately impacting disadvantaged groups such as women, racial/ethnic minorities, and sexual minorities (e.g., Ferrara, 2023). While perceiving discrimination can adversely affect targets by reducing their well-being (e.g., Bagci et al., 2020), and damaging their physical and mental health (e.g., Borrell et al., 2007; Landrine & Klonoff, 1996; Pascoe & Smart Richman, 2009; Schmitt et al., 2014; Williams et al., 2003), exposure to AI discrimination may also shape both disadvantaged and advantaged group members' willingness to take action against such unfair treatment. The current studies investigate this assumption and test the implications of AI discrimination (compared to human discrimination) for advantaged and disadvantaged group members' evaluation of the target group and the extent at which they are willing to fight for disadvantaged groups' rights.

A first step to understand reactions to AI discrimination is to understand how group members respond to discrimination experiences across status, irrespective of the source. One robust finding is that advantaged and disadvantaged groups perceive discrimination and group-based inequality differently (e.g., Knowles & Lowery, 2012). Due to power imbalances and structural limitations, disadvantaged or low-status groups are more frequently stigmatized and discriminated against (e.g., Link & Phelan, 2001) and the perception of discrimination is often a more internal, uncontrollable, and chronic social experience (Schmitt et al., 2002). In contrast, advantaged groups are less likely to perceive discrimination compared to disadvantaged groups, and they are less likely to challenge the status quo (e.g., Brown et al., 2022; Caricati & Owuamalam, 2020; Lowery et al., 2007; Sidanius & Pratto, 1999), and may even engage in collective action that aim to maintain the existing system favoring group-based inequality (e.g., Guvensoy et al., 2025; Osborne et al., 2019).

Among disadvantaged group members, seeking social change through collective action is one of the most effective responses to inequality (Ellemers et al., 1990; Wright et al., 1990). Collective action research suggests that there are many drivers of collective action behaviors, but one of the main motivations is perceived injustice and discrimination, often fueling the emotion of anger among disadvantaged group members (e.g., Van Zomeren et al., 2008). Indeed, research shows that people's recognition of group-based disadvantages drives people to challenge discrimination (Ellemers & Barreto, 2009; Wright et al., 1990; Wright & Tropp, 2002). For example, Uluğ and colleagues (2023) demonstrated that women who have witnessed gender discrimination through an experimental procedure were more willing to engage in collective action for gender equality. Thus, exposure to discrimination targeting one's own disadvantaged group is likely to increase disadvantaged group members' collective action intentions.

Moreover, both theoretical and empirical work suggests that perceiving discrimination or similar experiences of devaluation or exclusion targeting the ingroup may lead disadvantaged group members to identify with their ingroup more, thereby protecting themselves from the adverse effects of discrimination (Branscombe et al., 1999). Findings suggest that identifying with one's ingroup after perceiving group-based prejudice and discrimination may lead disadvantaged groups to evaluate their ingroup more positively (Dion & Earn, 1975). Thus, exposure to AI's discrimination against disadvantaged groups may potentially improve disadvantaged group members' attitudes toward the targeted group and their advocates (e.g., political groups such as feminists) and increase their willingness to engage in collective action for their equal treatment.

In contrast to disadvantaged groups, advantaged group members may either support equality struggles as allies or resist them to preserve the status quo (Kutlaca et al., 2022; Flood et al., 2021). Collective action research has traditionally portrayed advantaged groups as less supportive of equality, largely because they are motivated to maintain their privileges and perceive social change as a zero-sum loss (e.g., Sidanius & Pratto, 1999; Caricati & Owuamalam, 2020; Brown et al., 2022). Accordingly, advantaged groups are less likely to recognize discrimination against disadvantaged groups and may deny inequality when confronted with their own privilege (Knowles & Lowery, 2012; Phillips & Lowery, 2015), a pattern illustrated by findings showing that men and White individuals are less likely to view gender and racial inequalities as unjust (Miron et al., 2011).

Nevertheless, advantaged group members are also likely to respond to discrimination targeting disadvantaged group members in various ways. For example, research demonstrates that advantaged groups who perceive other groups as relatively deprived are more likely to support pro-equality policies and affirmative actions that help the disadvantaged (e.g., Mazzuca et al., 2022; Tougas & Veilleux, 1990; Veilleux & Tougas, 1989). Previous research has shown evidence that exposure to disadvantaged group members' discrimination may lead advantaged group members to display more positive attitudes toward the discriminated outgroup (Bagci et al., 2017) and greater motivation to engage in collective action supporting intergroup equality (Uluğ & Tropp, 2021), in the context of ethnic and racial relationships. Notably, some findings indicate that inequality, historical mistreatment, and conflict with disadvantaged groups can lead advantaged groups to feel more positively about the disadvantaged and reprimand their mistreatment, due to feelings such as guilt and shame (e.g., Allpress et al., 2010, 2014; Brown & Cehajic, 2008; Swim & Miller, 1999). Thus, it could be expected that advantaged groups would feel more positively about disadvantaged groups and their advocates, and would be more likely to engage in collective action

© Italian e-Learning Association

supporting intergroup equality after being exposed to the target group's discrimination.

Overall, perceiving discrimination and group-based inequalities may make both advantaged and disadvantaged groups more willing to engage in collective action for equality and improve their attitudes toward the disadvantaged and their advocates. Nevertheless, previous research has barely touched upon whether and how discrimination by AI would be evaluated by both advantaged and disadvantaged group members. Although responses to AI discrimination may parallel familiar intergroup processes, AI and related technologies are treated as novel social entities (Gray et al., 2025; Kahn & Shen, 2017), highlighting the importance of understanding the unique dynamics and consequences of this form of discrimination.

As AI systems become incorporated into domains that influence people's lives, such as finance, education, and job recruitment, concerns about AI discrimination have become more relevant than ever. However, since AI discrimination is only a recent concern, it is unclear whether it triggers the same reactions as human discrimination. One reason why AI's discriminatory behaviors may affect intergroup relationships differently than humans' discriminatory behaviors is because AI and similar technologies are not attributed as much mind (i.e., agentic and experiential capabilities) as humans (Gray et al., 2007). Therefore, one potential scenario is that people may perceive AI as less of a threat because, while AI and similar technologies are attributed some moral responsibility and intentionality, they are perceived to be less agentic compared to humans (Bigman et al., 2019; Floridi & Sanders, 2004; Geiselmann et al., 2023; Shank & DeSanti, 2018; Shank et al., 2021), overall making their discrimination relatively less intentional and thereby less threatening compared to human discrimination. This assumption also suggests that perceiving AI discrimination as less threatening may be also less likely to motivate both advantaged and disadvantaged group members' support and willingness to fight on behalf of disadvantaged group members' rights.

Supporting this suggestion, empirical evidence demonstrates that AI's discriminatory behaviors and moral violations are evaluated differently from humans' discriminatory behaviors and moral violations (Bigman et al., 2023; Shank & DeSanti, 2018). For instance, Bigman and colleagues (2023) compared the extent to which participants felt moral outrage (i.e., anger or disgust triggered by perceived violations of moral values) due to the discriminatory behaviors of algorithms and humans in a hiring setting. Across both racial and gender contexts, algorithms' discriminatory behaviors evoked less moral outrage in participants, and participants attributed less prejudice to the algorithm than they did to the human recruiter. Thus, even when a person and an AI system engage in the same discriminatory behavior, resulting in the same negative consequences, the behavior of AI discriminators are likely to be dismissed. However, the dismissal of AI discriminators may not necessarily indicate that their discriminatory behaviors are inconsequential since such biases can lead to the perpetuation of harmful stereotypes (e.g., Vlasceanu & Amodio, 2022) and make discrimination of particular disadvantaged groups more normative and legitimate.

Overall, AI's biases and discriminatory tendencies may lead to awkward, tense, or outrageous situations, perpetuating harmful stereotypes. While some work compared people's reactions to discrimination by AI's and humans (Bigman et al., 2023), no research investigated whether AI's discriminatory behaviors against disadvantaged groups affect people's attitudes toward the target group and their motivation to engage in collective action.

The current research uses gender inequality in two different countries as a context to investigate whether reading about AI's discriminatory behaviors against disadvantaged groups (i.e., women) influences people's collective action intentions for the disadvantaged group's equality, and their attitudes toward feminists (Study 1) and women (Study 2). Gender inequality is a particularly relevant context for examining AI discrimination because of its universality and commonality (United Nations Development Programme, 2023) and the marked gender imbalance among AI developers (Wajcman, 2010; Vargas-Solar, 2022).

Although gender discrimination is a global issue, reactions to gender inequality may vary across cultural contexts. Türkiye (Study 1) ranks relatively low on global gender equality indices, and gender discrimination against women is widely documented as structurally embedded in society (Kandiyoti, 2016; Özdemir Sarıgil, 2022; UNDP, 2023; World Economic Forum, 2024). At the same time, AI systems are increasingly integrated into sectors such as education and healthcare, and despite slower overall adoption compared to other countries, the Turkish population shows relatively high levels of AI use and familiarity (Eurostat, 2025; Maslej et al., 2025). Given Türkiye's persistent gender inequality alongside growing exposure to AI, it offers a relevant context for examining reactions to AI-driven gender bias. On the other hand, the United States (Study 2) ranks much higher in gender equality than Türkiye (World Economic Forum, 2025) and is home to several major AI research and development companies, including OpenAI, Google, and Meta, and AI technologies are widely deployed across multiple sectors (e.g., McElheran et al., 2024). Nevertheless, many documented cases of AI discrimination involve US-based companies or residents (e.g., Dastin, 2018; U.S. Equal Employment Opportunity Commission, 2022), providing two unique socio-cultural contexts for the investigation of the current research questions.

© Italian e-Learning Association

Based on the perceived discrimination literature, we expected that, irrespective of the source, perceived discrimination is likely to improve attitudes toward the target group and motivate both advantaged and disadvantaged group members to support women's rights. Considering that both advantaged (Batson et al., 1997, 2002; Swim & Miller, 1999; Vescio et al., 2003) and disadvantaged groups (Becker & Barreto, 2014; Dion & Earn, 1975; Osborne & Sibley, 2015; Uluğ et al., 2023) are likely to approve egalitarian policies and feel more positively about the targeted group and targets who confront prejudice, they may also feel positively about people who advocate for the disadvantaged (i.e., feminists). Therefore, we expected that:

- H1: Compared to participants in the control condition, participants in the human and AI discrimination conditions would both perceive greater threat to women, show more positive attitudes toward feminists (Study 1) and women (Study 2), and greater motivation to engage in collective action supporting women's rights.

Based on findings that demonstrate AI and similar technologies to be attributed less capacity for intentional behavior and prejudice as humans (Bigman et al., 2023; Gray et al., 2007; Shank & DeSanti, 2018), and since AI's discrimination evokes less moral outrage than humans' discrimination (Bigman et al., 2023), we expected that people would perceive AI discrimination as less threatening (harmful) to women than human discrimination:

- H2: Compared to the AI discrimination condition, participants in the human discrimination condition would perceive greater threat to women, show more positive attitudes toward feminists (Study 1) and women (Study 2), and greater motivation to engage in collective action supporting women's rights.

2. Study 1

Study 1 was conducted in Türkiye and examined both advantaged (men) and disadvantaged (women) group members' reactions to AI discrimination, compared to human discrimination and the control condition.

2.1 Materials and Method

2.1.1 Participants and Procedure

We recruited 350 participants through both convenience sampling and a research participation system at Sabanci University. A priori power analysis with G*Power (Faul et al., 2007) suggested at least 330 participants to reach 80% power with $\alpha = .05$ and a small effect size (estimated based on Bigman et al., 2023).

Six participants who did not identify as men or women were removed from all analyses as this study focuses on contrasting the advantaged (men) and disadvantaged groups (women). As the duration data were heavily skewed, we first performed a log10 transformation to identify outliers and then removed participants who took 3 SDs above average. Seven participants who took more than 109 minutes to complete the study were also removed from the data. The final sample consisted of 337 native Turkish-speaking adults ($M_{age} = 33.43$, $SD = 14.91$, 233 women). Participants reported having an average income ($M = 4.77$, $SD = 1.09$, ranging from 1 = *Low* to 7 = *High*), being moderately informed about AI systems ($M = 4.14$, $SD = 1.60$), and having an average amount of experience using AI systems ($M = 4.16$, $SD = 2.05$).

Participants were assigned to one of the three conditions: AI discriminator, human discriminator, or control condition. In the AI discriminator and human discriminator conditions, participants read a fake news report about AI systems or humans discriminating against women. In the human discriminator condition, the source of the discrimination was identified as "industry experts". Participants in the control condition read a fake news report on an unrelated topic. After reading the news report, participants responded to a questionnaire consisting of a manipulation check, scales about perceived threats toward women, collective action intentions for women's equal treatment, and their attitudes toward feminists, demographic questions, and a suspicion check. Students recruited through the course credit system were awarded course credits for their participation.

2.1.2 Materials

Fake news report. Fake news reports discussed either AI systems' discriminatory behaviors toward women, industry experts' discriminatory behaviors toward women, or an unrelated text about the agriculture sector. The reports were created based on examples of discrimination available from online sources and existing news articles that discuss AI's discriminatory behaviors (Dastin, 2018; Jonker & Rogers, 2024), and a team of researchers evaluated the texts to ensure the clarity of the writing and the manipulation (see Supplementary Materials).

Manipulation check. Participants rated the extent to which they perceived discrimination against women in the news report, on a scale of 1 (*Strongly Disagree*) to 7 (*Strongly Agree*).

Perceived threat toward women. The measure consisted of three items (e.g., '[Artificial intelligence systems/industry experts/the agriculture sector] can damage women's physical, mental, or emotional health'), all rated on a scale of 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). AI systems ($\alpha = .89$), industry experts ($\alpha = .91$), and the agriculture sector ($\alpha = .92$) all demonstrated good reliability.

© Italian e-Learning Association

Collective action intentions. Collective action intentions for women's equal treatment were measured with three items assessing their willingness to donate, attend a protest, and sign a petition for women's equal treatment, on a scale of 1 (*Very Unlikely)* to 7 (*Very Likely*). The scale was adapted from prior normative collective action measures used in the literature (e.g., Hoskin et al., 2019), and demonstrated good reliability ($\alpha = .73$).

Attitudes toward feminists. Attitudes toward feminists were measured with 6 bipolar items with a 7-point scale (e.g., 1 = *Contempt*, 7 = *Respect*) (e.g., Bagci et al., 2020; Wright et al., 1997), with higher scores indicating more positive attitudes ($\alpha = .92$).

Demographics. Participants indicated their age, gender, ethnicity, economic status, and knowledge about and experience with AI systems.

2.2 Results

2.2.1 Manipulation Check

A two-way ANOVA with Condition and Gender as predictors resulted in the significant effect of Condition ($F(2, 331) = 53.05$, $p < .001$, $\eta_p^2 = .243$). Tukey's HSD tests revealed that participants in the AI discriminator condition ($M = 5.80$, $SD = 1.63$) perceived discrimination against women to be higher than those in the control condition ($M = 3.70$, $SD = 2.02$; $p < .001$), but did not differ from those in the human discriminator condition ($M = 6.06$, $SD = 1.37$; $p = .468$). Moreover, Gender significantly influenced discrimination perceptions ($F(1, 331) = 10.32$, $p = .001$, $\eta_p^2 = .030$). Specifically, female participants ($M = 5.48$, $SD = 1.92$) perceived more discrimination against women than did male participants ($M = 4.64$, $SD = 1.99$). No interaction effect was found ($p = .607$).

2.2.2 Main Effects of Condition

A MANOVA with Condition as the predictor assessed whether there are significant differences across conditions regarding perceived threat, collective action intentions for women's equal treatment, and attitudes toward feminists. Multivariate tests showed that Condition had a significant effect ($F(6, 666) = 16.78$, $p < .001$, $\eta_p^2 = .131$), and the univariate tests revealed the effect of Condition on threat perceptions toward women ($F(2, 334) = 56.19$, $p < .001$, $\eta_p^2 = .252$). Tukey's HSD tests showed that perceived threat toward women was higher in the human discriminator condition ($M = 5.56$, $SD = 1.53$) compared to the AI discriminator ($M = 4.52$, $SD = 1.54$; $p < .001$) and control conditions ($M = 3.26$, $SD = 1.83$; $p < .001$). Similarly, perceived threat in the AI discriminator condition was significantly higher than control ($p < .001$). Collective action intentions for women's equal treatment ($p = .681$) and attitudes toward feminists ($p = .096$) did not significantly differ across conditions. These findings were replicated across the student sample and the community sample.

2.2.3 Exploratory Analyses

An exploratory MANOVA tested whether participants' gender predicted perceived threat to women, collective action intentions, and attitudes. The multivariate effects of Condition ($F(6, 660) = 13.07$, $p < .001$, $\eta_p^2 = .106$) and Gender were significant ($F(3, 329) = 24.55$, $p < .001$, $\eta_p^2 = .183$), but their interaction was not ($p = .138$). The unique effect of Condition on perceived threat toward women remained significant ($F(2, 331) = 41.98$, $p < .001$, $\eta_p^2 = .202$). Condition did not affect collective action intentions ($p = .605$) nor attitudes toward feminists ($p = .213$).

Gender significantly and uniquely affected perceived threat toward women ($F(1, 331) = 10.75$ $p = .001$, $\eta_p^2 = .031$), collective action intentions for women's equal treatment ($F(1, 331) = 32.91$, $p < .001$, $\eta_p^2 = .090$), and attitudes toward feminists ($F(1, 331) = 57.37$, $p < .001$, $\eta_p^2 = .148$). Female participants perceived significantly more threat ($M = 4.72$, $SD = 1.87$) than male participants ($M = 3.95$, $SD = 1.82$). Similarly, female participants' collective action intentions ($M = 5.96$, $SD = 1.06$) and attitudes toward feminists ($M = 5.28$, $SD = 1.33$) were significantly higher than male participants' collective action intentions ($M = 5.14$, $SD = 1.39$) and attitudes toward feminists ($M = 4.03$, $SD = 1.45$). We found no interaction effects (*p's* > .100).

Lastly, two exploratory mediation analyses were conducted via the Process Macro (Hayes, 2022; Model 4) to see whether perceived threat mediated the effect of condition (IV) on collective action intentions for women's equal treatment and attitudes toward feminists (DV). Indeed, perceived threat mediated the effect of condition on collective action intentions for women's equal treatment ($B = .10$, $SE = .04$, 95% CI [0.02, 0.18]). Moreover, the relative indirect effect of the human discriminator condition ($B = .22$, $SE = .11$, 95% CI [0.01, 0.45]) was stronger than that of the AI discriminator condition ($B = .12$, $SE = .06$, 95% CI [0.00, 0.26]), as indicated by the coefficients. Similarly, perceived threat mediated the effect of condition on attitudes toward feminists ($B = .23$, $SE = .09$, 95% CI [0.06, 0.43]). Again, the relative indirect effect of the human discriminator condition ($B = .39$, $SE = .14$, 95% CI [0.14, 0.67]) was stronger than that of the AI discriminator condition ($B = .22$, $SE = .08$, 95% CI [0.07, 0.39]), as indicated by the coefficients.

2.3 Discussion

The results suggest that, while discrimination by AI and humans was perceived as a potential threat to women (partial support for H1), AI discrimination was perceived as a lower threat compared to human discrimination. This finding is in line with existing evidence indicating that AI is believed to be less agentic (Gray et al., 2007) and evokes less moral outrage and is attributed less prejudice (Bigman et al., 2023). Unlike our expectations, however, collective action intentions or attitudes toward feminists remained unaffected by

© Italian e-Learning Association

the news report about discrimination. This finding was surprising, as prior research suggests that, in response to unequal treatment, both targets and non-targets become more willing to engage in collective action for disadvantaged groups (Mazzuca et al., 2022; Moscatelli et al., 2025), and report more positive attitudes toward discriminated groups (Bagci et al., 2017). While we did not find significant differences across conditions, we found that perceived threat mediated the effect of condition on collective action intentions as well as attitudes. Although the effect sizes are small, these results indicate that threat from human and AI discriminators relate to intergroup outcomes.

3. Study 2

Study 1 established that AI discrimination yields lower threat than human discrimination, yet the conclusion should not be drawn based on a single experiment. To evaluate the reliability and generalizability of the initial results and examine potential cross-cultural variation, Study 2 tested participants in the US. As discussed in the introduction, Türkiye ranks relatively low in global indices of gender equality and AI-related development. One possibility is that this context contributed to AI being perceived as less threatening to women in Study 1. The US, on the other hand, ranks higher on these dimensions, which may lead individuals to perceive AI systems as posing a greater threat to womanhood. While largely mirroring Study 1, Study 2 examined not only attitudes toward feminists (to replicate findings of Study 1) but also the more general group of women. The adjustment was made because the lack of attitude changes toward feminists in Study 1 might not have quite represented the targeted discriminated group (women); indeed, feminists have been previously associated with negative perceptions representing a highly marginalized social group (e.g., Jenen et al., 2009).

3.1 Materials and Method

3.1.1 Participants and Procedure

A total of 350 participants from the US were recruited via Prolific. As in Study 1, the data of participants who did not identify as men or women were removed (N = 8). The final sample included 342 participants (M_{age} = 39.71, SD = 12.61, 206 women). Similarly to Study 1, participants reported an average economic status (M = 4.30, SD = 1.35), adequate information about AI systems (M = 4.84, SD = 1.51), and average experience using AI systems in their daily lives (M = 4.34, SD = 1.96). The procedure of Study 2 was identical to that of Study 1.

3.1.2 Materials

Fake news reports. Three minor changes were made in the Study 1 news reports: an additional example of discrimination was included in the AI discriminator and human discriminator conditions, the human discriminators were changed from industry experts to "specialized officials," and the topic of the news report in the control condition was changed to the service industry. A team of researchers evaluated the texts to ensure clarity and consistency with the material from Study 1 (see Supplementary Materials).

Perceived threat to women. Due to the change in the topic of the news report for the control condition, the wording 'agriculture sector' was changed to 'service industry' in Study 2. The scales for AI systems (α = .92), specialized officials (α = .93), and the agriculture sector (α = .95) all demonstrated good reliability.

Collective action intentions. We added an item addressing collective action intentions for women's equal treatment through social media, which was adapted from previous research (e.g., Uluğ et al., 2023). Participants indicated their intentions to share and like social media content related to women's equal treatment (i.e., 'How likely are you to share and like related social media content for women's equal treatment?'), using the same 7-point scale (α = .90).

Attitudes toward women and feminists. In addition to measuring attitudes toward feminists, we used the same 6 bipolar statements to measure attitudes toward women, using a 7-point bipolar scale. Both attitudes toward women (α = .94) and feminists (α = .98) demonstrated good reliability.

3.2 Results

3.2.1 Manipulation Check

A Condition x Gender ANOVA revealed significant differences across the three conditions on the perceived discrimination against women (F(2, 336) = 72.79, p < .001, η_p^2 = .302). Tukey's HSD tests showed that the perceived discrimination against women was stronger in the AI discriminator condition (M = 5.24, SD = 1.84) than in the control condition (M = 3.11, SD = 1.86, p < .001), but not significantly different from the human discriminator condition (M = 5.76, SD = 1.44; p = .060). Unlike Study 1, no main effect of Gender was observed (p = .837). There were no significant interaction effects (p = .994).

3.2.2 Main Effects of Condition

A MANOVA with all dependent variables found the multivariate effect of Condition (F(8, 674) = 9.10, p < .001, η^2 = .097). As in Study 1, the univariate effect of Condition on perceived threat to women was significant (F(2, 339) = 30.16, p < .001, η_p^2 = .151). Tukey's HSD tests showed that the perceived threat toward women was stronger in the human discriminator condition than in the AI discriminator (p = .036) and control conditions (p < .001). Similarly, the AI discriminator condition perceived higher threat to women than the control condition (p < .001). We found no significant effect of Condition on collective action intentions for

© Italian e-Learning Association

women's equal treatment (p = .093), attitudes toward feminists (p = .089), and attitudes toward women (p = .168).

3.2.3 Exploratory Analyses

To explore whether gender influenced perceived threat, collective action intentions, and attitudes toward women and feminists, another MANOVA was conducted with Gender and Condition as predictors. The multivariate effects of Condition ($F(8, 668) = 9.01$, $p < .001$, $\eta_p^2 = .097$), Gender ($F(4, 333) = 4.20$, $p = .002$, $\eta_p^2 = .048$), and the interaction was significant ($F(8, 668) = 2.02$, $p = .042$, $\eta_p^2 = .024$). Consistent with the initial MANOVA, the current analysis revealed a significant difference in perceived threat toward women across conditions ($F(2, 336) = 28.10$, $p < .001$, $\eta_p^2 = .143$). Similarly, the univariate effect of Condition on attitudes toward feminists was significant ($F(2, 336) = 3.38$, $p = .035$, $\eta_p^2 = .020$). However, Tukey's HSD tests did not reflect this univariate effect, as there were no significant differences across the three conditions in participants' attitudes toward feminists (p's > .090). No significant univariate effect of Condition emerged for participants' collective action intentions for women's equal treatment (p = .055), and their attitudes toward women (p = .058).

Gender significantly affected participants' collective action intentions for women's equal treatment ($F(1, 336) = 13.27$, $p < .001$, $\eta_p^2 = .038$), and attitudes toward feminists ($F(1, 336) = 10.11$, $p = .002$, $\eta_p^2 = .029$). The means indicate that collective action intentions for women's equal treatment were higher for female participants (M = 5.26, SD = 1.62) than male participants (M = 4.55, SD = 1.87), and attitudes toward feminists showed the same pattern (M_F = 5.29, SD_F = 1.64; M = 4.72, SD = 1.73). Female and male participants did not differ in their perceived threat toward women (p = .088), and attitudes toward women (p = .053).

Moreover, an interaction between condition and gender predicted attitudes toward women ($F(2, 336) = 5.59$, p = .004, η_p^2 = .032). To investigate this interaction further, we conducted a follow-up univariate ANOVA on attitudes toward women. Tukey's HSD tests revealed that male participants in the human discriminator condition reported less favorable attitudes toward women (M = 5.67, SD = 1.25) than male participants in the control condition (M = 6.37, SD = .81, p = .006), female participants in the human discriminator condition (M = 6.39, SD = .94, p = .002), female participants in the AI discriminator condition (M = 6.32, SD = .85, p = .005), and female participants in the control condition (M = 6.25, SD = 1.19, p = .030). However, the difference was not significant between male participants in the human discriminator condition (M = 5.67, SD = 1.25) and male participants in the AI discriminator condition (M = 6.27, SD = .78, p = .061). We found no significant interaction effects for perceived threat to women (p = .429), collective action intentions for women's equal treatment (p = .209), and attitudes toward feminists (p = .056).

Lastly, we investigated whether perceived threat mediated the effect of condition (IV) on collective action intentions for women's equal treatment, attitudes toward feminists, and attitudes toward women (DV). Unlike Study 1, perceived threat did not mediate the effect of condition on collective action intentions for women's equal treatment (B = .10, SE = .06, 95% CI [-0.02, 0.23]), In contrast, perceived threat mediated the effect of condition on attitudes toward feminists (B = .19, SE = .06, 95% CI [0.07, 0.31]). The relative indirect effect of the human discriminator condition (B = .29, SE = .10, 95% CI [0.10, 0.49]) was stronger than that of the AI discriminator condition (B = .19, SE = .08, 95% CI [0.06, 0.37]), as indicated by the coefficients. Similarly, perceived threat mediated the effect of condition on attitudes toward women (B = .09, SE = .36, 95% CI [0.01, 0.16]). Again, the relative indirect effect of the human discriminator condition (B = .13, SE = .06, 95% CI [0.03, 0.25]) was stronger than that of the AI discriminator condition (B = .09, SE = .04, 95% CI [0.02, 0.18]).

3.3 Discussion

We replicated some findings from Study 1, but also uncovered some unique effects. First, Study 2 partially supported H1 and H2, demonstrating that both forms of discrimination trigger threats to women but AI discrimination was perceived as less threatening than human discrimination. These results align with Study 1 and the prior literature where AI systems were perceived as less threatening, presumably due to their lack of capacity for voluntary action (Bigman et al., 2023; Gray et al., 2007; Shank & DeSanti, 2018). Also replicating Study 1, no significant differences were found in collective action intentions for women's equal treatment and attitudes toward feminists and women (H3) in the hypothesized direction. Moreover, an exploratory MANOVA revealed a significant interaction between gender and condition, demonstrating that male participants who read about humans' discriminatory behaviors reported less favorable attitudes toward women, compared to male participants in the control condition and female participants. This pattern was not observed in Study 1 with Turkish participants and thus may be culturally specific, though further investigation is needed. This interaction can be explained by prior research demonstrating the advantaged group's willingness to oppose equality movements to 'protect' their privileges (e.g., Guvensoy et al., 2025; Van Laar et al., 2024). Lastly, in contrast to Study 1, the mediating effect of perceived threat emerged for attitudes but not for collective action intentions.

© Italian e-Learning Association

4. Discussion and Conclusions

The current studies examined whether and how discrimination by AI affects perceived threat toward the targeted group (women), attitudes toward the targeted and advocate groups (women and feminists), and collective action intentions for the targeted group. In two studies conducted in Türkiye and the US, we compared participants' responses to AI discrimination, human discrimination, and no discrimination. We hypothesized that, irrespective of the source, being exposed to gender discrimination would lead to increases in perceived threat to women, attitudes toward women and feminists, and motivation to engage in collective action for gender equality. We also expected that those reactions would be dampened when AI is the discriminator, given that AI is generally attributed less intention and agency, compared to humans (e.g., Bigman et al., 2023; Gray et al., 2007). Although attitudes and collective action intention remained unaffected, both discrimination conditions resulted in greater perceived threat to women, and discrimination was seen less threatening and harming when sourced by AI than humans.

These findings offer important implications regarding the potential risks of integrating AI into human life. While many recognize the benefits of AI for societal progress, its role in challenging – or reinforcing – existing hierarchical social structures has not received sufficient attention. Our studies demonstrated that discrimination was not seen as harmful when done by AI. Since discrimination by AI is not perceived as threatening as discrimination by humans, people may overlook the effects of its biased behaviors. In turn, AI systems may amplify existing unequal treatment of disadvantaged groups, and perhaps more critically, with such harmful outcomes insufficiently addressed, AI can normalize and further perpetuate social inequalities with limited accountability. Thus, it is critical for AI developers and policy makers to carefully consider the potential long-term risks of AI-driven discrimination.

On the other hand, collective action intentions for women's equal treatment and attitudes toward feminists and women did not differ across the three conditions. The findings may appear surprising because previous studies found reading about discrimination to result in positive intergroup outcomes (e.g., Bagci et al., 2017; Uluğ et al., 2023). Importantly, however, research also offers a more nuanced picture and shows that merely recognizing discrimination does not necessarily entail collective action intentions (Reimer et al., 2017), with recognition of one's privileges posing larger effects on willingness to act (e.g., Uluğ & Tropp, 2021). Research also highlights moral convictions and emotional reactions, like anger and sympathy, as key motivators of collective action in both advantaged (Iyer & Ryan, 2009; Van Zomeren et al., 2011) and disadvantaged groups (Iyer & Ryan, 2009; Van Zomeren et al., 2008, 2012) Likewise, perceived efficacy (i.e., the extent to which people believe in the impact of their actions) is another key motivator of collective actions (e.g., Van Zomeren et al., 2008). Taken together, we speculate that the specific scenario employed in the current work might not have been sufficient in eliciting emotional arousal or a sense of efficacy necessary to inspire action.

Notably, American male participants in the human discriminator condition reported less favorable attitudes toward women than male participants in the control condition and female participants. Research reports that advantaged groups show disinterest in collective action and support for the disadvantaged due to their zero-sum beliefs about equality (e.g., Van Laar et al. 2024). Closely imitating real-life circumstances, the human discriminator condition might have highlighted the differential treatment of men and women, leading enough male participants to show their willingness to maintain the favorable status quo. This pattern, however, was not observed in Türkiye. One speculative yet plausible account is that, compared to American men, Turkish men are less concerned about losing their privileges because the patriarchal system is perceived as rather stable (e.g., Bacchini et al., 2024).

Lastly, in both studies, male participants were in general less willing to engage in collective action for women's equal treatment and reported less favorable attitudes toward feminists. This gender difference aligns with existing findings in the literature that show women to be more motivated to defend their equality and feel more positively about feminists (e.g., Iyer & Ryan, 2009; Lemaster et al., 2015). While there were further differences across the two contexts in terms of how men and women responded to discrimination in general, these findings did not specifically relate to AI discrimination, therefore revealing that responses to AI discrimination were not substantially different across group status and country.

4.1 Limitations and Future Directions

Although the current studies offer unique contributions to literature by showing the potential effects of novel technologies on people's intergroup dynamics, its limitations must be acknowledged. First, although commonly used in research, self-report measures are susceptible to social desirability biases and thus might not have fully captured participants' genuine responses. For example, some participants may have felt uncomfortable reporting no intention to support women, especially during a time when femicides were being heavily discussed (Study 1). Participants may have reported stronger intentions to engage in collective action and positive attitudes than they actually intend or feel. Further research might be needed to use more implicit, as well as more behavioral measures to assess the outcome variables.

© Italian e-Learning Association

Another limitation concerns our limited understanding of why discrimination by AI versus humans resulted in different levels of threats toward women. AI is typically attributed less intention and agency than humans (e.g., Gray et al., 2007), but the present work did not directly test the mechanisms through which evaluations of human and AI perpetrators diverge. An alternative explanation, for example, is that AI discrimination was seen less consequential because it is assumed to be less prevalent than human discrimination. Further research is needed to scrutinize why AI discrimination was seen differently than human discrimination.

Finally, the political environment in Türkiye and the United States during data collection may have increased participants' baseline levels of collective action intentions. A month before the launch of Study 1, recent femicides were being addressed in Turkish media (Poyrazlar, 2024; We Will Stop Femicides Platform, 2024). Similarly, before the launch of Study 2, issues related to women's reproductive rights, including bans on women's emergency abortion access (Kruesi & Ungar, 2025; Lee, 2025), withholding payment for women's travel for abortion in the military (Kube, 2025) and defunding planned parenthood programs in multiple states (Chettiar, 2025; Ollstein, 2025) gathered attention. Some of this news, in addition to the unrest regarding the current government, increased public backlash, and led people to collective action for similar purposes (Deliso, 2025; King et al., 2025; Taylor, 2025). If participants were more prone to support collective action for gender equality at the time of the studies, the direct effects of discrimination on collective action intentions might have been less pronounced. Future research could replicate these studies in a less politically active climate or use different methods to manipulate discrimination, such as a simulated interaction between participants, an AI system, and a disadvantaged group member (i.e., target).

Overall, AI systems can reproduce biased and discriminatory outcomes that reinforce existing inequalities affecting disadvantaged groups. In this study, AI discriminators were perceived as less threatening than their human counterparts, yet neither increased collective action intentions nor improved attitudes toward the targeted group. These findings suggest that perceptions of discrimination may be obscured when harmful decisions are attributed to AI. Moreover, even when discrimination is recognized, awareness of discrimination alone is insufficient to mobilize support for social change. Further research is needed to examine how AI systems and their diverse forms of unfair treatment shape individual attitudes, behaviors, and social inequalities over the long term, and how society as a whole can promote awareness and effectively address these issues. Such work is crucial for understanding the cumulative societal consequences of discrimination in the age of AI and for informing ethical, regulatory, and policy responses.

References

Allen, R., & Masters, D. (2020). Artificial intelligence: The right to protection from discrimination caused by algorithms, machine learning and automated decision-making. *ERA Forum*, *20*(4), 585–598. https://doi.org/10.1007/s12027-019-00582-w

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Allpress, J. A., Barlow, F. K., Brown, R., & Louis, W. R. (2010). Atoning for colonial injustices: Group-based shame and guilt motivate support for reparation. *International Journal of Conflict and Violence*, *4*(1), 75–88. https://doi.org/10.4119/IJCV-2816

Allpress, J. A., Brown, R., Giner-Sorolla, R., Deonna, J. A., & Teroni, F. (2014). Two faces of group-based shame: Moral shame and image shame differentially predict positive and negative orientations to ingroup wrongdoing. *Personality and Social Psychology Bulletin*, *40*(10), 1270–1284. https://doi.org/10.1177/0146167214540724

Bacchini, E. A. M., Scheepers, D., Ellemers, N., & Naber, M. (2024). Traditionally advantaged group members' affective and physiological responses to social change. *Social Psychology*, *55*(6), 295–305. https://doi.org/10.1027/1864-9335/a000567

Bagci, S. C., Turnuklu, A., & Tercan, M. (2020). Positive intergroup contact decreases the likelihood that prejudicial attitudes become avoidant behavioral tendencies. *European Journal of Social Psychology*, *50*(3), 597–613. https://doi.org/10.1002/ejsp.2646

Bagci, S. C., Verkuyten, M., Koc, Y., Turnuklu, A., Piyale, Z. E., & Bekmezci, E. (2020). Being tolerated and being discriminated against: Links to psychological well-being through threatened social identity needs. *European Journal of Social Psychology*, *50*(7), 1463–1477. https://doi.org/10.1002/ejsp.2699

Bagci, S. C., Çelebi, E., & Karaköse, S. (2017). Discrimination towards ethnic minorities: How does it relate to majority group members' outgroup attitudes and support for multiculturalism. *Social Justice Research*, *30*(1), 1-22.

Batson, C. D., Chang, J., Orr, R., & Rowland, J. (2002). Empathy, attitudes, and Action: Can feeling for a member of a stigmatized group motivate one to help the group? *Personality and Social Psychology Bulletin*, *28*(12), 1656–1666. https://doi.org/10.1177/014616702237647

Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., Klein, T. R., & Highberger, L. (1997). Empathy and

© Italian e-Learning Association

attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology*, *72*(1), 105–118. https://doi.org/10.1037/0022-3514.72.1.105

Becker, J. C., & Barreto, M. (2014). Ways to go: Men's and women's support for aggressive and nonaggressive confrontation of sexism as a function of gender identification. *Journal of Social Issues*, *70*(4), 668–686. https://doi.org/10.1111/josi.12085

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, *23*(5), 365–368. https://doi.org/10.1016/j.tics.2019.02.008

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), 4–27. https://doi.org/10.1037/xge0001250

Borrell, L. N., Jacobs, D. R., Williams, D. R., Pletcher, M. J., Houston, T. K., & Kiefe, C. I. (2007). Self-reported racial discrimination and substance use in the coronary artery risk development in adults study. *American Journal of Epidemiology*, *166*(9), 1068–1079. https://doi.org/10.1093/aje/kwm180

Branscombe, N. R., Schmitt, M. T., & Harvey, R. D. (1999). Perceiving pervasive discrimination among African Americans: Implications for group identification and well-being. *Journal of Personality and Social Psychology*, *77*(1), 135–149. https://doi.org/10.1037/0022-3514.77.1.135

Brown, N. D., Jacoby-Senghor, D. S., & Raymundo, I. (2022). If you rise, I fall: Equality is prevented by the misperception that it harms advantaged groups. *Science Advances*, *8*(18), eabm2385. https://doi.org/10.1126/sciadv.abm2385

Brown, R., & Cehajic, S. (2008). Dealing with the past and facing the future: Mediators of the effects of collective guilt and shame in Bosnia and Herzegovina. *European Journal of Social Psychology*, *38*(4), 669–684. https://doi.org/10.1002/ejsp.466

Caricati, L., & Owuamalam, C. K. (2020). System justification among the disadvantaged: A triadic social stratification perspective. *Frontiers in Psychology*, *11*, 40. https://doi.org/10.3389/fpsyg.2020.00040

Chettiar, I. (2025, February 5). Evidence that the right's assault on women is here. *Newsweek*. https://www.newsweek.com/evidence-that-rights-assault-women-here-opinion-2026061

Crawford, K. (2016, June 25). Artificial intelligence's white guy problem. *The New York Times*. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

Darling, K. (2012). Extending legal rights to social robots. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2044797

Darling, K. (2015). "Who's Johnny?" Anthropomorphic framing in human-robot interaction, Integration, and Policy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2588669

Dastin, J. (2018, October 11). Insight – Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Deliso, M. (2025, February 6). Protesters across the US rally against Trump administration policies. *ABC News*. https://abcnews.go.com/US/anti-trump-protests-nationwide/story?id=118501194

Dion, K. L., & Earn, B. M. (1975). The phenomenology of being a target of prejudice. *Journal of Personality and Social Psychology*, *32*(5), 944–950. https://doi.org/10.1037/0022-3514.32.5.944

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *The SAGE Handbook of Prejudice, Stereotyping and Discrimination* (pp. 3–28). SAGE Publications Ltd. https://doi.org/10.4135/9781446200919.n1

Ellemers, N., & Barreto, M. (2009). Collective action in modern times: How modern expressions of prejudice prevent collective action. *Journal of Social Issues*, *65*(4), 749–768. https://doi.org/10.1111/j.1540-4560.2009.01621.x

Ellemers, N., Van Knippenberg, A., & Wilke, H. (1990). The influence of permeability of group boundaries and stability of group status on strategies of individual mobility and social change. *British Journal of Social Psychology*, *29*(3), 233–

© Italian e-Learning Association

246. https://doi.org/10.1111/j.2044-8309.1990.tb00902.x

Eurostat. (2025). *Usage of AI technologies increasing in EU enterprises*. https://ec.europa.eu/eurostat/fr/web/products-eurostat-news/-/ddn-20250123-3#:~:text=In%202024%2C%2013.5%25%20of%20enterprises,growth%20from%208.0%25%20in%202023.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Flood, M., Dragiewicz, M., & Pease, B. (2021). Resistance and backlash to gender equality. *Australian Journal of Social Issues*, *56*(3), 393–408. https://doi.org/10.1002/ajs4.137

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3848388

Geiselmann, R., Tsourgianni, A., Deroy, O., & Harris, L. T. (2023). Interacting with agents without a mind: The case for artificial agents. *Current Opinion in Behavioral Sciences*, *51*, 101282. https://doi.org/10.1016/j.cobeha.2023.101282

Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., & Tscheligi, M. (2015). Systematic analysis of video data from different human–robot interaction studies: A categorization of social signals during error situations. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00931

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Gray, K., Yam, K. C., Eng, A., Dillion, D., & Waytz, A. (2025). The psychology of robots and artificial intelligence. In D. Gilbert, S. Fiske, E. Finkel, & W. Mendes (Eds.), *Handbook of social psychology 6th Edition* (6th ed.). Situational Press. https://doi.org/10.70400/IKVM1257

Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review*, *47*(7), 1264–1279. https://doi.org/10.1108/OIR-08-2021-0452

Hayes, A. F. (2022). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (3rd ed.). The Guilford Press.

Hoskin, R. E., Thomas, E. F., & McGarty, C. (2019). Transnational contact and challenging global poverty: Intergroup contact intensifies (the right kind of) social identities to promote solidarity-based collective action for those low in social dominance. *Journal of Theoretical Social Psychology*, *3*(1), 23–34. https://doi.org/10.1002/jts5.24

Intahchomphoo, C., & Gundersen, O. E. (2020). Artificial intelligence and race: A systematic review. *Legal Information Management*, *20*(2), 74–84. https://doi.org/10.1017/S1472669620000183

Iyer, A., & Ryan, M. K. (2009). Why do men and women challenge gender discrimination in the workplace? The role of group status and in-group identification in predicting pathways to collective action. *Journal of Social Issues*, *65*(4), 791–814. https://doi.org/10.1111/j.1540-4560.2009.01625.x

Jenen, J., Winquist, J., Arkkelin, D., & Schuster, K. (2009). Implicit attitudes towards feminism. *Sex Roles*, *60*(1), 14-20. https://doi.org/10.1007/s11199-008-9514-3

Jonker, A., & Rogers, J. (2024, September 20). What is algorithmic bias? *IBM*. https://www.ibm.com/think/topics/algorithmic-bias

Kahn, P. H., & Shen, S. (2017). NOC NOC, who's there? A new ontological category (NOC) for social robots. In N. Budwig, E. Turiel, & P. D. Zelazo (Eds.), *New perspectives on human development* (1st ed., pp. 106–122). Cambridge University Press. https://doi.org/10.1017/CBO9781316282755.008

Kandiyoti, D. (2016). Locating the politics of gender: Patriarchy, neo-liberal governance and violence in Turkey. *Research and Policy on Turkey*, *1*(2), 103–118. https://doi.org/10.1080/23760818.2016.1201242

King, R. L., Moon, J., & Debushmann Jr, B. (2025, April 6). Anti-Trump protests held in cities across the US. *BBC*. https://www.bbc.com/news/articles/cz79ewg193ro

Knowles, E. D., & Lowery, B. S. (2012). Meritocracy, self-concerns, and whites' denial of racial inequity. *Self and Identity*, *11*(2), 202–222. https://doi.org/10.1080/15298868.2010.542015

Kruesi, K., & Ungar, L. (2025, March 6). What to know about the Idaho emergency abortion case dropped by the Trump administration. *PBS News*. https://www.pbs.org/newshour/health/what-to-know-about-the-idaho-emergency-abortion-case-dropped-by-the-trump-administration

© Italian e-Learning Association

Kube, C. (2025, January 31). Pentagon ends paying for travel for abortion and other reproductive care. *NBC News*. https://www.nbcnews.com/politics/national-security/pentagon-ends-policy-funding-travel-abortions-care-rcna190198

Kutlaca, M., Radke, H. R. M., & Becker, J. C. (2022). The impact of including advantaged groups in collective action against social inequality on politicized identification of observers from disadvantaged and advantaged groups. *Political Psychology*, *43*(2), 297–315. https://doi.org/10.1111/pops.12755

Landrine, H., & Klonoff, E. A. (1996). The schedule of racist events: A measure of racial discrimination and a study of its negative physical and mental health consequences. *Journal of Black Psychology*, *22*(2), 144–168. https://doi.org/10.1177/00957984960222002

Lee, C. (2025, March 7). Trump administration drops high-profile emergency abortion case, Leaving Advocates 'Devastated.' *TIME*. https://time.com/7265757/trump-lawsuit-abortion-idaho-ban-emergencies-activists-providers-respond/

Lemaster, P., Strough, J., Stoiko, R., & DiDonato, L. (2015). To have and to do: Masculine facets of gender predict men's and women's attitudes about gender equality among college students. *Psychology of Men & Masculinity*, *16*(2), 195–205. https://doi.org/10.1037/a0036429

Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. *Annual Review of Sociology*, *27*(1), 363–385. https://doi.org/10.1146/annurev.soc.27.1.363

Lowery, B. S., Knowles, E. D., & Unzueta, M. M. (2007). Framing inequity safely: Whites' motivated perceptions of racial privilege. *Personality and Social Psychology Bulletin*, *33*(9), 1237–1250. https://doi.org/10.1177/0146167207303016

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., … Oak, S. (2025). *Artificial intelligence index report 2025*. Stanford University Human Centered AI. https://hai.stanford.edu/ai-index/2025-ai-index-report

Mazzuca, S., Moscatelli, S., Menegatti, M., & Rubini, M. (2022). Men's reactions to gender inequality in the workplace: From relative deprivation on behalf of women to collective action. *Frontiers in Psychology*, *13*, 999750. https://doi.org/10.3389/fpsyg.2022.999750

McElheran, K., Li, J. F., Brynjolfsson, E., Kroff, Z., Dinlersoz, E., Foster, L., & Zolas, N. (2024). AI adoption in America: Who, what, and where. *Journal of Economics & Management Strategy*, *33*(2), 375–415. https://doi.org/10.1111/jems.12576

Miron, A. M., Warner, R. H., & Branscombe, N. R. (2011). Accounting for group differences in appraisals of social inequality: Differential injustice standards. *British Journal of Social Psychology*, *50*(2), 342–353. https://doi.org/10.1111/j.2044-8309.2010.02009.x

Moscatelli, S., Mazzuca, S., Menegatti, M., & Rubini, M. (2025). Women's participation in collective action for workplace gender equality: The role of perceived relative deprivation, resentment, and moral conviction. *Sex Roles*, *91*(4), 22. https://doi.org/10.1007/s11199-025-01573-7

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). Gender bias in AI: A review of contributing factors and mitigating strategies. *Proceedings of the 31st Australasian Conference on Information Systems (ACIS 2020)*, *27*. https://aisel.aisnet.org/acis2020/27

Ollstein, A. M. (2025, March 31). Trump admin cuts tens of millions from Planned Parenthood. *Politico*. https://www.politico.com/news/2025/03/31/trump-admin-cuts-tens-of-millions-from-planned-parenthood-00261763

Osborne, D., & Sibley, C. G. (2015). Opposing paths to ideology: Group-based relative deprivation predicts conservatism through warmth toward ingroup and outgroup members. *Social Justice Research*, *28*(1), 27–51. https://doi.org/10.1007/s11211-014-0227-1

Osborne, D., Jost, J. T., Becker, J. C., Badaan, V., & Sibley, C. G. (2019). Protesting to challenge or defend the system? A system justification perspective on collective action. European Journal of Social Psychology, 49(2), 244-269.

Özdemir Sarıgil, B. (2022). Combating gender-based discrimination against women in Turkey: Achievements and the challenges ahead. In G. Yılmaz (Ed.), *Non-discrimination in Turkey* (pp. 59–77). Springer International Publishing. https://doi.org/10.1007/978-3-031-08399-0_4

Pascoe, E. A., & Smart Richman, L. (2009). Perceived discrimination and health: A meta-analytic review. *Psychological Bulletin*, *135*(4), 531–554. https://doi.org/10.1037/a0016059

Phillips, L. T., & Lowery, B. S. (2015). The hard-knock life? Whites claim hardships in response to racial inequity. *Journal of Experimental Social Psychology*, *61*, 12–18. https://doi.org/10.1016/j.jesp.2015.06.008

© Italian e-Learning Association

Poyrazlar, E. (2024, October 10). Outrage over murders of women forces Turkey's Erdoğan to harden laws. *Politico*. https://www.politico.eu/article/outrage-murder-women-turkey-erdogan-feminicide-laws-ak-party/

Reimer, N. K., Becker, J. C., Benz, A., Christ, O., Dhont, K., Klocke, U., Neji, S., Rychlowska, M., Schmid, K., & Hewstone, M. (2017). Intergroup contact and social change: Implications of negative and positive contact for collective action in advantaged and disadvantaged groups. *Personality and Social Psychology Bulletin*, *43*(1), 121–136. https://doi.org/10.1177/0146167216676478

Schmitt, M., & Maes, J. (2002). Stereotypic ingroup bias as self-defense against relative deprivation: Evidence from a longitudinal study of the German unification process. *European Journal of Social Psychology*, *32*(3), 309–326. https://doi.org/10.1002/ejsp.91

Schmitt, M. T., Branscombe, N. R., Postmes, T., & Garcia, A. (2014). The consequences of perceived discrimination for psychological well-being: A meta-analytic review. *Psychological Bulletin*, *140*(4), 921–948. https://doi.org/10.1037/a0035754

Schmitt, M. T., Spears, R., & Branscombe, N. R. (2003). Constructing a minority group identity out of shared rejection: The case of international students. *European Journal of Social Psychology*, *33*(1), 1–12. https://doi.org/10.1002/ejsp.131

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Shank, D. B., North, M., Arnold, C., & Gamez, P. (2021). Can mind perception explain virtuous character judgments of artificial intelligence? *Technology, Mind, and Behavior*, *2*(3). https://doi.org/10.1037/tmb0000047

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and Oppression* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139175043

Swim, J. K., & Miller, D. L. (1999). White guilt: Its antecedents and consequences for attitudes toward affirmative action. *Personality and Social Psychology Bulletin*, *25*(4), 500–514. https://doi.org/10.1177/0146167299025004008

Taylor, A. (2025, February 18). 'Not My President's Day': Thousands gather in protest against Trump. *The Atlantic*. https://www.theatlantic.com/photo/2025/02/thousands-gather-protest-against-trump-musk/681721/

Tian, L., & Oviatt, S. (2021). A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction*, *10*(2), 1–32. https://doi.org/10.1145/3439720

Tougas, F., & Veilleux, F. (1990). The response of men to affirmative action strategies for women: The study of a predictive model. *Canadian Journal of Behavioural Science / Revue Canadienne*

Uluğ, Ö. M., Chayinska, M., & Tropp, L. R. (2023). Does witnessing gender discrimination predict women's collective action intentions for gender justice? Examining the moderating role of perceived female support. *Journal of Community & Applied Social Psychology*, *33*(2), 501–518. https://doi.org/10.1002/casp.2642

Uluğ, Ö. M., & Tropp, L. R. (2021). Witnessing racial discrimination shapes collective action for racial justice: Enhancing awareness of privilege among advantaged groups. *Journal of Applied Social Psychology*, *51*(3), 248–261. https://doi.org/10.1111/jasp.12731

United Nations Development Programme. (2023). *2023 Gender social norms index (GSNI): Breaking down gender biases: Shifting social norms towards gender equality*. United Nations Development Programme. https://hdr.undp.org/content/2023-gender-social-norms-index-gsni#/indicies/GSNI

U.S. Equal Employment Opportunity Commission. (2022). *EEOC sues iTutorGroup for age discrimination*. U.S. Equal Employment Opportunity Commission. https://www.eeoc.gov/newsroom/eeoc-sues-itutorgroup-age-discrimination

Uysal, M. S., Uluğ, Ö. M., Kanık, B., & Aydemir, A. (2022). "The liberation of LGBTQ+ will also liberate heterosexuals": Heterosexual feminist women's participation in solidarity-based collective action for LGBTQ+ rights. *European Journal of Social Psychology*, *52*(2), 377–390. https://doi.org/10.1002/ejsp.2799

Van Laar, C., Van Rossum, A., Kosakowska-Berezecka, N., Bongiorno, R., & Block, K. (2024). MANdatory – Why men need (and are needed for) gender equality progress. *Frontiers in Psychology*, *15*, 1263313. https://doi.org/10.3389/fpsyg.2024.1263313

Van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin*, *134*(4), 504–535. https://doi.org/10.1037/0033-2909.134.4.504

Van Zomeren, M., Postmes, T., & Spears, R. (2012). On conviction's collective consequences:

© Italian e-Learning Association

Integrating moral conviction with the social identity model of collective action. *British Journal of Social Psychology*, *51*(1), 52–71. https://doi.org/10.1111/j.2044-8309.2010.02000.x

Van Zomeren, M., Postmes, T., Spears, R., & Bettache, K. (2011). Can moral convictions motivate the advantaged to challenge social inequality?: Extending the social identity model of collective action. *Group Processes & Intergroup Relations*, *14*(5), 735–753. https://doi.org/10.1177/1368430210395637

Vargas-Solar, G. (2022). Intersectional study of the gender gap in STEM through the identification of missing datasets about women: A multisided problem. *Applied Sciences*, *12*(12), 5813. https://doi.org/10.3390/app12125813

Veilleux, F., & Tougas, F. (1989). Male acceptance of affirmative action for women: The results of alturistic or egoistical motives? *International Journal of Psychology*, *24*(1–5), 485–496. https://doi.org/10.1080/00207594.1989.10600060

Vescio, T. K., Sechrist, G. B., & Paolucci, M. P. (2003). Perspective taking and prejudice reduction: The mediational role of empathy arousal and situational attributions. *European Journal of Social Psychology*, *33*(4), 455–472. https://doi.org/10.1002/ejsp.163

Vlasceanu, M., & Amodio, D. M. (2022). Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences*, *119*(29), e2204529119. https://doi.org/10.1073/pnas.2204529119

Wajcman, J. (2010). Feminist theories of technology. *Cambridge Journal of Economics*, *34*(1), 143–152. https://doi.org/10.1093/cje/ben057

We Will Stop Femicides Platform. (2024). *We Will Stop Femicides Platform October 2024 report*. https://kadincinayetlerinidurduracagiz.net/veriler/3126/we-will-stop-femicides-platform-october-2024-report

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. *AI Now Institute*. https://ainowinstitute.org/publications/discriminating-systems-gender-race-and-power-in-ai-2

Williams, D. R., Neighbors, H. W., & Jackson, J. S. (2003). Racial/Ethnic discrimination and health: Findings from community studies. *American Journal of Public Health*, *93*(2), 200–208. https://doi.org/10.2105/AJPH.93.2.200

World Economic Forum. (2024). *Global gender gap report 2024*. World Economic Forum. https://www.weforum.org/publications/global-gender-gap-report-2024/

World Economic Forum. (2025). *Global gender gap report 2025*. World Economic Forum. https://www.weforum.org/publications/global-gender-gap-report-2025/

Wright, S. C., Aron, A., McLaughlin-Volpe, T., & Ropp, S. A. (1997). The extended contact effect: Knowledge of cross-group friendships and prejudice. *Journal of Personality and Social Psychology*, *73*(1), 73–90. https://doi.org/10.1037/0022-3514.73.1.73

Wright, S. C., Taylor, D. M., & Moghaddam, F. M. (1990). Responding to membership in a disadvantaged group: From acceptance to collective protest. *Journal of Personality and Social Psychology*, *58*(6), 994–1003. https://doi.org/10.1037/0022-3514.58.6.994

Wright, S. C., & Tropp, L. R. (2002). Collective action in response to disadvantage: Intergroup perceptions, social identification, and social change. In I. Walker & H. J. Smith (Eds.), *Relative deprivation: Specification, development, and integration* (pp. 200–236). Cambridge University Press.

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 21, No. 1 (2026), pp. 75-90

Beyond automation: the collaborative role of AI in countering Hate Speech. A scoping review

Antonella Cosola[a,1], Marinella Paciello[b], Alessandro Pollini[a], Francesca D'Errico[c]

[a]*Uninettuno Telematic International University – Rome (Italy)*

[b]*Sapienza, University of Rome – Rome (Italy)*

[c]*University of Bari "Aldo Moro"– Bari (Italy)*

(submitted: 31/3/2026; accepted: 23/7/2026; published: 17/8/2026)

Abstract

The unregulated proliferation of hate speech, defined as any expression of beliefs that incite hatred, discrimination, or violence against individuals or social groups (Fortuna & Nunes, 2018), in social media has called into question the efficacy of removal-based content moderation and has driven the need to explore alternative approaches. AI-generated counterspeech could be a scalable and cost-effective intervention (Ashida & Komachi, 2022). Counterspeech is defined as a direct response to hate speech aimed at refuting or undermining it (Wachs et al., 2023). Research in this growing field is fragmented across computational linguistics and social sciences, lacking systematic methods to define and evaluate effective counterspeech, revealing a gap between AI's technical capabilities and evidenced social impact.

This review aims to examine evidence from these two fields to identify the differing conceptualizations of counterspeech, map the collaborative approaches used for counterspeech generation and evaluation, and examine the alignment between stated research goals (e.g., reducing hate speech in digital environments) and the operational measures employed to identify them.

Only a few studies evaluate counterspeech in real-world interactive settings or measure long-term impact. We identify some key misalignments, such as the reliance on text metrics and human or artificial annotators as proxies for complex social interventions, instead of integrating psychological and behavioral measures. We propose a linked approach to counterspeech that links theoretical frameworks, procedures, measurements and research goals from the perspective of both psychology and computational linguistics to identify promising directions for future interdisciplinary collaboration and more socially meaningful AI-assisted interventions.

KEYWORDS: Counter-speech, Artificial Intelligence, Large Language Models, Human Augmentation.

DOI
https://doi.org/10.20368/1971-8829/1136378

CITE AS
Cosola, A., Paciello, M., Pollini, A., & D'Errico, F. (2026). Beyond automation: the collaborative role of AI in countering Hate Speech. A scoping review. *Journal of e-Learning and Knowledge Society*, *21*(1), 75-90.
https://doi.org/10.20368/1971-8829/1136378

1. Introduction

The choice to implement artificial intelligence (AI) in content moderation on social media is likely driven by the surge of negative polarizing content. Hate speech is defined as

> *"language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and can manifest itself with different linguistic styles, even in subtle forms or when humor is used" (Fortuna & Nunes, 2018, p.5).*

Deleting content and banning users tends to disperse hatred across less regulated platforms rather than resolving it (Horta Ribeiro et al., 2023), resulting in being now outdated in the current landscape. In recent

1 corresponding author - email: a.cosola@students.uninettunouniversity.net

© Italian e-Learning Association

years, therefore, researchers have begun to explore the potential of online counterspeech, defined as a direct response to hate speech aimed at refuting or undermining hate speech (Wachs et al. 2023). Counterspeech may be a more effective response to hateful content as it is not constrained by the single platform rules and doesn't undermine the principles of free expression (Zhu & Bhat, 2021).

Writing effective counterspeech is time-consuming and requires expertise, making human-only moderation costly (Chung et al., 2024; Sportelli et al., 2025). Large Language Models (LLMs) can be a solution, being easily applicable, less costly at scale (Ashida & Komachi, 2022), and preventing human strain from prolonged exposure to harmful content (Chung et al., 2024). Although AI-generated counterspeech can be a useful content moderation tool, it also presents some weaknesses, such as a lack of argumentative richness (Bonaldi et al., 2024), struggle with deeper reasoning (Mun et al., 2023), perceived inauthenticity by the users (Bar et al., 2024), and risk of generating hallucinated or non-factual information (Wang et al., 2024).

Critically, counterspeech evaluation still relies on traditional natural language processing (NLP) metrics. Also, LLM evaluators may favor AI-generated counterspeech (Zubiaga et al., 2024). Another critical issue is the absence of a metric that captures counterspeech quality or effectiveness, creating a misalignment with users' behavioral change. Computational research on AI-generated counterspeech is advancing rapidly, while psychosocial research lags behind. This misalignment limits insight into psychosocial outcomes, particularly as tool development lacks robust theoretical grounding (Chung et al., 2024). Consequently, unaddressed risks remain: fully automated approaches may promote diffusion of responsibility (Darley & Latané, 1968) and moral disengagement (Bandura, 1996), harming the social norms that bystander intervention relies on (Mathew et al., 2019).

Adopting a psychosocial lens, this review assesses how psychological and social dimensions are integrated into AI-generated counterspeech conceptualization, creation, and evaluation, ensuring a common object of study across terminological differences.

2. Review Methods

This review was conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines (Tricco et al., 2018) to ensure transparency and reproducibility. The protocol was not registered.

2.1 Search Strategy

The search was conducted in January 2026 on Scopus, Web of Science, and arXiv. Given the rapidly evolving nature of AI research, we included relevant pre-prints from the arXiv repository, even beyond the search date.

The search string, illustrated in Table 1, was designed to combine key concepts related to counterspeech (e.g., counterarguments, counternarratives) keywords meant to search the tools (e.g., AI, large language models) and terms related to human-AI collaboration (e.g., human-AI collaboration, human augmentation).

2.2 Selection Process

The selection process followed the PRISMA-Scr flowchart phases, as illustrated in Figure 1. References were initially imported into Excel for duplicate removal and the following screening process.

The initial search yielded 1184 records. After duplicate removal, 1100 remained. We excluded 730 studies published before January 1, 2017, as this marks the introduction of transformer models, which now underpin modern AI and Natural Language Processing.

Of the remaining, 608 were excluded during title and abstract screening as irrelevant, leaving 121 articles for full-text review. Two independent reviewers achieved moderate initial agreement (κ=0.608). After discussion with a third reviewer, we proceeded with the full-text assessment.

In the full-text assessment, an additional 60 studies were excluded due to a theoretical approach that was not equivalent to counterspeech because they presented a theoretical concept of counterspeech that did not align with the operational definition employed in this review, not generating counterspeech, being inadequate (book chapters or shared tasks overviews), or not being in the English language. The remaining 61 were further assessed: we excluded studies lacking human involvement in generation or counterspeech evaluation, those where researchers self-evaluated counterspeech, and kept studies with external participants as reviewers, such as crowdworkers, recruited evaluators, and real social media users (n=18). We also excluded one dataset evaluation paper, one due to a different theoretical conceptualization, and one non-English paper. Including 5 additional studies not returned by the keyword search, the final total is 45 papers. The full reference list is available in Appendix A.

For each included study, the following data were extracted:

a) bibliographic details and disciplinary affiliation;

b) operational definition of counterspeech used;

c) collaborative human-AI generation method;

d) evaluation methods.

© Italian e-Learning Association

Table 1 - Databases and citation indexes searched.

Database	Search query	Filters	Results
Scopus	TITLE-ABS-KEY ((“counterspeech” OR “counter speech” OR “counter narrative” OR “counter-narrative” OR “counter argument” OR “counterargument” OR “counter-argument” OR “counter reply” OR “counter-stereotype” OR “counter-aggression” OR “counter hate speech” OR “countering hate speech” OR “counter-terrorism” OR “counter-hate speech”) AND (“AI” OR “generative AI” OR “large language model” OR “LLM” OR “GPT” OR “natural language generation” OR “NLP” OR “language model” OR “artificial intelligence” OR “generation” OR “automation” OR “human-AI collaboration” OR “human augmentation”))	2000-2026	685
arXiv	“counterspeech” OR “counter speech” OR “counter narrative” OR “counter-narrative” OR “counter argument” OR “counterargument” OR “counter-argument” OR “counter reply” OR “counter-stereotype” OR “counter-aggression" OR “counter hate speech” OR “countering hate speech" OR “counter-terrorism” OR “counter-hate speech” AND “AI” OR “generative AI” OR “large language model” OR “LLM” OR “GPT” OR “natural language generation” OR “NLP" OR “language model” OR “artificial intelligence” OR “generation” OR “automation” OR “human-AI collaboration” OR “human augmentation”	Computer science (cs); 2000–2026	104
Web of Science	TS = ((“counterspeech" OR “counter speech” OR “counter narrative” OR “counter-narrative” OR “counter argument” OR “counterargument” OR “counter-argument” OR “counter reply” OR “counter-stereotype” OR “counter-aggression” OR “counter hate speech” OR “countering hate speech” OR “counter-terrorism” OR “counter-hate speech”) AND (“AI” OR “generative AI” OR “large language model” OR “LLM” OR “GPT” OR “natural language generation” OR “NLP” OR “language model” OR “artificial intelligence” OR “generation” OR “automation” OR “human-AI collaboration” OR “human augmentation”))	2000-2026	401

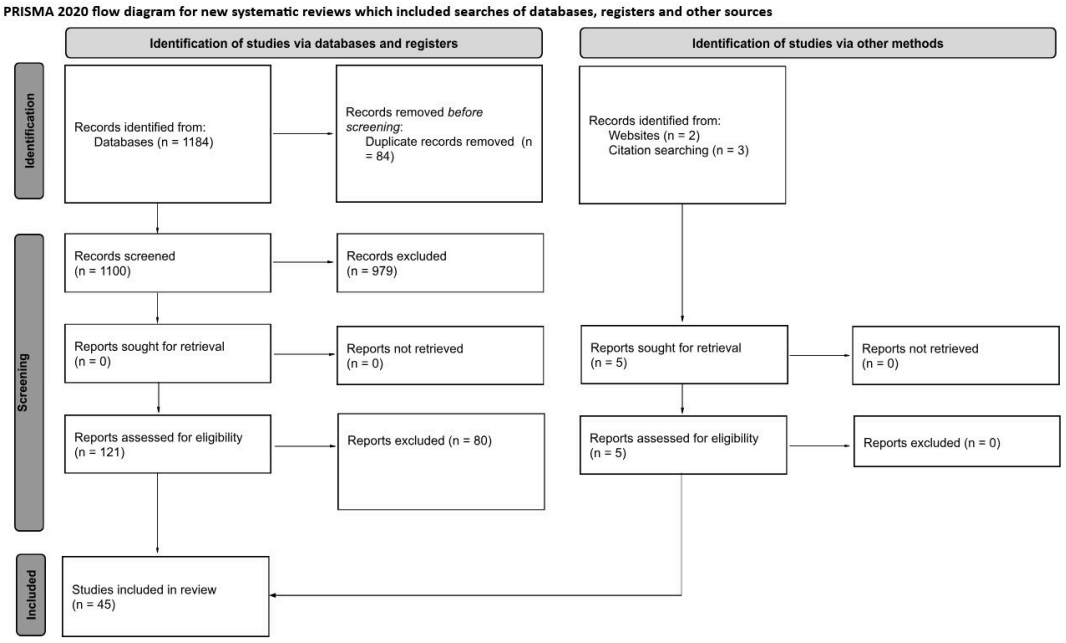


Figure 1 - PRISMA flow diagram showing the identification, screening, eligibility, and inclusion phases of the studies selected for this review.

© Italian e-Learning Association

We summarized and grouped the types of counterspeech conceptualizations, human-AI generation methods, evaluation methods, and broad findings.

2.3 Eligibility Criteria

Studies were included if they:

a) were published from January 1, 2017 onwards;
b) addressed counterspeech (or synonym terms, e.g., counternarratives and counter-arguments operationalized as counterspeech) as a direct response to counter or mitigate online hate speech;
c) employed computational or human-AI assisted methods for counterspeech generation and/or analysis, experimental/computational methods for evaluation; AI-generated counterspeech with human collaborators involved in at least one phase (development, testing, assessment, or writing);
d) included a systematic evaluation through computational metrics, human judgment, or behavioral/contextual measures;
e) were original research articles or conference proceedings.

Studies were excluded if they:

a) did not generate counterspeech, as the AI system was used only for hate speech/ counterspeech detection or classification;
b) employed AI only to generate counterspeech without human intervention;
c) conceptualized counterspeech as unrelated to online hate speech;
d) described shared task systems without humans employed in any phase of the counterspeech generation process;
e) were not empirical (theoretical papers, opinion pieces, editorials, book chapters, framework proposals without empirical validation, literature reviews, benchmark papers, and shared task proceedings);
f) were duplicate publications, only the most comprehensive or recent version was retained;
g) were not available as full-text;
h) were not written in the English language.

3. Results

This section provides a descriptive synthesis of the 45 included studies. The findings are organized according to the three objectives that guided this review: conceptualizations of counterspeech, collaborative generation methods, and evaluation approaches.

3.1 Counterspeech conceptualizations

As noted by Chung et al. (2024), computer science research often employs "counterspeech", "counternarratives", and "counterarguments" interchangeably, despite their distinct nuanced meanings in the social sciences. Benesch (2016) defines counternarratives as efforts to shift collective beliefs through institutional action, while counterspeech refers to spontaneous, citizen-based responses to hateful content (Garland et al., 2022). Counterarguments, in turn, challenge the factual or logical validity of specific claims and can function as a counterspeech strategy. Despite these distinctions, the three concepts share a common theoretical core as communicative acts responding to harmful discourse.

Many of the following taxonomies take as a base Benesch et al. (2016) original taxonomy, which takes into consideration eight non-exclusive counterspeech strategies. In Table 2, we list the main taxonomies found in the reviewed studies.

The absence of a shared taxonomy hinders cross-study comparison and prevents isolating which dimensions drive effectiveness.

3.2 Counterspeech generation methods: humans and AI collaboration

In this section, we focus on the collaborative generation strategies adopted, referred to as Human in the Loop (HITL), that reject the notion that counterspeech can or should be fully delegated to AI.

By mapping this technical landscape, we aim to identify dominant practices and emerging innovations.

In the studies reviewed, HITL manifests in different configurations. The majority of them take humans into account, reviewing AI-generated responses before publication. A smaller number of studies involve humans as trainers (Bengoetxea et al., 2024; Bonaldi et al., 2022; Fanton et al., 2021; Mun et al., 2023), providing feedback that shapes model behavior over time.

In the Author-Reviewer pipeline, an LLM generates counterspeech candidates while human experts filter and refine them (Ashida & Komachi, 2022; Fanton et al., 2021; Wang et al., 2024). This collaborative approach has been shown to drastically reduce the time experts need to create training datasets, from approximately 480 seconds per pair to 49-76 seconds (Tekiroğlu et al., 2020).

Two studies adopted HITL approaches to develop counterspeech datasets in non-English languages.

© Italian e-Learning Association

Table 2 - Counterspeech taxonomies. The classification aligns with the multidimensional framework proposed by Wang et al. (2026), which differentiates between sentence type, tone, persuasion strategies (factual information and logical reasoning) and affective strategies (influencing emotions, social bonds, peripheral cues).

Strategy	Definition	Sub-strategies	Representative studies
Cognitive	Aim to influence beliefs through factual information and logical reasoning	Presentation of facts; correction; counter-examples; pointing out contradictions; explanations; targeting stereotypes (alternate groups, alternate qualities, counterexamples, external factors, broadening, general denouncing)	Benesch et al. (2016); Fraser et al. (2023); Hassan and Alikhani (2023); Mun et al. (2023)
Affective	Aim to influence emotions, empathy, and social identification	Empathy; identification with the target; emphasizing positive qualities; emotional support	Benesch et al. (2016); Bilewicz et al. (2021); Fraser et al. (2023)
Normative	Appeal to social norms, shared values, and potential consequences	Denouncing speech; invoking norms; warning of consequences; descriptive and prescriptive norms	Benesch et al. (2016); Bilewicz et al. (2021); Fraser et al. (2023); Gupta et al. (2024)
Dialogic	Engage the speaker through interaction and reflective exchange	Critical questions; probing; clarification; dialogic contrast	Hassan and Alikhani (2023); Fraser et al. (2023)
Rhetorical	Rely on communicative style and expressive choices rather than content	Humor; tone; use of visual media; rhetorical framing; rhetorical appeal (logos, ethos, pathos)	Alyahya and Aldayel (2025); Benesch et al. (2016); Fraser et al. (2023); Mathew et al. (2019)
Transformative	Reformulate or replace harmful content rather than directly opposing it	Alternative speech; microinterventions (making harm visible, disarming, educating, seeking support)	Ashida and Komachi (2022); Lee et al. (2024)

Prasannan et al. (2025) used commercial models to generate Malayalam hate-counterspeech pairs, with two rounds of validation by native speakers: first for naturalness and coherence, then for contextual appropriateness and ethical sensitivity. Sahoo et al. (2024) developed a Hindi, Indian, and English counternarrative dataset where annotators first wrote and validated counternarratives, then an mGPT model generated additional instances, which were validated and modified before inclusion.

Chung and Bright (2024) developed an iterative adversarial collection procedure to improve counterspeech generation models. The authors conducted four rounds of data collection. In each round, a model (DynaCounter) was used to generate counterspeech, and its outputs were examined to identify failure cases. To generate adversarial examples (challenging cases where the model performed poorly), the researchers collaborated with ten trained members of a civil society organization specializing in online abuse countering.

Recent studies have explored collaborative interventions. Wu et al. (2025) first prompted the model to describe the hate speech content it was responding to, ensuring accurate understanding. A human reviewer then evaluated this description and corrected it if not accurate. In the second phase, the generated counterspeech was reviewed and manually revised by human collaborators, which could modify expressions, adjust wording, or rephrase entire sentences to better align with communicative goals. A co-writing approach was designed by Ding et al. (2025). They introduce CounterQuill, a model that educates users through reflection and structured support. This system is organized in three sessions: learning about hate and counter speech; brainstorming, identifying harmful patterns and generating initial ideas; co-writing, where users refine their counterspeech, with the system offering suggestions and revisions that users can accept, modify, or reject.

There is also a shared consensus that humans should also be employed as safeguards (Bonaldi et al., 2024; Hassan & Alikhani, 2023). In these studies, humans are usually employed to monitor experiments in real-time, ensuring that interventions stay within ethical guidelines and do not escalate hostility (Bär et al., 2024; Podolak et al., 2024).

© Italian e-Learning Association

3.3 Counterspeech evaluation methods

The generation strategies reviewed above reveal increasingly sophisticated technologies, yet sophistication does not guarantee effectiveness. In the following section, we map the evaluation landscape, distinguishing between automatic metrics, human judgments, LLM evaluations, and psychosocial measures.

3.3.1 Linguistic Metrics

A significant portion of the studies we reviewed relies on automatic linguistic metrics to evaluate AI-generated counterspeech. These metrics allow researchers to assess large volumes of generated text without the time and cost of human evaluation. In Table 3, we outline relevant metrics used to assess the quality of the generated counterspeech. Automatic measures can address structural properties of text such as length, grammaticality, vocabulary size, and expansion. However, these measures are not specific to counterspeech evaluation and were not listed.

Automatic metrics enable large-scale evaluation, but face significant limitations. BLEU and ROUGE correlate poorly with human judgments in creative tasks (Chung et al., 2024) and assume a "gold standard" reference (problematic for counterspeech, where multiple valid responses can exist). Diversity metrics capture lexical variation but not its communicative function. Toxicity detectors may miss subtle harm or over-flag legitimate content. Most critically, no automatic metric addresses whether counterspeech achieves its intended social effects.

3.3.2 Human Evaluation Metrics

Beyond automatic metrics, a subset of studies incorporates human evaluation, where trained annotators or crowdworkers rate counterspeech on critical dimensions, acting as proxies for real social media users. The assessed measures are presented in Table 4.

Similar to automated metrics, human evaluation metrics that focused on general text quality were excluded from this review, as they do not directly measure counterspeech properties.

This heterogeneity in human evaluation practices mirrors the methodological fragmentation already documented, reinforcing the need for shared standards that enable systematic comparison across studies.

Table 3 - Automatic linguistic metrics for counterspeech.

Dimension	Description	Representative Metrics	Relevance
Semantic metrics	Measure semantic overlap between generated counterspeech and human-written reference texts	BLEU, ROUGE, METEOR, GLEU; BERTScore (Zhang et al., 2019; Hengle et al., 2024); BARTScore (Hengle et al., 2025); MoverScore; Semantic Similarity (Gupta et al., 2023)	Indicates how closely AI output approximates natural human language
Linguistic quality and readability	Assess grammatical correctness, coherence, and accessibility of the generated text	GRUEN (Zhu & Bhat, 2021); Flesch Reading Ease (Cima et al., 2025)	Influences perceived credibility, clarity, and the likelihood of message comprehension by target audiences
Diversity and novelty	Evaluate lexical and structural variety to avoid repetitive or generic outputs	Self-BLEU; Entropy; Novelty (Wang & Wan, 2018; Tekiroglu et al., 2022; Fanton et al., 2022; Gupta et al., 2023); Repetition Rate (Zubiaga et al., 2024)	Repetitive language may reduce perceived authenticity and persuasiveness in interpersonal or public health interventions
Toxicity and safety	Detect harmful, aggressive, or toxic content in generated counterspeech	Toxicity scores (Hengle et al., 2024; Wang et al., 2024b; Chung and Bright, 2024); Safety Guardrails (Bonaldi et al., 2024)	Essential for ensuring that counterspeech does not unintentionally perpetuate harm or escalate conflict
Effectiveness	Quantify the extent to which counterspeech counters hateful content through refutation, factual support, or reduction in hate intensity	Success Rate of Countering (Jiang et al., 2025); Pro/Con scores, Argument Quality (Hengle et al., 2024); PD-Score (Hengle et al., 2025); Hate Mitigation Value (Wang et al., 2024b)	Aim to determine whether AI-generated messages can meaningfully reduce hate or change attitudes

© Italian e-Learning Association

Table 4 - Human evaluation metrics of AI-generated counterspeech.

Dimension	Definition	Representative Ratings	Relevance
Offensiveness	Whether the generated response contains content that may cause harm or be perceived as offensive	Offensiveness (Ashida & Komachi, 2022)	Most generated text rated as safe, though variability in human perception exists; some outputs deemed offensive
Aggressiveness	Degree of confrontational, inflammatory, or hostile content, including personal attacks	Aggressiveness (Hengle et al., 2025); emotional tone assessment (Prasannan et al., 2025)	Low aggressiveness is prioritized to avoid escalating interactions; cultural appropriateness matters (e.g., Malayalam community context)
Politeness / Suitability	Adherence to social conventions of respect and courtesy; whether the response follows civil rules and avoids attacking the hater personally	Politeness (Song et al., 2024); Suitability (Bonaldi et al., 2024); Suitableness (Chung et al., 2020, 2021)	High politeness serves as a proxy for safe, socially acceptable responses and may reduce defensive reactions from recipients
Relevance	Whether the response addresses the correct topic and target group identified in the hate speech	Relevance (Bonaldi et al., 2024); contextual relevance (Hengle et al., 2024; Kumar et al., 2025)	High relevance increases persuasiveness; attention regularization techniques have shown improvements
Specificity	How well the response engages with the specific content of the hate speech, including logical strength, precision, and depth	Cogency (Bonaldi et al., 2024); correspondence (Lee et al., 2024); relatedness, specificity, richness, coherence (Baez Santamaria et al., 2024)	Specific, detailed responses are more likely to be perceived as thoughtful and effective than generic ones
Stance	Degree to which the response explicitly or implicitly opposes the hateful content	Stance: agreeing, neutral, disagreeing (Ashida & Komachi, 2022)	Most AI models demonstrate a clear stance against hate, which is essential for counterspeech interventions
Informativeness	Extent to which the response provides facts, data, or contextual information that challenges hate speech claims	Informativeness (Ashida & Komachi, 2022; Bonaldi et al., 2024; Chung et al., 2020, 2021); phrase guidance (Lee et al., 2024)	High informativeness does not guarantee factual accuracy; hallucinations (factual errors) remain a concern
Factuality	The accuracy and reliability of real-world data, statistics, and verifiable evidence incorporated into a counterspeech response	Truthfulness (Wang et al., 2024); Factuality (Wilk et al., 2025)	Factual inconsistencies or errors can compromise the authority of the source and lower the perceived persuasiveness of counterspeech
Effectiveness	The actual or perceived social, psychological, or persuasive impact of the counterspeech intervention	Effectiveness (Baez Santamaria et al., 2024; Zheng et al., 2023); persuasiveness (Wilk et al., 2025); improvement in speaker's hateful perception (Lee et al., 2024); argumentative effectiveness (Hengle et al., 2024; Kumar et al., 2025)	Effectiveness is conceptualized in multiple ways (emotional, persuasive, argumentative), reflecting its complexity as a primary outcome of interest for psychological interventions

3.3.3 LLM-as-a-judge

The limitations of conventional evaluation methods have prompted researchers to explore an alternative: using LLMs as evaluators (referred to as LLM-as-a-judge), which employs models to assess generated text directly. It achieves stronger alignment with human assessments than traditional metrics (Hengle et al., 2025), while offering the scalability of automated approaches. Commercial models can be used for this purpose (Zubiaga et al., 2024), but new research adopts specialized evaluation models, including PandaLM (Bennie et al., 2024) and JudgeLM (Zhu et al., 2023).

Chung et al. (2024) have criticized the validity and reliability of using LLMs as NLG evaluators. Specifically, there is an issue with how automated metrics are evaluated. In response, Hengle et al. (2025) developed CSeval, a dataset of human judgments on counterspeech from five AI models across four dimensions: contextual relevance, aggressiveness, argument coherence, and suitableness. Using these ratings, they built Auto-CSEval, an LLM-based evaluator employing Chain of Thought (CoT) prompting. This method consists of asking the model to reason step-by-step about each quality dimension

© Italian e-Learning Association

before producing a score. Auto-CSEval outperformed both traditional metrics and other LLM evaluators in correlation with human judgments.

While LLM judges show improved human alignment, research identifies several critical biases: other works (Bennie et al., 2025) underline how LLM judges like JudgeLM (Zhu et al., 2023) frequently assign higher scores to AI-generated responses than to human-preferred or human-edited ones. They suggest that JudgeLM may favor AI-generated text due to stylistic and syntactic features, rather than substantive quality such as logical or factual accuracy.

3.3.4 Psychosocial Metrics

Unlike evaluation, psychosocial metrics investigate the outcomes of generated counterspeech. The following paragraph presents the examined measures.

- User engagement. Studies agree that counterspeech presence increases discussion engagement. Podolak et al. (2024) tested the effectiveness of LLM-generated (GPT-4) counterspeech on Twitter. The counterspeech was relevant and accurate in responses, since the model could retrieve information from 23 verified articles. Interventions with LLM-generated responses significantly reduced user engagement with the hate speech tweets. For original tweets that had received at least ten views, engagement decreased by over 20. The study of Wu et al. (2025) found, across three studies with 809 participants, that people were more willing to engage when counterspeech was present, regardless of the source. Empathy-based counterspeech significantly increased user engagement intention compared to fact-based counterspeech. Moreover, participants were more willing to engage with AI-generated counterspeech when the AI identity was not disclosed. When AI identity was revealed, trust in both human and AI counterspeakers decreased.
- Post deletion and toxic language. Bär et al. (2024) measured behavioural changes of 2,664 users after being randomly assigned to contextualized and non-contextualized counterspeech on Twitter/X. Post deletion rate tracked whether the author of the original hate speech removed their post; future toxicity levels measured whether the perpetrator adopts less toxic language in subsequent posts, offering a longitudinal indicator of possible educational or deterrent effects. Non-contextualized counterspeech employing a warning of consequences strategy significantly reduced online hate speech. Contextualized counterspeech was proven ineffective and even backfired. The authors attribute the backfire to perceived deception from the undisclosed AI identity, not the counterspeech strategy itself.
- Reduction of verbal aggression. Bilewicz et al. (2021) estimated the impact of counterspeech interventions on users' engagement in personal attacks or acts of verbal violence within a Reddit community. The study found a significant reduction in verbal aggression following the intervention: the proportion of verbally aggressive comments decreased from pre to post-intervention. Norm-inducing, disapproval, and empathizing counterspeech strategies were all effective in reducing verbal aggression compared to the control group. However, the norm-inducing intervention did not differ significantly in effectiveness from either the disapproval or the empathizing interventions.
- Counterspeech participation. Wang et al. (2026) exposed 52 bystanders to their chatbot Civilbot, designed according to their three-dimensional framework (sentence type, tone, strategic intent). It was perceived as credible and normatively appropriate, but its persuasiveness was limited by superficial reasoning that was too AI-like. Cognitive strategies outperformed affective ones on perceived quality and acceptance, with positive tone amplifying this effect, while negative tone risked normalizing aggression and raising ethical concerns (Cecconi et al., 2020). Questions could stimulate reflection when genuine, but became hostile if rhetorical. Participants distinguished between their expectations of Civilbot (objective, normative) and their own human behavior (using a negative tone for emotional venting). Mismatches between context and strategy weakened the impact.
- Persuasiveness. Alyahya & Aldayel (2025) found that counterspeech employing reason as a persuasion mode is associated with obtaining more supportive replies than those using emotion or credibility. Interestingly, in their study, humans tend to lean towards reason, while machine-generated counterspeech tends to exhibit an emotional persuasion mode. Mun et al. (2023) found that strategy effectiveness varies: providing alternative, higher-quality information was chosen as convincing twice as frequently as appealing to a different group's perspective, even though both require similar levels of reasoning. Broadening was the preferred strategy, while generic denouncement was the least preferred. Cima et al. (2025) demonstrated that contextualized counterspeech significantly outperformed generic counterspeech in terms of persuasiveness.

Notably, Hong et al. (2024) developed conversation outcome classifiers that predict what happens after counterspeech is deployed. Conversation incivility captures the overall tone of the discussion, based on the number of civil and uncivil comments (Poggi et al.,

© Italian e-Learning Association

2013) and the unique authors contributing each. Hater reentry tracks whether the original hate speech author returns to the conversation and whether their reentry is hateful or non-hateful. However, this analysis was conducted retrospectively on a benchmark dataset of Reddit conversations and not in a real intervention.

A closer look at psychosocial findings reveals both convergences and divergences. Studies in real contexts capture actual behaviors, whereas controlled context studies can face an intention-behavior gap. Additionally, real context studies can measure long-term network dynamics, where counterspeech could dampen the spread of negativity across a user's neighboring accounts.

Lastly, almost all studies assume that the target of counterspeech is the hater, but several findings (Bilewicz et al., 2021; Wu et al., 2026) suggest that the main effect may be on bystanders, implying designs targeted within the community rather than on individual change.

4. Discussion and conclusions

Counterspeech is a social intervention aimed at countering hate speech; its effectiveness depends on prosocial effects on both the hater and bystanders, such as changes in behavior, norms, attitudes, and empathy toward victims. When AI generates such interventions, key questions arise: what makes counterspeech effective? How does an artificial sender affect perception? And what happens to digital communities when machines act as social actors? However, reducing the effectiveness to the machine alone is an oversimplification.

The preceding sections have presented our findings in response to the three objectives that guided this review: conceptualizations of counterspeech, collaborative approaches to generation and evaluation, and alignment between goals and measures. In this discussion, we synthesize the findings in our review, highlight research gaps, and propose future directions.

The explorative findings of this review reveal a theoretical and methodological fragmented picture. Conceptualizations vary widely; taxonomies often blend tone, rhetorical strategies, and normative functions. Rather than treating this heterogeneity as a limitation alone, we propose that future research could leverage these dimensions to be distinguished and operationalized. This review proposed a six-dimension taxonomy (see Table 2). However, this taxonomy is purely illustrative. Moving forward, we argue that the field would benefit from a collective effort to establish a shared conceptual framework. Generation methods differ across studies, and evaluation metrics are often incomparable, limiting synthesis. Automated and human metrics typically focus on formal text properties, failing to capture psychological and behavioral outcomes.

Human evaluations attempt to fill this gap, but are usually applied to decontextualized texts; only a few studies assess counterspeech in real-world contexts (Bär et al., 2024; Bilewicz et al., 2021; Podolak et al., 2024). There is a structural disconnect between LLM capabilities (now more context-aware and nuanced) and psychosocial implementation. Traditional metrics correlate poorly with human judgment (Bonaldi et al., 2024; Cima et al., 2025), reflecting a systematic misalignment between objectives and measures. To address the latter, researchers have turned to LLMs as evaluators, achieving better scalability and correlation with human judgment (Hengle et al., 2025). However, validity concerns remain (Chung et al., 2024). JudgeLM often favors AI-generated over human-preferred responses (Bennie et al., 2025), suggesting susceptibility to stylistic cues rather than substantive quality. Future directions should include defining and reliably measuring effective counterspeech through integrated computational and psychosocial expertise, and expanding non-English datasets for culturally aware evaluation.

We argue that computational metrics alone are insufficient for evaluating counterspeech effectiveness. They should be complemented by psychosocial measures that can target different mechanisms of change (normative, cognitive or affective), therefore strategy-specific. Moreover, LLM judges cannot replace human evaluators as they can't assess psychosocial effectiveness. We therefore recommend a hybrid pipeline where LLMs serve as initial evaluators and humans (ideally real users) remain the decisive standard.

Research on the social identity of chatbots as counterspeakers remains scarce. Recent studies (for instance, Wang et al., 2026) show differing expectations for AI versus human counterspeakers, but systematic evaluation of whether users accept AI-generated counterspeech is lacking, leaving its effectiveness unclear. Moreover, identity transparency is also an issue. Wu et al. (2025) found that nondisclosure increases engagement, while disclosure reduces it. Bär et al. (2024) observed that revealing AI involvement led to backfire effects and conversational aggression. Nondisclosure may boost short-term engagement but risks eroding trust; disclosure may provoke hostility. Future research should address this ethical problem by investigating how different degrees of transparency can be successfully implemented in real life context.

Human in the Loop (HITL) approaches vary: humans as reviewers filtering AI outputs (Ashida & Komachi, 2022; Fanton et al., 2021), humans as trainers providing feedback to improve models over time (Bengoetxea et al., 2024; Bonaldi et al., 2022), as in an Author-Reviewer pipeline (Tekiroğlu et al., 2020; Wu

© Italian e-Learning Association

et al., 2025), and leveraging human expertise to identify model weaknesses and generate targeted training data (Chung & Bright, 2024). Genuine collaboration in counterspeech writing has only been employed by Ding et al. (2025) with their chatbot CounterQuill. The most common human role is still that of a database curator, reviewer, or proofreader. There is a lack of collaboration evaluation in terms of the psychological and educational effectiveness of counterspeech. We align with Chung et al. (2024) in stating that counterspeech generation tools should be deployed as suggestion tools to assist humans rather than replace them.

Future research should prioritize systematizing effective counterspeech through a unified interdisciplinary framework, essential for moving from fragmented efforts toward cumulative knowledge. Unexplored outcomes also deserve attention, including cognitive and affective changes, across platforms and communities. Additionally, research should examine how users perceive AI counterspeakers across contexts and cultures, which identities enhance or reduce effectiveness, how human-AI dynamics shape community norms, and how to balance transparency with impact, all questions essential for ethically grounded system design. A priority should be collaborative alternatives to full automation: AI systems that support human empowerment, serving as tools for augmentation rather than substitution. As AI takes on social roles in social media, we must promote not only technical progress but also morally responsible and informed use.

Finally, we acknowledge several limitations of this study. Given the rapid evolution of AI research, we chose to include preprints. Although these have not yet undergone formal peer review, we carefully examined their content and relevance for the purposes of this scoping review. Furthermore, the substantial heterogeneity in conceptualizations, generation methods, and evaluation metrics across the 45 included studies precluded any quantitative synthesis or meta-analysis. Consequently, our findings should be interpreted as a mapping of the field rather than as a definitive assessment of effectiveness.

References

Ashida, M., & Komachi, M. (2022). Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions. In K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, & Z. Talat (A c. Di), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) (pp. 11–23). Association for Computational Linguistics.

Baez Santamaria, S., Gomez Adorno, H., & Markov, I. (2024). Contextualized Graph Representations for Generating Counter-Narratives against Hate Speech. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (A c. Di), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 7664–7674). Association for Computational Linguistics.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. Journal of Personality and Social Psychology, 71, 364–374.

Bär, D., Maarouf, A., & Feuerriegel, S. (2024). Generative AI may backfire for counterspeech (arXiv:2411.14986). arXiv.

Benesch, S. (2016). Counterspeech on Twitter: A Field Study.

Bennie, M., Zhang, D., Xiao, B., Cao, J., Liu, C. X., Meng, J., & Tripp, A. (2025). PANDA -- Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset (arXiv:2501.00697). arXiv.

Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. Aggressive Behavior, 47(3), 260–266.

Bonaldi, H., Damo, G., Ocampo, N. B., Cabrio, E., Villata, S., & Guerini, M. (2024). Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering (arXiv:2410.03466). arXiv.

Bonaldi, H., Dellantonio, S., Tekiroğlu, S. S., & Guerini, M. (2022). Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. In Y. Goldberg, Z. Kozareva, & Y. Zhang (A c. Di), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 8031–8049). Association for Computational Linguistics.

Cecconi, C., Poggi, I., & D'Errico, F. (2020). Schadenfreude: malicious joy in social media interactions. Frontiers in Psychology, 11, 558282.

Cima, L., Miaschi, A., Trujillo, A., Avvenuti, M., Dell'Orletta, F., & Cresci, S. (2025). Contextualized Counterspeech: Strategies for Adaptation, Personalization, and Evaluation. WWW - Proc. ACM Web Conf., 5022–5033.

Chung, Y.-L., Abercrombie, G., Enock, F., Bright, J., & Rieser, V. (2024). Understanding Counterspeech for Online Harm Mitigation. ArXiv.org.

Chung, Y.-L., & Bright, J. (2024). On the Effectiveness of Adversarial Robustness for Abuse Mitigation

© Italian e-Learning Association

with Counterspeech. In K. Duh, H. Gomez, & S. Bethard (A c. Di), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 6988–7002). Association for Computational Linguistics.

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology, 8(4, Pt.1), 377–383.

Ding, X., Ping, K., Gunturi, U. S., Carik, B., Stil, S., Wilhelm, L. T., Daryanto, T., Hawdon, J., Lee, S. W., & Rho, E. H. (2025). Designing Human-AI Collaboration to Support Learning in Counterspeech Writing (arXiv:2410.03032). arXiv.

Fanton, M., Bonaldi, H., Tekiroğlu, S. S., & Guerini, M. (2021). Human-in-the-Loop for Data Collection: A Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In C. Zong, F. Xia, W. Li, & R. Navigli (A c. Di), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 3226–3240). Association for Computational Linguistics.

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, 51(4), 1–30.

Garland, J., Ghazi□Zahedi, K., Young, J. G., Hébert□Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter□speech online. EPJ Data Science, 11(3):3.

Hassan, S., & Alikhani, M. (2023). DisCGen: A Framework for Discourse-Informed Counterspeech Generation (arXiv:2311.18147). arXiv.

Hengle, A., Padhi, A. K., Bandhakavi, A., & Chakraborty, T. (2025). CSEval: Towards Automated, Multi-Dimensional, and Reference-Free Counterspeech Evaluation using Auto-Calibrated LLMs. In L. Chiruzzo, A. Ritter, & L. Wang (A c. Di), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 5402–5419). Association for Computational Linguistics.

Hengle, A., Padhi, A., Singh, S., Bandhakavi, A., Akhtar, M. S., & Chakraborty, T. (2024). Intent-conditioned and Non-toxic Counterspeech Generation using Multi-Task Instruction Tuning with RLAIF. In K. Duh, H. Gomez, & S. Bethard (A c. Di), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 6716–6733). Association for Computational Linguistics.

Hong, L., Luo, P., Blanco, E., & Song, X. (2024). Outcome-Constrained Large Language Models for Countering Hate Speech (arXiv:2403.17146). arXiv.

Horta Ribeiro, M., Hosseinmardi, H., West, R., & Watts, D. J. (2023). Deplatforming did not decrease Parler users' activity on fringe social media. PNAS Nexus, 2(3).

Kumar, A., Bandhakavi, A., & Chakraborty, T. (2025). Counterspeech the ultimate shield! Multi-Conditioned Counterspeech Generation through Attributed Prefix Learning (arXiv:2505.11958). arXiv.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherje, A. (2019). Thou shalt not hate: Countering Online Hate Speech (arXiv:1808.04409). arXiv.

Mun, J., Allaway, E., Yerukola, A., Vianna, L., Leslie, S.-J., & Sap, M. (2023). Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language (arXiv:2311.00161). arXiv.

Podolak, J., Łukasik, S., Balawender, P., Ossowski, J., Piotrowski, J., Bakowicz, K., & Sankowski, P. (2024). LLM generated responses to mitigate the impact of hate speech. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (A c. Di), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 15860–15876). Association for Computational Linguistics.

Poggi, I., D'Errico, F., & Vincze, L. (2013). Comments by words, face and body. Journal on Multimodal User Interfaces, 7(1), 67-78.

Prasannan, P., Kumaresan, P. K., Rajiakodi, S., Subalalitha, C. N., & Chakravarthi, B. R. (2025). Counter-speech generation for homophobic and transphobic social media content in Malayalam. Social Network Analysis and Mining, 15(1), 87.

Sahoo, N. R., Beria, G. P., & Bhattacharyya, P. (2024). IndicCONAN: A Multilingual Dataset for Combating Hate Speech in Indian Context. Proceedings of the AAAI Conference on Artificial Intelligence, 38(20), 22313–22321.

Sportelli, C., Cicirelli, P. G., Paciello, M., Corbelli, G., & D'Errico, F. (2025). "Let's Make the Difference!" Promoting Hate Counter□Speech in Adolescence Through Empathy and Digital Intergroup Contact. Journal of Community & Applied Social Psychology, 35(1), e70028

Tekiroğlu, S. S., Chung, Y.-L., & Guerini, M. (2020). Generating Counter Narratives against Online Hate

© Italian e-Learning Association

Speech: Data and Strategies. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (A c. Di), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1177–1190). Association for Computational Linguistics.

Tricco, AC, Lillie, E, Zarin, W, O'Brien, KK, Colquhoun, H, Levac, D, Moher, D, Peters, MD, Horsley, T, Weeks, L, Hempel, S et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018,169(7):467-473. doi: 10.7326/M18-0850

Wachs, S., Valido, A., Espelage, D. L., Castellanos, M., Wettstein, A., & Bilz, L. (2023). The relation of classroom climate to adolescents' countering hate speech via social skills: A positive youth development perspective. Journal of Adolescence, 95(6), 1127–1139.

Wang, H., Pan, Y., Song, X., Zhao, X., Hu, M., & Zhou, B. (2024). F^2RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (A c. Di), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 4457–4470). Association for Computational Linguistics.

Wang, M., Ma, S., Li, N., Zhang, P., Li, C., Gu, N., & Lu, T. (2026). Echoes of Norms: Investigating Counterspeech Bots' Influence on Bystanders in Online Communities (arXiv:2603.03687). arXiv.

Wu, C., Wang, Y., Zhang, Y., Wang, H., & Pang, Y. (2025). Confront hate with AI: How AI-generated counter speech helps against hate speech on social Media? Telematics and Informatics, 101, 102304.

Zhu, W., & Bhat, S. (2021). Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech (arXiv:2106.01625). arXiv.

Zhu, L., Wang, X., & Wang, X. (2023, October 26). JudgeLM: Fine-tuned Large Language Models are Scalable Judges. arXiv.Org.

Zubiaga, I., Soroa, A., & Agerri, R. (2024). A LLM-based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (A c. Di), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 9572–9585). Association for Computational Linguistics.

© Italian e-Learning Association

Appendix A

Alyahya, G., & Aldayel, A. (2025). *Hatred Stems from Ignorance! Distillation of the Persuasion Modes in Countering Conversational Hate Speech* (arXiv:2403.15449). arXiv. https://doi.org/10.48550/arXiv.2403.15449

Ashida, M., & Komachi, M. (2022). Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions. In K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, & Z. Talat (Eds.), *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (pp. 11–23). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.woah-1.2

Baez Santamaria, S., Gomez Adorno, H., & Markov, I. (2024). Contextualized Graph Representations for Generating Counter-Narratives against Hate Speech. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 7664–7674). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.450

Bär, D., Maarouf, A., & Feuerriegel, S. (2024). *Generative AI may backfire for counterspeech*

(arXiv:2411.14986). arXiv. https://doi.org/10.48550/arXiv.2411.14986

Bengoetxea, J., Chung, Y.-L., Guerini, M., & Agerri, R. (2024). Basque and Spanish Counter Narrative Generation: Data Creation and Evaluation. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2132–2141). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.192/

Bennie, M., Zhang, D., Xiao, B., Cao, J., Liu, C. X., Meng, J., & Tripp, A. (2025). *PANDA -- Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset* (arXiv:2501.00697). arXiv. https://doi.org/10.48550/arXiv.2501.00697

Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, *47*(3), 260–266. https://doi.org/10.1002/ab.21948

Bonaldi, H., Attanasio, G., Nozza, D., & Guerini, M. (2023). Weigh Your Own Words: Improving Hate Speech Counter Narrative Generation via Attention Regularization. In Y.-L. Chung, H. Bonaldi, G. Abercrombie, & M. Guerini (Eds.), *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)* (pp. 13–28). Association for Computational Linguistics. https://aclanthology.org/2023.cs4oa-1.2/

Bonaldi, H., Damo, G., Ocampo, N. B., Cabrio, E., Villata, S., & Guerini, M. (2024). *Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering* (arXiv:2410.03466). arXiv. https://doi.org/10.48550/arXiv.2410.03466

Bonaldi, H., Dellantonio, S., Tekiroğlu, S. S., & Guerini, M. (2022). Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 8031–8049). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.549

Chen, G., Cheng, L., Luu, A. T., & Bing, L. (2024). Exploring the Potential of Large Language Models in Computational Argumentation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp.

2309–2330). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.126 Chung, Y.-L., & Bright, J. (2024). On the Effectiveness of Adversarial Robustness for Abuse Mitigation with

Counterspeech. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 6988–7002). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.386

Chung, Y.-L., Tekiroğlu, S. S., & Guerini, M. (2020). Italian Counter Narrative Generation to Fight Online Hate Speech. In J. Monti, F. Dell'Orletta, & F. Tamburini (Eds.), *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)* (pp. 82–88). CEUR Workshop Proceedings. https://aclanthology.org/2020.clicit-1.15/

Chung, Y.-L., Tekiroğlu, S. S., & Guerini, M. (2021). Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 899–914. https://doi.org/10.18653/v1/2021.findings-acl.79

© Italian e-Learning Association

Cima, L., Miaschi, A., Trujillo, A., Avvenuti, M., Dell'Orletta, F., & Cresci, S. (2025). Contextualized Counterspeech: Strategies for Adaptation, Personalization, and Evaluation. *WWW - Proc. ACM Web Conf.*, 5022–5033. https://doi.org/10.1145/3696410.3714507

Das, M., Pandey, S., Sethi, S., Saha, P., & Mukherjee, A. (2024). Low-Resource Counterspeech Generation for Indic Languages: The Case of Bengali and Hindi. In Y. Graham & M. Purver (Eds.), *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 1601–1614). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-eacl.111

Ding, X., Ping, K., Gunturi, U. S., Carik, B., Stil, S., Wilhelm, L. T., Daryanto, T., Hawdon, J., Lee, S. W., & Rho, E. H. (2025). *Designing Human-AI Collaboration to Support Learning in Counterspeech Writing* (arXiv:2410.03032). arXiv. https://doi.org/10.48550/arXiv.2410.03032

Doğanç, M., & Markov, I. (2023). From Generic to Personalized: Investigating Strategies for Generating Targeted Counter Narratives against Hate Speech. In Y.-L. Chung, H. Bonaldi, G. Abercrombie, & M. Guerini (Eds.), *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)* (pp. 1–12). Association for Computational Linguistics. https://aclanthology.org/2023.cs4oa-1.1/

Fanton, M., Bonaldi, H., Tekiroğlu, S. S., & Guerini, M. (2021). Human-in-the-Loop for Data Collection: A Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 3226–3240). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2021.acl-long.250

Furman, D., Torres, P., Rodríguez, J., Letzen, D., Martinez, M., & Alemany, L. (2023). High-quality argumentative information in low resources approaches improve counter-narrative generation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2942–2956). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.194

Gupta, R., Desai, S., Goel, M., Bandhakavi, A., Chakraborty, T., & Akhtar, Md. S. (2023). Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5792–5809). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.318

Hassan, S., & Alikhani, M. (2023). *DisCGen: A Framework for Discourse-Informed Counterspeech Generation* (arXiv:2311.18147). arXiv. https://doi.org/10.48550/arXiv.2311.18147

Hengle, A., Padhi, A. K., Bandhakavi, A., & Chakraborty, T. (2025). CSEval: Towards Automated, Multi-Dimensional, and Reference-Free Counterspeech Evaluation using Auto-Calibrated LLMs. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5402–5419). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.279

Hengle, A., Padhi, A., Singh, S., Bandhakavi, A., Akhtar, M. S., & Chakraborty, T. (2024). Intent-conditioned and Non-toxic Counterspeech Generation using Multi-Task Instruction Tuning with RLAIF. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 6716–6733). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.374

Hong, L., Luo, P., Blanco, E., & Song, X. (2024). *Outcome-Constrained Large Language Models for Countering Hate Speech* (arXiv:2403.17146). arXiv. https://doi.org/10.48550/arXiv.2403.17146

Jiang, S., Tang, W., Chen, X., Tang, R., Wang, H., & Wang, W. (2025). ReZG: Retrieval-augmented zero-shot counter narrative generation for hate speech. *Neurocomputing*, *620*, 129140. https://doi.org/10.1016/j.neucom.2024.129140

Kumar, A., Bandhakavi, A., & Chakraborty, T. (2025). *Counterspeech the ultimate shield! Multi-Conditioned Counterspeech Generation through Attributed Prefix Learning* (arXiv:2505.11958). arXiv. https://doi.org/10.48550/arXiv.2505.11958

© Italian e-Learning Association

Lee, S., Jung, D., Park, C., Lee, S., & Lim, H. (2024). *Alternative Speech: Complementary Method to Counter-Narrative for Better Discourse* (arXiv:2401.14616). arXiv. https://doi.org/10.48550/arXiv.2401.14616

Mun, J., Allaway, E., Yerukola, A., Vianna, L., Leslie, S.-J., & Sap, M. (2023). *Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language* (arXiv:2311.00161). arXiv. https://doi.org/10.48550/arXiv.2311.00161

Piot, P., & Parapar, J. (2025). Decoding Hate: Exploring Language Models' Reactions to Hate Speech. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 973–990. https://doi.org/10.18653/v1/2025.naacl-long.45

Podolak, J., Łukasik, S., Balawender, P., Ossowski, J., Piotrowski, J., Bakowicz, K., & Sankowski, P. (2024). LLM generated responses to mitigate the impact of hate speech. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 15860–15876). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.931

Prasannan, P., Kumaresan, P. K., Rajiakodi, S., Subalalitha, C. N., & Chakravarthi, B. R. (2025). Counter-speech generation for homophobic and transphobic social media content in Malayalam. *Social Network Analysis and Mining*, *15*(1), 87. https://doi.org/10.1007/s13278-025-01507-x

Saha, P., Datta, A., Jana, A., & Mukherjee, A. (2024). *CrowdCounter: A benchmark type-specific multi-target counterspeech dataset* (arXiv:2410.01400). arXiv. https://doi.org/10.48550/arXiv.2410.01400

Sahoo, N. R., Beria, G. P., & Bhattacharyya, P. (2024). IndicCONAN: A Multilingual Dataset for Combating Hate Speech in Indian Context. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22313–22321. https://doi.org/10.1609/aaai.v38i20.30237

Song, X., Mamidisetty, S., Blanco, E., & Hong, L. (2024). *Assessing the Human Likeness of AI-Generated Counterspeech* (arXiv:2410.11007). arXiv. https://doi.org/10.48550/arXiv.2410.11007

Tekiroğlu, S. S., Bonaldi, H., Fanton, M., & Guerini, M. (2022). Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: A Comparative Study. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 3099–3114). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.245

Tekiroğlu, S. S., Chung, Y.-L., & Guerini, M. (2020). Generating Counter Narratives against Online Hate Speech: Data and Strategies. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1177–1190). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.110

Wang, H., Pan, Y., Song, X., Zhao, X., Hu, M., & Zhou, B. (2024). F^2RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4457–4470). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.255

Wang, H., Tian, Z., Song, X., Zhang, Y., Pan, Y., Tu, H., Huang, M., & Zhou, B. (2024). Intent-Aware and Hate-Mitigating Counterspeech Generation via Dual-Discriminator Guided LLMs. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 9131–9142). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.800/

Wang, M., Ma, S., Li, N., Zhang, P., Li, C., Gu, N., & Lu, T. (2026). *Echoes of Norms: Investigating Counterspeech Bots' Influence on Bystanders in Online Communities* (arXiv:2603.03687). ArXiv. https://doi.org/10.48550/arXiv.2603.03687

Wilk, B., Shomee, H. H., Maity, S. K., & Medya, S. (2025). Fact-based Counter Narrative Generation to Combat Hate Speech. *Proceedings of the ACM on Web Conference 2025, WWW '25*, 3354–3365. https://doi.org/10.1145/3696410.3714718

Wu, C., Wang, Y., Zhang, Y., Wang, H., & Pang, Y. (2025). Confront hate with AI: How AI-generated counter speech helps against hate speech on social Media? *Telematics and Informatics*, *101*, 102304. https://doi.org/10.1016/j.tele.2025.102304

© Italian e-Learning Association

Zheng, Y., Ross, B., & Magdy, W. (2023). What Makes Good Counterspeech? A Comparison of Generation Approaches and Evaluation Metrics. In Y.-L. Chung, H. Bonaldi, G. Abercrombie, & M. Guerini (Eds.), *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)* (pp. 62–71). Association for Computational Linguistics. https://aclanthology.org/2023.cs4oa-1.5/

Zhu, W., & Bhat, S. (2021). *Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech* (arXiv:2106.01625). arXiv. https://doi.org/10.48550/arXiv.2106.01625

Zubiaga, I., Soroa, A., & Agerri, R. (2024). A LLM-based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 9572–9585). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.559

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026), pp. 91-104

From euphemism to extremism: emerging alphabets of hate in digital academic spaces in Eastern Indonesia

Remerta Noni Naatonis[1], Sumarlin Sumarlin, Jimi Asmara, Yosep Jacob Latuan, Dewi Anggraini, Menhya Snae, Heni Heni, Erna Rosani Nubatonis, Sadam Muhammad Natzir, Ewaldus Kase

STIKOM Uyelindo – Kupang (Indonesia)

(submitted: 5/3/2026; accepted: 23/5/2026; published: 28/8/2026)

Abstract

This study examines the emergence and escalation of hate speech in digital academic spaces across Eastern Indonesia through a mixed-methods approach integrating critical discourse analysis and artificial intelligence-based natural language processing. Drawing on a corpus of 4,200 posts and comments collected from institutional LMS platforms, WhatsApp/Telegram channels, and YouTube comment sections across four provinces (Papua, Maluku, Nusa Tenggara Timur, and Sulawesi), the study finds that 38.4% of the total sample contains some form of hate speech, with implicit euphemistic hate speech (IHS-E) consistently outnumbering explicit hate speech (EHS) at an average ratio of 1.5:1 across all platforms and provinces. Critical discourse analysis identifies five primary linguistic strategies constituting the local "alphabet of hate": ethnic euphemism and regional nickname deployment, covert religious dog-whistling, ironic praise, dehumanizing metaphor, and socioeconomic stereotyping, each operating in province-specific intersectional configurations. Computational evaluation using a combined pipeline of fine-tuned IndoBERT, DeepHate, and a community-validated local language lexicon classifier achieves a macro-averaged F1 of 0.859, with the local lexicon proving superior to transformer-based models alone in detecting IHS-E. A documented five-stage escalation trajectory confirms that the movement from euphemism to extremism follows an identifiable and intervenable pattern. This study contributes a culturally responsive hate speech detection framework and provides an empirical foundation for more inclusive digital academic governance in Eastern Indonesia.

KEYWORDS: Digital Hate Speech, Alphabet of Hate, Eastern Indonesia, Natural Language Processing, Critical Discourse Analysis.

DOI
https://doi.org/10.20368/1971-8829/1136316

CITE AS
Naatonis, R.N., Sumarlin, S., Asmara, J., Latuan, Y.J., Anggraini, D., Snae, D., Heni, H., Nubatonis, E.R., Natzir, S.M., & Kase, E. (2026). From euphemism to extremism: emerging alphabets of hate in digital academic spaces in Eastern Indonesia. *Journal of e-Learning and Knowledge Society*, *22*(1), 91-104. https://doi.org/10.20368/1971-8829/1136316

1. Introduction

The rapid spread of digital communication platforms has dramatically reshaped how public discourse unfolds. While these shifts have opened meaningful avenues for academic exchange, they have equally provided fertile ground for new and evolving forms of hateful expression. In educational settings across Indonesia, particularly in the eastern regions where ethnic and cultural diversity runs deep, digitally mediated hate speech has emerged as a genuine threat to inclusive learning environments and institutional credibility. The blending of long-standing offline social prejudices with the anonymizing and amplifying power of online platforms has given rise to what some researchers call an "alphabet of hate", a shifting, coded vocabulary through which hostility toward marginalized communities is spread and normalized (Alorainy et al., 2021; Baider, 2023).

Hate speech, broadly speaking, refers to any communicative act that demeans, threatens, or stirs hostility against individuals or groups on the basis of characteristics such as ethnicity, religion, gender, or

[1] corresponding author - email: reyheka@gmail.com

© Italian e-Learning Association

regional identity. It has long been recognized as deeply corrosive to social life (Alkomah & Ma, 2022; Castellanos et al., 2023). Its digital forms, however, present considerably more complex challenges. The problem is not simply one of scale or speed, though harmful content does travel faster and further online than ever before. It is also that perpetrators have grown increasingly sophisticated in their use of language, finding ways to sustain impact while avoiding detection. In academic digital spaces such as learning management systems, institutional forums, university-affiliated social media groups, and student communication channels, these dynamics carry particular weight, given that such environments are expected to foster critical thinking and respectful engagement (Nascimento et al., 2023; Schäfer et al., 2024).

Eastern Indonesia, covering provinces such as Papua, Maluku, Nusa Tenggara Timur, and Sulawesi, presents a uniquely layered sociolinguistic context for studying hate speech online. The region's multiethnic and multilingual character, shaped by histories of colonial stratification, interreligious friction, and geographic isolation, has created conditions in which identity-based hostility can take root and spread through digital channels (Hayaty et al., 2020; Titley, 2020). At the same time, growing internet access and smartphone use across these provinces have brought previously disconnected communities into contact with both regional and global streams of hateful content. Much of this content circulates through local language varieties, euphemistic expressions, and culturally specific references that standard hate speech detection systems are ill-equipped to catch (Kovala et al., 2023; Nikunen, 2023).

Research on automated hate speech detection has come a long way in identifying explicit and overt forms of online hostility, especially in widely spoken languages such as English, Spanish, and Arabic (Al-Makhadmeh & Tolba, 2020; Cao et al., 2020; del Arco et al., 2021). Models built on deep learning architectures and transformer-based systems like BERT have shown real promise in classifying hate speech at scale (Alatawi et al., 2021; Khan et al., 2021). Yet a stubborn gap remains. These systems consistently underperform when faced with hate speech that works through indirection, through euphemism, irony, coded phrasing, and cultural dog-whistles, particularly in low-resource and regional language settings (Charitidis et al., 2020; Casula et al., 2021).

The path from veiled hostility to outright extremism in online hate discourse rarely happens all at once. More often, it unfolds gradually, as language that initially seems ambiguous or even benign takes on more explicitly hateful meanings through repeated use within particular online communities (Norocel et al., 2022; Titley, 2020). In academic digital spaces, this kind of escalation is frequently enabled by structural blind spots: content moderation tools calibrated for majority languages, educators and administrators who lack the training to recognize culturally embedded signals of hate, and platform governance frameworks that err on the side of free expression rather than harm prevention (Baider, 2023; Schäfer et al., 2024). The outcome is an environment where hate speech does not merely exist in the margins but is, in a very real sense, sustained by the very conditions that structure digital academic life.

This study draws theoretically on three strands of scholarship that have largely evolved in isolation from one another, yet whose convergence is precisely what understanding hate speech in non-Western, multilingual, and postcolonially situated settings demands. The first strand is computational hate speech detection research, which has made genuine strides in classification accuracy for high-resource languages, but has faced recurring criticism for treating hate speech as a surface-readable phenomenon, one that can be reliably identified through lexical lookups or syntactic patterns alone (Alkomah & Ma, 2022; Nascimento et al., 2023). The second is critical discourse analysis, a tradition that grounds itself in the ideological weight of language and that has long attended to how apparently innocuous or ambiguous utterances can, through indirection, irony, and intertextual layering, reproduce structures of domination and exclusion (Norocel et al., 2022; Etaywe & Zappavigna, 2023). The third strand, rarely foregrounded in hate speech research, is the postcolonial and decolonial critique of knowledge production. This tradition insists that the analytical categories, evaluative standards, and methodological assumptions built into mainstream social science carry the imprint of their Western origins, and may require substantial renegotiation before they can meaningfully illuminate societies shaped by different histories of language, identity, and political violence (Kovala et al., 2023; Titley, 2020).

In Eastern Indonesia, the boundaries between ethnic, religious, regional, and class identities are historically contingent and politically charged in ways that diverge fundamentally from the identity formations assumed in most Western hate speech research. In this setting, the decolonial perspective is not an optional theoretical supplement but a methodological requirement. The present study brings these three strands together by operationalizing the concept of the "alphabet of hate" as a category that is at once computationally detectable, discursively interpretable, and culturally embedded, producing an analytically pluralist framework whose insights can travel beyond the Eastern Indonesian context from which they emerge.

Against this backdrop, the present study investigates the linguistic and semiotic mechanisms through which hate speech emerges, evolves, and escalates within digital academic spaces in Eastern Indonesia. Specifically, it examines how locally inflected forms of hateful language rooted in regional vernaculars, ethnic stereotyping, religious discrimination, and intersecting axes of marginalization constitute an emergent "alphabet of hate" that is simultaneously legible to in-group members and opaque to automated detection systems

© Italian e-Learning Association

and institutional gatekeepers alike (Alkomah & Ma, 2022; Nascimento et al., 2023). In doing so, this study contributes to a growing interdisciplinary literature at the intersection of computational linguistics, critical discourse analysis, education technology, and regional studies, with direct implications for the design of culturally responsive hate speech detection systems and inclusive digital pedagogy.

The significance of this inquiry is underscored by the limited scholarly attention given to hate speech, digital education, and regional language dynamics in the Indonesian context. While some studies have addressed hate speech detection in Bahasa Indonesia and certain local languages (Hayaty et al., 2020; Khan et al., 2021), empirical research focused specifically on Eastern Indonesia's academic digital spaces remains scarce. This gap is particularly concerning given the region's history of ethno-regional inequality and the vulnerability of marginalized student populations. As Castellanos et al. (2023) and Schäfer et al. (2024) demonstrate, hate speech in educational environments causes significant psychological harm, weakens social cohesion, and erodes the conditions necessary for genuine academic inquiry.

The conceptual and methodological tools currently dominating hate speech research, including transformer-based detection pipelines, annotation taxonomies, and models of escalatory discourse, have been developed almost entirely within Western, high-resource-language contexts. English, German, Spanish, and Arabic have attracted the bulk of scholarly investment, and the assumptions embedded in these frameworks do not transfer cleanly into settings shaped by different linguistic histories and political conflicts (Alkomah & Ma, 2022; Kovala et al., 2023).

The contribution of this study therefore runs deeper than filling a geographic gap in the literature. Digital hate speech in Eastern Indonesia takes distinct forms, follows particular escalation pathways, and requires detection strategies that are grounded in local context. Importing Western frameworks wholesale inevitably produces a distorted picture. By anchoring the analysis in Eastern Indonesia's sociolinguistic and historical realities, this study advances a model in which cultural embeddedness and local knowledge are central to the analytical apparatus. In doing so, it contributes to the broader project of decolonizing hate speech research, treating Eastern Indonesia not merely as a site for applying theory developed elsewhere, but as a generative context capable of producing genuinely new theoretical insights.

2. Procedures and methods

2.1 Research Design

This study adopts an exploratory-descriptive mixed-methods design that integrates qualitative Critical Discourse Analysis (CDA) with computational Natural Language Processing (NLP) techniques. The design is grounded in a social constructivist epistemological framework, which treats hate speech not as a static linguistic artifact but as a socially constructed practice whose meaning is shaped by power relations, cultural identity, and institutional dynamics (Baider, 2023; Nikunen et al., 2025). This stance is particularly appropriate for the Eastern Indonesian academic context, where the emergence of the "alphabet of hate" is deeply tied to historically rooted structures of ethno-regional inequality, religious differentiation, and geographic marginalization.

The mixed-methods approach responds to a dual analytical challenge: the need for large-scale, systematic detection of hate speech patterns across a diverse digital corpus, and the need for interpretive depth in understanding how locally inflected euphemisms, coded language, and cultural dog-whistles evade conventional detection systems (Nascimento et al., 2023; Alkomah & Ma, 2022). Figure 1 presents the overall six-phase research procedure.

2.2 Data Sources and Corpus Construction

Primary data were collected from three categories of digital academic spaces directly affiliated with higher education institutions across Eastern Indonesia. These three categories correspond precisely to the digital spaces identified in the introduction as sites where hate speech dynamics carry heightened significance: online learning management systems and institutional forums, student communication channels used for academic coordination, and academic video content platforms. Geographic coverage encompassed four provinces Papua, Maluku, Nusa Tenggara Timur (NTT), and Sulawesi selected on the basis of their ethnocultural diversity and documented exposure to digital hate discourse (Hayaty et al., 2020; Titley, 2020). Table 1 presents the full distribution of data sources.

Purposive sampling guided the selection of specific groups, channels, and comment threads based on thematic relevance to identity-based discourse and preliminary indicators of hate content identified in a scoping phase. A total corpus of 4,200 posts and comments was compiled across a twelve-month data collection period. Table 2 provides a methodological justification for the adequacy of this corpus size against established benchmarks in the hate speech detection literature.

All data collection procedures adhered strictly to the ethical guidelines of the Association of Internet Researchers (AoIR). User identities were comprehensively anonymised prior to analysis. Content directly identifiable to specific individuals beyond what was necessary for linguistic analysis was permanently excluded from the research database. Platform administrators were notified of the research activity where feasible. Institutional ethical clearance was obtained prior to the commencement of data collection (Alorainy et al., 2021; Casula et al., 2021).

© Italian e-Learning Association

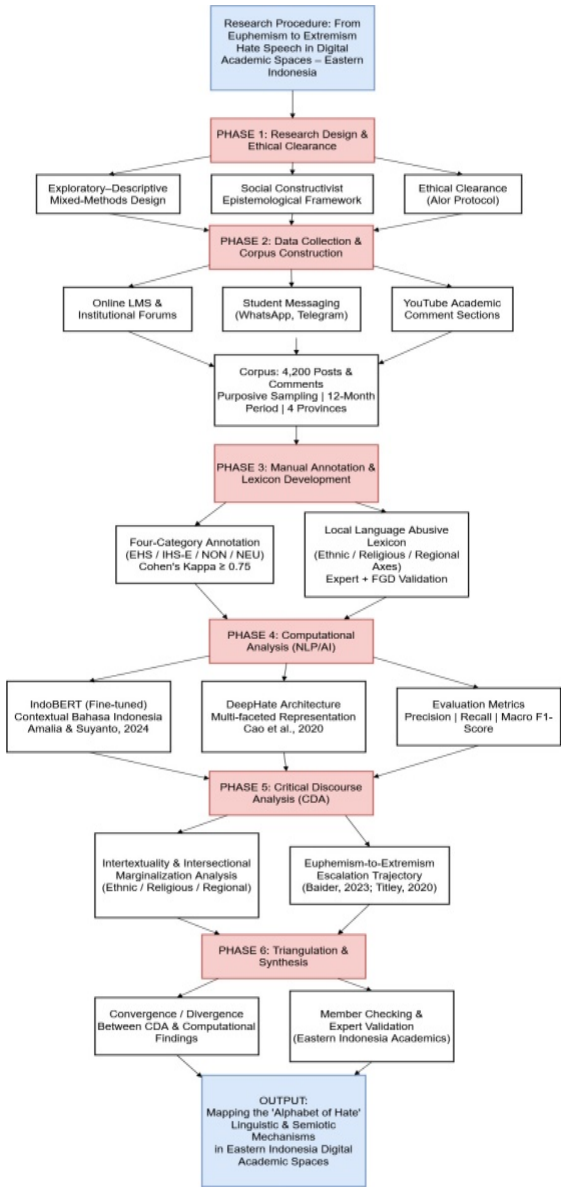


Figure 1 - Six-Phase Research Procedure: From Euphemism to Extremism in Eastern Indonesia Digital Academic Spaces.

2.3 Manual Annotation and Local Language Lexicon Development

Phase 3 comprised two parallel tracks of data preparation. The first track involved manual annotation of a stratified sample of 1,200 data points by three trained annotators with documented competence in Bahasa Indonesia and at least one Eastern Indonesian regional language. The annotation scheme was specifically designed to operationalise the euphemism-to-extremism trajectory that constitutes the core analytical focus of this study, distinguishing between four mutually exclusive categories. Crucially, and in alignment with the introduction's emphasis on the intersection of ethnic, religious, and regional axes of marginalisation, the scheme explicitly incorporates these intersectional dimensions as discriminative features within each category, as detailed in Table 3.

Inter-annotator agreement was measured using Cohen's Kappa coefficient, with a minimum acceptance threshold of kappa >= 0.75, consistent with standards applied in comparable hate speech annotation studies (Paasch-Colberg et al., 2021; Rieger et al., 2021). The IHS-E category, covering implicit and euphemistic hate speech, received particular methodological attention given its centrality to the research problem, as it represents the critical threshold in the escalatory trajectory from covert to explicit hate. In cases of annotator disagreement, a fourth expert adjudicator with specialization in Eastern Indonesian sociolinguistics was engaged for final resolution.

The second track involved constructing a domain-specific abusive words lexicon for Eastern Indonesian regional languages, directly addressing the detection gap identified earlier regarding locally coded hate language. The lexicon was built through expert consultation with regional linguists from Papua, Maluku, NTT, and Sulawesi, complemented by focus group discussions with students from those provinces. It specifically maps euphemistic codewords, culturally embedded slurs, regionally specific derogatory expressions, and identity-based dog-whistles that fall below the detection threshold of standard Indonesian-language hate speech systems (Hayaty et al., 2020; Khan et al., 2021).

2.4 Computational Analysis

Phase 4 applied three complementary computational models to the full 4,200-item corpus, using the 1,200 annotated instances as training and validation data in a 70:15:15 split. The models were selected on the principle of methodological complementarity: each contributes a distinct representational capability that, in combination, provides broader detection coverage particularly for the IHS-E category than any single model alone. Table 4 summarises the models and their characteristics.

IndoBERT, a BERT-based transformer model pre-trained on large-scale Indonesian language corpora, was fine-tuned on the annotated dataset to improve contextual sensitivity to Bahasa Indonesia and regional linguistic patterns (Amalia & Suyanto, 2024). Hyperparameter optimization was performed through grid search, with early stopping applied to prevent overfitting. The DeepHate architecture (Cao et al., 2020) was deployed as a comparative model, offering simultaneous representation across sentiment, topical, and semantic dimensions, making it particularly suited to detecting implicit hate speech where no single dimension alone is sufficient for reliable classification. The local language lexicon classifier served as a rule-based supplementary layer, capturing region-specific euphemisms and coded hate markers documented during Phase 3 lexicon development that fall outside the vocabulary distributions of both BERT and DeepHate training corpora (Al-Makhadmeh & Tolba, 2020; Alsafari et al., 2020).

© Italian e-Learning Association

Table 1 - Data Sources and Corpus Distribution by Platform and Province.

No	Data Source	Platform Type	Province Coverage	Posts / Comments
1	Official & Unofficial University Social Media Groups affiliated with Higher Education Institutions	Online LMS & Institutional Forums	Papua, Maluku, NTT, Sulawesi	1,540
2	Student Communication Channels Used for Academic Purposes	WhatsApp, Telegram	Papua, Maluku, NTT, Sulawesi	1,780
3	Comment Sections on Academic Video Content Uploaded by Universities and Lecturers	YouTube	Papua, Maluku, NTT, Sulawesi	880
	TOTAL	3 Platform Types	4 Provinces	4,200

Table 2 - Corpus Size Justification Against Established Benchmarks in Hate Speech Detection Research.

Criterion	Standard / Benchmark	This Study	Assessment	Reference
Minimum corpus for NLP model training	>= 1,000 annotated instances	1,200 annotated	Sufficient	Amalia & Suyanto (2024)
Total corpus for frequency analysis	>= 3,000 posts recommended	4,200 total	Exceeds benchmark	Nascimento et al. (2023)
Platform diversity	>= 2 platform types for cross-platform validity	3 platform types	Exceeds benchmark	Alkomah & Ma (2022)
Geographic coverage	Multi-regional for generalisability	4 provinces	Strong coverage	Alorainy et al. (2021)

Table 3 - Multi-Layer Annotation Scheme with Intersectional Axes (aligned with Introduction).

Category	Code	Description	Intersectional Axes Covered
Explicit Hate Speech	EHS	Direct, overt derogatory language targeting protected group characteristics	Ethnicity, religion, regional identity, gender
Implicit Hate Speech (Euphemistic)	IHS-E	Coded or euphemistic language conveying hate through indirection, cultural dog-whistles, or irony primary target category aligned with research title	Regional nicknames used pejoratively, coded religious stereotypes, ethnically inflected irony
Offensive but Non-Hate	OFN	Profane or uncivil language that does not target protected characteristics	General profanity, situational frustration, non-identity-based insults
Neutral / Academic	NEU	Normal academic discourse with no offensive or hateful content	Assignment discussions, peer feedback, course-related questions

Table 4 - Computational Models Employed for Hate Speech Detection and Their Characteristics.

Model	Architecture	Training Data	Key Strength
IndoBERT (Fine-tuned)	Transformer (BERT-based)	Indonesian corpus + annotated dataset	Deep contextual understanding of Bahasa Indonesia and regional language patterns
DeepHate	Multi-faceted CNN+LSTM	Multi-source social media data	Simultaneous sentiment, topical, and semantic representation suited to IHS-E detection
Local Lexicon Classifier	Rule-based + Word Embedding	Eastern Indonesia local language corpus (FGD-validated)	Captures region-specific euphemisms and coded hate terms invisible to generic models

© Italian e-Learning Association

Model performance was evaluated using precision, recall, and macro-averaged F1-score. Per-class performance on the IHS-E category was designated as the primary evaluation criterion, reflecting the centrality of implicit and euphemistic hate speech to the research problem and the disproportionate difficulty of this classification task relative to explicit hate speech categories (Alatawi et al., 2021; Ahmed & Lin, 2022).

2.5 Critical Discourse Analysis (CDA) Framework

Alongside the computational pipeline, Phase 5 applied a Critical Discourse Analysis (CDA) framework to a purposively selected sub-corpus of 320 texts classified as containing explicit or implicit hate speech by at least two of the three computational models. The CDA framework addressed three analytical dimensions: the dehumanization mechanisms through which hateful language constructs ethnic, religious, and regional outgroups (Abdalla et al., 2021; Etaywe & Zappavigna, 2023); the intertextual links between individual hate instances and broader extremist discourse circulating in the Indonesian digital public sphere (Norocel et al., 2022; Nikunen et al., 2021); and the escalatory progression from euphemistic formulations to explicitly extremist expressions (Titley, 2020; Baider, 2023).

The CDA analysis proceeded through iterative close reading and theoretical integration, ensuring that interpretive claims remained grounded in specific textual evidence. Particular attention was given to how Eastern Indonesian regional identity markers, vernacular religious references, and ethnically inflected stereotypes function as vehicles for coded hate expression that is legible to in-group members yet opaque to outside observers (Elfrida & Pasaribu, 2023; Tahir & Ramadhan, 2024). This focus directly operationalizes the concept of the "alphabet of hate" as a locally evolved, culturally embedded semiotic repertoire.

2.6 Triangulation, Validity, and Reliability

Phase 6 integrated the findings of the computational analysis (Phase 4) and the CDA (Phase 5) through systematic triangulation, following the methodological logic articulated by Baider (2023) and Nikunen et al. (2025), who argue that technical and humanistic approaches to online hate research are mutually indispensable. Points of convergence between the two analytical perspectives were treated as strongly grounded findings, while points of divergence were treated as productive sites for deeper investigation, particularly in cases where computational classifiers and discourse analysts reached different conclusions on implicit hate speech.

Validity was strengthened through multiple mechanisms. Internal validity was reinforced through method triangulation, multi-source data collection, and member checking with five academic informants from Eastern Indonesian institutions, who reviewed a representative sample of findings and assessed cultural accuracy and interpretive plausibility. External validity was maintained through the geographic and demographic diversity of the sample across four provinces and three platform types. Construct validity was ensured by grounding the annotation scheme and CDA categories in established theoretical frameworks responsive to the research questions. Reliability of the annotation instrument was confirmed through Cohen's Kappa coefficient, while computational model reliability was assessed through cross-validation procedures. All research protocols were approved by the institutional ethics committee and complied with AoIR digital research ethics standards (Charitidis et al., 2020).

3. Results

This section presents the empirical findings of the study on the dynamics of hate speech in digital academic spaces in Eastern Indonesia. The results are organized systematically, including the outcomes of corpus annotation, the level of inter-annotator reliability, the performance of computational models in detecting hate speech, regional variations in the prevalence of hate speech, the linguistic strategies that constitute the local "alphabet of hate", and the escalation patterns of discourse from euphemistic expressions toward more explicit and extreme forms as identified through Critical Discourse Analysis (CDA).

3.1 Corpus Annotation Outcomes

Analysis of the 4,200-item corpus yielded a total of 1,614 instances (38.4%) classified as containing some form of hate speech, comprising 641 instances of Explicit Hate Speech (EHS, 15.3%) and 973 instances of Implicit Hate Speech of the euphemistic type (IHS-E, 23.2%). The remaining 2,286 instances were classified as either Offensive Non-Hate (OFN, 672 instances, 16.0%) or Neutral/Academic (NEU, 1,914 instances, 45.6%). Table 5 and Figure 2 present the full distribution by platform and category.

A notable platform effect was observed. YouTube comment sections displayed the highest combined hate speech rate at 45.0%, followed by WhatsApp and Telegram channels at 40.0%, and Online LMS and Institutional Forums at 32.9%. WhatsApp and Telegram channels contained the highest absolute count of IHS-E instances (n=445), suggesting that the perceived privacy of closed messaging environments may lower inhibitions against euphemistic hate expression. YouTube comment sections, by contrast, exhibited the highest proportion of explicit hate speech relative to IHS-E, a pattern likely attributable to the more anonymous and adversarial discourse norms that characterize open video platform comment sections. These figures are further contextualized by YouTube's dominant role as a communication medium among Indonesian youth. Indonesia consistently ranks among the countries with the highest YouTube viewership globally, and university-age users dedicate substantially

© Italian e-Learning Association

more time to the platform than to any comparable digital service. This trend is especially pronounced in Eastern Indonesia, where limited broadband infrastructure has historically restricted access to bandwidth-heavy applications, and where YouTube's mobile-friendly design has made it a primary venue for youth social interaction. This saturation helps explain both the high volume of academic content uploaded by universities and lecturers, and the correspondingly elevated rate of hate speech accumulating in comment sections beneath it.

3.2 Inter-Annotator Reliability

Inter-annotator agreement across all three annotator pairs and all four categories yielded an overall mean Cohen's Kappa of 0.838, exceeding the minimum acceptance threshold of 0.75 established in the methodology.

Table 6 presents the per-category Kappa scores.

As anticipated by the research design, the IHS-E category produced the lowest Kappa scores across all annotator pairs (range: 0.75 to 0.77), reflecting the inherent interpretive difficulty of classifying euphemistic and coded hate expressions that lack explicit lexical markers. This finding provides direct empirical support for the core premise of the study: that implicit hate speech operating through cultural coding and linguistic indirection represents a classification challenge that cannot be resolved by surface-level linguistic analysis alone. The relatively lower but still substantial agreement on IHS-E is consistent with results reported in comparable annotation studies and confirms the appropriateness of the intersectional CDA approach as a complementary analytical layer.

Table 5 - Corpus Annotation Results by Platform and Category (n=4,200).

Category	Code	LMS & Forums (n=1,540)	WhatsApp Telegram (n=1,780)	YouTube (n=880)	TOTAL (n=4,200)
Neutral / Academic	NEU	788 (51.2%)	783 (44.0%)	343 (39.0%)	1,914 (45.6%)
Offensive Non-Hate	OFN	246 (16.0%)	285 (16.0%)	141 (16.0%)	672 (16.0%)
Implicit Hate Speech (Euphemistic)	IHS-E	308 (20.0%)	445 (25.0%)	220 (25.0%)	973 (23.2%)
Explicit Hate Speech	EHS	198 (12.9%)	267 (15.0%)	176 (20.0%)	641 (15.3%)
Combined Hate (IHS-E + EHS)	-	506 (32.9%)	712 (40.0%)	396 (45.0%)	1,614 (38.4%)

Table 6 - Inter-Annotator Agreement (Cohen's Kappa) by Category.

Category	Annotator Pair	Cohen's Kappa	Agreement Level
NEU (Neutral/Academic)	A1-A2 / A1-A3 / A2-A3	0.91 / 0.89 / 0.90	Almost Perfect
EHS (Explicit Hate Speech)	A1-A2 / A1-A3 / A2-A3	0.88 / 0.86 / 0.87	Almost Perfect
OFN (Offensive Non-Hate)	A1-A2 / A1-A3 / A2-A3	0.81 / 0.79 / 0.80	Substantial
IHS-E (Implicit / Euphemistic)	A1-A2 / A1-A3 / A2-A3	0.77 / 0.75 / 0.76	Substantial (lowest - confirms IHS-E difficulty)
Overall (Mean Kappa)	All pairs	0.838	Almost Perfect

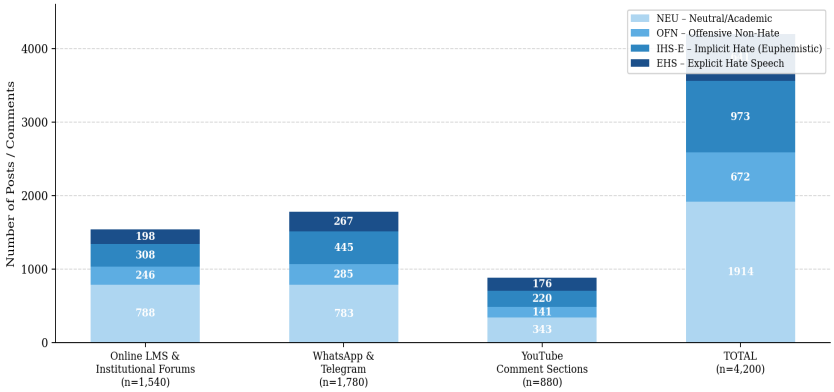


Figure 2 - Corpus Label Distribution by Platform and Category (n=4,200).

© Italian e-Learning Association

3.3 Computational Model Performance

Table 7 and Figure 3 present the F1-scores achieved by each of the three individual models and the combined pipeline across all four classification categories.

The combined pipeline achieved the highest macro-averaged F1-score of 0.859, outperforming all individual models. Across individual models, IndoBERT attained the highest macro-averaged F1 (0.830), followed by DeepHate (0.809) and the Local Lexicon Classifier (0.745). Critically, performance patterns on the IHS-E category diverged substantially from overall ranking: the Local Lexicon Classifier, while weakest on macro-average, achieved the second-highest IHS-E F1-score (0.758), surpassing IndoBERT (0.712) on this specific category. This finding confirms that the region-specific abusive lexicon developed in Phase 3 captures a distinct and complementary class of euphemistic hate signals that transformer-based models trained on general Indonesian corpora fail to detect. The combined pipeline's IHS-E F1 of 0.783 represents a meaningful improvement over any single model, validating the methodological complementarity approach adopted in this study.

3.4 Provincial Variation in Hate Speech Prevalence

Figure 4 presents the provincial distribution of IHS-E and EHS prevalence rates as a percentage of each province's sub-corpus. Papua exhibited the highest rates of both IHS-E (26.4%) and EHS (18.7%), followed by Maluku (IHS-E: 22.1%; EHS: 14.2%), NTT (IHS-E: 20.8%; EHS: 13.5%), and Sulawesi (IHS-E: 18.3%; EHS: 11.9%).

Across all four provinces, IHS-E consistently exceeded EHS in prevalence, with an average IHS-E-to-EHS ratio of approximately 1.5:1. This persistent pattern across geographically and ethnically diverse provinces constitutes a robust empirical finding: in Eastern Indonesian academic digital spaces, covert, euphemistic hate speech is systematically more prevalent than overt, explicit hate speech, a relationship that has significant implications for both automated detection strategies and institutional governance responses.

3.5 Linguistic Strategies Constituting the Local Alphabet of Hate

CDA analysis of the 320-item sub-corpus identified five primary linguistic strategies through which the IHS-E alphabet of hate operates in Eastern Indonesian academic digital spaces. Table 8 summarizes these strategies by frequency and primary provincial locus.

The most prevalent strategy, accounting for 32.1% of all IHS-E instances, was the pejorative deployment of ethnic euphemisms and regional nicknames. In the Papua sub-corpus, this strategy manifested primarily through the ironic or dismissive use of Papuan ethnic labels in academic contexts, functioning as coded markers of intellectual or civilizational inferiority without ever deploying explicit slurs. The second most frequent strategy, religious dog-whistling (25.4%), was concentrated in Maluku and NTT, reflecting the historically charged interreligious dynamics of these

Table 7 - Model Performance Summary: F1-Scores by Category.

Model	EHS F1	IHS-E F1	OFN F1	NEU F1	Macro Avg F1
IndoBERT (Fine-tuned)	0.884	0.712	0.801	0.921	0.830
DeepHate	0.861	0.688	0.779	0.908	0.809
Local Lexicon Classifier	0.743	0.758	0.612	0.867	0.745
Combined Pipeline	0.901	0.783	0.819	0.934	0.859

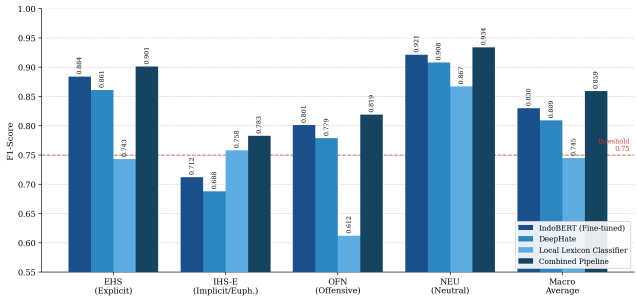


Figure 3 - F1-Score by Model and Category (IHS-E = primary evaluation criterion).

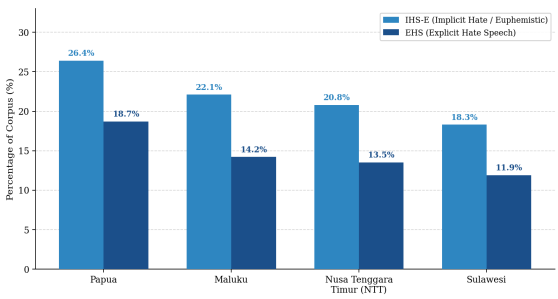


Figure 4 - Prevalence of IHS-E and EHS by Province (%).

© Italian e-Learning Association

provinces. Ironic praise (19.4%) appeared most prominently in Sulawesi, where its deployment as a contemptuous rhetorical device frequently targeting students from other islands was documented across multiple platform types. Dehumanizing metaphor (13.8%) and socioeconomic stereotyping (9.3%) rounded out the five-strategy taxonomy, each operating in distinct intersectional configurations across the provinces.

3.6 The Euphemism-to-Extremism Escalation Trajectory

Figure 5 maps the distribution of the full corpus across the five-stage escalation trajectory derived from the CDA analysis, and Figure 6 presents the intersectional axes of hate speech across provinces as identified in the CDA sub-corpus.

The escalation trajectory reveals a characteristic funnel pattern: the majority of posts remain at Stage 1 (Neutral/Ambiguous, n=1,914), with progressively smaller but still substantial populations advancing to Stage 2 (Euphemistic Coding, n=589), Stage 3 (Implicit Hostility, n=384), Stage 4 (Overt Denigration, n=288), and Stage 5 (Explicit Extremism, n=353). The slight uptick in volume between Stage 4 and Stage 5 suggests that once the threshold of overt denigration is crossed within a thread or channel, escalation to explicit extremism is facilitated by a form of discursive momentum, a finding with implications for early-warning intervention design.

The heatmap in Figure 6 reveals that ethnic identity is the dominant axis of hate speech in Papua (38.2%) and Sulawesi (31.6%), while religious reference dominates in Maluku (38.6%) and NTT (41.3%). Regional origin constitutes a secondary but consistently significant axis across all provinces, ranging from 22.3% in Maluku to 31.5% in Papua. Gender and socioeconomic status are present across all provinces but at notably lower levels, suggesting that while these axes are operative, they function primarily as secondary rather than primary vehicles for hate speech in the academic digital contexts studied.

4. Discussion

The findings of this study generate a rich set of insights that both confirm and extend existing scholarship on hate speech in digital spaces. This discussion is structured around five interrelated thematic concerns: the overall prevalence and severity of hate speech in Eastern Indonesian academic digital spaces; the diagnostic significance of IHS-E's dominance over EHS; the performance of computational detection systems in low-resource and culturally specific contexts; the intersectional structure of the local alphabet of hate; and the implications for institutional governance and future research.

Table 8 - Linguistic Strategies in IHS-E: Frequency and Provincial Distribution (n = 973).

Linguistic Strategy	Description	Frequency (n=973 IHS-E)	Primary Province
Ethnic Euphemism & Regional Nicknames	Pejorative use of regional labels and ethnic nicknames disguised as informal reference	312 (32.1%)	Papua
Religious Dog-Whistle & Coded Allusion	Indirect reference to religious affiliation as marker of distrust or inferiority	247 (25.4%)	Maluku, NTT
Ironic Praise & Sarcastic Commendation	Surface-level positive framing that communicates contempt through irony	189 (19.4%)	Sulawesi
Dehumanizing Metaphor	Non-literal comparisons reducing outgroup members to animals or objects	134 (13.8%)	Papua, Maluku
Socioeconomic Stereotyping	Coded class or development-level references tied to regional or ethnic identity	91 (9.3%)	All provinces

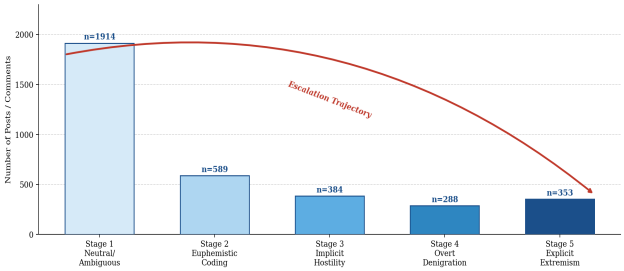


Figure 5 - Distribution of Posts Across the Five-Stage Euphemism-to-Extremism Escalation Trajectory.

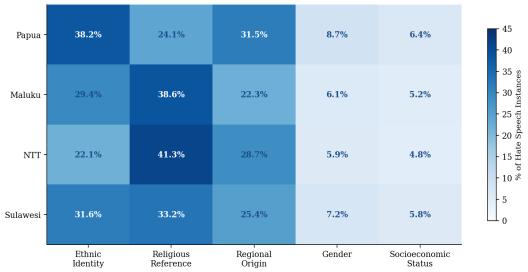


Figure 6 - Intersectional Axes of Hate Speech by Province (%) Based on CDA Sub-Corpus (n=320).

© Italian e-Learning Association

4.1 Prevalence, Severity, and the Academic Digital Space as a Site of Risk

The finding that 38.4% of the sampled corpus contained some form of hate speech represents a markedly high prevalence rate that demands serious institutional attention. While direct cross-study comparison is complicated by differences in corpus composition, platform type, and annotation methodology, this figure is broadly consistent with, and in several respects exceeds, prevalence rates reported in comparable studies of online hate in educational and youth-oriented contexts (Castellanos et al., 2023; Chekol et al., 2023). The fact that hate speech was documented across all three platform types and all four provinces rules out the possibility that these findings reflect isolated, platform-specific anomalies, and instead suggests a systemic pattern of digital hate embedded in the fabric of academic digital interaction in Eastern Indonesia.

This finding resonates with Schäfer et al. (2024), who demonstrate that online hate speech in educational environments is associated not only with psychological harm to its direct targets but also with deteriorating perceptions of social cohesion among all members of the digital community, including bystanders. Similarly, Castellanos et al. (2023) document the heightened vulnerability of adolescent and young adult populations in educational contexts to the affective and psychological impacts of hate speech. In the Eastern Indonesian context, these harms are likely to be amplified by the pre-existing conditions of ethno-regional inequality and geographic marginalization documented in the introduction, creating a compounding risk environment in which digital hate reinforces and extends offline patterns of exclusion (Titley, 2020; Nikunen, 2023).

The platform effect observed, specifically the higher combined hate speech rate on YouTube comment sections (45.0%) relative to LMS forums (32.9%), aligns with the broader literature on the role of platform architecture in shaping discourse norms. Rossini (2022) argues that incivility and intolerance in online political talk are partly a function of affordances that enable anonymity and reduce accountability. Casula et al. (2021) similarly document how platform moderation failures on Twitter allowed violent and hateful content to persist despite formal policy violations. The relative openness and anonymity of YouTube comment sections, compared to the institutionally affiliated LMS environments, appears to function as a structural facilitator of more extreme hate expressions, a finding with direct implications for platform-specific governance strategies (Baider, 2023). It is worth underscoring that YouTube's substantial presence in this dataset is not simply a product of how we sampled. It reflects, rather, a sociological reality: YouTube is the dominant digital platform for young Indonesians in terms of daily active engagement, and that dominance is felt with particular force across the provinces of Eastern Indonesia, where it serves as a primary conduit for information-seeking and peer interaction among university students alike (Statcounter, 2024). Recognizing this reality does two things analytically: it validates the decision to include YouTube as a data source in the first place, and it helps explain why hate speech in YouTube comment sections carries governance urgency that exceeds what its raw share of the corpus alone might suggest.

4.2 The Dominance of Implicit over Explicit Hate Speech: Implications for Detection

Perhaps the most theoretically significant finding of this study is the consistent dominance of IHS-E over EHS across all platforms and all provinces, with an average ratio of approximately 1.5:1. This pattern constitutes strong empirical evidence for the central argument of the research: that in Eastern Indonesian academic digital spaces, hate speech overwhelmingly operates through covert, euphemistic, and culturally coded linguistic strategies rather than through overt slurs and explicit threats. This finding directly challenges the implicit assumptions of many first-generation hate speech detection systems, which were calibrated primarily against lexically explicit content and consequently lack the sensitivity to detect the more pervasive implicit form (Alkomah & Ma, 2022; Alorainy et al., 2021).

The computational results reinforce this point. The underperformance of both IndoBERT (IHS-E F1: 0.712) and DeepHate (IHS-E F1: 0.688) on the IHS-E category, despite their strong performance on EHS (F1s of 0.884 and 0.861 respectively), demonstrates that even state-of-the-art transformer-based models trained on Indonesian data struggle with the detection of euphemistic hate speech rooted in regional cultural knowledge. This aligns with the findings of Nascimento et al. (2023), who identify content-based implicit hate speech as the primary frontier challenge in the field, and with Ahmed and Lin (2022), who argue that active learning approaches to hate speech detection must prioritize hard-to-classify instances precisely because they represent the dominant mode of online hate in many contexts. The superior performance of the Local Lexicon Classifier on IHS-E (F1: 0.758) relative to both deep learning models underscores the irreplaceable value of culturally grounded, community-validated lexical resources in low-resource language hate speech detection (Hayaty et al., 2020; Khan et al., 2021).

The combined pipeline's improvement on IHS-E (F1: 0.783) through the integration of all three models demonstrates that methodological complementarity is not merely a desirable feature but a practical necessity for hate speech detection in culturally and linguistically specific contexts. This finding supports the call by Al-Makhadmeh and Tolba (2020) for ensemble approaches to hate speech detection and reinforces the case for hybrid pipelines that combine the contextual strengths of transformer models with the cultural precision of domain-specific lexicons (Alsafari et al., 2020; Al-Hassan & Al-Dossari, 2022).

© Italian e-Learning Association

4.3 The Local Alphabet of Hate: CDA Findings and Their Theoretical Significance

The five-strategy taxonomy of IHS-E linguistic mechanisms identified through CDA constitutes the core contribution of this study to understanding regionally specific hate speech repertoires. The concept of the "alphabet of hate," operationalized here as a locally evolved, culturally embedded semiotic inventory, extends existing frameworks beyond the Western and high-resource language contexts in which most theoretical development has occurred (Norocel et al., 2022; Kovala et al., 2023). The prevalence of ethnic euphemism and regional nickname deployment as the dominant IHS-E strategy in Papua (32.1%) reflects the dehumanization mechanisms theorized by Abdalla et al. (2021) and Etaywe and Zappavigna (2023), wherein ostensibly neutral ethnic labels are deployed in contexts that communicate contempt. This logic operates through a locally specific semiotic register that remains invisible to general-purpose detection systems and legible only to those with intimate knowledge of the region's ethnocultural dynamics.

The dominance of religious dog-whistling in Maluku (38.6%) and NTT (41.3%) reflects historically rooted interreligious tensions, demonstrating how post-Reformasi crisis narratives are digitally recycled and normalized through covert discursive strategies in academic spaces (Titley, 2020; Castellanos et al., 2023). Equally significant is the ironic praise strategy identified in Sulawesi, wherein surface-level positive framing communicates contempt through socially shared understanding of its true meaning. As Nikunen et al. (2025) and Baider (2023) observe, this exploitation of communicative ambiguity allows speakers to simultaneously assert deniability and deliver a hate message, posing serious challenges for content moderation frameworks that require clear evidence of harmful intent. Together, these findings underscore the urgent need for context-sensitive, culturally informed approaches to hate speech detection and governance.

4.4 The Escalation Trajectory and Early Warning Implications

The five-stage escalation trajectory documented in Figure 5 provides the most direct empirical instantiation of the study's core theoretical argument: that the movement from euphemism to extremism in digital hate discourse is not random or discontinuous but follows a patterned, progressive logic that can, in principle, be detected and interrupted at early stages. The funnel pattern, with the largest population at Stage 1 and progressively smaller populations at Stages 2 through 4 before a slight uptick at Stage 5, suggests that most hateful discourse initiates at the level of ambiguous or euphemistic expression and that only a minority of instances escalate to fully explicit extremism.

This trajectory bears a structural resemblance to the patterns of online extremist radicalization documented by Rieger et al. (2021) in fringe communities on 8chan, 4chan, and Reddit, and to the affective escalation dynamics analyzed by Nikunen et al. (2021) in the context of far-right social media communities. The crucial difference is that the present findings document this trajectory not in explicitly extremist or political communities but in ostensibly academic digital spaces, where institutional oversight is typically less vigilant and where the normalization of early-stage euphemistic hate may proceed largely undetected (Hiaeshutter-Rice & Hawkins, 2022; Ridwanullah et al., 2025). The slight volumetric uptick between Stage 4 and Stage 5 is particularly noteworthy and warrants further investigation: it suggests that, once the threshold of overt denigration is crossed, the inhibition against fully explicit hate expression is substantially reduced, creating a critical intervention window at the Stage 3 to Stage 4 transition that policy and platform governance should prioritize.

4.5 Intersectionality, Geographic Variation, and Implications for Governance

The heatmap findings presented in Figure 6 reveal that the alphabet of hate in Eastern Indonesian academic digital spaces is not monolithic but differentially structured by province. Papua and Sulawesi are characterized by ethnic identity as the primary axis, while Maluku and NTT are characterized by religious reference, with regional origin functioning as a secondary but consistent axis across all provinces. This geographic differentiation carries direct governance implications, as effective institutional and platform responses cannot adopt a one-size-fits-all approach but must be calibrated to the specific intersectional configurations of hate operating in each provincial context (Norocel et al., 2022; Paasch-Colberg et al., 2021). Current detection approaches tend to treat target categories as discrete rather than intersectionally combined, yet the most harmful forms of online hate in the Indonesian context often operate precisely at the intersection of ethnic, religious, and regional identities, exploiting the compound vulnerability of targets belonging to multiple marginalized categories simultaneously (Elfrida & Pasaribu, 2023; Tahir & Ramadhan, 2024).

From a governance perspective, the high prevalence of hate speech across institutionally affiliated digital spaces, combined with its predominantly covert and culturally coded character, suggests that current institutional responses focused on explicit content violations are inadequate. Higher education institutions in Eastern Indonesia require governance frameworks that incorporate culturally competent moderation trained to recognize locally specific hate repertoires, early-warning systems capable of detecting escalation markers before full extremism emerges, and educational interventions that build digital literacy and counter-speech capacities among students, with particular attention to the intersectional dynamics that render certain student populations disproportionately

© Italian e-Learning Association

vulnerable (Schäfer et al., 2024; Castellanos et al., 2023; Coats, 2021).

5. Conclusions

This study empirically confirms that hate speech is a widespread, systemic, and predominantly covert phenomenon in Eastern Indonesian digital academic spaces. Of the 4,200-item corpus analysed, 38.4% contained hate speech, with implicit euphemistic hate speech (IHS-E) consistently exceeding explicit hate speech (EHS) at a ratio of 1.5:1 across all platforms and provinces. The five linguistic strategies identified through critical discourse analysis, namely ethnic euphemism and regional nickname deployment, covert religious dog-whistling, ironic praise, dehumanizing metaphor, and socioeconomic stereotyping, collectively constitute an alphabet of hate that is intersectionally structured and province-specific in its configuration. Computational evaluation demonstrates that a combined pipeline integrating IndoBERT, DeepHate, and a community-validated local language lexicon classifier achieves the strongest performance (macro F1: 0.859), validating the necessity of hybrid approaches that complement transformer model contextual strength with the cultural precision of locally grounded lexical resources in low-resource language contexts.

The documented five-stage escalation trajectory confirms that the movement from euphemism to extremism follows an identifiable and intervenable pattern, with the critical intervention window located at the Stage 3 to Stage 4 transition before discourse reaches full extremism. Taken together, this study makes a dual contribution: theoretically, by operationalizing the concept of the "alphabet of hate" as an empirically verifiable analytical category in the Eastern Indonesian context; and practically, by providing the methodological and empirical foundation for the development of culturally responsive detection systems, locally sensitive moderation frameworks, and digital academic governance policies that are more equitable and inclusive for all students regardless of their ethnic, religious, or regional background.

References

Abdalla, M., Ally, M., & Jabri-Markwell, R. (2021). Dehumanisation of 'outgroups' on Facebook and Twitter: Towards a framework for assessing online hate organisations and actors. *SN Social Sciences*, *1*(9), 1-28. https://doi.org/10.1007/s43545-021-00240-4

Ahmed, U., & Lin, J. C. W. (2022). Deep explainable hate speech active learning on social-media data. *IEEE Transactions on Computational Social Systems*, 1–11. https://doi.org/10.1109/TCSS.2022.3165136

Alatawi, H. S., Alhothali, A. M., & Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. *IEEE Access*, 9, 106363–106374. https://doi.org/10.1109/ACCESS.2021.3100435

Al-Hassan, Ar., & Al-Dossari, H. (2022). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*, 28(6), 1963–1974. https://doi.org/10.1007/s00530-020-00742-w

Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. *Information (Switzerland)*, 13(6), 273. https://doi.org/10.3390/info13060273

Al-Makhadmeh Z., Tolba A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*, 102(2), 501–522.

Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. (2021). Challenges of hate speech detection in social media. *SN Computer Science*, *2*(2), Article 98. https://doi.org/10.1007/s42979-021-00457-3

Alsafari S., Sadaoui S., Mouhoub M. (2020). Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, 19, 100096.

Amalia, Fadila & Suyanto, Yohanes. (2024). Offensive Language and Hate Speech Detection using BERT Model. IJCCS (Indonesian Journal of Computing and Cybernetics Systems). 18. 10.22146/ijccs.99841.

Baider, F. (2023). Accountability issues, online covert hate speech, and the challenge of content moderation. *Politics and Governance*, *11*(2), 249–260. https://doi.org/10.17645/pag.v11i2.6465

Cao R., Lee R. K. W., Hoang T. A. (2020). *Deephate: Hate speech detection via multi-faceted text representations* [Conference session]. In: 12th ACM Conference on Web Science, WebSci '20. ACM (p. 1120). https://doi.org/10.1145/3394231

Castellanos, M., Wettstein, A., Wachs, S., Kansok-Dusche, J., Ballaschk, C., Krause, N., & Bilz, L. (2023). Hate speech in adolescents: A binational study on prevalence and demographic differences. *Frontiers in Education*, *8*, 1-14. https://doi.org/10.3389/feduc.2023.1076249

© Italian e-Learning Association

Casula P., Anupam A., Parvin N. (2021). *We found no violation!: Twitter's violent threats policy and toxicity in online discourse* [Conference session]. C&T '21: Proceedings of the 10th International Conference on Communities & Technologies Wicked Problems in the Age of Tech, C&T '21. ACM (p. 151159). https://doi.org/10.1145/3461564.3461589.

Charitidis P., Doropoulos S., Vologiannidis S., Papastergiou I., Karakeva S. (2020). Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, 17, 100071.

Chekol, M. A., Moges, M. A., & Nigatu, B. A. (2023). Social media hate speech in the walk of Ethiopian political reform: Analysis of hate speech prevalence, severity, and natures. *Information Communication and Society*, *26*(1), 218-237. https://doi.org/10.1080/1369118X.2021.1942955

Coats, S. (2021). 'Bad language' in the Nordics: Profanity and gender in a social media corpus. *Acta Linguistica Hafniensia*, *53*(1), 22-57. https://doi.org/10.1080/03740463.2021.1871218

del Arco F. M. P., Molina-Gonzlez M. D., Urea-Lpez L. A., Martn-Valdivia M. T. (2021). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166, 114120. https://doi.org/10.1016/j.eswa.2020.114120

Elfrida, R., & Pasaribu, A. N. (2023). Hate speech on social media: A case study of blasphemy in Indonesian context. *English Review: Journal of English Education*, *11*(2), 433-440. https://doi.org/10.25134/erjee.v11i2.7909

Etaywe, A., & Zappavigna, M. (2023). The role of social affiliation in incitement: A social semiotic approach to far-right terrorists' incitement to violence. *Language in Society,* 1-26. https://doi.org/10.1017/s0047404523000404

Hayaty M., Adi S., Hartanto A. D. (2020). Lexicon-based indonesian local language abusive words dictionary to detect hate speech in social media. *Journal of Information Systems Engineering and Business Intelligence*, 6(1), 9–17.

Hiaeshutter-Rice, D., & Hawkins, I. (2022). The language of extremism on social media: An examination of posts, comments, and themes on Reddit. *Frontiers in Political Science*, *4*, Article 805008. https://doi.org/10.3389/fpos.2022.805008

Khan M. M., Shahzad K., Malik M. K. (2021). Hate speech detection in roman urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(1), 1–19. https://doi.org/10.1145/3414524

Kovala U., Saresma T., Vaarakallio T. (2023). The Finns Party: From catch-all populism to radical anti-immigration discourse. In Deringer L., Ströbel L. (Eds.), *International discourses of authoritarian populism* (pp. 129–145). Routledge.

Nascimento, F. R. S., Cavalcanti, G. D. C., & Da Costa-Abreu, M. (2023). Exploring automatic hate speech detection on social media: A focus on content-based analysis. *SAGE Open*, *13*(2), 1–24. https://doi.org/10.1177/21582440231181311

Nikunen K. (2023). Affective temporalities of digital hate cultures. In Lünenborg M., Röttger-Rössler B. (Eds.), *Affective formation of publics* (pp. 173–190). Routledge.

Nikunen K., Hokka J., Nelimarkka M. (2021). Affective practice of soldiering: How sharing images is used to spread extremist and racist ethos on Soldiers of Odin Facebook site. *Television & New Media*, 22(2), 166–185.

Nikunen, K., Haara, P., Kosonen, H., Knuutila, A., Pöyhtäri, R., & Saresma, T. (2025). Contested meanings of hate speech and the post-truth condition on digital platforms. *Social Media + Society*, *11*(2), 1–13. https://doi.org/10.1177/20563051251341794

Nikunen, K., Haara, P., Kosonen, H., Knuutila, A., Pöyhtäri, R., & Saresma, T. (2025). Contested meanings of hate speech and the post-truth condition on digital platforms. *Social Media + Society*, *11*(2), 1–13. https://doi.org/10.1177/20563051251341794

Norocel C., Saresma T., Lähdesmäki T., Ruotsalainen M. (2022). Performing "us" and 'other': Intersectional analyses of right-wing populist media. *European Journal of Cultural Studies*, 25(3), 897–915. https://doi.org/10.1177/1367549420980002

Paasch-Colberg S., Strippel C., Trebbe J., Emmer M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, 9(1), 171–180. https://doi.org/10.17645/mac.v9i1.3399

Pettersson K., Norocel O. C. (2024). Vernacular constructions of the relationship between freedom of speech and (potential) hate

© Italian e-Learning Association

speech: The case of Finland. *European Journal of Social Psychology*, 54(3), 701–714. https://doi.org/10.1002/ejsp.3045

Ridwanullah, A. O., Sule, S. Y., Usman, B., & Abdulsalam, L. U. (2025). Politicization of hate and weaponization of Twitter/X in a polarized digital space. *Journal of Asian and African Studies*, *60*(2), 1–18. https://doi.org/10.1177/00219096241230500

Rieger D., Kümpel A. S., Wich M., Kiening T., Groh G. (2021). Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media + Society*, 7(4), 1–14. https://doi.org/10.1177/20563051211052906

Rossini P. (2022). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3), 399–425. https://doi.org/10.1177/0093650220921314

Schäfer S., Sülflow M., Reiners L. (2022). Hate speech as an indicator for the state of the Society: Effects of hateful user comments on perceived social dynamics. *Journal of Media Psychology*, 34, 3–15. https://doi.org/10.1027/1864-1105/a000294

Schäfer, S., Rebasso, I., Boyer, M. M., & Planitzer, A. M. (2024). Can we counteract hate? Effects of online hate speech and counter speech on the perception of social groups. *Communication Research*, *51*(4), 412–436. https://doi.org/10.1177/00936502231201091

Tahir, I., & Ramadhan, M. G. F. (2024). Hate speech on social media: Indonesian netizens' hate comments of presidential talk shows on YouTube. *LLT Journal: A Journal on Language and Language Teaching*, *27*(1), 1–20. https://doi.org/10.24071/llt.v27i1.8180

Titley G. (2020). Racism and media, in a moment of crisis and opportunity. *Ethnic and Racial Studies*, 43(13), 2386–2394. https://doi.org/10.1080/01419870.2020.1789189

© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026)

Evolving teacher interests in innovative learning environments: impacts of exposure and implications for professional development

Amelia R. Granda-Pinan[a,1], Óscar R. Lozano[b]

[a]*University of Valencia, Dept. of Comparative Education and History of Education – Valencia (Spain)*

[b]*University of Valencia, Dept. of Didactics of Experimental Sciences – Valencia (Spain)*

(submitted: 08/10/2024; accepted: 10/03/2026; published: 5/5/2026)

Abstract

Teachers' practice is increasingly oriented towards designing and working in Innovative Learning Environments (ILE). In-service teachers' professional development is being designed coherently, sometimes allowing them to experience learning situations in such environments. Two surveys were designed, and validated, to answer to two main questions concerning teachers' interests and how they evolve once they know ILEs and have this training experience. To get all the information needed, one survey was filled before taking part in the training experience, and the other one after it. 255 answers were received and analyzed to extract several conclusions. Results indicate that teachers, after the training, became increasingly open to embracing more active, participatory methods, integrating more innovative, immersive and interactive technology. They also showed greater interest in how to zone and plan meaningful learning experiences in such environments. Furthermore, the main topics teachers wish to receive further training on relate to designing appropriate learning experiences and selecting suitable methodologies. These findings suggest that engaging teachers in a training experience within an ILE serves as a catalyst for professional development in these topics. The study has implications for both educational administrations and school leaders promoting ILEs.

KEYWORDS: Learning Environment, Educational Innovation, Teacher Training, Interest, Educational Strategies.

DOI
https://doi.org/10.20368/1971-8829/1136075

CITE AS
Granda-Pinan, A.R., & Lozano, O.R. (2026). Evolving teacher interests in innovative learning environments: impacts of exposure and implications for professional development. *Journal of e-Learning and Knowledge Society*, *22*(1). https://doi.org/10.20368/1971-8829/1136075

1. Introduction

Teachers' motivation plays a crucial role in processes of educational transformation. Their willingness to participate in professional development after their daily teaching responsibilities, to diversify their teaching methods and resources and to overcome the fear of failure largely depends on their intrinsic motivation. Educational administrations can support this transformation through different measures, such as establishing supportive policies, providing additional resources and, importantly, offering high-quality professional development that guides teachers through the steps required to implement successful changes in their practice.

In this context, an important question arises: what constitutes high-quality teacher training? What factors effectively influence teachers' motivation to further develop their professional practice? Effective professional development depends on several key elements. First, it requires the guidance of experienced trainers. Second, it should incorporate innovative content, maybe including resources that can be adapted to different educational contexts. Finally, it should take place in an stimulating environment that encourages reflection and pedagogical change.

This study is focused on the latter element: the environment, posing the following question: how does the environment affect teachers' motivation or interests to learn? Does the environment enhance or diminish teachers' motivation towards exploring certain topics in depth? Despite extensive research on teacher education (e.g., Revista Interuniversitaria de Formación del Profesorado; Revista Electrónica Interuniversitaria de Formación del Profesorado), the impact of learning environments on teachers' interests remains

[1] corresponding author - email: amelia.granda@uv.es

© Italian e-Learning Association

underexplored. This omission is regrettable given the significant amount of money that government and public institutions invest in professional development, and more recently in creating specific labs to foster specific learning, namely European Schoolnet's Future Classroom Lab. From a theoretical standpoint, whether having a training experience in such a cared environment positively affects teachers' motivation or not will give clues about how this will work. From a practical standpoint, it supports or not the creation or use of such environments to enhance and foster the desired educational transformation.

Addressing these theoretical and practical concerns, this study aimed to analyze whether there were differences in teachers' training interests after having experienced a two-hour session, carefully designed in terms of the methodology used, the resources facilitated, the work developed and the space in which it takes place (for further information see Annex I). The session was developed in a carefully designed space divided in several areas, each of them focused on skills teachers should help students develop, using activities focused on the development on these skills (Annex II).

To be able to know any possible variation, teachers underwent a test before starting the experience and once it was finished. We focus on teachers' interests on topics related to creating an Innovative Learning Environment in their school, as this is one of the main training programs developed lately by different national and international institutions.

We contribute to professional development literature in three ways. First, when gathering teachers' interests on training, we are able to highlight which topics are of great interest and which ones are not. This helps authorities and schools when designing training programs and academics when identifying topics to explore further. Second, we identify the impact that an experience in a training lab, with specific characteristics, has on teachers' motivations. This helps to support the creation of such environments aiming for professional development experiences and to investigate the impact that such realities have on teachers' interests and therefore in their future. Third, we advance Innovative Learning Environment research by assessing these two issues, which is of great interest to researchers and practitioners worldwide.

We decided to focus on what is known as Innovative Learning Environments (ILEs) as we noticed an international tendency on training teachers in this topic and an increasing interest among teachers and schools to implement them. Although there is not a consensus concerning its definition, we have chosen the one given by the group of professors from University of Melbourne who have been studying them for a decade. Professors Blannin, Mahat, Cleveland, Morris and Imms (2020) define these environments as highly flexible spaces equipped with intentional furniture and ubiquitous technology, used in innovative ways to facilitate student-centered learning experiences. Research evidence indicates that such environments promote student-centered pedagogies (Byers et al., 2018b; Cleveland, 2016; Granda-Pinan & Rojo-Bofill, 2024; Jorion et al., 2016), as teachers tend to involve students in more active, collaborative and creative learning activities (Byers et al., 2018a).

To train teachers on this topic, several institutions have created labs in which they built a space as an example of an ILE. Some examples are the Future Classroom Lab in Brussels (European Schoolnet, 2024) or Aula del Futuro in Spain (INTEF, 2024). These training labs have been designed to promote reflection but not to serve as fixed models to replicate in schools. Research states that teachers in contact with one of such spaces, both in a lab or in a school, feel compelled to modify their teaching approaches (Byers et al., 2014), feel more motivated to work in a more innovative way. This leads us to pose the research questions: 1) How does teachers' interest evolve when they get in contact for the first time with this training environment? 2) What are the topics that become more interesting for teachers once they delve into Innovative Learning Environments?

When designing training related to this topic, three main aspects are considered. First, the most important topic is related to student-centered teaching methodologies. There is not one approach or methodology more important than others, but all of them can fit in it, depending on teachers' and students' needs and interests. Some of these methodologies are identified by Blázquez (2016): Service Learning, Project-based learning, Problem-based learning or Cooperative Learning.

Second, training is designed concerning digital competence and use of digital devices. In such learning environments, digital space extends and complements the physical space (Gonzalez-Mohino et al., 2023; Granda-Pinan et al., 2024; Rivera-Vargas et al., 2024). Therefore, training on digital use is basic. Once more, there is no basic technology to be implemented in it, although some international institutions (Educause, 2023; ODITE, 2022) have identified some which are or will be essential in education, such as Robots, Computational thinking, Programming, Virtual Reality, Artificial Intelligence or 3D printing.

Third, literature emphasizes the importance of training teachers in how to exploit the space itself (French et al., 2022). The space that it is posed in doesn't have some fixed physical characteristics, but some essential definitory ones. Innovative learning spaces are designed under the values and principles of flexibility, openness, interconnectivity and activity (García-Mongue et al., 2023; King et al., 2015; Mahat et al., 2018; Radcliffe, 2009), fostering interaction among students, teachers, knowledge and resources (Mahat et al., 2018). From this perspective and in line with studies such as the one developed by Nelson et al. (2023), we

© Italian e-Learning Association

identified as possible training themes the impact of aspects such as colors, furniture or sound in learning, or the design and use of different learning zones.

These three big topics have been considered to develop the tool to gather teachers' interests in training, but we decided there was a lack of didactics. Teachers not only need to know the existence, characteristics and operation of these resources, but also require knowledge about how to integrate them into the curriculum and, finally, how to use them in their teaching and learning process (Suárez et al., 2010). It is then suggested that there may be an interest in understanding how to design learning situations, how to temporize them, how to tutor and mediate in them or how to apply evaluation methods appropriate for this pedagogy. Therefore, this is the fourth aspect considered in our study.

This study is part of a broader line of research on teachers' professional development in Innovative Learning Environments (ILEs). Some contextual and preliminary aspects of this research were previously presented by the authors in a related publication (Lozano et al., 2024). However, the present paper extends that work by specifically analysing how teachers' training interests evolve after participating in a training experience within an ILE, based on pre- and post-survey data.

2. Materials and Methods

This study adopted a quantitative, non-experimental, descriptive, comparative, and correlational research design, using a questionnaire as the main data collection instrument (McMillan & Schumacher, 2001). This approach was considered appropriate because the research was conducted in a natural, real-world context (Losada López & López-Feal Ramil, 2003).

2.1. Process of development and validation of the instrument

Two questionnaires were specifically designed for this study and administered before and after the training session, as no existing instruments met the research objectives. Participants were informed of the study's purpose and provided informed consent.

The content validation process followed the specifications of Lawshe (1975), with the modifications introduced by Tristán-López (2008). Thirteen experts in the field (secondary school teachers, university teachers in Education and Educational Sciences, inspectors, and professional development advisors) evaluated each item in terms of clarity, relevance and importance. Both questionnaires were structured in four dimensions, including multiple-choice items or Likert-scale items (1 - no interest, 5 - A lot of interest). These dimensions include: A) Demographic variables: ten items (four included in the post-questionnaire) were used to collect information of the participants' personal profile and their prior experience with ILEs; B) Methodologies and strategies, where respondents chose how interested they are on twelve different approaches; C) Digital Technology, gathering interest on ten items; D) Educative Spaces, made of three items, and E) Implementation on ILEs, in which respondents reflect on their interest on seven different aspects related to didactics.

2.2. Sample

The sample comprised 225 teachers from different educational stages (see Table 1). A non-probabilistic convenience sampling strategy was employed. The questionnaire was administered in paper format to teachers who requested a visit to the Aula del Futuro, a teacher training laboratory located in Valencia (Spain), between April and July 2023. These visits were advertised on the official professional development institution's website and were intended for teaching staff from non-university educational institutions. All active teaching staff were eligible to complete the survey under the sole condition that they had requested a visit to the mentioned Aula del Futuro. The randomness was, therefore, semi-restricted to the teaching staff requesting the activity (full visit).

Table 1 - Distribution of respondents' background characteristics (N=225).

Characteristics	Frequency (n)	Percentage (%)
Gender		
• Female	177	78.7
• Male	43	19.1
• Lost	5	2.2
Stage		
• Infant and Primary Education	166	73.80
• Secondary Education	51	22.7
• Lost	8	3.6
Age		
• 20-30	26	11.6
• 31-40	83	36.9
• 41-50	70	31.1
• 51-60	44	19.6
• Lost	2	0.9
Years of teaching experience		
• 0-10	91	40.4
• 11-20	77	34.2
• More than 20	57	25.3

Only participants with no prior contact with innovative learning spaces were included. The visit was guided by

© Italian e-Learning Association

specialized staff from the professional development network under the jurisdiction of the Generalitat Valenciana, always following the same structure, prepared beforehand by a group of experts on ILEs with the aim of providing teachers with a rich and active experience on what it means to learn in such scenario (Annex I).

2.3. Data analysis

The statistical package IBM SPSS 28 was used to analyse the obtained data. Both questions are answered describing the data found. After verifying the assumptions of normality and homogeneity of variances, several non-parametric tests were applied to the independent variables, including the Mann-Whitney U test and the Kruskal-Wallis H test, as well as Pearson's bivariate correlations. These analyses made it possible to examine the relationships between general variables in Dimension A (educational stage, age, etc.) and the research variables included in Dimensions B, C, D and E.

3. Results

We have organized the results found according to the two research questions posed in the introduction.

3.1 Research question 1: How does teachers' interest evolve when they get in contact for the first time with this training environment?

The answer to this question depends on the dimension studied.

Dimension B: Methodologies and strategies

Table 2 compiles the number, frequencies (in parentheses) and mean of responses of teachers' interests in the methodologies and strategies consulted, pre and post training experience.

Answers given pre and post training experience were compared using Wilcoxon test. Results are gathered in Table 3.

Dimension C: Digital technology

The second dimension gathered information about 10 digital tools or strategies that can be implemented in schools. Table 4 shows the results, pre and post training experience, for each item.

We can observe that teachers' interest increases in some of these aspects after having the learning experience in the lab, and others decrease (Table 4). However, we only found significance through the Wilcoxon test for the variation found on items 6, 7 and 8.

Dimension D: Educative Spaces

Three aspects are studied in this dimension, encountering the following frequencies (table 6).

Although we can observe that there is a tendency of an increase of interest on the three aspects, Wilcoxon test (table 7) supports that there are differences statistically speaking on dimension three.

Dimension E: Implementation on ILEs

The last dimension studied focused on didactics. Table 8 shows the interest expressed by teachers pre and post visit.

In Table 8 we observe that, in general, teachers feel more motivated towards learning these aspects after having a learning experience in an Innovative Learning Environment. The only aspect towards which interest decreased was the one related to scheduling a learning process (c3), although there are no differences statistically significant, so it could have been a matter of chance. There were differences for items 2, 4 and 5, which means that teachers' interest in learning how to design and program learning experiences in ILEs, how to manage coexistence and participation and how to select and organize the appropriate methodology or strategy to be developed in an ILE to achieve learning increases.

3.2 Research question 2: What are the topics that become more interesting for teachers once they delve into Innovative Learning Environments?

To answer this question, we have used the information found on tables 2, 4, 6 and 8 to create table 10, which compiles the means, number of positive answers and percentage of them for every item after the visit. With it, we can describe that teachers feel more interested in Dimension E, being the items with higher means the ones referring to "Design and programming of learning experiences in environments other than the traditional classroom (e2)" (x=3,68), "Selection and organization of the appropriate methodology or strategy to promote learning (e5)" (x=3,66) and "Design, programming, and creation of projects or experiences that promote learning (e1)" (x=3,64).

Focusing on Dimension B, teachers are more interested in "Cooperative learning (b4)" (x=3,62) and "Project-based learning (b1)" (x=3,61). In fact, this last one is the one who got the highest percentage of positive answers.

Teachers expressed having a high interest in Dimension D, being the last item the most interesting, the one referring to "Zoning of useful learning environment (d3)" (x=3,55).

Last, we can observe that interest on Dimension C is lower than the other ones. The most interesting item for teachers in this dimension has been "Virtual reality (c6)" with a mean x=3,28.

© Italian e-Learning Association

Table 2 - Interest in Dimension B items (Frequencies), pre and post training experience (N=225).

Item		SCALE 1-4 (1=No interest, 4=A lot of interest, NC = no answer)					Mean
		1	2	3	4	NC	
b1 Project-based learning	Pre	4 (1,8)	17 (7,6)	79 (35,1)	112 (49,8)	13 (5,8)	3,41
	Post	2 (0,9)	7 (3,1)	54 (24,0)	126 (56,0)	36 (16,0)	3,61
b2 Problem-based learning	Pre	5 (2,2)	37 (16,4)	67 (29,8)	78 (34,7)	38 (16,9)	3,17
	Post	4 (1,8)	19 (8,4)	74 (32,9)	85 (37 8)	43 (19,1)	3,32
b3 Inquiry-based learning	Pre	5 (2,2)	30 (13,3)	77 (34,2)	68 (30,2)	45 (20,0)	3,16
	Post	4 (2,8)	17 (7,6)	74 (32,9)	85 (37,8)	45 (20,0)	3,33
b4 Cooperative learning	Pre	4 (1,8)	13 (5,8)	57 (25,3)	133 (59,1)	18 (8,0)	3,54
	Post	2 (0,9)	7 (3,1)	51 (22,7)	128 (56,9)	37 (16,4)	3,62
b5 Flipped classroom	Pre	12 (5,3)	39 (17,3)	59 (26,2)	70 (31,1)	45 (20,0)	3,04
	Post	10 (4,4)	31 (13,8)	67 (29,8)	71 (31,6)	46 (20,4)	3,11
b6 Design Thinking	Pre	10 (4,4)	35 (15,6)	58 (25,8)	70 (31,1)	52 (23,1)	3,09
	Post	11 (4,9)	33 (14,7)	68 (30,2)	69 (30,7)	44 (19,6)	3,08
b7 Visual Thinking	Pre	10 (4,4)	33 (14,7)	56 (24,9)	77 (34,2)	49 (21,8)	3,14
	Post	9 (4,0)	29 (12,9)	71 (31,6)	72 (32,0)	44 (19,6)	3,14
b8 Personalised learning	Pre	4 (1,8)	21 (9,3)	51 (22,7)	99 (44,0)	50 (22,2)	3,40
	Post	3 (1,3)	25 (11,1)	65 (28,9)	87 (38,7)	45 (20,0)	3,31
b9 Service learning	Pre	9 (4,0)	32 (14,2)	50 (22,2)	83 (36,9)	51 (22,7)	3,19
	Post	7 (3,1)	30 (13,3)	60 (26,7)	81 (36,0)	47 (20,9)	3,21
b10 Gamification	Pre	5 (2,2)	26 (11,6)	57 (25,3)	105 (46,7)	32 (14,2)	3,36
	Post	5 (2,2)	16 (7,1)	59 (26,2)	102 (45 3)	43 (19,1)	3,42
b11 - Corners or ambiences learning	Pre	4 (1,8)	20 (8,9)	56 (24,9)	110 (48,9)	35 (15,6)	3,43
	Post	2 (0,9)	18 (8,0)	53 (23,6)	111 (49,3)	41 (18,2)	3,48
b12 - Universal Design of Learning	Pre	8 (3,6)	14 (6,2)	51 (22,7)	116 (51,6)	36 (16,0)	3,46
	Post	2 (0,9)	16 (7,1)	53 (23,6)	111 (49,3)	43 (19,1)	3,50

Table 3 - Wilcoxon test comparing answers given pre and post training experience.

Items	N	Neg Ranges	Pos Ranges	Draws	Z	p
b1	180	8	42	130	-4,588	<,001
b2	156	18	38	100	-2,743	,006
b3	148	12	36	100	-3,617	<,001
b4	174	15	30	129	-2,383	,017
b5	150	17	36	97	-2,143	,032
b6	145	29	35	81	-,616	,538
b7	147	30	33	84	-,314	,753
b8	145	30	23	92	-,922	,356
b9	146	28	33	85	-,980	,327
b10	164	21	30	113	-1,079	,280
b11	158	16	31	111	-1,711	,087
b12	161	17	28	116	-1,736	,083

© Italian e-Learning Association

Table 4 - Interest in Dimension C items (Frequencies), pre and post training experience (N=225).

Item		SCALE 1-4 (1=No interest, 4=A lot of interest, NC = no answer)					Mean
		1	2	3	4	NC	
c1 Robotics	Pre	13 (5,8)	46 (20,4)	55 (24,4)	64 (28,4)	47 (20,9)	2,96
	Post	15 (6,7)	47 (20,9)	51 (22,7)	66 (29,3)	46 (20,4)	2,94
c2 Programming	Pre	9 (4,0)	52 (23,1)	57 (25,3)	70 (31,1)	37 (16,4)	3,00
	Post	16 (7,1)	42 (18,7)	58 (25,8)	65 (28,9)	44 (19,6)	2,95
c3 Computational thinking	Pre	11 (4,9)	59 (26,2)	56 (24,9)	52 (23,1)	47 (20,9)	2,84
	Post	16 (7,1)	52 (23,1)	61 (27,1)	50 (22,2)	46 (20,4)	2,81
c4 Audiovisual technology	Pre	8 (3,6)	23 (10,2)	82 (36,4)	77 (34,2)	35 (15,6)	3,20
	Post	7 (3,1)	21 (9,3)	77 (34,2)	79 (35,1)	41 (18,2)	3,24
c5 - Document/ Presentation/Infographics	Pre	13 (5,8)	28 (12,4)	67 (29,8)	79 (35,1)	38 (16,9)	3,13
	Post	7 (3,1)	28 (12,8)	59 (26,2)	85 (37,8)	46 (20,4)	3,24
c6 Virtual Reality	Pre	7 (3,1)	37 (16,4)	68 (30,2)	69 (30,7)	44 (19,6)	3,10
	Post	7 (3,1)	27 (12,0)	57 (25,3)	92 (40,9)	42 (18,7)	3,28
c7 Augmented Reality	Pre	10 (4,4)	39 (17,3)	61 (27,1)	73 (32,4)	42 (18,7)	3,08
	Post	9 (4,0)	31 (13,8)	56 (24,9)	83 (36,9)	46 (20,4)	3,19
c8 3D Printing	Pre	17 (87,6)	44 (19,6)	55 (24,4)	62 (27,6)	47 (20,9)	2,91
	Post	14 (6,2)	35 (15,6)	62 (27,6)	69 (30,7)	45 (20,0)	3,03
c9 Scanning	Pre	25 (11,1)	48 (21,3)	53 (23,6)	47 (20,9)	52 (23,1)	2,71
	Post	15 (6,7)	51 (22,7)	65 (28,9)	43 (19,1)	51 (22,7)	2,78
c10 Artificial Intelligence	Pre	12 (5,3)	34 (15,1)	51 (22,7)	76 (33,8)	52 (23,1)	3,10
	Post	11 (4,9)	31 (19,8)	53 (23,6)	74 (32,9)	56 (24,9)	3,12

Table 5 - Wilcoxon test comparing answers given pre and post training experience.

Items	N	Neg Ranges	Pos Ranges	Draws	Z	p
c1	179	28	25	96	-,555	,579
c2	181	28	32	97	-,157	,875
c3	179	32	34	82	-,317	,751
c4	184	30	38	90	-,316	,752
c5	179	29	40	86	-1,007	,314
c6	183	23	45	84	-2,997	,003
c7	179	20	40	93	-2,192	,028
c8	180	23	39	87	-2,011	,044
c9	174	30	36	77	-,587	,557
c10	169	24	36	84	-1,637	,102

© Italian e-Learning Association

Table 6 - Interest in Dimension D items (Frequencies), pre and post training experience.

Item		SCALE 1-4 (1=No interest, 4=A lot of interest, NC = no answer)					Mean
		1	2	3	4	NC	
d1 - Characteristics of learning spaces (colours, light, sound…)	Pre	4 (1,8)	22 (9,8)	65 (28,9)	104 (46,2)	30 (13,3)	3,38
	Post	1 (0,4)	17 (7,6)	62 (27,6)	108 (48,0)	37 (16,4)	3,47
d2 - Furniture in useful learning spaces	Pre	4 (1,8)	15 (6,7)	62 (27,6)	110 (48,9)	34 (15,1)	3,46
	Post	1(0,4)	16 (7,1)	58 (25,8)	110 (48,9)	40 (17,8)	3,50
d3 - Zoning of useful learning environment	Pre	3 (1,3)	22 (9,8)	52 (23,1)	114 (50,7)	34 (15,1)	3,45
	Post	1 (0,4)	14 (6,2)	52 (23,1)	119 (52,9)	39 (17,3)	3,55

Table 7 - Wilcoxon test comparing answers given pre and post training experience.

Items	N	Neg Ranges	Pos Ranges	Draws	Z	p
d1	188	20	35	106	-1,327	,184
d2	185	24	38	93	-1,322	,186
d3	186	22	40	94	-2,107	,035

Table 8 - Interest in Dimension D items (Frequencies), pre and post training experience.

Item		SCALE 1-4 (1=No interest, 4=A lot of interest, NC = no answer)					Mean
		1	2	3	4	NC	
e1 - Design, programming, and creation of projects or experiences that promote learning	Pre	1 (0,4)	8 (3,6)	59 (26,2)	133 (59,1)	24 (10,7)	3,61
	Post	1 (0,4)	9 (4,0)	47 (20,9)	131 (58,2)	37 (16,4)	3,64
e2 - Design and programming of learning experiences in environments other than the traditional classroom	Pre	2 (0,9)	8 (3,6)	61 (27,1)	125 (55,6)	29 (12,9)	3,58
	Post	0 (0,0)	9 (4,0)	42 (18,7)	137 (60,9)	37 (16,4)	3,68
e3 - Scheduling of a learning process	Pre	2 (0,9)	22 (9,8)	67 (29,8)	106 (47,1)	28 (12,4)	3,41
	Post	3 (1,3)	22 (9,8)	60 (26,7)	100 (44,4)	40 (17,8)	3,39
e4 - Management of coexistence and participation in innovative learning environments	Pre	1 (0,4)	16 (7,1)	65 (28,9)	114 (50,7)	29 (12,9)	3,49
	Post	2 (0,9)	11 (4,9)	51 (22,7)	124 (55,1)	37 (16,4)	3,58
e5 - Selection and organization of the appropriate methodology or strategy to promote learning	Pre	1 (0,4)	12 (5,3)	64 (28,4)	122 (54,2)	26 (11,6)	3,54
	Post	0 (0,0)	8 (3,6)	47 (20,9)	130 (57,8)	40 (17,8)	3,66
e6 - Design, organization, and development of co-teaching	Pre	2 (0,9)	18 (8,0)	55 (24,4)	121 (53,8)	29 (12,9)	3,51
	Post	4 (1,8)	14 (6,2)	43 (19,1)	124 (55,1)	40 (17,8)	3,55
e7 - Design of learning assessment: what, how, and when to evaluate	Pre	1 (0,4)	17 (7,6)	64 (28,4)	122 (54,2)	21 (9,3)	3,50
	Post	3 (1,3)	11 (4,9)	53 (23,6)	119 (52,9)	39 (17,3)	3,55

Table 9 - Wilcoxon test comparing answers given pre and post training experience.

Items	N	Neg Ranges	Pos Ranges	Draws	Z	p
e1	188	15	24	129	-,837	,403
e2	188	12	35	117	-3,242	,001
e3	185	20	27	115	-,306	,759
e4	188	13	34	117	-2,709	,007
e5	185	11	32	119	-2,514	,012
e6	185	17	28	117	-1,353	,176
e7	186	21	22	124	-,039	,969

© Italian e-Learning Association

Table 10 - Compilation of means, number of positive answers (Num) and percentage of them (%) post training experience.

Dimension B			Dimension C			Dimension D			Dimension E		
Item	**Mean**	**Num (%)**	**Item**	**Mean**	**Num (%)**	**Item**	**Mean**	**Num (%)**	**Item**	**Mean**	**Num (%)**
b1	3,61	180 (80,0)	c1	2,94	117 (52,0)	d1	3,47	170 (75,6)	e1	3,64	178 (79,1)
b2	3,32	159 (70,7)	c2	2,95	123 (54,7)	d2	3,50	168 (74,7)	e2	3,68	179 (79,6)
b3	3,33	159 (70,7)	c3	2,81	111 (49,3)	d3	3,55	171 (76,0)	e3	3,39	160 (71,1)
b4	3,62	179 (79,6)	c4	3,24	156 (69,3)				e4	3,58	175 (77,8)
b5	3,11	138 (61,4)	c5	3,24	144 (64,0)				e5	3,66	177 (78,7)
b6	3,08	137 (60,9)	c6	3,28	149 (66,2)				e6	3,55	167 (74,2)
b7	3,14	143 (63,6)	c7	3,19	139 (61,8)				e7	3,55	172 (76,5)
b8	3,31	152 (67,6)	c8	3,03	131 (58,3)						
b9	3,21	141 (62,7)	c9	2,78	108 (48,0)						
b10	3,42	161 (71,5)	c10	3,12	127 (56,5)						
b11	3,48	164 (72,9)									
b12	3,50	164 (72,9)									

4. Discussion and Conclusions

Throughout this study, the two research questions posed in the Introduction were addressed. The first question sought to explore how teachers' interests evolve when engagement with an Innovative Learning Environment (ILE) occurs for the first time. Specifically, the aim was to examine how exposure to such environments, through a visit to a lab designed with ILE characteristics and a hands-on learning experience, impacts teachers' perceptions and enthusiasm for adopting new educational practices. This immersive experience was carefully crafted to not only familiarize teachers with ILEs but also to challenge their traditional views on pedagogy and space, offering them an opportunity to reflect on how learning environments shape educational outcomes. It should be noted that the training provided was relatively brief (two hours in duration), which may influence the results. In addition, the study aimed to measure short-term interest rather than actual behavioral change. Further research could further explore the potential effects of this type of training over a longer period.

Regarding teaching methodologies and strategies (Dimension B), the results of the Wilcoxon test indicate a significant increase in teachers' interest in innovative, student-centered approaches, such as Project-based learning, Problem-based learning, Inquiry-based learning, Cooperative Learning and Flipped Classroom. These findings suggests that teachers become more receptive to adopting active, participatory methodologies or strategies that promote students' critical thinking and problem-solving skills.

In Dimension C, which focuses on digital technology, the results reveal a heightened interest in training related to Virtual Reality, Augmented Reality and 3D printing. These technologies, which can be considered as peripheral to the core of educational practice, seem to be seen by educators as powerful enablers of more innovative, immersive and interactive learning experiences. This finding is consistent with global trends in education, where the integration of technology is increasingly considered essential for preparing students to navigate a rapidly evolving digital landscape (Educause, 2023; ODITE, 2022).

In relation to training in spatial design and organization (Dimension D), a notable increase in interest in the zoning of learning spaces was observed. Teachers became more engaged with the idea of configuring learning environments to maximize comfort, flexibility and adaptability. This reflects a growing awareness that physical space is not a neutral backdrop to learning but can be strategically designed to enhance collaboration, creativity and student involvement.

© Italian e-Learning Association

In Dimension E, which relates to didactics in ILEs, our findings indicate a strong increase in teachers' interest in learning how to design and program learning experiences in ILEs, how to manage coexistence and participation and how to select and organize the appropriate methodology or strategy to be developed in an ILE to achieve learning. This suggests that teachers are eager to gain practical knowledge that can help them implement innovative teaching strategies in these new environments.

These findings allow us to conclude that engaging teachers in a training experience within an ILE serves as a catalyst for further professional development, and provide clear indications of the content that should be prioritized in future training initiatives. The exposure to innovative environments stimulates a curiosity and motivation to delve deeper into specific pedagogical approaches, particularly those that align with the characteristics of ILEs, characteristics that are already in some of the laws structuring education in different countries. Educational authorities or professional development institutions can use this information to design targeted training programs that meet teachers' evolving needs, ensuring that such experiences are not isolated, but rather integrated into broader, continuous professional learning pathways.

Our second research question sought to uncover which specific topics become more interesting for teachers once they delve into Innovative Learning Environments. Results showed that the three topics of greatest interest are in Dimension E: "Design and programming of learning experiences in environments other than the traditional classroom (e2)" (x=3,68), "Selection and organization of the appropriate methodology or strategy to promote learning (e5)" (x=3,66) and "Design, programming, and creation of projects or experiences that promote learning (e1)" (x=3,64). Thus, teachers feel the need to receive training on how to implement an ILE. This is consistent with previous research indicating that teachers often struggle to identify the pedagogical strategies needed to create effective innovative learning environments (Cleveland, 2016) and do not readily transition to more active pedagogical approaches (Beery et al., 2013), pointing to the need for professional development that helps translate innovative pedagogical principles into classroom practice. Our findings reaffirm the importance of structured, targeted professional development in supporting teachers as they navigate this transition from more traditional settings.

In addition to these key findings, the most relevant topics in each dimension were identified, providing clear guidance on the content to be addressed in future professional development programs. For Dimension B, "Cooperative learning (b4)" (x=3,62) and "Project-based learning (b1)" (x=3,61) emerged as the most important. In Dimension D, "Zoning of useful learning environment (d3)" (x=3,55) attracted the highest level of interest, while in Dimension C, "Virtual reality (c6)" (x=3,28) stood out.

In general terms, greater interest was observed in didactic issues. However, it should be noted that the most frequently selected item, both in frequency (80%) and in total amount of positive answers, is "Project based learning", with interest shown by 180 out of 225 teachers. It is interesting to observe this, as it is a topic that has been widely promoted and covered in professional development over the past decade. The sustained interest in this methodology may reflect its versatility and relevance across various subjects and educational contexts. It should also be noted that some other aspects received almost the same number of positive responses (Cooperative learning and the three items in dimension E previously commented, e2, e5 and e1).

It is also worth highlighting that all aspects showing statistically significant differences are related to an increase in interest, suggesting that teachers' engagement with ILEs generally fosters a positive response towards further professional development. However, some areas saw a slight decrease in interest when comparing pre- and post-visit results. Specifically, Robotics, Programming and Computational Thinking in Dimension C; Design Thinking and Personalized learning in Dimension B, and Scheduling of a learning process in Dimension E all saw a decline. Although these declines were not statistically significant, they warrant further investigation to understand whether they represent temporary fluctuations or reflect deeper concerns about the practicality or relevance of these topics in ILEs.

Finally, an interesting disconnect regarding Artificial Intelligence (AI) was observed. Despite the growing public and educational interest in AI, it did not emerge as a major area of focus for teachers. This could be due to the fact that our training experience did not explicitly address AI, meaning that teachers may not have been prompted to consider its implications within the context of ILEs. Nevertheless, given the increasing role of AI in shaping future learning technologies, it would be of interest to consider developing further studies on the topic.

This study has several limitations that may affect the generalizability and interpretation of the findings. The sample size and limited geographic scope may restrict the representativeness of the teaching population, and the use of self-reported data may introduce biases such as social desirability. Additionally, the study measures only short-term shifts in interest, leaving the long-term impact of ILE exposure unclear. It also does not account for the practical challenges teachers may face in implementing innovative practices in real-world settings. Finally, the controlled nature of the ILE training lab may not fully capture the diversity of ILE experiences in various educational contexts, limiting the applicability of the results.

© Italian e-Learning Association

Despite these limitations, our study also shows several strengths, including its ability to highlight significant shifts in teachers' interests after exposure to an Innovative Learning Environment, the use of rigorous statistical methods to analyze changes across multiple pedagogical dimensions and its contribution to the growing body of research on how educational spaces and technologies influence teaching practices. Additionally, the study provides actionable insights for designing more targeted professional development programs and underscores the importance of aligning professional development with the evolving demands of 21st-century education.

In conclusion, our study highlights the transformative and motivational potential of ILEs in reshaping teachers' professional interests and guiding them towards more innovative, student-centered educational practices. While certain topics, such as Project-Based Learning and Cooperative Learning, continue to dominate teacher interest, the growing emphasis on technology and spatial design underscores the evolving nature of pedagogy in the 21st century. The findings also stress the importance of targeted professional development in equipping teachers with the tools and strategies necessary to navigate these new educational landscapes effectively.

Acknowledgements

The authors wish to express sincere gratitude to the Coordinator of the Aulas Transformadoras program of the Regional Ministry of Education, Culture, Universities and Employment for commissioning this study and facilitating access to the sample. Their support and collaboration were essential to the successful completion of this research.

This research received no external funding.

References

Beery T. A., Shell D., Gillespie G., & Werdman E. (2013). The impact of learning space on teaching behaviors. *Nurse Education in Practice, 13*(5), 382-387. https://doi.org/10.1016/j.nepr.2012.11.001

Blannin J., Mahat M., Cleveland B., Morris J., & Imms W. (2020). Teachers as Embedded Practitioner-Researchers in Innovative Learning Environments. *Center for Educational Policy Studies Journal, 10*(3), 99-116. https://doi.org/10.26529/cepsj.887

Blázquez D. (2016). *Métodos de enseñanza en educación física. Enfoques innovadores para la enseñanza de competencias* [Teaching methods in physical education. Innovative approaches to teaching skills]. Inde.

Byers T., Imms W., & Hartnell-Young E. (2014). Making the case for space: The effect of learning spaces on teaching and learning. *Curriculum and Teaching, 29*(1), 5–19. https://doi.org/10.7459/ct/29.1.02

Byers T., Imms W., & Hartnell-Young E. (2018a). Comparative analysis of the impact of traditional versus innovative learning environment on student attitudes and learning outcomes. *Studies in Educational Evaluation, 58*, 167–177. https://doi.org/10.1016/j.stueduc.2018.07.003

Byers T., Imms W., & Hartnell-Young E. (2018b). Evaluating teacher and student spatial transition from a traditional classroom to an innovative learning environment. *Studies in Educational Evaluation, 58*, 156–166. https://doi.org/10.1016/j.stueduc.2018.07.004

Cleveland B. (2016). Addressing the spatial to catalyse socio-pedagogical reform in middle years education. In K. Fisher (Ed.), *The translational design of schools* (pp. 27–49). Sense Publishers.

Educause. (2023). *Horizon report*. https://library.educause.edu/resources/2023/5/2023-educause-horizon-report-teaching-and-learning-edition

European Schoolnet (27 august 2024). *Future Classroom Lab*. https://fcl.eun.org/

French R., Mahat M., Kvan T., & Imms W. (2022). Viewing the transition to innovative learning environments through the lens of the burke-litwin model for organizational performance and change. *Journal of Educational Change, 23*, 115-130. https://doi.org/10.1007/s10833-021-09431-5

García Mongue A., Pellicer Iborra C., Jiménez Ruiz M., & Gómez González F. J. (2023). El espacio en el trabajo por proyectos: el caso de los proyectos del Modelo Rubik. *Revista Electrónica Interuniversitaria de Formación del Profesorado, 26*(2), 37–56. https://doi.org/10.6018/reifop.559421

Gonzalez-Mohino M., Rodriguez-Domenech M., Callejas-Albiñana A., & Castillo-Canalejo A. (2023). Empowering Critical Thinking: The Role of Digital Tools in Citizen Participation. *Journal of New Approaches in Educational Research, 12*(2), 258-275. https://doi.org/10.7821/naer.2023.7.1385

Granda-Pinan, A.R., Roda-Segarra, J., & Mengual-Andrés, S. (2024). Espacios digitales de aprendizaje como prolongación del espacio físico en el contexto de los Entornos Innovadores de Aprendizaje. In Morales Cevallos, M.B., Marín-Marín, J. A., Berbel Oller, P. y Villegas Castro, A.

© Italian e-Learning Association

S. (Eds.), *Desafíos de la educación contemporánea: perspectivas formativas para una sociedad digital* (pp. 86-103). Dykinson. https://doi.org/10.2307/jj.17381582

Granda-Pinan, A.R. & Rojo-Bofill, L.M. (2024). Los Entornos Innovadores de Aprendizaje como respuesta a los retos educativos del siglo XXI. Research in Education and Learning Innovation Archives, 32, 22-35. https://doi.org/10.7203/realia.32.27803INTEF (9 September 2024). *Aula del Futuro*. https://auladelfuturo.intef.es/

Jorion N., Taeyaerts D., & Jeanes W. (2016). Promoting collaboration using team based classroom design. *Creative Education, 07*(05), 724–729. https://doi.org/10.4236/ce.2016.75076

King E., Joy M., Foss J., Sinclair J., & Sitthiworachart, J. (2015). Exploring the impact of a flexible, technology-enhanced teaching space on pedagogy. *Innovations in Education and Teaching International, 52*(5), 522-535. http://dx.doi.org/10.1080/14703297.2014.896222

Lawshe C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Losada López J.L., & López-Feal Ramil R. (2003). *Métodos de investigación en ciencias humanas y sociales* [Research methods in human and social sciences]. Thomson-Paraninfo.

Lozano, Ó. R., Granda-Pinan, A. R., & Alameda-Villarrubia, A. (2024). Training on Innovative Learning Environments: Identifying Teachers' Interests. *Education Sciences, 14*(6), 601. https://doi.org/10.3390/educsci14060601

Mahat M., Bradbeer C., Byers T., & Imms W. (2018). *Innovative Learning Environments and Teacher Change: Defining key concepts.* University of Melbourne, LEaRN.

Mcmillan J.H., & Schumacher S. (2001). *Research in education: A conceptual introduction.* Longman

Nelson D., Bonem E., & Fitzsimmons J. (2023). Instructor Perceptions of Teaching in a New Active Learning Building. *Journal of Learning Spaces, 12*(1). files.eric.ed.gov/fulltext/EJ1400289.pdf

ODITE (2022*). Informe ODITE sobre tendencias educativas 2022*. https://ciberespiral.org/es/informe-odite-tendencias-educativas/

Radcliffe D. (2009). *A pedagogy-space-technology (PST) framework for designing and evaluating learning places. In Learning spaces in higher education: Positive outcomes by design*. Proceedings of the Next Generation Learning Spaces 2008 Colloquium. University of Queensland: Brisbane (pp. 11-16).

Rivera-Vargas P., Calderón-Garrido D., Jacovkis J., & Parcerisa L. (2024). Exploring student and family concerns and confidence in BigTech digital platforms in public schools. *Journal of New Approaches in Education Research 13*(5). https://doi.org/10.1007/s44322-023-00003-4

Suárez J, Maiz F., & Meza Pardo M. (2010) Inteligencias múltiples: una innovación pedagógica para potenciar el proceso enseñanza aprendizaje. *Investigación y Postgrado, 25*(1), 81-94.

Tristán-López A. (2008). Modificación al modelo de lawshe para el dictamen cuantitativo de la validez de contenido de un instrumento objetivo. *Avances en medición, 6*(1), 37-48.

© Italian e-Learning Association

Annex I

Description of the activity developed during the visit to the Innovative Learning Environment

BACKGROUND
The Future Classroom Lab model promotes the use of active methodologies for skills development supported by digital technologies in flexible spaces. We propose adopting this participatory approach in the design of this guided tour of the Future Classroom Lab (FCL), Aula del Futuro (AdF) in Spanish, creating a route through the six zones that also considers accessibility in the teaching and learning process.
GOAL
Understanding the FCL as a flexible space where active methodologies supported by digital technology can be used in the teaching-learning processes, taking into account the principles of accessibility.
LENGTH
90'-120'

0	**BEFORE ENTERING THE CLASSROOM**		
- Welcome. Explanation of the objective of the visit: to understand the FCL as an ILE. *- Creating groups of 4 visitors (heterogeneous)* ***5'***			
1	**EXPLORE -** Search | Investigate | Discover		
Focus	***Resources***	***Purpose of the space***	***Activity***
Space for computational thinking experimentation	Virtual reality glasses, mobile devices, robotics kits...	Support exploration and experimentation with analysis and understanding of environments.	*The visiting teaching group is invited to tour the classroom and try to discover its principles, resources and possibilities.* *Reflection sheet for the entire tour.* ***5'***
2	**INTERACT -** Ask | Collaborate | Debate		
Focus	***Resources***	***Purpose of the space***	***Activity***
Space for exchange and collaborative work.	Mobile tables and chairs. Interactive whiteboard, OER for teaching and research staff...	Support interactivity, collaboration and the development of debates together with formative assessment.	*Each group is asked to think of 4 words that they identify with the* AdF. *Each person will focus on one of the words agreed upon in the group.* *They hand out tablets (material control) or use their mobile phone to create a collaborative word cloud that they will see access to on the whiteboard.* ***10'***
3	**RESEARCH -** Examine | Question | Analyze | Find out		
Focus	***Resources***	***Purpose of the space***	***Activity***
Space for observation and resolution of problems or challenges.	Flexible furniture. Laptops and *Flip pannel*.	To support the development of active research and cross-cutting projects with the promotion of critical thinking.	*Multilevel activity:* *The groups carry out a small research on the Internet to find out what the Classroom of the Future is, what the project and the model consist of.* *- Basic level: Research the AdF INTEF site (or AT site if published) and watch the videos.* *- Medium level: Investigate the INTEF and FCL and AdF sites, watching the videos.* *- Advanced Level: No guide provided* *- Extension: Recognize Fake News about the FCL that appears on social networks or in the press and correct this media misinformation.* ***10'***

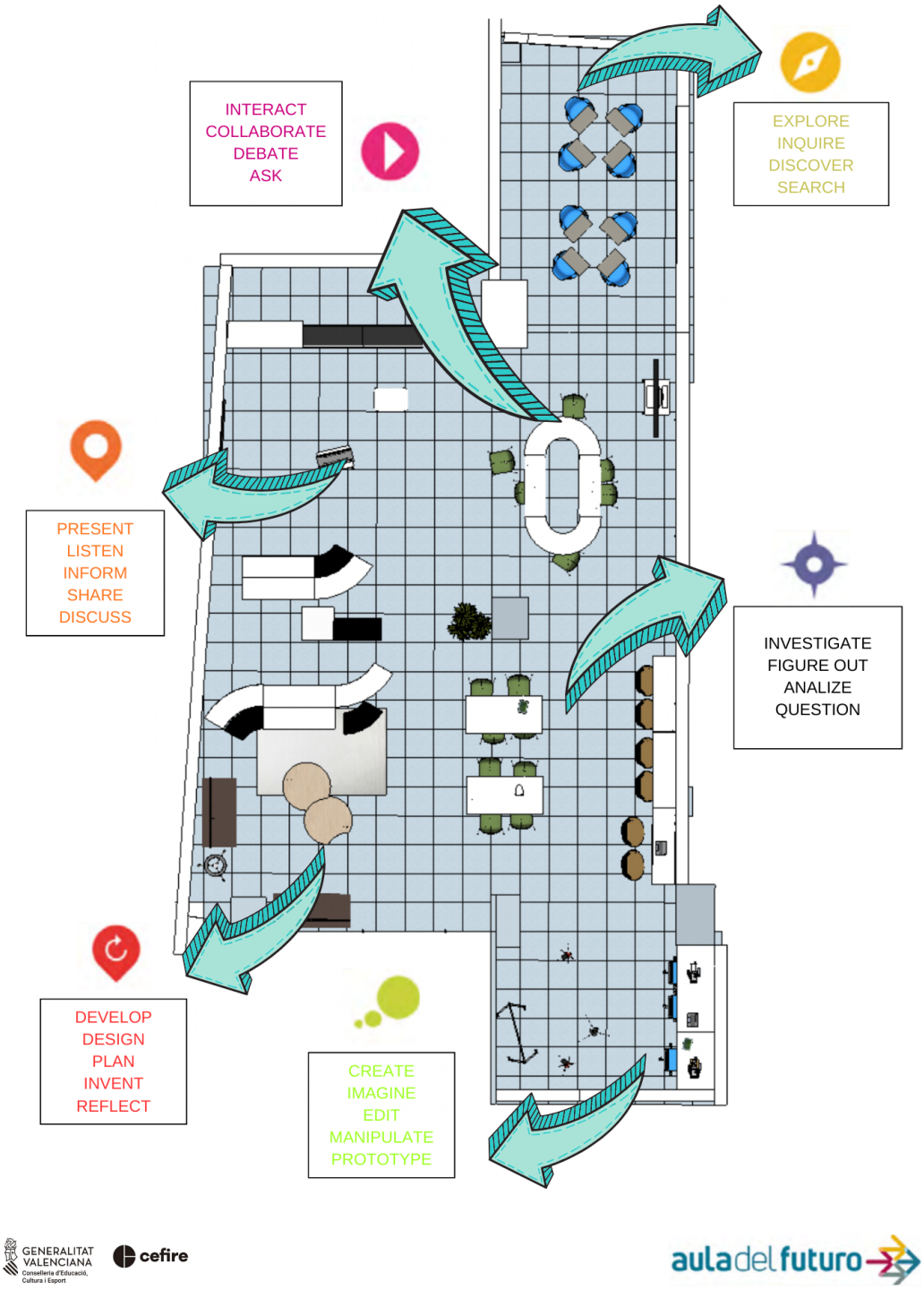
© Italian e-Learning Association

4	**DEVELOP-** Plan | Design | Invent		
Focus	***Resources***	***Purpose of the space***	***Activity***
Space dedicated to informal learning and reflection.	Soft furniture, study corners, etc.	Support self-expression and self-directed, personalized learning.	*The groups develop a base script for their final product: a video or podcast of approximately two minutes in length that explains what the FCL is.* *Materials provided: Script template for video and script template for podcast in print and digital.* ***15'***
5	**CREATE -** Imagine | Manipulate | Edit		
Focus	***Resources***	***Purpose of the space***	***Activity***
Space for creation and production.	Chroma, cameras, computers, 3D printers, SW video and audio editing...	Support initiative and creativity and develop social skills through PBL in collaborative groups.	*Creation of the final product (video or podcast) to explain what the Future Classroom Lab is.* ***15'***
6	**PRESENT -** Show | Share | Listen | Inform		
Focus	***Resources***	***Purpose of the space***	***Activity***
Space for initial or final presentations.	Reconfigurable furniture Interactive displays.	Support the sharing of results by encouraging active listening and co-evaluation.	*The groups present their final products. At the end, the dynamization team shows its presentation and explanation of the FCL.* *You can answer some questions briefly or start a colloquium.* *Now, they are asked to recreate a word cloud to contrast the initial vision with the final one.* ***15'***
	DEVELOP- Plan | Design | Invent		
Focus	***Resources***	***Purpose of the space***	***Activity***
Space dedicated to informal learning and reflection.	Soft furniture, study corners, etc.	Support self-expression and self-directed, personalized learning.	*Each participant returns to the reflection space to fill in the* AdF *visit sheet and write a short text about what they now know the* AdF *is. Give the sheet to the dynamic team and save your concluding text in a virtual repository or pen drive.* ***10'***

© Italian e-Learning Association

Annex II

The Aula del Futuro Teacher Training Lab (Valencia, Spain).



© Italian e-Learning Association

JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026)

How can educational videos with annotations help students in online learning environments?

Riad Bourbia[a,b], Samia Drissi[c], Yacine Lafifi[b,1], Sevinc Gulsecen[d]

[a]*Computer Science Department, University of Souk-Ahras – Souk-Ahras (Algeria)*

[b]*LabStic laboratory, 8 May 1945 University of Guelma – Guelma (Algeria)*

[c]*LiM Laboratory, Faculty of Science and Technology, University of Souk-Ahras – Souk-Ahras (Algeria)*

[d]*Informatics Department, University of Istanbul, Istanbul (Turkey)*

(submitted: 22/12/2024; accepted: 1/5/2026; published: 23/5/2026)

Abstract

In recent years, many new technologies have been used to enhance the learning outcomes of learners in online learning. Video annotation is one of these new technologies which is an innovative way to make learning more interactive and engaging. It can accommodate different learning styles by turning passive video-watching into an active learning experience. This research aims to evaluate the effectiveness of an e-Learning platform incorporating a Video Annotated Technique. Learners can pinpoint challenging parts of educational videos and then receive personalised explanations from their teachers. The primary purpose of these annotations is to clarify and provide thorough explanations of complex topics, ultimately improving understanding and making preparation for final examination tests easier. This paper presents a new personalization approach by managing the learners' annotations. Furthermore, it offers a video annotation tool incorporated into an e-Learning environment. This tool was tested on a sample of students, and the experimental results demonstrated a significant positive impact on learners' performance.

KEYWORDS: Online Learning Environments, Video-Based Learning, Video Annotation, Assisting Learners, Learning Personalization.

DOI
https://doi.org/10.20368/1971-8829/1136133

CITE AS
Bourbia, R., Drissi, S., Lafifi, Y., & Gulsecen, S. (2026). How can educational videos with annotations help students in online learning environments? *Journal of e-Learning and Knowledge Society*, *22*(1).
https://doi.org/10.20368/1971-8829/1136133

1. Introduction

Traditional teaching methods, often instructor-centered and based on linear content, no longer meet the needs of today's learners, who are seeking more interactive and immersive learning experiences (Stigler et al., 2015; Bower, 2017). The rise of online learning has amplified this need, creating a demand for pedagogical approaches that not only convey information but also actively and meaningfully engage students (Laurillard, 2013).

The adoption of video in education, commonly referred to as video-based learning (VBL), represents a significant advancement in modern pedagogy. This approach leverages both visual and auditory channels to facilitate knowledge acquisition and retention, thereby offering a more immersive and engaging learning experience (Chui et al., 2016; Brame, 2016; Lam et al., 2021a; Lange & Costley, 2020; Baber et al., 2022; Arora, 2023; Navarrete et al., 2025). The rapid transition to online learning, particularly during the COVID-19 pandemic, has further emphasized the importance of video-based learning (Dong, Cao, & Li, 2020; Lam et al., 2021a; Ahn et al., 2025).

Digital video platforms such as YouTube and Vimeo are profoundly reshaping contemporary society, attracting billions of users and becoming increasingly integrated into educational practices. According to Statista (2023), 93% of internet users consume online video content, with 26% specifically engaging with educational videos. In this context, a substantial proportion of topics covered in high school science curricula is now available on YouTube, often in

[1] corresponding author - email: laf_yac@yahoo.fr

© Italian e-Learning Association

multiple formats reflecting diverse pedagogical approaches (Alder et al., 2025).

This expansion is further illustrated by the rise of large-scale online learning platforms such as Coursera, which provide courses and recorded lectures from numerous universities worldwide. By the end of 2025, Coursera had reached approximately 197 million registered learners (Coursera, 2026), highlighting the growing demand for flexible and accessible learning solutions.

Educational videos offer considerable pedagogical potential. They enhance the quality of demonstrations, support students in preparing for practical and laboratory activities, and foster motivation as well as sustained engagement in learning (Voronkin, 2019). More broadly, the integration of recorded video lectures constitutes a central driver of online learning, as it increases flexibility while improving the affordability and accessibility of education for diverse audiences (Ahn et al., 2025).

The growing need for dynamic pedagogical approaches, combined with technological advances, has led to the development of Video Annotation Tools (VAT). These tools are increasingly recognized as essential for enhancing online learning by addressing several challenges identified in recent research. One major issue concerns maintaining learner engagement and attention, particularly in the context of lengthy videos, which may negatively affect learning effectiveness (Hughes et al., 2019; Tseng, 2021). This challenge is especially pronounced in environments where interaction with educational content remains largely passive (Guo, Kim, & Rubin, 2014; Hui & Campbell, 2018; Dong et al., 2020; Vandenberg & Magnuson, 2021). Video Annotation Tools transform this dynamic by promoting active learner participation and enabling direct interaction with content at key moments, thereby enhancing knowledge retention (Fiorella & Mayer, 2016; Love et al., 2014; Lam et al., 2021b; Von Wachter et al., 2023).

Empirical studies further support the effectiveness of these tools. For instance, Cassano et al. (2024) examine the impact of a video annotation tool on student engagement during the autonomous learning phase within a flipped classroom model. Their findings indicate a significant improvement in student engagement, particularly in the emotional dimensions, while transforming a traditionally passive activity into an active learning experience.

Moreover, video annotations play a critical role in managing cognitive load, a key factor in effective learning. According to Sweller, Ayres, and Kalyuga (2011), learning is optimized when cognitive load is appropriately regulated. By segmenting information and directing learners' attention to essential elements, video annotations help reduce cognitive overload and promote deeper understanding and more effective knowledge assimilation.

In addition, video annotations support collaborative learning by enabling learners to insert notes, questions, and comments directly within video content. This functionality fosters deeper understanding and critical reflection (Mirriahi et al., 2018; West, 2020; Von Wachter et al., 2023), while also enhancing communication and practical skills (Pérez-Torregrosa et al., 2017; Susantini et al., 2019; Chui et al., 2021; Kim Chau Leung et al., 2021). This interactive approach not only increases motivation but also contributes to the development of an active and engaged learning community (Kay & Kletskin, 2012).

Recent empirical evidence further reinforces these findings. For example, a qualitative study conducted by Rasenberg et al. (2025) on medical students using a Video Feedback System (VFS) during clinical internships shows that learners use annotations in a structured manner for self-assessment and feedback.

In this context, the present study introduces SANED_anot+, a research-oriented video annotation system designed to enhance learner support within an online learning environment through the SANED platform (Bourbia et al., 2024). SANED_anot+ is specifically developed to support learners facing difficulties by enabling them to identify and annotate complex segments of educational videos. In response, instructors enrich these segments with detailed explanations, thereby providing targeted support. These personalized annotations are subsequently shared with all learners, promoting collaborative learning and facilitating targeted assistance and collective knowledge construction.

To effectively address this issue, we must consider several essential research inquiries that require thoughtful and comprehensive responses. So, two main questions might be addressed:

- What is the impact of the video annotation system on the performance of low-scoring learners during pre-test assessment?
- How does the video annotation system affect learners' performance with high pre-test scores?

The document is organised as follows: Section 2 reviews how video annotation can assist learners with difficulties in e-Learning.

The proposed approach is described in Section 3. Section 4 introduces the development tool, while Section 5 presents an experiment conducted in a higher education institution, presenting significant findings and initiating a discussion. Finally, the general conclusion, the work limitations, and the future work are highlighted in Section 6.

© Italian e-Learning Association

2. Literature Review for video annotation as a way to support e-Learning

As defined by Higuera and colleagues (2014), video annotations refer to valuable data explicitly associated with the spatiotemporal segments of the video. Annotation Video Tools allow additional insightful commentary without modifying the video (Khurana and Chandak, 2013; Chatti, 2016).

According to Yousef's research team (2015), annotations have been found to enhance communication, promote in-depth analysis of video content, and provide valuable feedback.

In the following, we explore a variety of video annotation tools and summarise their applicability and limitations while emphasising the critical distinctions with the proposed video annotation tool. A set of academic, commercial, and open-access tools was selected to ensure a comprehensive approach. Innovative and advanced teaching approaches foster Educational research tools in stimulating learning environments. Commercial solutions offer reliability and ease of use, crucial for large-scale implementation and quick user adoption. In the end, open-access tools provide extended access to quality educational resources, usually free or cheap.

Seven video annotation systems were selected for this analysis due to their potential to help learners and innovative features. These include VideoAnt https://ant.umn.edu/ (Hosack, 2010; Van Drom, 2018), a free web application for video annotation developed by the University of Minnesota, EdPuzzle https://edpuzzle.com/ (EdPuzzle, Rahayu, 2022), a free platform for educational video lessons, CourseMapper https://www.uni-due.de/soco/research/projects/coursemapper.php (Chatti et al., 2016) a collaborative video annotation and analytics platform, GoReact https://get.goreact.com/ (Hager et al., 2020) a video feedback software, TRAVIS GO https://app.travis-go.org/ (Klug, 2021) a web application for simple and collaborative video annotation, ViDeX (Fong et al., 2018) a video player for personalising educational videos, and FEVA (Shrestha, 2023) an event video annotation tool.

All these tools offer various ways to annotate videos, from direct interaction with the content to more organised and collaborative approaches. They enable users, whether teachers or students, to interact with video content more successfully.

A deeper analysis of these video annotation tools reveals that each offers distinct features that suit specific educational environments. With EdPuzzle, teachers can design interactive lessons by embedding questions and comments within videos, increasing

Table 1 - Comparison of Basic Technical Features of selected VATs.

Basic technical features		VideoAnt	EdPuzzle	CourseMapper	GoReact	Travis GO	ViDeX	FEVA
License	Open Source							√
	Proprietary	√	√	√	√	√	√	
	Commercial				√			
	Open Access	√	√			√		√
	Maintained	√	√	√	√		√	√
	Outdated							
Platform	Web	√	√	√	√	√	√	√
	Native							
Annotation Approach	Manual	√	√	√	√	√	√	√
	Automatic							
Context of use	Educative	√	√	√	√	√	√	
	Research							√
Annotation Type	Note	√	√	√	√	√	√	
	Question	√	√	√	√			
	External Resource	√	√	√	√			
Input Type	Textual	√	√	√	√	√	√	
	Audio		√		√			√
	Image				√			
	Video				√			√
Annotation Rating	Teachers			√				
	Peers	√		√				

© Italian e-Learning Association

student engagement and understanding of the content. CourseMapper innovates by integrating annotations into a mind-map view, providing a better structure for the course material, and offering visual analysis tools to help learners identify the most viewed and annotated parts of videos. Travis GO, on the other hand, focuses primarily on the teaching and analysis of audiovisual media, offering features relevant to video annotation in an educational context. As a versatile tool, Go React facilitates in-depth video analysis and collaborative feedback through synchronised annotations.

FEVA distinguishes itself with its distinctive design, facilitating rapid and precise video annotation through an intuitive interface and streamlined settings. In contrast, ViDeX is positioned as a video player designed specifically for video-based learning, encouraging active viewing, the sharing of annotations between students and teachers, and the identification of consensus annotations based on the 'wisdom of crowds', which refers to the idea that the collective intelligence of a group can often be more accurate and reliable than that of an individual.

In addition, the literature review has enabled us to draw up a list of advanced features of VATs that aim to support learning and help learners. The chosen systems were examined in detail, with particular attention paid to the advanced features (Table 6). These criteria include video annotation capability, collaboration and user interaction facilitation, integration with other educational designs such as LMSs, and compatibility with various content formats. The analysis also assesses usability, accessibility, personalising the learning experience, and providing targeted help.

Finally, the criteria consider support for teacher feedback, assessment of the usefulness of annotations, and sharing and discussion functions, all are essential for a practical and interactive e-Learning environment. Table 6 provides a detailed comparative summary between our developed tool, SANED_Anot+, and the various systems selected for this study.

The study intends to assess the efficacy of video annotation tools in assisting learners with learning disabilities in online platforms. The research reveals several aspects to be taken into account. We found that annotations, mainly initiated by teachers, aim to clarify the course content. However, they are often applied uniformly without considering cognitive variations and students' preferences in terms of learning styles. This one-size-fits-all approach does not effectively address the specific needs of learners with difficulties, which poses a significant problem regarding comprehension.

Moreover, there is an evident lack of adaptation and personalisation of video annotations in response to the requests of students facing challenges. While specific solutions offer responses to inquiries, they fail to rectify the issue by providing tailored video annotations that cater to the particular requirements of these students. Ultimately, assessing video annotations based on indicators like "like" or "dislike" buttons is entirely subjective. This evaluation method does not effectively utilise feedback to improve the tools for students with difficulties.

A novel approach for providing learners with video annotation is being considered to overcome the shortcomings of the previously described works. The main objectives of this approach and the associated tool are presented in the following section.

3. The proposed video annotation module

The research presented in this article is an extension of that shown in the SANED project (Bourbia et al., 2024), which aims to provide support for learners engaged in online learning. SANED is an e-Learning platform that uses the traces left by learners to detect those who have difficulties and provides them with immediate and personalised assistance.

The system additionally assists teachers in structuring learning contents hierarchically, implementing prerequisites, and tailoring teaching resources to match the learner's cognitive level and preferred learning style. SANED dynamically maintains and refines the learner's cognitive profile to adapt resources to their current state as they progress through the learning process.

SANED consists of three distinct subsystems, along with a newly added module, SANED_Anot+, specifically dedicated to the annotation process (see Figure 1). The first subsystem focuses on learning, offering teachers and learners essential tools for effectively managing learning contents and assessment resources.

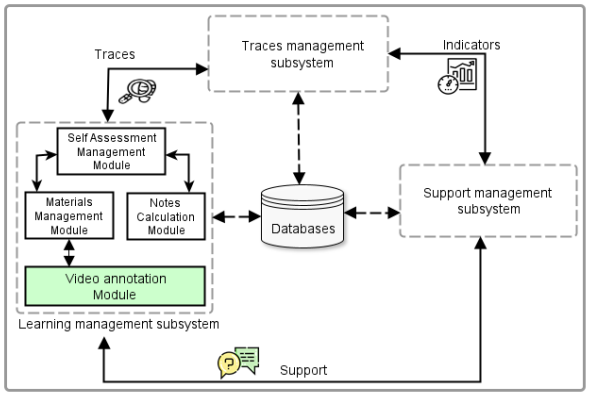


Figure 1- Overview of the General Architecture of SANED and the SANED_Anot+ Module.

The second subsystem collects information regarding the different activities and traces of the learners.

© Italian e-Learning Association

Subsequently, it furnishes the third subsystem with the essential data required to identify learners encountering difficulties and suggest appropriate forms of assistance.

The integration of SANED_Anot+ is designed to significantly enhance the learning experience by fostering dynamic and personalized interactions between teachers and learners with educational video content. This video annotation module fulfills several key objectives:

- Personalization of Learning: Provide learners with personalized educational support by allowing the addition of comments, questions, and supplementary explanations directly on the videos, tailored to their specific needs and queries.
- Enhancement of Engagement: Encourage active participation from learners by involving them in the analysis of video content, which strengthens their engagement and understanding of the subject matter.
- Facilitation of Real-Time Assistance: Enable teachers to address learners' questions accurately and specifically through annotations synchronized with particular moments in the videos.
- Optimization of Communication and Feedback: Improve communication between teachers and learners through ongoing dialogue based on annotations, facilitating immediate and constructive feedback.

3.1 Functionalities of the Video Annotation Module

The main functionalities of the video annotation module are summarized below.

- Learner-Teacher Interaction: Learners can highlight sections of the video they find complex and submit questions or comments, to which teachers can respond directly through additional annotations.
- Synchronized Annotations: Teachers can add annotations at specific moments in the videos to clarify concepts, provide further explanations, or pose questions designed to stimulate learners' thinking.
- Collaborative Assessments: Learners can evaluate the annotations, thereby contributing to a collaborative learning environment where educational content is continually enhanced by user feedback.

3.2 The Integration Process into the SANED System

All components of the video annotation module are integrated coherently within a structured framework, aimed at ensuring the overall effectiveness of the system.

The internal structure of the annotation module can be depicted as follows:

- Centralized Management Module: The video annotation module is incorporated into SANED's learning management subsystem, enabling centralized management of annotations and facilitating access to resources for both teachers and learners.
- Unified User Interface: The annotation interface is designed to seamlessly integrate with the rest of the SANED learning environment, ensuring a smooth transition between the various functionalities offered by the system.
- Data Security: The database plays a crucial role within the SANED system by centralizing and securing all data collected from the various subsystems. Specifically, it stores all video annotations and interactions between learners and teachers. By consolidating this information in a central location, the database not only ensures the integrity and consistency of the data but also facilitates access for all system users.

4. Presentation of SANED_Anot+ tool

This tool offers learners an easy-to-use interface for watching educational videos and allows them to interact with their teacher for help when they face difficulties understanding certain parts of the video content. When making a request (4), students pinpoint the moments in the video (3) where they require assistance. All the posted comments are displayed chronologically in the annotation feed (2). This function allows users to easily follow the conversation and see the progression of observations over time. Additionally, the chronological order helps maintain the context and relevance of each comment within the feed.

After learners make their requests, the teacher can offer feedback and support to help them understand the complex parts of the video content. He can add different annotation types, such as text, graphics, or a link to an external resource, to enrich the learning experience (Figure 3). Personalised feedback from the teacher demonstrates a real commitment to learners, offering genuine assistance under challenging situations. This approach improves learner motivation by providing reassurance and support on e-Learning platforms without making them feel alone, isolated, or unsupported.

In addition, with this user-friendly User Interface, Learners can evaluate the annotation by assigning a rating using a star rating system, or they can provide feedback through comments if they believe it has been helpful to them. All these features ensure that learners have an efficient and engaging learning process. It is important to note that SANED_Anot+ is currently a

© Italian e-Learning Association

prototype developed and deployed within the context of this study and is not yet publicly accessible.

5. Experiment: results and discussion

An experiment was conducted at the Computer Science Department of Guelma University in Algeria. The experiment evaluates the effectiveness of a tool called SANED_Anot+ (Supporting Online Learning through Annotated Educational Videos).

The experiment aims to determine how the video annotation system affects students with low pre-test and high pre-test scorers. To achieve these goals, it is essential to establish the appropriate metrics, collect relevant data, and perform statistical analysis to assess the critical relationships and differences between crucial groups or variables. The results of this experiment provide a direct answer to the research questions posed in the introduction section.

5.1 Participants

The experiment involved the participation of sixty-nine (69) third-year undergraduate students. These students were enrolled in "the Introduction to Cryptography and Computer Security" course and had access to the video course content.

5.2 Methodology

The course is organized into several instructional units, each delivered through recorded video lectures. Each video focuses on a specific concept or topic, thereby ensuring a progressive and structured coverage of the entire curriculum. This organization promotes a clear segmentation of the content, facilitating its assimilation by learners. Furthermore, this format supports flexible and individualized learning, allowing students to access the materials at their own pace, review the sequences as often as needed, and pay particular attention to the concepts they perceive as most challenging.

The experimental procedure consists of a sequence of steps followed by the participants. Initially, students are asked to watch the course videos to build a foundational understanding of the subject. A preliminary test is then administered to assess their initial level of comprehension. Learners who encounter difficulties are invited to submit their questions.

Subsequently, students are given access to enriched video content, including teacher-provided annotations that address their specific needs. These annotations aim to clarify complex concepts and provide targeted support. The experiment concludes with a final test designed to evaluate learners' progress and overall understanding. In addition, a post-experiment survey

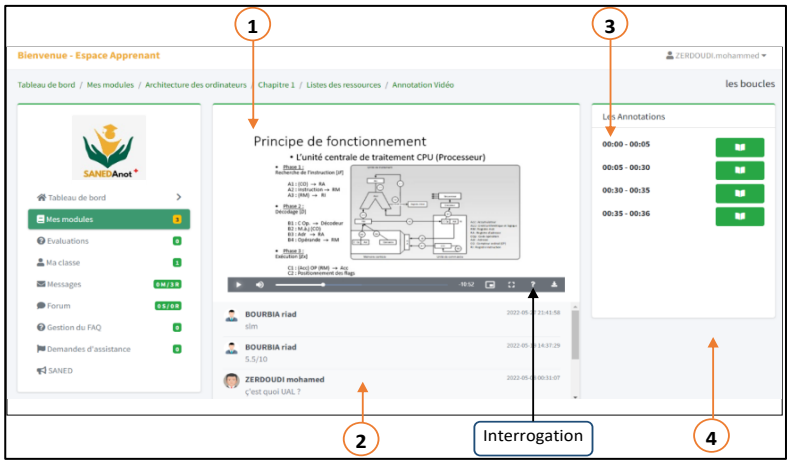


Figure 2 - A screenshot of SANED_Anot+ for a learner.

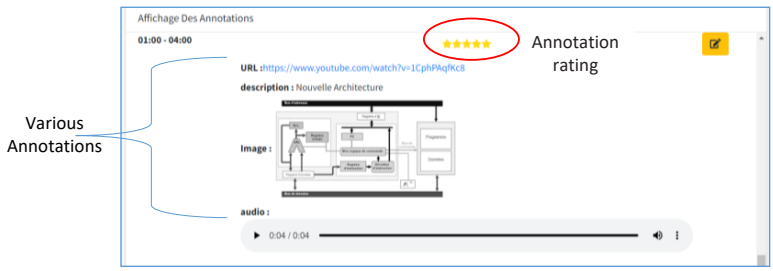


Figure 3 - Used annotation types.

© Italian e-Learning Association

is conducted to collect participants' perceptions and feedback regarding their learning experience. The questionnaire gathers learners' opinions on the impact of the video annotation system on their exam preparation. However, the results of this survey are not included in the present study and will be addressed in future work. Finally, the developed system is accessible to students through a simple enrollment process via any web browser.

5.3. Results and discussion

This part aims to demonstrate the effectiveness of our underlying system by evaluating its performance in improving the cognitive profiles of the learners.

This section will evaluate how the system influences students with learning difficulties. We present the results derived from a practical evaluation of our system using the following research hypotheses:

- **Null hypothesis H^10**: using the video annotation system does not affect the performance of learners with low pre-test scores.
- **Alternative hypothesis H^11:** using the video annotation system improves learners' performance with low pre-test scores.

To evaluate these hypotheses, a pre-test was administered to all learners, after which students who obtained scores below the average were selected (Table 2).

Table 2 - Number of participants involved in the experiment.

Number of students who took part in the pre-test	Number of students who passed the pre-test	Number of students who failed the pre-test
69	26	43

The learners who did not pass the pre-test (n = 43) subsequently used the SANED_Anot+ platform to address their learning difficulties and better prepare for the final assessment. A paired-samples Student's t-test was conducted to compare pre-test and post-test scores for learners with low initial performance. This test is appropriate for assessing differences between two related measurements obtained from the same participants. All statistical analyses were performed using SPSS (Statistical Package for the Social Sciences).

Prior to the analysis, the assumption of normality was assessed on the difference scores using the Shapiro–Wilk test. The results indicated no significant deviation from normality, supporting the use of the paired t-test. As shown in Table 3, the mean score from the pre-test (M = 2.8140, SD = 0.95757) to the post-test (M = 5.2791, SD = 2.06243). The mean difference between the two conditions was 2.46512 (SD = 2.23978), with a 95% confidence interval ranging from 1.77581 to 3.15442. The paired t-test indicated that this difference was statistically significant, $t(42) = 7.217$, $p < 0.001$, at the 0.05 significance level ($\alpha = 0.05$), suggesting that learners' performance improved following the intervention.

Table 3 - Paired T-test results for low-scoring learners in the pre-test.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Post-test Scores	5,2791	43	2,06243	,31452
	Pre-test Scores	2,8140	43	,95757	,14603

Paired Samples Test

		Paired Differences			
					95% Confidence Interval of the Difference
		Mean	Std. Deviation	Std. Error Mean	Lower
Pair 1	Post-test Scores - Pre-test Scores	2,46512	2,23978	,34156	1,77581

Paired Samples Test

		Paired Differences			
		95% Confidence Interval of the Difference			
		Upper	t	df	Sig. (2-tailed)
Pair 1	Post-test Scores - Pre-test Scores	3,15442	7,217	42	,0000000071

Situation	N	Mean	SD	t-score	p-value
Pre-test	43	2,8140	0,95757	7,217	,0000000071
Post-test	43	5,2791	2,06243		

© Italian e-Learning Association

These findings suggest that the use of the SANED_Anot+ platform may have contributed to improving the performance of learners with low pre-test scores. Accordingly, the null hypothesis (H^10) was rejected, while the alternative hypothesis (H^11) was supported.

However, these results should be interpreted with caution. The analysis does not identify the specific factors underlying the observed improvement, nor does it account for potential variability in learners' responses to the intervention. To strengthen the conclusions, further analyses, including subgroup investigations, are planned.

While the results for low-performing learners (Table 3) indicate a significant improvement following the intervention, it is important to examine whether similar effects can be observed among students with higher initial performance.

This analysis focuses on a subgroup of students who achieved a passing score on the pre-test (i.e., students scoring five or more out of ten). It should be noted that the maximum score in Algerian universities is twenty. The aim is to assess whether the intervention produced similar effects for students with higher initial performance.

To this end, the following hypotheses were formulated:

- **Null hypothesis H^20:** The utilisation of the video annotation system has no impact on the performance of students who have achieved high pre-test scores. ·
- **Alternative hypothesis H^21:** The utilisation of the video annotation system improves the performance of students who have achieved high scores on their pre-tests. ·

The data were analysed using SPSS software, and the results are presented in Table 4. The obtained p-value ($p = 0.064$) exceeds the conventional significance threshold ($\alpha = 0.05$), indicating that there is insufficient statistical evidence to reject the null hypothesis.

Consequently, the results do not provide significant evidence that the use of the video annotation system improved the performance of students who initially achieved high scores. Similarly, the alternative hypothesis (H^21) is not supported. Overall, these findings suggest no statistically significant difference between pre-test and post-test scores for this group, although a slight increase in post-test scores can be observed. To further examine the differences between learner groups, an additional analysis was conducted, and the results are presented in Table 5.

An independent-samples t-test was performed to compare the mean improvement (difference scores) between Group 1 (G1) and Group 2 (G2). Prior to the analysis, the assumption of homogeneity of variances was assessed using Levene's test. The results indicated that this assumption was met ($F = 0.947$, $p = 0.334$),

Table 4 - Paired T-test results for high-scoring learners in the pre-test.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Post-test Scores	6,1538	26	1,75937	,34504
	Pre-test Scores	5,5000	26	,81240	,15933

Paired Samples Test

		Paired Differences			
					95% Confidence Interval of the Difference
		Mean	Std. Deviation	Std. Error Mean	Lower
Pair 1	Post-test Scores - Pre-test Scores	,65385	1,71912	,33715	-,04052

Paired Samples Test

		Paired Differences	t	df	Sig. (2-tailed)
		95% Confidence Interval of the Difference			
		Upper			
Pair 1	Post-test Scores - Pre-test Scores	1,34822	1,939	25	,064

Situation	N	Mean	SD	t-score	p-value
Pre-test	26	5,5000	,81240	1,939	,064
Post-test	26	6,1538	1,75937		

© Italian e-Learning Association

Table 5 - Independent T-test results for comparison between the two groups.

Groups Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Difference	G1	43	2,4651	2,23978	,34156
	G2	26	,6538	1,71912	,33715

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Difference	Equal variances assumed	,947	,334	3,538	67	,001	1,81127	,51200	,78931	2,83323
	Equal variances not assumed			3,774	63,092	,000	1,81127	,47993	,85223	2,77031

allowing the use of the results under the assumption of equal variances. Descriptive statistics showed that Group 1 exhibited a higher mean improvement (M = 2.4651, SD = 2.23978) compared to Group 2 (M = 0.6538, SD = 1.71912). The independent t-test revealed a statistically significant difference between the two groups, t(67) = 3.538, p = 0.001. The mean difference was 1.81127, with a 95% confidence interval ranging from 0.78931 to 2.83323.

These findings suggest that the improvement observed in Group 1 was greater than that of Group 2. However, this result should be interpreted with caution, as the difference may be influenced by initial performance levels and other contextual factors not controlled for in the analysis.

Overall, these findings suggest that SANED_Anot+ may contribute to supporting students experiencing learning difficulties by providing targeted guidance and assistance, particularly for those with lower initial performance. However, these results should be interpreted with caution. While the intervention appears to benefit certain learner profiles, its effectiveness is not uniform across all groups. Furthermore, to the best of our knowledge, no video annotation systems have been reported in the literature within a context directly comparable to ours, particularly with regard to targeted support for struggling learners. This lack of comparable studies limits our ability to benchmark our findings against existing approaches and to draw broader generalizations.

The Table 6 provides a comparative overview of the key features of SANED_Anot+ and the tools discussed in Section 2.

6. Conclusion and future work

In recent years, increasing attention has been devoted to the use of annotations to enhance the quality of video-based online learning. Traditional video-based instructional approaches often suffer from limitations such as a teacher-centered design and a lack of interactivity and personalization. In this context, video annotation has emerged as a promising approach to enrich the learning experience by transforming passive content into interactive and multimodal educational resources. Such annotations may include explanatory texts, subtitles, comments, embedded questions, or links to additional materials, thereby fostering deeper engagement and understanding.

This study introduced SANED_Anot+, a video annotation prototype designed to enhance learners' engagement and academic performance. The system enables the customization of annotations according to learners' needs, allowing teachers to provide targeted explanations and personalized support, while also enabling learners to interact with the content through comments and feedback. By promoting active participation, the platform encourages a shift from passive viewing to a more interactive learning process, while offering learners a sense of guidance and continuous support.

The experimental evaluation aimed to assess the effectiveness of the proposed approach. The results suggest that SANED_Anot+ may contribute to improving the performance of learners experiencing difficulties, particularly through the provision of targeted feedback. More specifically, learners with low initial performance demonstrated significant improvements in their post-test results, highlighting the potential of the system to support struggling students and reduce learning disparities.

© Italian e-Learning Association

Table 6 - Comparison of advanced features of selected VATs to support learners with SANED_Anot+.

Analysis criteria	VideoAnt	EdPuzzle	CourseMapper	GoReact	Travis GO	ViDeX	FEVA	SANED_Anot+
Enhancement of Comprehension (Rey et al. 2020 ;)		√	√	√		√		√
Information retention (Chatti et al. 2016; Hager et al. 2020)			√	√				√
Reflection (Pérez-Torregrosa et al., 2017; Leung et al., 2021; Von Wachter et al., 2023)	√	√	√	√		√		
Accessible and flexible	√	√	√	√	√	√	√	√
Interaction and communication (Chatti et al. 2016; Lai et al. 2019; Leung et al. 2021; Onat et al. 2023 ; Von Wachter et al. 2023)	√	√	√	√	√	√	√	√
Feedback (Ardley et al. 2020 ; Leung et al. 2021 ; Lam et al. 2021 ; Von Wachter et al. 2023)	√	√	√	√				√
Personalisation of learning (Chatti et al. 2016)			√					√
Self-regulation (Mirriahi et al. 2018)			√					√
Engagement (Mirriahi et al. 2018 ; Rey et al., 2020 ; Lam et al. 2021)	√	√	√	√	√	√		√
Active learning (Kalboussi et al. 2015)	√	√	√	√	√	√		√
Collaboration (Chatti et al. 2016 ; Klug et al. 2021)	√		√		√		√	
Annotation sharing (Chatti et al. 2016; Hager et al. 2020)	√	√	√		√	√	√	√
Allow integration with LMS. (Chatti et al. 2016 ; Hager et al. 2020 ; Rahayu et al. 2022)		√	√	√				
Analytics (Chatti et al. 2016)		√	√					

Yes: ✓, no: blank

In contrast, no statistically significant improvement was observed for learners who had already achieved high pre-test scores, suggesting that the impact of the system may vary depending on learners' initial performance levels.

Despite these promising findings, several limitations should be acknowledged. First, the study was conducted within a single course and a single institution, which may limit the generalizability of the results to other educational contexts. Second, the effectiveness of SANED_Anot+ relies heavily on teachers' involvement in creating and managing annotations, which may increase workload and raise concerns regarding scalability. In addition, the system's effectiveness may depend on the level of learners' active participation, which can vary significantly across individuals.

Future work will focus on extending the evaluation to more diverse learning contexts and larger populations. Furthermore, the integration of artificial intelligence techniques will be explored to automate and optimize the annotation process. In particular, predictive analytics and machine learning approaches could be leveraged to generate adaptive and personalized feedback, thereby supporting teachers in tailoring their instructional strategies to individual learner needs.

References

Adler, M.V., Madsen, J., Hedberg, J., Steinberg, R.S., & Parra, L.C. (2025). Effect of explanation videos on learning: The role of attention and academic performance. *Education and Information Technologies*, *30*, 11797 - 11825. DOI:10.1007/s10639-024-13292-9

Ahn, D., & Chan, J.C. (2025). What Drives Student Engagement and Learning in Video Lectures? An Investigation of Instructor Visibility, Playback Speed, and Student Preferences. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.70026

Ardley, JN, & Hallare, M. (2020). The Feedback Cycle: Lessons Learned With Video Annotation Software During Student Teaching. *Journal of Educational Technology Systems*, 49, 112 - 94. DOI:10.1177/0047239520912343

Arora, M. (2023). Creating videos: a pedagogic tool for 21st century teachers of India. *Journal of e-Learning and Knowledge Society*, *19* (3), 82-86. https://doi.org/10.20368/1971-8829/1135856

Baber, A., & Fawad, B. (2022). The Impact of Educational Videos on the Academic Performance of University Students in Distance Learning.

© Italian e-Learning Association

Journal of Positive School Psychology 2022, Vol. 6, No. 12, 1233-1249. URL: https://journalppw.com/index.php/jpsp/article/view/14933

Bower, M. (2017). Design of Technology-Enhanced Learning: Integrating Research and Practice. DOI:10.1108/9781787141827

Bourbia, R., Drissi, S., & Lafifi, Y. (2024). How to provide intelligent assistance to learners in e-learning environments. *International Journal of Technology Enhanced Learning.* 16(1), 74-101. DOI:10.1504/ijtel.2024.135433

Brame, C.J. (2016). Effective Educational Videos: Principles and Guidelines for Maximising Student Learning from Video Content. CBE Life Sciences Education, 15. https://doi.org/10.1187/cbe.16-03-0125

Cassano, G., Di Blas, N., & Mataresi, A. (2024). Enhancing Learning Engagement in the Flipped Classroom using a Video-Annotation Tool. *Electronic Journal of e-Learning*. DOI:10.34190/ejel.22.9.3259

Chatti, M. A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Yousef, A. M. F., & Schroeder, U. (2016). Video annotation and analytics in Course Mapper. Smart Learning Environments, 3(1). https://doi.org/10.1186/s40561-016-0035-1

Chiu, P., Chen, H., Huang, Y. M., Liu, C. J., Liu, M. C., & Shen, M. (2016). A video annotation learning approach to improve the effects of video learning. Innovations In Education And Teaching International, 55(4), 459-469. https://doi.org/10.1080/14703297.2016.1213653

Coursera, (2026). "About coursera." https://about.coursera.org/.

Dong, C., Cao, S., & Li, H. (2020). Young children's online learning during COVID-19 pandemic: Chinese parents' beliefs and attitudes. Children and Youth Services Review, 118, 105440. https://doi.org/10.1016/j.childyouth.2020.105440

Edpuzzle [Internet]. Edpuzzle: About us [cited 15 Nov. 2023]. Available from: https://edpuzzle.com/

Fiorella, L., & Mayer, R.E. (2016). Eight Ways to Promote Generative Learning. Educational Psychology Review, 28, 717-741. DOI:10.1007/S10648-015-9348-9

Fong, M., Dodson, S., Zhang, X., Roll, I., & Fels, S.S. (2018). ViDeX: A Platform for Personalizing Educational Videos. Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries.JCDL'18. DOI:10.1145/3197026.3203865

Guo, P.J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: an empirical study of MOOC videos. Proceedings of the first ACM conference on Learning @ scale conference. DOI:10.1145/2556325.2566239

Hager, K.D., Fiechtl, B.J., & Gunn, S.L. (2020). Assessing Student Performance Using Video Recordings in Field-Based Experiences. DOI:10.26077/60EE-875F

Higuera, C.D.L., Prié, Y., Canellas, C.M., & Riou, M. (2014). Enrichissement de vidéos pédagogiques et apprentissage actif : le projet COCO. Informatique et MOOC, 24 juin 2014, Paris, https://www.societe-informatique-de-france.fr/wp-content/uploads/2014/05/2014-06-j-info-mooc-cmorais-canellas.pdf (accessed on September 2023)

Hosack, B. (2010). VideoANT: Extending Online Video Annotation beyond Content Delivery. TechTrends, 54, 45-49.

Hughes, C., Costley, J., & Lange, C. (2019). The effects of multimedia video lectures on extraneous load. Distance Education,40 (1), 54–75.

Hui, B., & Campbell, R. (2018). The discrepancy between learning and practicing digital citizenship. Journal of Academic Ethics, 16(2), 117-131

Kalboussi, A., Omheni, N., Mazhoud, O., & Kacem, A.H. (2015). An Interactive Annotation System to Support the Learner with Web Services Assistance. 2015 IEEE 15th International Conference on Advanced Learning Technologies,

Kay, R.H., & Kletskin, I. (2012). Evaluating the use of problem-based video podcasts to teach mathematics in higher education. Comput. Educ., 59, 619-627. DOI:10.1016/j.compedu.2012.03.007

Khurana, K., & Chandak, M.B. (2013). Study of Various Video Annotation Techniques. *Int. J. Adv. Res. Comput. Commun.* Eng. 2(1), 909–914 (2013)

Klug, D., & Schlote, E. (2021): Designing a Web Application for Simple and Collaborative Video Annotation That Meets Teaching Routines and Educational Requirements. DOI: 10.18420/ecscw2021_ep15

Lai, C., Chen, L., Yen, Y., & Lin, K. (2019). Impact of video annotation on undergraduate nursing students' communication performance and commenting behaviour during an online peer-assessment activity. *Australasian Journal of Educational Technology, 36*, 71-88. DOI:10.14742/ajet.4341

Lam, N. C. C., & Habil, H. (2021. a). The Use of Video Annotation in Education: A Review. Asian

© Italian e-Learning Association

Journal Of University Education, 17(4), 84. https://doi.org/10.24191/ajue.v17i4.16208

Lam, C. N. C., & Habil, H. (2021.b). Enriching Student Learning Through Video-annotated Peer Feedback Activity: A Guide. *International Journal of Academic Research in Progressive Education and Development*, 10(3), 46–60. DOI:10.6007/IJARPED/v10-i3/10712

Lange, C., & Costley, J. (2020). Improving online video lectures: learning challenges created by media. *International Journal of Educational Technology in Higher Education*, 17, 1–18

Laurillard, D. (2002). Rethinking University Teaching 2nd Edition: A conversational framework for the effective use of learning technologies. DOI:10.4324/9780203304846

Leung, K. C. C., & Shek, M. P. M. (2021). Adoption of video annotation tool in enhancing students' reflective ability level and communication competence. Coaching : An International Journal Of Theory, Research And Practice, 14(2), 151-161. https://doi.org/10.1080/17521882.2021.1879187

Love, B., Hodge, A., Grandgenett, N., & Swift, A. W. (2014). Student learning and perceptions in a flipped linear algebra course. *International Journal of Mathematical Education in Science and Technology, 45*(3), 317–324

Mirriahi, N., Joksimovíc, S., Gašević, D., & Dawson, S. (2018). Effects of instructional conditions and experience on student reflection : a video annotation study. *Higher Education Research And Development*, *37*(6), 1245-1259. DOI: 10.1080/07294360.2018.1473845

. Navarrete, E., Nehring, A., Schanze, S., Ewerth, R., & Hoppe, A. (2025). A closer look into recent video-based learning research: A comprehensive review of video characteristics, tools, technologies, and learning effectiveness. *International Journal of Artificial Intelligence in Education*, *35*(4), 1631-1694. http://dx.doi.org/ 10.1007/s40593-025-00481-x

Onat, O. & Gulsecen, S. (2023). Reducing perceived transactional distance in distance education: The impact of the chatbot. International Journal of Technology in Education and Science (IJTES), 7(4), 483-499. https://doi.org/10.46328/ijtes.511

Pérez-Torregrosa, A.B., Díaz-Martín, C., & Ibáñez-Cubillas, P. (2017). The Use of Video Annotation Tools in Teacher Training. Procedia - Social and Behavioral Sciences, 237, 458-464. https://doi.org/10.1016/j.sbspro.2017.02.090

Rahayu, E.M., & Bhaskoro, P. (2022). Interactive Media Edpuzzle and Its Implementation in Teaching Vocabulary in New Normal Era. Jo-ELT (Journal of English Language Teaching) Fakultas Pendidikan Bahasa & Seni Prodi Pendidikan Bahasa Inggris IKIP.

Rasenberg, E. M. C., Brand, G. M., & Ellemieke, B. P. T., et al. (2025). How do undergraduate medical students use the annotation option of a video feedback system when recording consultations with real patients? A qualitative document analysis. *BMC Medical Education,* **25**, 809. https://doi.org/10.1186/s12909-025-07405-2

Rey, F.J., Galán, R.P., Robles, V.C., & Cebrián de la Serna, M. (2020). The video guides and social tagging with the methodology of video annotation. 2020 X International Conference on Virtual Campus (JICV), 1-3.DOI:10.1109/JICV51605.2020.9375642

Shrestha, S., Sentosatio, W., Peng, H., Fermuller, C., & Aloimonos, Y. (2023). FEVA: Fast Event Video Annotation Tool. ArXiv, abs/2301.00482.

Statista, (2023). DataReportal Survey; GWI. URL: https://www.statista.com/statistics/1254810/top-video-content-type-by-global-reach/ (accessed on October 2023)

Stigler, J. W., Geller, E. H., & Givvin, K. B. (2015). Zaption: A Platform to Support Teaching, and Learning about Teaching, with Video. *Journal of E-Learning and Knowledge Society*, *11*(2). https://doi.org/10.20368/1971-8829/1042

Susantini, E., Indana, S., Isnawati, I., & Sayitri, S.D. (2019). Developing Instructional Video to EnhanceBiology Pre-Service Teachers' Metacognitive Skills. Proceedings of the Mathematics, Informatics, Science, and Education International Conference (MISEIC 2019). DOI: 10.2991/miseic-19.2019.35

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory. Springer Science & Business Media.

Tseng, SS. (2021). The influence of teacher annotations on student learning engagement and video watching behaviors. Int J Educ Technol High Educ 18, 7 (2021). https://doi.org/10.1186/s41239-021-00242-5

Van Drom A. (2018). 3 Plateformes en ligne pour créer des contenus vidéo interactifs pour les étudiants. URL : https://eductive.ca/ressource/3-plateformes-en-ligne-pour-creer-des-contenus-video-interactifs-pour-les-etudiants/ (accessed on October 2023)

Vandenberg, S., & Magnuson, M. (2021). A comparison of student and faculty attitudes on the use of Zoom, a video conferencing platform: A mixed methods study. Nurse Education in Practice, 54. https://doi.org/10.1016/j.nepr.2021.103138

© Italian e-Learning Association

Voronkin, O. (2019). Educational Video in the University: Instruments, Technologies, Opportunities and Restrictions. International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications. https://ceur-ws.org/Vol-2387/20190302.pdf

Von Wachter, J., & Lewalter, D. (2023). Video Annotation as a Supporting Tool for Video-based Learning in Teacher Training – A Systematic Literature Review. International Journal Of Higher Education, 12(2), 1. https://doi.org/10.5430/ijhe.v12n2p1

Yousef, A. M. F., Chatti, M. A., Danoyan, N., Thüs, H., & Schroeder, U. (2015). Video-Mapper: A Video Annotation Tool to Support Collaborative Learning in MOOCs. Proceedings of the Third European MOOCs Stakeholders Summit EMOOCs 2015. pp. 131-140.

© Italian e-Learning Association

JOURNAL OF e-LEARNING AND KNOWLEDGE SOCIETY

www.je-lks.org

VOLUME 22 | ISSUE NO. 1 | APRIL 2026

AN INTERNATIONAL AND OPEN ACCESS JOURNAL
BY THE ITALIAN E-LEARNING ASSOCIATION (SIE-L)

www.sie-l.it

ISSN (online) 1971 - 8829 | ISSN (paper) 1826 - 6223