Jurnal Audience
JURNAL ILMU KOMUNIKASI
Vol 09 No. 02 : 2026

P-ISSN: 2620-8393
E-ISSN: 2685-8010

Examining Hate Speech and Social Media in Indonesia: A Systematic Literature Review

Ujaran Kebencian dan Media Sosial di Indonesia: Tinjauan Pustaka

Gemilang Bayu Ragil Saputra[1], Nurliana[2], Dwi Rohma Wulandari[3], Dyah Fitria Kartika Sari[4], Kudratullah[5]

[1,3,4,5]Department of Communication Studies, Faculty of Social and Political Sciences, Tadulako University, Palu, Indonesia
[2]Department of Communication Studies, Faculty of Social and Political Sciences, Abdul Azis Lamadjido University, Palu, Indonesia

[1,3,4,5]Jl. Soekarno Hatta Km. 9. Kota Palu, Indonesia
Email: [1]gbragils@untad.ac.id

Received : June 28, 2026 ; Revised: July 28, 2026; Accepted: August 25, 2026

Abstract

Hate speech and social media are increasingly prominent topics for scholars in Indonesia, yet prior reviews on the subject tend to remain within a single discipline. Building on the digital boom triggered by the Covid-19 pandemic, we systematically mapped the thematic trends of hate speech and social media research in Indonesia published between 2020 and 2025. We queried our corpus from ScienceDirect and utilised R, RStudio, and the Bibliometrix package as the primary tools to analyse the corpus. Our findings show that this remains a rapidly growing field (CAGR of 45.4% per year), with hate speech and offensive content detection forming its conceptual core, and deep learning and language models. Dominant methods cover deep learning and language models such as BERT and its Indonesian derivative. The thematic evolution shifted from situational triggers of Covid-19 pandemics to long-term structural issues such as hate speech itself and disinformation. Interpreting these patterns through Computer-Mediated Communication (CMC) and the Social Identity model of Deindividuation Effects (SIDE) as primary lenses, with Social Identity Theory (SIT) as a secondary lens, we argue that platform affordances and anonymity intensify group-based hostility in ways the literature has yet to explicitly theorise, particularly within Indonesia's distinct religious and linguistic context.

Kata Kunci: Clustering; Hate Speech; Social Media; Systematic Literature Review; Thematic Mapping

Abstrak

Ujaran kebencian dan media sosial semakin menjadi topik yang menonjol bagi para akademisi di Indonesia, namun tinjauan sebelumnya tentang subjek ini cenderung tetap dalam satu disiplin ilmu. Berdasarkan ledakan digital yang dipicu oleh pandemi Covid-19, kami secara sistematis memetakan tren tematik penelitian ujaran kebencian dan media sosial di Indonesia yang diterbitkan antara tahun 2020 dan 2025. Kami meng-query korpus kami dari ScienceDirect dan menggunakan R, RStudio, dan paket Bibliometrix sebagai alat utama untuk menganalisis korpus tersebut. Temuan riset menunjukkan bahwa ini tetap menjadi bidang yang berkembang pesat

(CAGR sebesar 45,4% per tahun), dengan deteksi ujaran kebencian dan konten ofensif membentuk inti konseptualnya, serta pembelajaran mendalam dan model bahasa. Metode dominan mencakup pembelajaran mendalam dan model bahasa seperti BERT dan turunannya dalam bahasa Indonesia. Evolusi tematik bergeser dari pemicu situasional pandemi Covid-19 ke isu-isu struktural jangka panjang seperti ujaran kebencian itu sendiri dan disinformasi. Menginterpretasikan pola-pola ini melalui Komunikasi Berbasis Komputer (CMC) dan model Identitas Sosial dari Efek Deindividuasi (SIDE) sebagai lensa utama, dengan Teori Identitas Sosial (SIT) sebagai lensa sekunder, kami berargumen bahwa kemampuan platform dan anonimitas memperburuk permusuhan berbasis kelompok dengan cara yang belum secara eksplisit dijelaskan dalam literatur, terutama dalam konteks agama dan bahasa yang khas di Indonesia.

Keywords: Klustering; Media Sosial; Pemetaan Tematik; Tinjauan Pustaka Sistematis; Ujaran Kebencian

1. Introduction

Digital transformation has reconfigured The advances of information and communication technologies, in a measurable manner, are shifting and affecting how humans interact and inform each other which also includes how we shape public opinions (AlKhudari et al., 2024; Ausat, 2023; Shah, 2024; Swastiningsih et al., 2024). Thus, it can be said that social media became one of the most significant products of the digital revolution of our civilisation. By using the internet and social media specifically, public space was expanded to not only physical but digital as well. If we look at the global populace of social media, the existing users reached 5,24 billion with 4 percent or 200 million new users annually. The numbers were also reflected in Indonesia, as a similar trend was also happening. As of 2025, there were 143 million social media users in Indonesia. A number which exceeds half of the total population (We Are Social & Meltwater, 2025). Thus, the complex and massive scale of social media made it a recreational, entertainment, and communication space for personal use as well as the main infrastructure of today's public discourse, political opinion and social identity battlegrounds (Literat & Kligler-Vilenchik, 2019; Muhammad, 2024; Shah, 2024).

The significance of social media in our contemporary social life is academically proven to be connected to the negatives of social life. When a social space was duplicated and migrated into digital space, the social dynamics will also be migrated into the aforementioned space (Abdel-Aziz et al., 2016; Hurova & Shkurov, 2023; Kääntä, 2025). This includes conflict potentials, prejudices, and intergroup hostility. Social media is structurally different in characteristics if we compare it with conventional media. The issue of anonymity, speed of information, content personalisation and algorithms, and the lack of direct moderation collectively creates a condition in which hate speech can be used widely and uncontrollably (Bioglio & Pensa, 2022; Giumetti & Kowalski, 2022; Jaidka et al., 2021; Yeh & Swinehart, 2022).

The use of hate speech in social media brought forth consequences for individuals, groups, and all of the users as a collective. The repeated exposure of hate speech on an individual will impact said individual psychologically, from anxiety and emotional pressure to trauma (Aldamen, 2023; Hassim et al., 2024; Shmulewitz et al., 2025; Stahel & Baier, 2023). As a collective, hate speech can deepen social polarisation, which erodes group cohesion and triggers intolerance and structural discrimination (Almeida et al., 2023; Blanco-Herrero et al., 2024; Giraud et al., 2025). If we look at political contexts, studies show that hate speech is often used as a political instrument to mobilise support, construct identity boundaries, and influence public policy discourse (Kentmen-Cin, 2025; Romero-Rodríguez et al., 2023; Solovev & Pröllochs, 2023). It can be said that hate speech has the potential to threaten social and democratic stability. In more extreme cases, hate speech is contributing to real-world conflict escalation. For example, communal conflict through hate-motivated violence against a particular group (Allwood et al., 2022; Ceccato & Simões Agapito, 2024; Chetty & Alathur, 2018).

Responding to the urgency of widespread hate speech in social media and considering the current state of knowledge includes extreme lexical diversity, assymetric conceptual structure, and theme clustering, we set on exploring the thematic trends of the studies on hate speech and social media. While systematic literature reviews on this topic are not something new, more often than not it focuses on one discipline separately. Other considerations include the Covid-19 pandemics with forced global digitalization rapidly.

Based on the described data on the existing condition, we aim to explore and map the thematic trends of studies on hate speech and social media in Indonesia in a more specific manner between 2020 and 2025. Through this exploration and trend mapping, we hope that this study can comprehensively illustrate the direction of development of social media and hate speech studies in the last five years, as well as serve as a reference for researchers, policymakers, and platform developers in formulating strategic steps to reduce the prevalence and negative impact of hate speech in the digital space.

2. Theoretical Framework

Considering the impacts, we ought to understand what hate speech is. Hate speech is a verbal and nonverbal expression that is demeaning, attacking, or discriminating against individuals or groups based on certain attributes, such as race, ethnicity, religion, gender, disability, or sexual orientation (Ghenai et al., 2025; Hietanen & Eddebo, 2023; Vergani et al., 2024). In practice, the boundaries between freedom of expression, social critiques, satire, and hate speech are a difficult discourse to simply resolve. The simple but complex definition of hate speech needs to be addressed, as clear conceptual boundaries means more accurately applicable.

The United Nations defines hate speech as any form of communication that attacks or uses derogatory and discriminatory language against a person or group based on their inherent identity (United Nations, 2019). This definition not only gives more flexibility for any social and cultural context but also opens broad interpretation among scholars, thus creating a dynamic discussion on the conceptualisation.

Aside from hate speech, social media as a medium of communication also needs to be understood. Social media as a digital platform operates on different logic compared to conventional media

(Bennett & Segerberg, 2023; Karanasios et al., 2024; Ren et al., 2024; Walters, 2022). Social media was built upon user participation, social networking, and content personalisation (Avalle et al., 2024; Gerbaudo, 2026; Kamboj & Sharma, 2023; Miranda et al., 2023; Salma et al., 2024; Teepapal, 2025). This instance can be seen with the use of algorithm-based content recommendations that maximise users' participation and social media use which pushes emotional, provocative, and controversial content to be more massively distributed (Gerbaudo, 2026; Salma et al., 2024; Teepapal, 2025). This characteristic made social media a structural actor that shapes interaction patterns and discourses instead of acting as a passive channel of communication (Laor, 2022; Rodilosso, 2024; Törnberg, 2022; Waworuntu et al., 2022). One aspect to note is that social media is not a homogenised single entity but consists of different platforms with different cultures and mechanisms.

The understanding of hate speech and social media can be intertwined in a more comprehensive framework. The connection between hate speech and social media is not a coincidental relation but a structural one where technical designs of the social media platform facilitate, accelerate, and augment the impacts of hate speech (Almeida et al., 2023; Barth et al., 2023; Maarouf et al., 2024; Solovev & Pröllochs, 2023).

Computer-Mediated Communication (CMC) is a theoretical thinking within Communication Studies that examines how digital media structurally alter the ways messages are produced, disseminated, and interpreted. Key concepts in CMC include "reduced social cues" or the diminution of nonverbal social signals, the hyperpersonal model (Walther, 1996), which explains how mediated communication can actually intensify certain affects compared to face-to-face interaction, and platform affordances (Boyd, 2010) which elaborate on technical properties such as persistence, scalability, searchability, and replicability that shape user behavior. Social media operates on a different logic than conventional media and acts as a structural actor shaping interaction patterns which was anchored in CMC.

The Social Identity model of Deindividuation Effects (SIDE) (Reicher et al., 1995; Spears, 2017; Vilanova et al., 2017) is a specific development within CMC explaining that anonymity and the scarcity of personal identity cues in digital spaces do not erase user identity; rather, they heighten the salience of group identity when group norms are prominent in the interaction context. This challenges the long-held assumption that anonymity leads to a loss of self of classic Zimbardo-style deindividuation. SIDE argues that anonymity shifts the focus from personal identity to social or group identity, causing behavior to become more polarized in alignment with group norms, including hostile behavior toward out-groups. This framework is directly relevant to anonymity, information speed, content personalization and algorithms, and the lack of direct moderation as conditions that facilitate the widespread and uncontrolled spread of hate speech. Notably, (Jaidka et al., 2021) study explicitly examines network affiliation under conditions of deindividuation.

The literature on hate speech and social media generally follows two paths that rarely intersect. The first path consists of computational detection research (machine learning or deep learning-based hate speech classification) which focuses on model accuracy rather than explaining the social origins of the phenomenon. The second path comprises communication and social research addressing the impacts of hate speech such as polarization, psychological trauma, and conflict escalation. But rarely explicitly linking these to CMC or SIDE mechanisms that explain how platform architectures generate such conditions. This gap is precisely why the literature including the previous systematic literature reviews mentioned in Introduction, tends to focus on disciplines in isolation. The computational field examines what can be detected, while the social field examines the impacts.

Social Identity Theory (SIT) (Jost & Sidanius, 2016) explains that individuals derive part of their self-concept from group membership and tend to engage in in-group and out-group categorization to maintain collective self-esteem that also includes the derogation of other groups. SIT serves as a secondary lens explaining the psychosocial motives behind hate speech, while CMC & SIDE explains the medium-related mechanisms that facilitate and intensify these motives in the digital space. This combination aligns with the literature cited and elaborated in Introduction regarding social polarization, the erosion of group cohesion, and political instrumentalization (Almeida et al., 2023; Blanco-Herrero et al., 2024; Giraud et al., 2025; Kentmen-Cin, 2025; Romero-Rodríguez et al., 2023; Solovev & Pröllochs, 2023).

Since our study is a bibliometric systematic literature review (SLR) rather than a direct content analysis of hate speech, CMC, SIDE and SIT are not used to analyze individual texts. Instead, they serve as an interpretive framework for interpreting the thematic patterns and cluster structures generated by Bibliometrix. In other words, the clusters, node centralities, and thematic evolution are interpreted as indicators of the extent to which existing literature has or has not empirically operationalized CMC, SIDE and SIT.

3. Research Method

This study uses a Systematic Literature Review (SLR) approach to identify, evaluate, and synthesise relevant academic papers through systematic, transparent, and replicable steps. This approach differs from a narrative literature review, which tends to be more selective and subjective. SLR enforces a query protocol, inclusion and exclusion criteria, and explicit data screening. Thus, both the process and the results can be retraced and retested. We choose this approach based on the aims of the study, which is to explore and map the thematic trends of studies focusing on hate speech and social media in Indonesia from 2020 to 2025.

To ensure the systematisation of the review process, we chose Preferred Reporting Items for Systematic Reviews and Meta-Analyses or PRISMA, as the primary framework. PRISMA was chosen as its framework provided a standardised structure for data identification, screening, eligibility assessment, and inclusion. The standardised structure allows more transparent and accountable reporting (Page, McKenzie, Bossuyt, Boutron, Hoffmann, Mulrow, Shamseer, Tetzlaff, & Moher, 2021; Page, McKenzie, Bossuyt, Boutron, Hoffmann, Mulrow, Shamseer, Tetzlaff, Akl, et al., 2021).

Data were collected from Scopus' academic database (ScienceDirect). ScienceDirect was chosen for its scope and position as a reputable database index. A query was conducted using Boolean keywords "social media" AND "hate speech" AND "indonesia" to ensure the collected data discussed all three keywords. Data were limited to publications in 2020 through 2025. This limitation considers the pandemic, which globally changed the digital landscape. Our initial query resulted in 128 entries.

Collected data then forwarded to the screening stage of PRISMA. In this stage, we focused on deduplicating the collected data and evaluating the completeness of the metadata to ensure that one entry was evaluated once and no entry was affected by missing metadata.

We excluded 4 entries for duplication and incompleteness of the metadata, resulting in 124 entries to be forwarded into the eligibility assessment stage.

We performed two stages of eligibility assessment. First is document type exclusion. From 124 entries, we excluded 37 entries as non-research articles, including review articles, encyclopedias, book chapters, data articles, discussion, editorials, mini reviews, and others. This step is crucial as this study only focuses on empirical findings. Thus, other article entries that are summaries, opinions, or not the result of direct research are not relevant for inclusion in the analysis corpus. The second step of eligibility assessment is done by filtering the publication based on access status. We excluded 44 entries which were not open access. The aforementioned step ensured that the corpus consisted of open access entries. Thus, data extraction and synthesis can be done fully without limitation to the full-text articles. Every data entry that fulfilled all the criteria was then forwarded to the inclusion stage in the corpus. We employ tool-assisted analysis using R, RStudio, and the Bibliometrix package (Aria & Cuccurullo, 2017, 2026).

4. Findings and Discussion

Our findings showed that studies on social media and hate speech experienced an upward trend (see table 1 and fig. 2). If we look at the compound annual growth rate, we can see that average annual growth is 45,4%. It indicates a rate that is considered very high. From this initial finding, the growth can be an early indicator that this topic is a rapidly developing research field.

Table 1. Annual scientific production

Publication Year	Articles
2020	2
2021	5
2022	2
2023	9
2024	12
2025	13
Total	**43**

We then identified the most frequent words by identifying bigram phrases in abstracts and phrases in keywords. We identified 1.921 unique phrases from the abstracts and 196 unique phrases from the keywords. Figure 3 presents the 10 most frequent phrases inferred from the abstracts. In figure 3, we can see that "social media" significantly dominates with 53 occurrences, more than four times higher compared to “hate speech” (14).

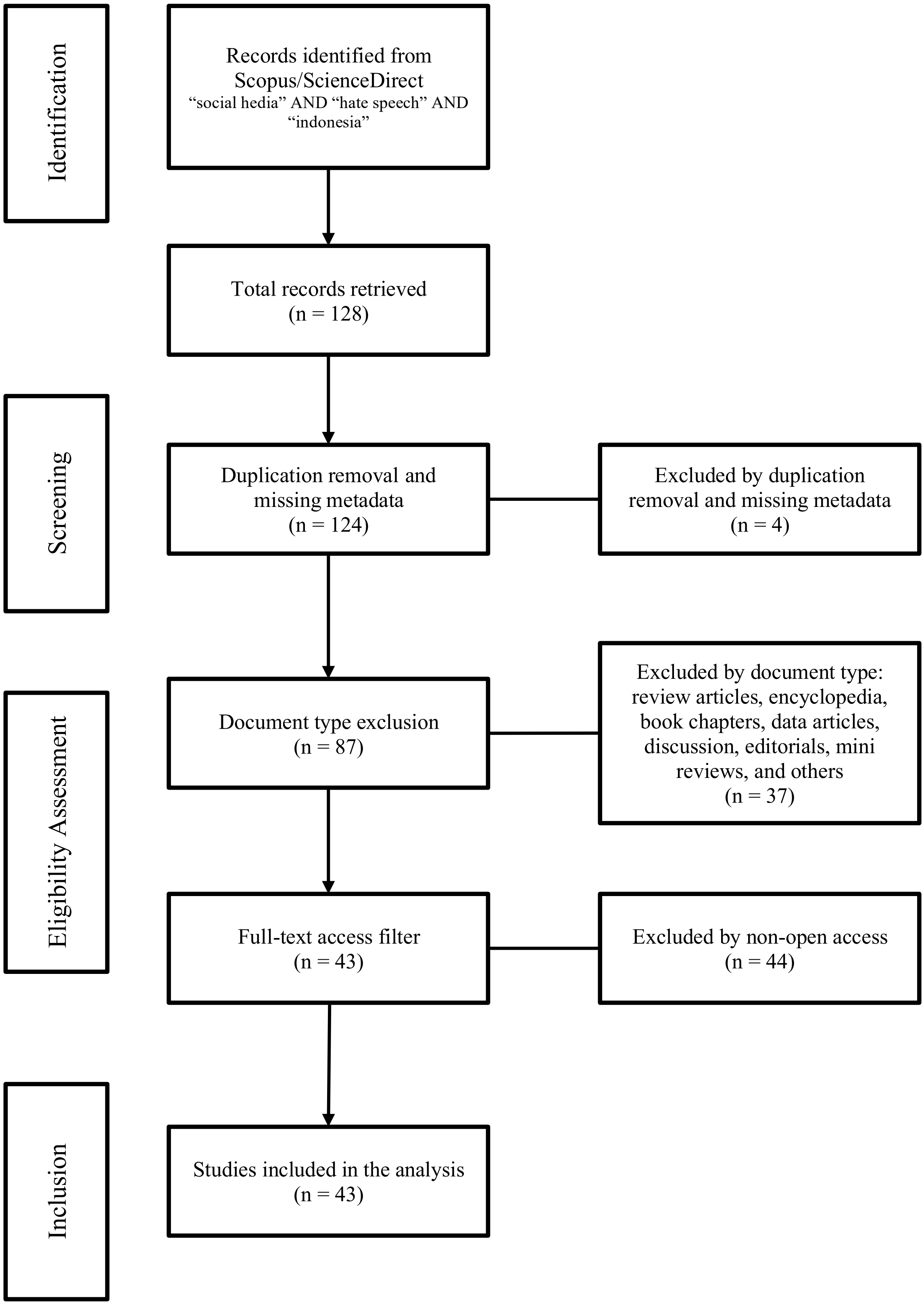

Figure 1. PRISMA diagram

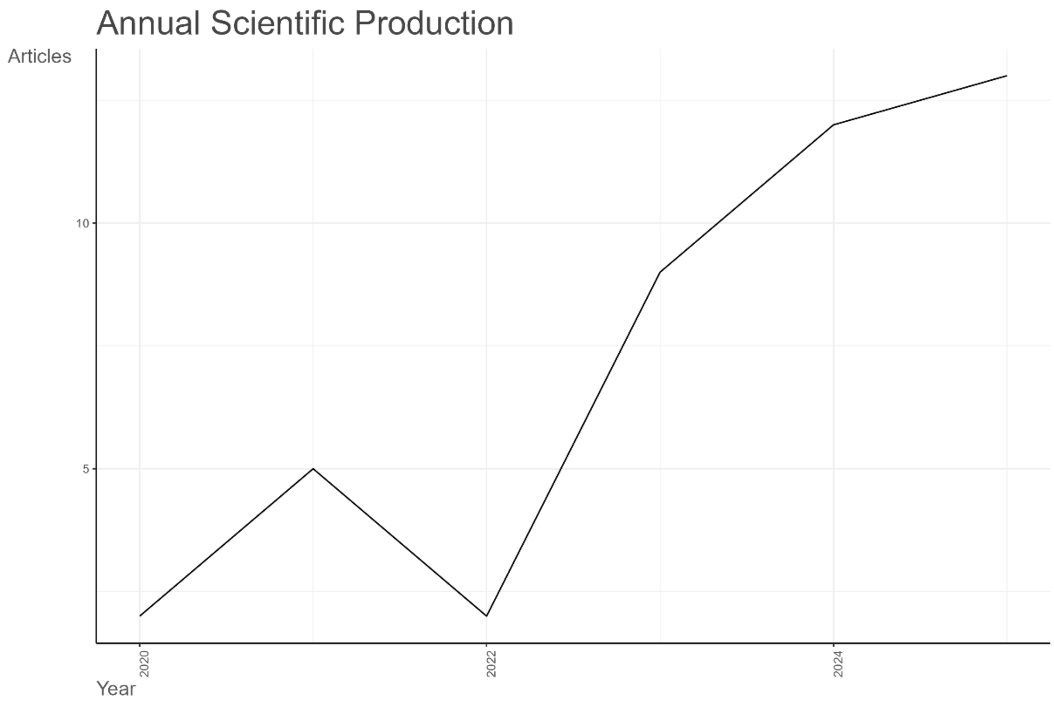

Figure 2. Annual scientific production diagram

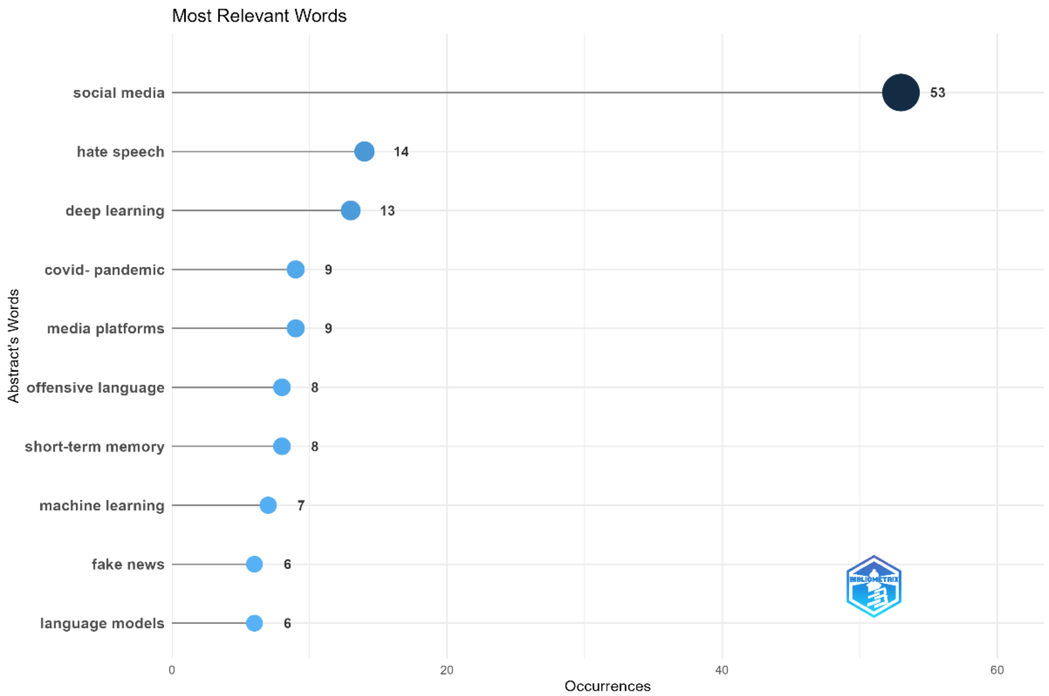

Figure 1. Most relevant words by frequency, inferred from abstract

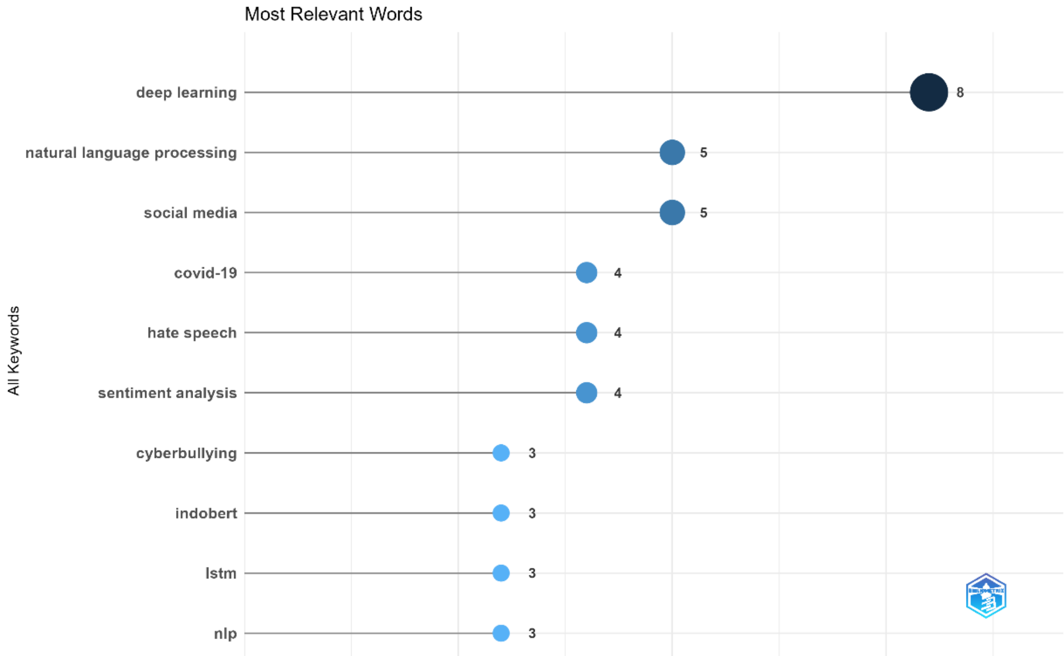

Figure 3. Most relevant words by frequency, inferred from keywords

The emergence of technical terms such as short-term memory which refers to Long Short-Term Memory or LSTM, affirmed the emergence of deep learning in studies concerning social media and hate speech. Uniquely, we found the term pp lending (peer to peer lending) which refers to an online lending service that is booming in Indonesia. The long tail of 1.921 unique phrases has a very low frequency of emergence (1 to 3), which points out the lexical diversity and broad interdisciplinary topics.

In contrast to the previous finding which was based on words in the abstract, we found that the top position of the most frequent words is “deep learning” (8) instead of “social media”. It indicates that research approach is more often be used as a primary identity or topic compared to the object of the study. The emergence of the term “indobert”, which is an Indonesian variant of the BERT model, confirmed the contribution of language model-based studies in this corpus. The long tail of the 196 phrases we identify also consisted of singleton keywords which portray the interdisciplinary nature of the corpus.

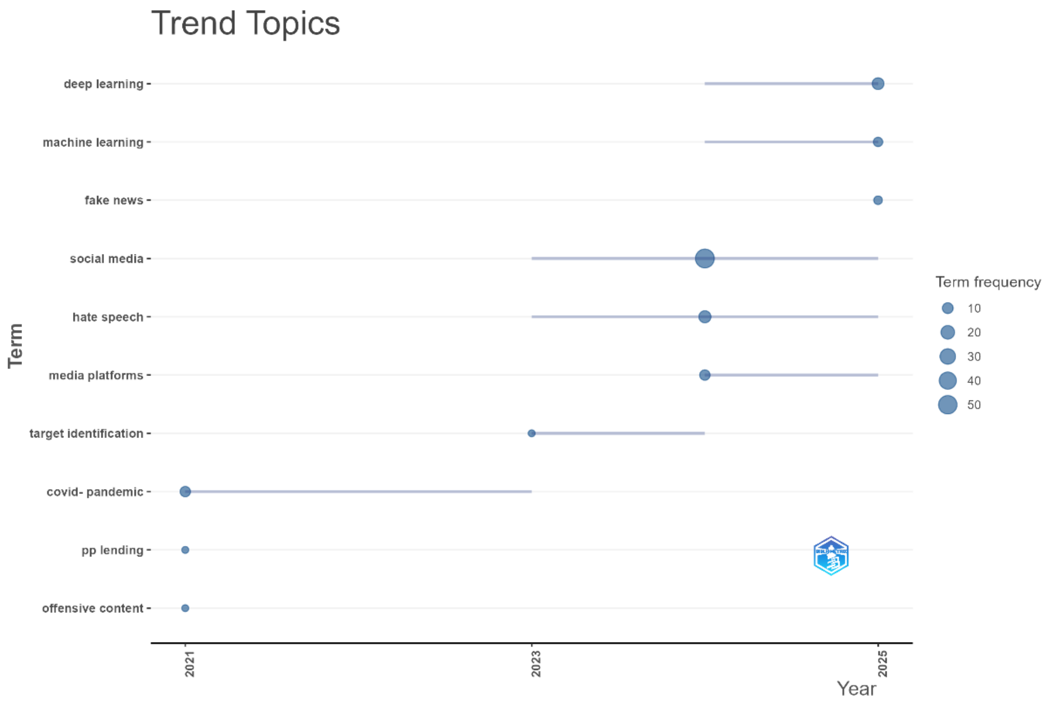

Figure 4. Trend topics spread (2020-2025)

Figure 5 shows a thematic shift in three phases. The first phase (2020-2023) is dominated by the Covid-19 pandemic issue, peer-to-peer lending, and offensive content. This phase indicates that the first wave of studies was triggered by the Covid-19 context and the socioeconomic problems which accompany the pandemic, especially illegal lending practices due to economic issues. The median phase (2023-2024) is marked by the emergence of focused topics such as target identification, social media, and hate speech. While the latest phase (2024-2025) shows another focus shift to media platforms, deep learning, machine learning, and fake news.

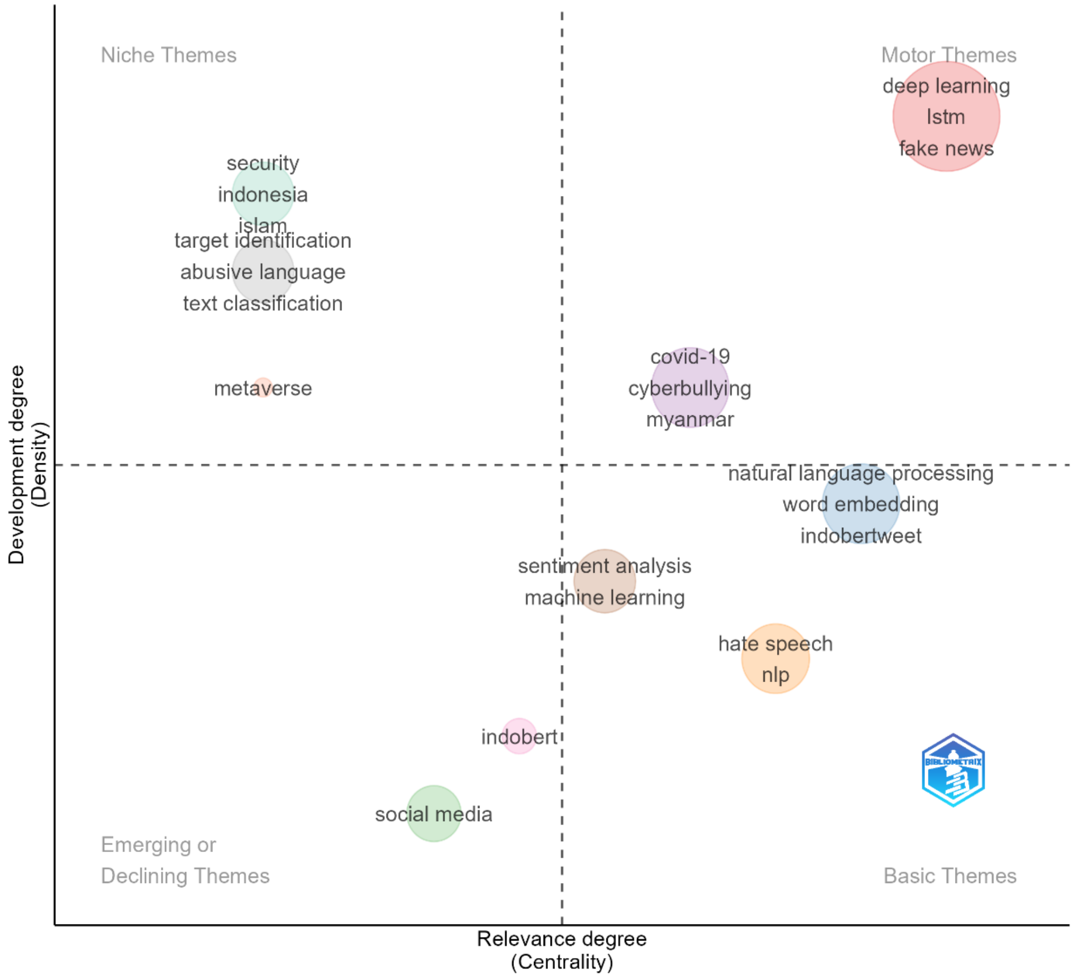

Figure 5. Thematic mapping

From the identified words and phrases, we map our findings in a co-occurrence network in figure 6. Our mapping showed that there are three active clusters with different degrees of centrality. Cluster 1 (red) is the biggest with 16 nodes. The highest centrality node is social media (betweenness = 288,80; closeness = 0,032; PageRank = 0,204), followed by Covid pandemic (betweenness = 27,0), media platforms (betweenness = 11,0), and hate speech (PageRank = 0,050). Other nodes in this cluster are offensive language, target identification, labelled data, media activity, online harassment, social networks, southeast asian, study investigates, bidirectional encoder, classification model, encoder representations, and hateful speech.

Cluster 2 (blue) is the second biggest cluster with 9 nodes. This cluster points to the node deep learning as the most central node of this cluster (betweenness = 38,86; PageRank = 0,074). This cluster also indicates the relations to more technical topics with its correlation on language models such as short-term memory (LSTM), learning models, abusive language, language detection, learning techniques, and sentiment analysis.

While cluster 3 (green) is the smallest, it's also the most focused on machine learning and natural language processing (NLP) specifically. The aforementioned argument is signified by the existence of the node machine learning followed by language models, natural language, and language processing. The structure of these three clusters shows a fairly neat separation between the objects of study (social media and hate speech in Cluster 1), specific/advanced methods (deep learning in Cluster 2), and scientific foundations (classical machine learning or NLP in Cluster 3).

We also identified 10 thematic clusters which were distributed into four groups: motor themes, basic or transversal themes, niche themes, and emerging or declining themes. This distribution can be seen in figure 7 and table 2. Motor themes indicate two primary themes with high centrality and density, which are deep learning and Covid-19. Deep learning and Covid-19 became themes that have matured and become driving forces of the structure of the field.

Table 2. Thematic clustering details

Cluster label	Words by association	Occurrence / frequencey
1. deep learning	deep learning, lstm, cnn, fake news, neural networks, youtube	19
2. natural language processing	natural language processing, indobertweet, word embedding	9
3. social media	social media	5
4. covid-19	covid-19, cyberbullying, myanmar	9
5. hate speech	hate speech, nlp	7
6. sentiment analysis	sentiment analysis, machine learning	6
7. indobert	indobert	3
8. target identification	abusive language, target identification, text classification	6
9. security	indonesia, islam, security	6
10. metaverse	metaverse	2

Basic themes exhibit three topics: natural language processing, hate speech, and sentiment analysis. These topics have high centrality but low density. Thus, this theme indicates important themes which closely related to other themes but not yet developed in depth.

Niche themes encompass three primary topics: target identification, security, and metaverse. The topics within niche themes have low centrality but high density. This indicates that the themes are developing in depth but isolated and specific.

Emerging and declining themes cover two topics: social media and indobert. Both topics have low centrality and low density. It indicates that the aforementioned themes are still in the early stages or are on the fringes of the field.

If we look at all of the presented data as a narrative unit, we found consistent emerging patterns that empower one another. The annual scientific production data shows the growth of publications in 2022. The annual peak (13) in 2025 strongly indicates that the topics of hate speech and social media were right at the peak of their growth rate at the time this data was collected. It implies that specific topics such as the detection of hate speech and offensive content based on deep learning in the context of Indonesia and Southeast Asia are active and receptive to new findings. Although the projected models may also imply a plateau of growth rate following a life cycle pattern commonly found in research fields that grow from niche topics to scientific maturity, as is common with technology and scientific trends that follow the innovation diffusion curve. We ought to be careful in interpreting the model as it was extrapolated from a limited amount of corpus (43 entries) and time frame (2020-2025). The publication dip in 2022, flanked by spikes in 2021 and 2023, likely reflects a natural lull following the initial wave of research related to the Covid-19 pandemic.

The topic may be rediscovered and reinforced by a massive wave of deep learning and language model-based research since 2023. In other words, the annual scientific production figure (fig. 2) should not be read as an absolute limit, but rather as a projection of the saturation point based on the growth pattern observed so far. If new methodological breakthroughs such as the adoption of large generative language models or geographic expansion of the scope emerge, the actual growth curve could potentially deviate significantly from this projection.

The findings based on the most frequent words (both abstract and keywords) and co-occurrence network consistently confirm the existence of three conceptual layers. The first layer is the object of study, which includes social media, hate speech, and offensive content. The second layer covers the methods and approaches of dominant research, such as deep learning and its derivatives such as LSTM, CNN, and BERT. This layers emphasize on the methodological contributions instead of the subject of the studies. The third layer includes more general academical foundation and technical sciences such as machine learning and NLP, which forms separate cluster that acted as a conceptual bridge between the object (first layer) and method (second layer).

This three-layer pattern is also in line with our thematic mapping. Deep learning emerges as a motor theme with high centrality and density, which indicates that this method is not only often discussed but has also matured internally with many closely related sub-techniques, such as LSTM, CNN, and neural networks. Deep learning also became the central point and bridging of topics. Uniquely, the social media cluster emerges as the most frequent in the entire corpus but became one of the emerging themes on the thematic mapping. This occurrence can be explained as the term “social media” acted as an umbrella term which appeared at almost all points in the documents within the corpus without correlating with other subgroups or clusters. Thus, social media became a term which is spread thinly in most clusters other than forming its own dense cluster. This phenomenon is common to generic terms in bibliometric thematic map analysis.

One of the most distinctive findings of this corpus is the representation of research based on the context of Indonesia and its surrounding region. The emergence of BERT derivative terms (indobert and indobertweet) and peer-to-peer lending (which is relevant to fintech regulation in Indonesia) showed that this corpus has a strong local and regional research content. Indonesian language-based content detection and cybersecurity issues also emerge from the utilisation of NLP and algorithms in content moderation. The emergence of the term Islam also depicted the topic of religion moderation in social media, which is considered a sensitive topic in Indonesia.

This pattern suggests that the field of hate speech and offensive content detection research is no longer dominated by English-language case studies alone, but has evolved into multilingual and multicultural studies that address the needs of content moderation in the Global South. This became a relevant trend given that global social media platforms frequently face criticism for their limited capacity to moderate content in non-English languages.

Trend topics (fig. 5) provide a coherent and historically meaningful overview of thematic evolution. The first wave of research in this corpus was largely driven by the social impact of the Covid-19 pandemic, from the surge in online interactions, the increase in offensive content in virtual public spaces, to the issue of illegal online lending that was rampant during that period. The transition to the 2023–2024 phase signals a shift from contextual-situational issues (the Covid-19 pandemic) to more permanent structural issues (hate speech as an ongoing phenomenon on digital platforms). The latest phase of 2024–2025 shows that cutting-edge research in this field is increasingly oriented towards two things: (a) methodological refinement through the adoption of more sophisticated deep learning techniques, and (b) expanding the scope of the problem from just hate speech to disinformation and fake news as a related threat of equal importance.

This three-phase evolution pattern is consistent with findings on the thematic map (fig. 7), where the Covid-19 cluster is classified as a driving theme with high centrality and density rather than a neglected, outdated theme. This indicates that the pandemic issue is not merely a momentary trigger, but has been structurally integrated into the conceptual framework of the field. This phenomenon is likely because the pandemic served as a starting point for long-term studies on online behaviour, mental health, and disinformation, the impact of which remains relevant for discussion in the subsequent period.

The basic theme cluster of natural language processing, hate speech, and sentiment analysis is the most promising area for new researchers. We inferred that high centrality indicates that this area is highly relevant and closely connected to almost all other themes, while low density indicates that the area is not yet conceptually saturated, which means that there is still room for significant theoretical and methodological deepening.

The niche theme clusters represent deeply developed yet isolated subfields that are suitable for researchers seeking highly specialised studies such as cybersecurity related to hate speech in the metaverse. Finally, the emerging theme cluster is worth noting in particular. The low density of indobert (33.33, ranked 2 out of 10) as a specific Indonesian-language technology indicates that IndoBERT-based NLP research is still in its early stages of development and has the potential to be an important growth area given the pressing need for Indonesian-language content moderation, which continues to grow with the growth of social media users in Indonesia.

We found that the field growth serves as an indicator of the institutionalization of the CMC space. The growth trend in publications we observed which characterized by a Compound Annual Growth Rate (CAGR) of 45.4% and a peak in output in 2025 (13 articles) is not merely a quantitative measure of academic interest. We see the results as reflection the deepening institutionalization of social media as a space for computer-mediated communication (CMC) and a primary arena for the production and circulation of public discourse in Indonesia. Within the CMC framework, this sharp rise in research activity aligns with the expanding affordances of digital platforms, specifically their capacity for audience reach (scalability), the retention of interaction records (persistence), and the facilitation of content replication (replicability). This in turn structurally create greater scope for problematic communication phenomena, such as hate speech, to emerge and be documented. Thus, this surge in academic interest parallels the expanding scale of the CMC phenomenon itself, rather than representing a mere passing trend.

In our findings, we also found that the centrality of social media serves as evidence of structural affordance, rather than merely reflecting word frequency. Co-occurrence network findings reveal that the "social media" node holds the highest centrality within Cluster 1 far surpassing the "hate speech" node and acts as a bridge to concepts ranging from offensive language, target identification, and online harassment to social networks. Viewed through CMC, this pattern aligns with the argument that the medium is not a passive backdrop but a structural actor shaping interaction patterns among participants. This high linguistic centrality reflects how the literature consistently positions platform affordances rather than just the content of the speech itself as the nexus bridging various derivative phenomena such as online harassment and target identification. However, it is important to note that "social media" is classified as an emerging or declining theme on the thematic map, characterized by low centrality and density. This paradox suggests that the term is frequently employed as a generic setting rather than being operationalized as a specific CMC variable such as anonymity affordances or algorithmic personalization, implying that the analytical potential of the CMC perspective within this corpus remains underutilized.

The three-phase thematic evolution represents a shift from situational to structural deindividuation (SIDE). The pattern of thematic evolution we identified which moving from an initial phase of 2020–2023 that was dominated by the Covid-19 pandemic and illegal online lending, to an intermediate phase of 2023–2024 that focused on hate speech as a persistent issue, and finally to the most recent phase of 2024–2025 which was expanding into disinformation and deep learning can be interpreted through the logic of the Social Identity model of

Deindividuation Effects (SIDE). SIDE posits that the effect of anonymity on group behavior is contextual. SIDE intensifies when specific group norms become salient in a given social situation, rather than remaining constant. The initial phase of our corpus, driven by the economic and health pressures of the pandemic including the proliferation of illegal online lending cases, can be viewed as a state of acute situational deindividuation, wherein a collective crisis temporarily heightened the salience of specific group norms. The shift in discourse toward structural issues during 2023–2024 further confirmed by the classification of the Covid-19 cluster as a motor theme characterized by high centrality and density, rather than a fading theme indicates that the literature began to recognize hate speech not merely as a fleeting crisis response, but as a pattern embedded in the platform architecture itself. This aligns precisely with the core argument of SIDE which explains that the structure of the medium, rather than just the immediate context sustains the salience of group identity over time.

The four-quadrant thematic map structure reflects the gap between technical operationalization and explanations grounded in Social Identity Theory (SIT). Classifying themes into motor themes such as deep learning and Covid-19; basic themes such as NLP, hate speech, and sentiment analysis; niche themes such as target identification, security, and metaverse; and emerging such as social media and IndoBERT reveals patterns best interpreted through the lens of SIT. The "target identification" theme within the niche cluster which characterized by high density but low centrality essentially represents a technical operationalization of a core SIT concept. However, its isolation from the core clusters suggests that research developing computational target identification techniques has yet to sufficiently link findings back to theoretical explanations regarding why specific groups become targets of hate speech. A similar pattern appears in the basic themes cluster. The high centrality yet low density of NLP, hate speech, and sentiment analysis indicate that while these themes are recognized as important across nearly all subfields, they lack deep conceptual exploration. This leaves significant room for future research to explicitly link automated detection with the SIT framework, rather than stopping at mere classification accuracy.

The Indonesian context presents a unique, underexplored intersection of CMC, SIDE, and SIT. Distinctive findings from this corpus such as the emergence of terms like IndoBERT, issues surrounding peer-to-peer online lending linked to Indonesian fintech regulations, and themes of religious content moderation underscore that hate speech dynamics in Indonesia cannot be fully explained by the English-language research that has long dominated the field. When viewed through the lens of SIT, the emergence of religious moderation themes reveals that in-group andn out-group categorization based on religious identity serves as a primary axis of hate speech in Indonesia's digital sphere, a group dynamic distinct from the dominant identity axes such as race, ethnicity, gender found in Western literature. Meanwhile, the status of IndoBERT as an emerging theme with low density indicates that while technical capabilities for detecting Indonesian-language hate speech are evolving, research linking these capabilities to local platform affordances (CMC) and local group identity dynamics (SIT) such as how platforms like WhatsApp or X differ from Western platforms in facilitating hate speech related to ethnicity, religion, race, and inter-group relations (SARA) remains very limited. This represents the most productive avenue for future research that contextually integrates these three theoretical lenses.

5. Conclusion

When the findings above are viewed holistically, a consistent three-tiered conceptual pattern emerges from the keyword and co-occurrence network analyses comprising the layer of the object of study (social media, hate speech, offensive content), the methodological layer (deep learning and its derivatives), and the technical foundation layer (ML, NLP). This pattern reveals that existing literature is far more advanced in developing methodological capabilities than in achieving theoretical maturity from a Communication Science perspective. CMC explains how platform affordances structurally facilitate the spread of hate speech. SIDE explains how anonymity and minimal moderation amplify—rather than obscure or the salience of group identities. While SIT explains the psychosocial motives underlying the categorization and derogation of out-groups. Yet, these three theories are rarely operationalized explicitly or in an integrated manner within the analyzed corpus. The consistently low density observed in themes with high centrality serves as a quantitative indicator of this gap. While these themes are widely discussed, they have not yet been deeply analyzed through coherent analytical frameworks. Given the limitations regarding corpus size (43 documents) and the observation period (six years), we recommend that future research explicitly operationalize CMC and SIDE as primary frameworks and SIT as a secondary framework. This approach would not only help deepen the exploration of fundamental thematic clusters that remain wide open for study but also bridge the gap between technical detection capabilities and a structural-social understanding of why hate speech continues to proliferate in Indonesia's digital landscape.

Several limitations need to be addressed to allow for a fair interpretation of these results. First, the corpus analysed is relatively small (43 documents). Thus, statistical models used still have a reasonable degree of uncertainty given the limited number of historical data points (only 6 years of observation). Second, data anomalies were also found in two documents with extreme citation counts (2.022 and 2.023 citations) in the queried data that deviate substantially from the citation distribution of other documents, which showed generally 0–26 citations. Such anomalies are common due to metadata recording errors in source databases, such as incorrectly merging citation data from different entries.

References

Abdel-Aziz, A. A., Abdel-Salam, H., & El-Sayad, Z. (2016). The role of ICTs in creating the new social public place of the digital era. *Alexandria Engineering Journal*, *55*(1), 487–493. https://doi.org/10.1016/j.aej.2015.12.019

Aldamen, Y. (2023). Xenophobia and Hate Speech towards Refugees on Social Media: Reinforcing Causes, Negative Effects, Defense and Response Mechanisms against That

Speech. *Societies*, *13*(4), 83. https://doi.org/10.3390/soc13040083

AlKhudari, M. N., Abduljabbar, O. J., Al Manaseer, A. M., & AL-Omari, M. S. (2024). The role of social media in shaping public opinion among Jordanian university students. *Journal of Infrastructure Policy and Development*, *8*(8), 5489. https://doi.org/10.24294/jipd.v8i8.5489

Allwood, M., Ghafoori, B., Salgado, C., Slobodin, O., Kreither, J., Waelde, L. C., Larrondo, P., & Ramos, N. (2022). Identity-based hate and violence as trauma: Current research, clinical implications, and advocacy in a globally connected world. *Journal of Traumatic Stress*, *35*(2), 349–361. https://doi.org/10.1002/jts.22748

Almeida, P., Pereira, J., & Candido, D. (2023). Online hate speech on social media in Portugal: extremism or structural racism? *Social Identities*, *29*(5), 419–435. https://doi.org/10.1080/13504630.2024.2324277

Aria, M., & Cuccurullo, C. (2017). bibliometrix : An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Aria, M., & Cuccurullo, C. (2026). *Science Mapping Analysis: A Primer with Biblioshiny*. McGraw-Hill.

Ausat, A. M. A. (2023). The Role of Social Media in Shaping Public Opinion and Its Influence on Economic Decisions. *Technology and Society Perspectives (TACIT)*, *1*(1), 35–44. https://doi.org/10.61100/tacit.v1i1.37

Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., & Quattrociocchi, W. (2024). Persistent interaction patterns across social media platforms and over time. *Nature*, *628*(8008), 582–589. https://doi.org/10.1038/s41586-024-07229-y

Barth, N., Wagner, E., Raab, P., & Wiegärtner, B. (2023). Contextures of hate: Towards a systems theory of hate communication on social media platforms. *The Communication Review*, *26*(3), 209–252. https://doi.org/10.1080/10714421.2023.2208513

Bennett, W. L., & Segerberg, A. (2023). The logic of connective action: digital media and the personalization of contentious politics. In *Handbook of Digital Politics* (pp. 287–338). Edward Elgar Publishing. https://doi.org/10.4337/9781800377585.00030

Bioglio, L., & Pensa, R. G. (2022). Analysis and classification of privacy-sensitive content in social media posts. *EPJ Data Science*, *11*(1), 12. https://doi.org/10.1140/epjds/s13688-022-00324-y

Blanco-Herrero, D., Amores, J. J., & Arcila-Calderón, C. (2024). Inequalities and the media: spread and reception of hate speech against migrants and refugees in social media. In *Research Handbook on Inequalities and Work* (pp. 211–227). Edward Elgar Publishing.

https://doi.org/10.4337/9781800886605.00025

Ceccato, V., & Simões Agapito, L. S. (2024). Hate-motivated crimes in Brazil: an overview of crimes against LGBTQI+ people. *Safer Communities*, *23*(3), 233–243. https://doi.org/10.1108/SC-12-2023-0052

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118. https://doi.org/10.1016/j.avb.2018.05.003

Gerbaudo, P. (2026). TikTok and the algorithmic transformation of social media publics: From social networks to social interest clusters. *New Media & Society*, *28*(3), 1019–1036. https://doi.org/10.1177/14614448241304106

Ghenai, A., Noorian, Z., Moradisani, H., Abadeh, P., Erentzen, C., & Zarrinkalam, F. (2025). Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users. *Information Processing & Management*, *62*(3), 104079. https://doi.org/10.1016/j.ipm.2025.104079

Giraud, E. H., Poole, E., de Quincey, E., & Richardson, J. E. (2025). Learning from online hate speech and digital racism: From automated to diffractive methods in social media analysis. *The Sociological Review*, *73*(6), 1388–1407. https://doi.org/10.1177/00380261241305260

Giumetti, G. W., & Kowalski, R. M. (2022). Cyberbullying via social media and well-being. *Current Opinion in Psychology*, *45*, 101314. https://doi.org/10.1016/j.copsyc.2022.101314

Hassim, M. N., Mohamad Nasir, N. N. A., & Zamri, N. A. K. (2024). Hate Speech In The Digital Age: A Study in Terms of Impact And Social Implications. *International Journal of Law, Government and Communication*, *9*(38), 01–12. https://doi.org/10.35631/IJLGC.938001

Hietanen, M., & Eddebo, J. (2023). Towards a Definition of Hate Speech—With a Focus on Online Contexts. *Journal of Communication Inquiry*, *47*(4), 440–458. https://doi.org/10.1177/01968599221124309

Hurova, I. V., & Shkurov, Y. V. (2023). Man in Digitized Urban Socio-Cultural Space. *Anthropological Measurements of Philosophical Research*, (24), 75–87. https://doi.org/10.15802/ampr.v0i24.295321

Jaidka, K., Zhou, A., Lelkes, Y., Egelhofer, J., & Lecheler, S. (2021). Beyond Anonymity: Network Affordances, Under Deindividuation, Improve Social Media Discussion Quality. *Journal of Computer-Mediated Communication*. https://doi.org/10.1093/jcmc/zmab019

Kääntä, L. (2025). Digital space. In *Research Handbook on Social Interaction* (pp. 167–178). Edward Elgar Publishing. https://doi.org/10.4337/9781035317264.00014

Kamboj, S., & Sharma, M. (2023). Social media adoption behaviour:

Consumer innovativeness and participation intention. *International Journal of Consumer Studies*, *47*(2), 523–544. https://doi.org/10.1111/ijcs.12848

Karanasios, S., Adrot, A., Hayes, P., & Cooper, V. (2024). The interplay of logics and social media: An organisational field perspective. *Information Systems Journal*, *34*(1), 80–124. https://doi.org/10.1111/isj.12472

Kentmen-Cin, C. (2025). Hate Speech on Social Media: A Systemic Narrative Review of Political Science Contributions. *Social Sciences*, *14*(10), 610. https://doi.org/10.3390/socsci14100610

Laor, T. (2022). My social network: Group differences in frequency of use, active use, and interactive use on Facebook, Instagram and Twitter. *Technology in Society*, *68*, 101922. https://doi.org/10.1016/j.techsoc.2022.101922

Literat, I., & Kligler-Vilenchik, N. (2019). Youth collective political expression on social media: The role of affordances and memetic dimensions for voicing political views. *New Media & Society*, *21*(9), 1988–2009. https://doi.org/10.1177/1461444819837571

Maarouf, A., Pröllochs, N., & Feuerriegel, S. (2024). The Virality of Hate Speech on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–22. https://doi.org/10.1145/3641025

Miranda, S., Trigo, I., Rodrigues, R., & Duarte, M. (2023). Addiction to social networking sites: Motivations, flow, and sense of belonging at the root of addiction. *Technological Forecasting and Social Change*, *188*, 122280. https://doi.org/10.1016/j.techfore.2022.122280

Muhammad, S. H. (2024). The Role of Social Media in Shaping Political Discourse and Public Opinion in the Digital Age. *Mahogany Journal De Social*, *1*(3), 156–169. https://doi.org/10.37899/mjds.v1i3.111

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Ren, J., Dong, H., Popovic, A., Sabnis, G., & Nickerson, J. (2024). Digital platforms in the news industry: how social media platforms impact traditional media news viewership. *European Journal of Information Systems*, *33*(1), 1–18. https://doi.org/10.1080/0960085X.2022.2103046

Rodilosso, E. (2024). Filter Bubbles and the Unfeeling: How AI for Social Media Can Foster Extremism and Polarization. *Philosophy & Technology*, *37*(2), 71. https://doi.org/10.1007/s13347-024-00758-4

Romero-Rodríguez, L. M., Castillo-Abdul, B., & Cuesta-Valiño, P. (2023). The Process of The Transfer of Hate Speech to Demonization and Social Polarization. *Politics and Governance*, *11*(2). https://doi.org/10.17645/pag.v11i2.6663

Salma, B., Fatima, T., Sara, A., & Merieme, B. (2024). *Artificial Intelligence in Social Media: From Content Personalization to User Engagement* (pp. 45–52). https://doi.org/10.1007/978-3-031-65018-5_5

Shah, S. S. (2024). The Role of Social Media in Shaping Public Opinion on Environmental Issues. *Premier Journal of Environmental Science*. https://doi.org/10.70389/PJES.100002

Shmulewitz, D., Levitin, M. D., Skvirsky, V., Vider, M., Lev-Ran, S., & Mikulincer, M. (2025). Exposure to online hate speech is positively associated with post-traumatic stress disorder symptom severity. *Scientific Reports*, *15*(1), 29869. https://doi.org/10.1038/s41598-025-16168-1

Solovev, K., & Pröllochs, N. (2023). Moralized language predicts hate speech on social media. *PNAS Nexus*, *2*(1). https://doi.org/10.1093/pnasnexus/pgac281

Stahel, L., & Baier, D. (2023). Digital Hate Speech Experiences Across Age Groups and Their Impact on Well-Being: A Nationally Representative Survey in Switzerland. *Cyberpsychology, Behavior, and Social Networking*, *26*(7), 519–526. https://doi.org/10.1089/cyber.2022.0185

Swastiningsih, S., Aziz, A., & Dharta, Y. (2024). The Role of Social Media in Shaping Public Opinion: A Comparative Analysis of Traditional vs. Digital Media Platforms. *The Journal of Academic Science*, *1*(6), 620–626. https://doi.org/10.59613/fm1dpm66

Teepapal, T. (2025). AI-driven personalization: Unraveling consumer perceptions in social media engagement. *Computers in Human Behavior*, *165*, 108549. https://doi.org/10.1016/j.chb.2024.108549

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, *119*(42). https://doi.org/10.1073/pnas.2207159119

United Nations. (2019). *United Nations Strategy and Plan of Action on Hate Speech*.

Vergani, M., Perry, B., Freilich, J., Chermak, S., Scrivens, R., Link, R., Kleinsman, D., Betts, J., & Iqbal, M. (2024). Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech, and hate incidents: A systematic review. *Campbell Systematic Reviews*, *20*(2). https://doi.org/10.1002/cl2.1397

Walters, P. (2022). Reclaiming Control: How Journalists Embrace Social Media Logics While Defending Journalistic Values. *Digital Journalism*, *10*(9), 1482–1501. https://doi.org/10.1080/21670811.2021.1942113

Waworuntu, E. C., Mandagi, D. W., & Pangemanan, A. S. (2022). 'I See It, I Want It, I Buy It': The Role of Social Media Marketing in Shaping Brand Image and Gen Z's Intention to Purchase Local Product. *Society*, *10*(2), 351–369. https://doi.org/10.33019/society.v10i2.463

We Are Social, & Meltwater. (2025). *Digital 2025 Global Overview Report*.

.

Yeh, E., & Swinehart, N. (2022). Social media literacy in L2 environments: navigating anonymous user-generated content. *Computer Assisted Language Learning*, *35*(8), 1731–1753. https://doi.org/10.1080/09588221.2020.1830805