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

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

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).

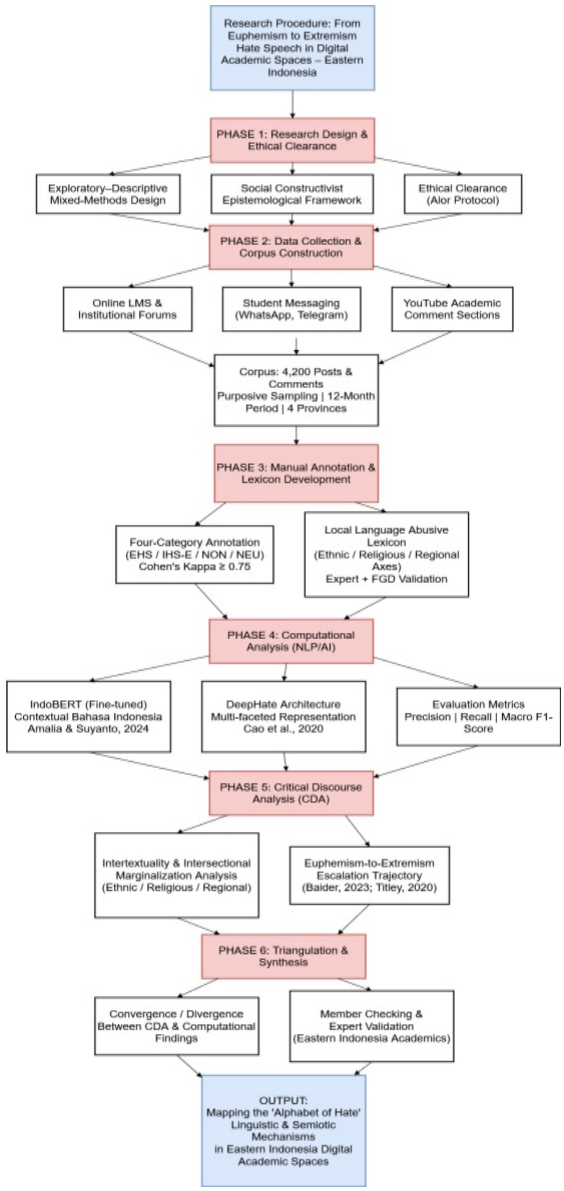

Figure 1 - Six-Phase Research Procedure: From Euphemism to Extremism in Eastern Indonesia Digital Academic Spaces.

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).

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

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

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

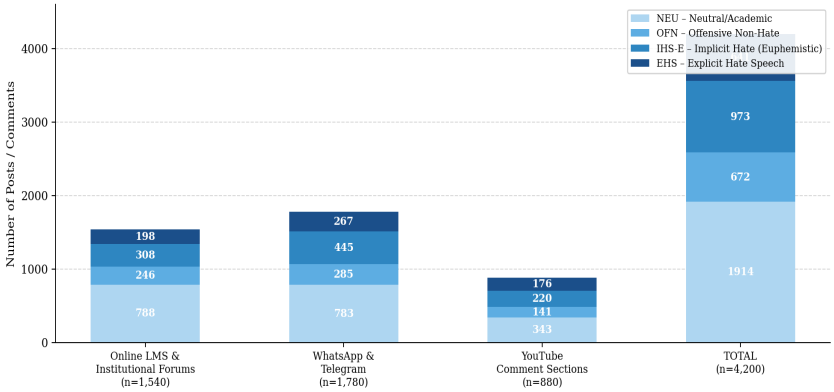

Figure 2 - Corpus Label Distribution by Platform and Category (n=4,200).

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

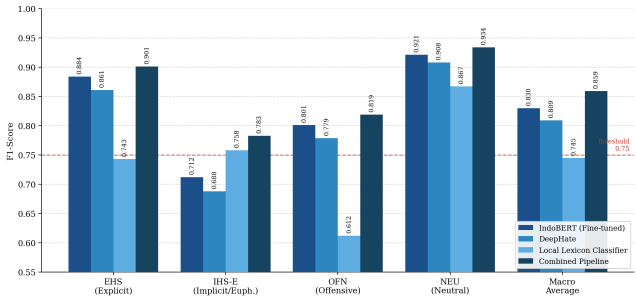

Figure 3 - F1-Score by Model and Category (IHS-E = primary evaluation criterion).

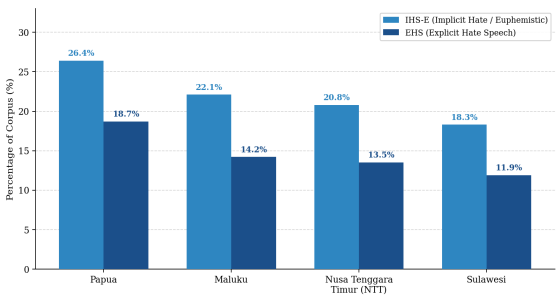

Figure 4 - Prevalence of IHS-E and EHS by Province (%).

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

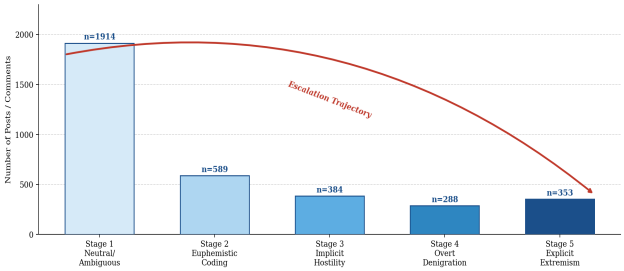

Figure 5 - Distribution of Posts Across the Five-Stage Euphemism-to-Extremism Escalation Trajectory.

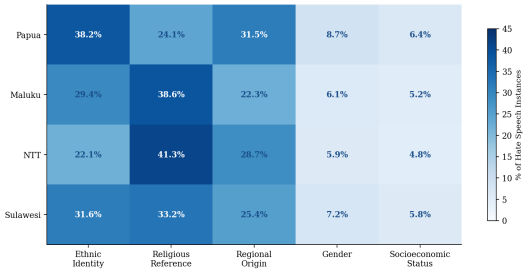

Figure 6 - Intersectional Axes of Hate Speech by Province (%) Based on CDA Sub-Corpus (n=320).

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).

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

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

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

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