

Article

Artificial Intelligence Approaches to Hate Speech Detection: A Bibliometric Analysis of Scholarly Development and Governance-Oriented Research Structures

Nikos Koutsoupas  and Marios Nosios * 

Department of International and European Studies, University of Macedonia, 54636 Thessaloniki, Greece; nk@uom.edu.gr

* Correspondence: mnosios@uom.edu.gr

Abstract

The governance of hate-related online communication increasingly relies on artificial intelligence, yet the scientific landscape linking computational detection methods with regulatory and ethical frameworks remains fragmented. This study provides a systematic bibliometric analysis of scholarship on the application of artificial intelligence to hate speech detection in digital environments, focusing on literature that examines hate speech and online hate through artificial intelligence, machine learning, deep learning, natural language processing, and other automated detection techniques. A dataset of 2137 publications indexed in Scopus between 2013 and 2026 was constructed and analyzed using the bibliometrix package in R. Descriptive indicators, thematic mapping, keyword co-occurrence analysis, citation structures, and temporal trend analysis were employed to examine the field's conceptual organization, methodological evolution, and publication dynamics. The results reveal rapid annual growth, strong interdisciplinary collaboration, and a research structure dominated by language-processing methodologies, with natural language processing, machine learning, and deep learning constituting the central analytical infrastructure. Temporal patterns indicate a progression from dataset construction and feature engineering toward neural architectures, transformer models, and increasing attention to multilingual challenges. Overall, the findings indicate that the field remains predominantly oriented toward the development of scalable computational detection systems, while governance-related concerns, such as transparency, accountability, and linguistic inclusivity, emerge as structurally secondary, albeit increasingly salient, dimensions within the broader AI-centered research landscape.

Keywords: artificial intelligence; machine learning; hate speech; UNESCO; algorithmic governance; digital regulation; scientometrics



Academic Editor: Sudeshna Roy

Received: 4 March 2026

Revised: 23 June 2026

Accepted: 10 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The regulation of hate speech occupies a structurally contested position within international human rights law, as it stands at the normative intersection of equality protection and freedom of expression (Heinze, 2016; Brown, 2015). Foundational instruments such as the International Covenant on Civil and Political Rights (ICCPR) simultaneously enshrine the protection of expression under Article 19 and impose an obligation to prohibit advocacy of hatred constituting incitement under Article 20 (Alkiviadou, 2018). Similarly, the International Convention on the Elimination of All Forms of Racial Discrimination

(ICERD) establishes more explicit duties regarding racist propaganda and the prohibition of discriminatory organizations (Kapelańska-Pręgowska & Pucelj, 2023). Scholarly analysis consistently emphasizes that these provisions do not yield a singular, operational definition of hate speech; rather, they articulate a contextual threshold framework contingent upon intent, likelihood of harm, speaker authority, and prevailing social circumstances. This approach is most systematically reflected in the Rabat Plan of Action's multi-factor incitement test (Vidgen & Derczynski, 2020). The resulting doctrinal indeterminacy acquires heightened significance in the context of digital governance, where speech crosses jurisdictions with divergent legal traditions and regulatory capacities, thereby complicating the translation of international norms into platform moderation standards and automated detection architectures (Gorwa et al., 2020; Fortuna & Nunes, 2018).

Digital communication infrastructures further intensify these tensions because platform architectures do not merely host content but actively structure its visibility, amplification, and perceived harm (Gillespie, 2018; Matamoros-Fernández, 2017). Empirical research on social media governance demonstrates that algorithmic ranking systems, engagement-driven recommendation mechanisms, and network clustering dynamics can amplify emotionally polarizing or identity-targeted content, thereby facilitating coordinated harassment and large-scale exposure (Cinelli et al., 2021; M. H. Ribeiro et al., 2020). In contrast to traditional broadcast media, digital speech is transnational, persistent, and frequently anonymized, rendering jurisdictional attribution and enforcement mechanisms increasingly complex (DeNardis & Hackl, 2015). Moreover, linguistic variability, coded discourse, and culturally embedded idioms complicate classification efforts, as the semantic valence of potentially harmful expression often depends on contextual pragmatics rather than lexical indicators alone (Davidson et al., 2017; Waseem, 2016). These structural characteristics explain why contemporary governance of online hate speech increasingly relies on hybrid regulatory configurations that combine formal legal obligations, platform self-regulation, and international normative guidance rather than relying exclusively on criminal prohibition (Jhaver et al., 2018; Chandrasekharan et al., 2017). Within this broader context, the present study investigates the evolution of scholarship at the intersection of hate speech, online hate, artificial intelligence, machine learning, deep learning, natural language processing, and automated detection technologies.

2. Literature Review

2.1. International and Regional Governance Frameworks

The governance of online hate speech has increasingly been shaped by multilateral initiatives and regional regulatory instruments that emphasize procedural operationalization rather than purely substantive definition (N. Suzor et al., 2018). The United Nations Strategy and Plan of Action on Hate Speech adopts an intentionally expansive policy conception that encompasses both unlawful incitement and lawful yet harmful expression, thereby foregrounding prevention, education, and coordinated institutional responses (Gagliardone et al., 2015). The Council of Europe's Recommendation CM/Rec (2022)16 similarly conceptualizes hate speech as a layered phenomenon requiring proportionate legal, administrative, and pedagogical interventions (Buyse, 2014). The European Union's Digital Services Act represents a more binding procedural model, imposing structured obligations on large online platforms regarding transparency reporting, systemic risk assessment, user complaint mechanisms, and the mitigation of harmful dissemination dynamics (Frosio & Geiger, 2023). Academic commentary frequently characterizes this regulatory evolution as a movement away from direct speech prohibition toward process-oriented accountability regimes that regulate decision-making structures and governance procedures rather than prescribing substantive outcomes in advance (Peitz, 2025).

2.2. Computational Detection Methods and Their Limitations

The movement from legal definition toward procedural governance also creates a conceptual tension for automated detection research. Surveys of the field have repeatedly identified definitional inconsistency as a persistent limitation in computational hate speech studies (Fortuna & Nunes, 2018; Schmidt & Wiegand, 2017). For instance, research has shown that hate speech is frequently difficult to distinguish from offensive language in automated classification tasks (Davidson et al., 2017). Related research also demonstrates that dataset construction and annotation practices shape what is ultimately classified as hateful content (Waseem & Hovy, 2016). Consequently, studies categorized as hate speech detection research do not always examine hate speech in the narrower legal or human rights sense. Instead, they often evaluate computational approaches for identifying broader forms of harmful online communication, depending on the categories embedded in the dataset and the objectives of the classification task.

Alongside these governance developments, the technical literature on automated hate speech detection has undergone a significant methodological transformation (Schmidt & Wiegand, 2017). Early computational approaches relied predominantly on supervised machine learning models, such as support vector machines and logistic regression, employing manually engineered lexical and syntactic features (Nobata et al., 2016). Contemporary systems increasingly deploy deep neural architectures, particularly transformer-based pretrained language models such as BERT and multilingual derivatives, which enable contextual semantic modeling through self-attention mechanisms and large-scale pre-training across heterogeneous textual corpora (Devlin et al., 2019). These architectures have substantially improved performance on benchmark toxicity datasets by capturing long-range dependencies and contextual word meaning, thereby enabling more nuanced differentiation between superficially similar utterances with distinct pragmatic implications (Zhang et al., 2018). Nonetheless, empirical research consistently demonstrates that high benchmark accuracy does not necessarily translate into robust real-world moderation performance, especially within multilingual or culturally heterogeneous environments (Arango et al., 2022; Röttger et al., 2021).

A central technical challenge concerns dataset bias and representational asymmetry (Mehrabi et al., 2021). Hate speech corpora often overrepresent English-language Western social contexts, leading to systematic performance degradation for under-resourced languages and dialectal communities (Joshi et al., 2020). Research on dialectal variation, including African American English and other sociolects, reveals that models often misclassify identity-linked linguistic features as toxicity indicators due to spurious correlations embedded within training data (Sap et al., 2019; Blodgett et al., 2016). This phenomenon raises profound concerns about equality, as automated moderation systems may disproportionately flag expressions from already marginalized communities, thereby reproducing structural discrimination under the guise of algorithmic neutrality. UNESCO's ethical principles of inclusiveness and fairness intersect directly with these challenges. However, within the AI-focused scholarly corpus examined in this study, the explicit translation of UNESCO's AI ethics framework into concrete machine learning implementation practices does not emerge as a dominant or consolidated line of inquiry (Floridi et al., 2018).

Explainability and accountability constitute an additional critical interface between technical capability and normative governance. Machine learning classifiers, particularly deep neural networks, are frequently criticized for opacity, as their internal representations are often difficult to interpret (Lipton, 2018). Explainable AI methodologies, including local feature attribution techniques and structured model documentation frameworks, seek to mitigate this opacity by providing post hoc explanatory mechanisms or standardized reporting of model limitations (Mitchell et al., 2019; M. T. Ribeiro et al., 2016). However,

technical transparency alone cannot resolve the fundamental statistical trade-offs inherent in automated moderation. Detection systems necessarily balance false positives, which risk suppressing legitimate political or identity-based expression, against false negatives, which permit harmful content to remain visible (Gorwa, 2019). Such trade-offs cannot be eliminated through algorithmic refinement alone; rather, they necessitate normative determinations concerning acceptable risk allocation, procedural safeguards, and thresholds for human review (Selbst et al., 2019; Ananny & Crawford, 2016). Within the scope of the present bibliometric analysis, transparency-related concepts, including explainability, interpretability, and XAI, are therefore examined in terms of their relative prominence within the AI-centered scholarly corpus, rather than as independently analyzed doctrinal constructs.

2.3. Human Rights Implications and the Governance Gap

The technical limitations of automated detection become especially significant when such systems are embedded within platform moderation infrastructures. Their deployment raises broader human rights considerations regarding due process, transparency, and accountability in corporate governance (N. P. Suzor, 2019). Scholars of platform regulation caution that large-scale automated filtering may have chilling effects on lawful expression when users lack meaningful avenues for appeal or clarity about moderation criteria (York & Zuckerman, 2019; Penney, 2016). Simultaneously, platforms often resist comprehensive transparency into their proprietary models, invoking trade secret protections and the risk of adversarial manipulation, thereby perpetuating an accountability deficit (Burrell, 2016). UNESCO's governance framework addresses this structural tension by emphasizing human-in-the-loop oversight, institutional accountability mechanisms, and continuous system auditing, rather than relying exclusively on automated decision-making (Cath et al., 2018). This orientation reflects an emerging consensus that AI moderation should operate as a decision-support infrastructure embedded within deliberative institutional processes rather than as an autonomous enforcement authority (Rahwan, 2018).

Beyond detection and sanctioning, UNESCO's policy paradigm foregrounds preventive strategies grounded in education and social resilience (Grizzle et al., 2014). Media and information literacy programmes, counter-speech initiatives, and intercultural dialogue strategies are conceptualized as structural interventions targeting the socio-political drivers of hate speech rather than its symptomatic manifestations alone (Mihailidis & Viotty, 2017). This preventive orientation reconceptualizes AI not merely as a punitive filtering instrument but as a potential analytical and pedagogical resource capable of identifying emergent discourse patterns, supporting early-warning monitoring frameworks, and informing targeted digital citizenship initiatives (Taddeo & Floridi, 2018). Such an expanded analytical framework repositions hate speech governance from a narrow focus on classification accuracy toward long-term societal capacity building, while also highlighting the need to integrate technical moderation systems more explicitly within broader accountability and human rights frameworks.

2.4. Research Gaps

Existing scholarship has generated substantial insights into the legal, governance, and computational dimensions of hate speech moderation, while review studies have concentrated primarily on conceptual definitions, methodological approaches, and technical advances within the field (Schmidt & Wiegand, 2017; Fortuna & Nunes, 2018). By contrast, comparatively limited attention has been devoted to systematically mapping the intellectual structure, thematic evolution, and knowledge base of research at the intersection of hate speech and artificial intelligence. Consequently, the manner in which technological innovation, governance challenges, and emerging scholarly priorities converge within

this interdisciplinary domain remains insufficiently understood. These shortcomings are particularly apparent in relation to the incorporation of normative governance principles into AI-driven research and practice. More specifically, the extent to which UNESCO's AI ethics principles are translated into machine learning implementation remains inadequately specified, particularly with respect to operational auditing procedures, multilingual data governance, and rights-sensitive evaluation frameworks. Similarly, empirical evidence assessing whether UNESCO-informed governance approaches improve moderation fairness or reduce discriminatory outcomes remains limited. In addition, the continued dominance of high-resource languages in natural language processing research raises unresolved concerns regarding equitable protection across linguistically diverse populations. Collectively, these limitations highlight the need for scholarship that moves beyond parallel legal and technical perspectives and advances more integrated frameworks connecting human rights principles with the design, evaluation, and governance of AI-mediated content moderation systems.

3. Methods

This study employs a systematic bibliometric approach to examine the scientific landscape surrounding artificial intelligence methodologies for hate speech detection and their relevance to governance-oriented frameworks. A structured dataset was constructed using the Scopus database, selected for its extensive multidisciplinary coverage across computer science, the social sciences, decision sciences, and the humanities (Baas et al., 2020). The database was chosen to ensure comprehensive indexing of both computational research and policy-oriented scholarship addressing digital content moderation while providing standardized citation metadata suitable for bibliometric mapping, citation analysis, and thematic network construction. Although specialized technical databases provide important coverage of computer science and engineering research, the use of Scopus as a single source ensured methodological consistency, comprehensive multidisciplinary coverage, and a uniform metadata structure throughout the analytical process (Schotten et al., 2017). To curate the dataset, a targeted search query was developed. The search query used is as follows:

TITLE-ABS-KEY (("hate speech" OR "online hate" OR "digital hate") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "natural language processing" OR "automated content moderation" OR "algorithmic detection")) AND (LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "DECI") OR LIMIT-TO (SUBJAREA, "ARTS") OR LIMIT-TO (SUBJAREA, "SOCI")) AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "cp")) AND (LIMIT-TO (LANGUAGE, "English")).

The temporal scope encompassed publications from 2013 to 2026, with 2013 representing the earliest publication year retrieved by the search query and marking the emergence of indexed scholarship on AI-based hate speech detection within the dataset. The screening and selection procedures adhered to the PRISMA 2020 guidelines to ensure transparency throughout the identification, eligibility assessment, and inclusion stages (Page et al., 2021). Duplicate records, non-English publications, and materials falling outside the defined thematic scope were excluded (see Figure 1).

Following the construction and refinement of the dataset, the analytical process progressed through successive bibliometric stages to capture both the field's structural composition and developmental trajectory. Descriptive statistics were first generated to evaluate annual scientific output, citation performance, document typologies, authorship distribution, and patterns of collaboration. These indicators facilitated the identification of growth phases, shifts in publication formats, and evolving configurations of collective research activity. Particular emphasis was placed on accelerating publication rates over time and on

the distribution between conference proceedings and journal articles, given the domain's technological orientation. All data processing and quantitative analyses were conducted using the bibliometrix package in R, thereby ensuring systematic metadata management and analytical reproducibility (Aria & Cuccurullo, 2017).

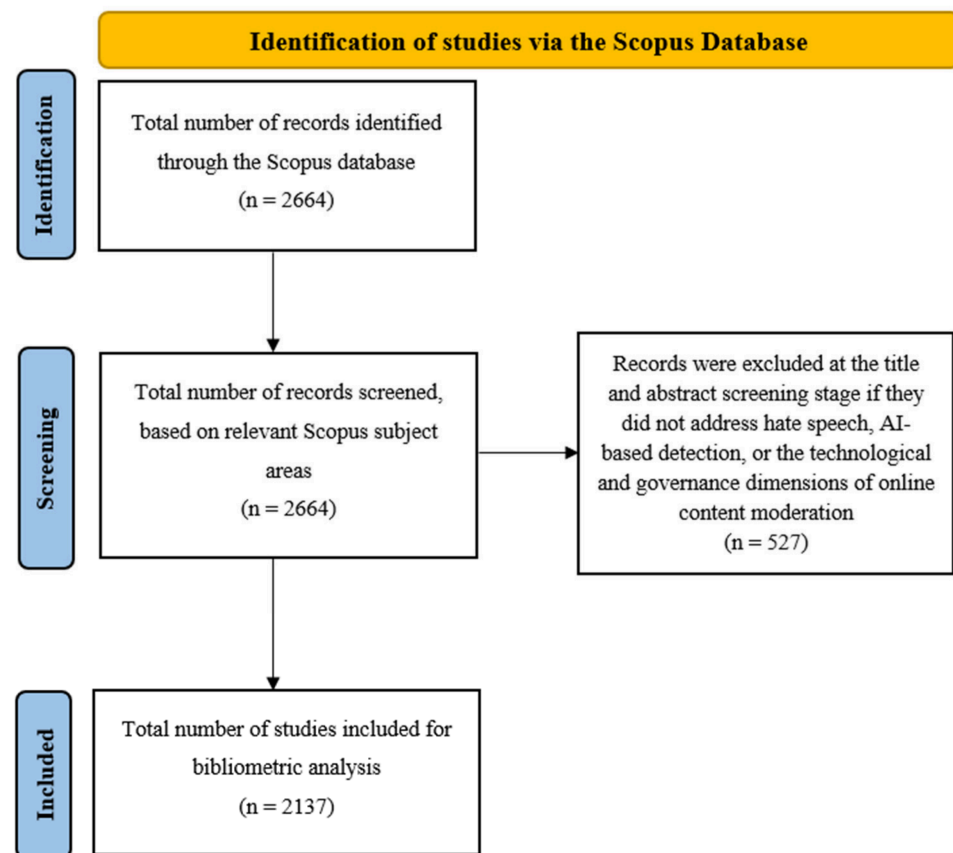


Figure 1. PRISMA 2020 flow diagram of the selection process.

Subsequent stages focused on mapping the conceptual and intellectual architecture of the research landscape. Frequency analyses and co-occurrence networks were constructed to identify dominant thematic clusters and structural interconnections among recurrent research elements. This approach enabled the detection of thematic concentrations and clarified how methodological innovations and problem-oriented research trajectories interact within the broader domain. Science mapping techniques were used to visualize thematic positioning based on centrality and density, thereby distinguishing foundational, consolidated, specialized, and emerging areas of inquiry. Temporal analyses further traced the evolution of research priorities, illuminating shifts in methodological orientation and the diffusion of advanced modeling approaches. Complementary network analyses, including collaboration structures and citation patterns, supported the interpretation of the field's institutional organization and knowledge flows.

Collectively, these procedures provide a coherent analytical framework for examining the structural dynamics and conceptual development of AI applications in hate speech research within a multidisciplinary context. At the same time, certain methodological constraints delimit the scope of interpretation. The search strategy was deliberately designed to capture literature located at the intersection of hate speech and AI methodologies; by incorporating computational descriptors such as “machine learning,” “deep learning,” and “natural language processing,” it structurally prioritizes technically oriented scholarship. Consequently, regulatory or doctrinal analyses that address hate speech governance with-

out explicit reference to AI-related terminology may be underrepresented. Furthermore, inclusion was based on the presence of the selected search terms in titles, abstracts, or keywords; therefore, hate speech was not necessarily the sole focus of every retrieved publication. In addition, the restriction to English-language publications necessarily shapes the proportional representation of scholarship produced in non-Anglophone regulatory contexts. Additionally, records for 2025–2026 reflect the state of database indexing at the time of data extraction; associated bibliometric indicators should therefore be interpreted in light of the evolving nature of database coverage. Finally, while bibliometric analysis enables the identification of patterns in publication output, citation structures, and thematic associations, it does not directly measure regulatory effectiveness or institutional impact; references to governance implications in this study thus pertain to the structural positioning of such concepts within the research landscape rather than to an empirical evaluation of policy outcomes.

4. Results

The results section presents the empirical findings of the bibliometric analysis, identifying the structural themes, methodological trajectories, and publication patterns that characterize contemporary research on computational hate speech detection. These findings are systematically organized around thematic mapping, keyword distribution, temporal trends, citation structures, and descriptive statistical indicators.

4.1. Descriptive Statistics

In Table 1, the descriptive statistics are depicted for the bibliometric dataset covering the period 2013–2026, synthesizing key indicators of publication output, citation patterns, authorship configuration, and document typology. The corpus comprises 2137 documents distributed across 933 distinct sources, revealing considerable dispersion across journals, conference proceedings, and edited volumes. An annual growth rate of 34.69 percent signals pronounced expansion, indicating that scholarly activity in this domain has accelerated sharply over the past decade. The relatively low average document age of 3.34 years further confirms the field's recency, underscoring its ongoing dynamism and developmental fluidity. This temporal concentration closely corresponds to the broader proliferation of artificial intelligence and machine learning methodologies in recent years. An average of 16.28 citations per document suggests moderate citation impact, characteristic of a domain that is both active and consolidating, rather than one that has reached paradigmatic stability. The presence of 10,926 cited references points to extensive intertextual engagement and a dense network of methodological and conceptual cross-referencing.

Regarding content descriptors, the dataset contains 5427 Keywords Plus entries and 3178 author-provided keywords. The disparity between these figures reflects the interplay between database-generated indexing expansion and author-driven conceptual framing. The high overall volume of keywords indicates substantial thematic diversity, suggesting methodological pluralism and ongoing conceptual experimentation within the research landscape.

Authorship patterns reveal that 5812 authors participated, yet only 89 produced single-authored documents. Although 99 publications are single-authored, the average of 3.74 co-authors per document points to a predominantly collaborative research culture. This collaborative orientation is further reinforced by an international co-authorship rate of 24.66 percent, indicating the existence of transnational research networks. The technical complexity inherent in computational modelling, dataset curation, and algorithmic evaluation likely necessitates interdisciplinary collaboration that integrates expertise in linguistics, computer science, data science, and social analysis.

Table 1. Descriptive Statistics.

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan	2013:2026
Sources (Journals, Books, etc.)	933
Documents	2137
Annual Growth Rate %	34.69
Document Average Age	3.34
Average citations per doc	16.28
References	10,926
DOCUMENT CONTENTS	
Keywords Plus (ID)	5427
Author's Keywords (DE)	3178
AUTHORS	
Authors	5812
Authors of single-authored docs	89
AUTHORS COLLABORATION	
Single-authored docs	99
Co-Authors per Doc	3.74
International co-authorships %	24.66
DOCUMENT TYPES	
article	634
conference paper	1503

The distribution of document types demonstrates a clear predominance of conference papers (1503) over journal articles (634). This imbalance reflects the publication norms characteristic of computer science and artificial intelligence, where conference proceedings frequently serve as the primary venue for disseminating cutting-edge findings. The prominence of conference outputs implies rapid methodological iteration and comparatively short research cycles, consistent with the pace of technological innovation in machine learning and natural language processing.

Ultimately, these descriptive indicators portray a rapidly expanding, highly collaborative, and methodologically agile field. The concentration of recent publications, the density of thematic descriptors, and the dominance of conference-based dissemination collectively signal strong alignment with computational research traditions and underscore the emergent yet consolidating character of automated hate speech analysis (see Table 1).

4.2. Annual Scientific Production

The graph in Figure 2 shows the annual number of published articles from 2013 to 2026, thereby tracing the longitudinal expansion of computational research on hate-related content. The temporal distribution reveals a pronounced transformation from marginal scholarly activity to sustained, high-volume production. The initial phase, spanning 2013 to 2016, is marked by extremely limited output: one article in 2013, one in 2015, and two in 2016. This sparse production indicates that computational approaches to hate speech had not yet coalesced into a distinct and rapidly advancing research domain. The low

publication density suggests exploratory efforts dispersed across adjacent fields rather than a consolidated scholarly trajectory. A modest inflection point becomes visible in 2017, with seven published articles. Although still limited in scale, this increase signals the early aggregation of thematic and methodological attention.

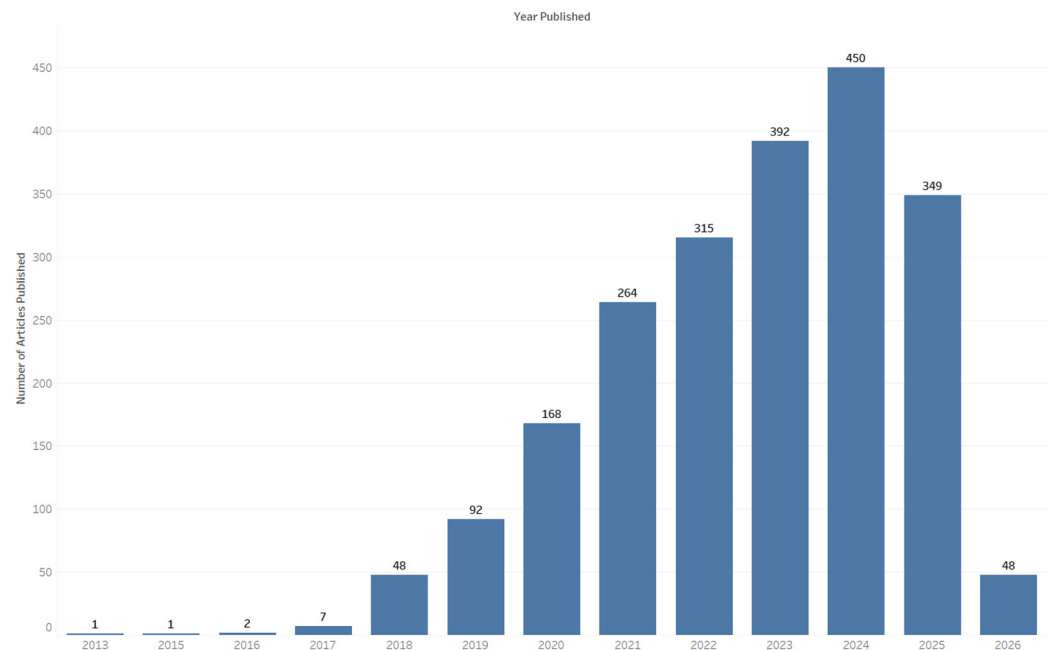


Figure 2. Annual Scientific Production.

A decisive structural shift emerges in 2018, when annual output rises sharply to 48 articles, followed by continued expansion to 92 in 2019. This period coincides with the consolidation of neural network architectures and the widespread diffusion of deep learning techniques within natural language processing. The rapid adoption of these modelling paradigms appears to have catalyzed publication activity, lowering technical barriers and enabling scalable detection experiments across an increasingly available set of social media datasets.

From 2020 onward, the trajectory enters a phase of accelerated, near-exponential growth. Publication output increases to 168 articles in 2020 and to 264 in 2021, representing a near-doubling within two years. Expansion continued in 2022, with 315 articles, and intensified further in 2023, with 392 publications. The peak occurs in 2024, reaching 450 articles, the highest annual production recorded in the dataset. The steep upward curve between 2019 and 2024 reflects not merely incremental growth but a structural transformation of the field. During this period, computational hate speech analysis evolves from a specialized subdomain into a central and highly institutionalized research priority, attracting sustained scholarly investment across computer science, linguistics, and interdisciplinary governance studies.

A partial decline is observable in 2025, with output decreasing to 349 articles. However, this figure remains substantially above all pre-2021 levels and does not indicate a substantive contraction of research activity. Rather, it may reflect indexing delays, database coverage dynamics, or the beginning of stabilization following rapid expansion. The 2026 value of 48 articles is clearly incomplete and most plausibly represents a partial reporting year; it should therefore not be interpreted as evidence of a downturn.

Overall, the annual production pattern delineates three analytically distinct stages: an exploratory phase prior to 2017, characterized by minimal output; a consolidation phase between 2018 and 2019, marked by steady growth and methodological alignment; and an

acceleration phase from 2020 to 2024, defined by exponential expansion. The magnitude and velocity of growth in the latter period underscore that computational analysis of hate speech has become a core and rapidly evolving research agenda, driven by advances in machine learning and natural language processing and reinforced by intensifying societal and regulatory attention to online harms (see Figure 2).

4.3. Most Relevant Author Keywords

Figure 3 presents the most salient author keywords, ranked by frequency of occurrence in the dataset. The distribution is markedly concentrated around a limited set of highly recurrent terms. With 740 occurrences, the expression ‘hate speech’ overwhelmingly dominates the corpus, signaling strong thematic coherence and minimal dispersion toward adjacent or loosely affiliated phenomena. This concentration demonstrates that the dataset is not diffused across broader categories of online harm but is instead explicitly anchored in a clearly delimited conceptual object. In descending order of frequency, natural language processing (536 occurrences), machine learning (473), and deep learning (444) follow closely behind. The proximity among these methodological descriptors suggests that computational modelling does not merely supplement the field but constitutes its epistemic foundation. Their prominence indicates that algorithmic approaches function as the principal analytical infrastructure through which the phenomenon is interrogated and operationalized.

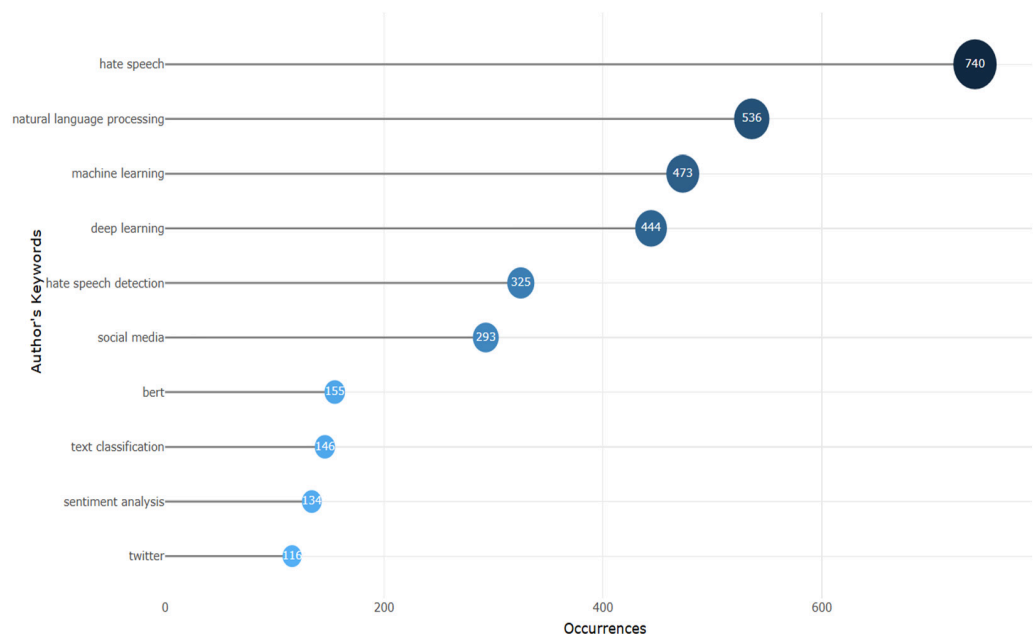


Figure 3. Most Relevant Author Keywords.

A second tier of keywords reflects increasing technical specificity and task orientation. The term ‘hate speech detection’ appears 325 times, indicating a notable shift from conceptual framing to applied system development. Although less frequent than the core phenomenon and the broader methodological categories, it is present in substantial numbers, indicating that a significant proportion of the literature is explicitly dedicated to operational detection and classification tasks. The recurrence of social media (293 occurrences) further situates the empirical focus within platform-mediated communication environments, indicating that digital platforms serve as the primary data source and application context for algorithmic interventions.

More specialized methodological terms appear less frequently: the BERT model is mentioned 155 times, text classification 146 times, and sentiment analysis 134 times. The

relatively limited visibility of specific architectures compared to broader methodological paradigms suggests a preference for general computational frameworks over particular technical implementations. Twitter, with 116 occurrences, represents a common empirical focus, indicating that datasets are concentrated around a limited number of platforms.

Ultimately, the keyword distribution outlines a research landscape shaped by the intersection of a clearly defined social phenomenon and computational language-processing techniques. The prevalence of methodological descriptors among the most frequent terms indicates that the literature is primarily structured around algorithmic detection frameworks, with digital communication platforms as the primary empirical environment for the development, testing, and deployment of these systems. Equally revealing is the absence of governance-oriented concepts among the most prominent keywords. Terms associated with regulation, accountability, transparency, platform governance, and human rights do not emerge as dominant thematic markers, suggesting that the field remains principally oriented toward technical detection challenges, while broader governance concerns occupy a comparatively peripheral position within the research landscape (see Figure 3).

4.4. Trend Topics

The trend topics graph in Figure 4 traces the temporal evolution of selected author keywords from 2019 to 2025. Horizontal lines denote the duration of prominence for each term, while bubble size represents relative frequency. Taken together, these visual cues illuminate shifts in methodological orientation and the progressive consolidation of thematic priorities within the field.

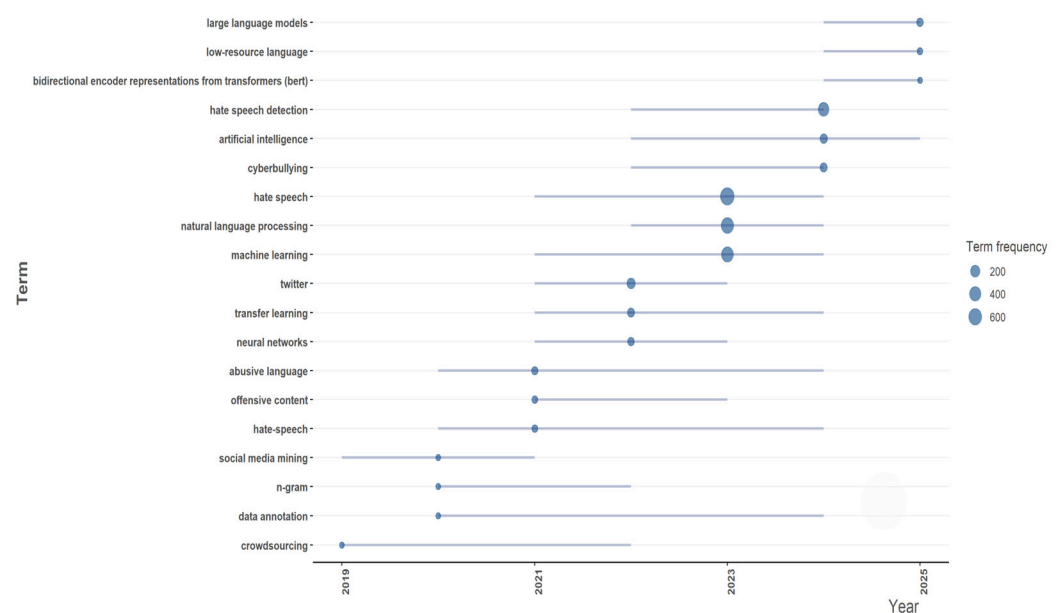


Figure 4. Trend Topics.

The earliest phase, spanning approximately 2019 to 2021, is marked by the prominence of terms such as crowdsourcing, data annotation, n-gram, and social media mining. These keywords reflect an initial stage primarily focused on infrastructure development: constructing labeled datasets, refining annotation protocols, and extracting data from platform-based environments. The visibility of crowdsourcing and annotation during this period indicates that substantial effort was directed toward assembling supervised learning corpora to support computational classification tasks. Likewise, the presence of n-gram signals relies on conventional text-representation techniques, characteristic of

pre-transformer approaches grounded in feature engineering rather than contextualized semantic modelling.

Beginning around 2021, a discernible methodological transition emerges. Terms such as neural networks, transfer learning, and machine learning expand both in temporal span and frequency, signaling a shift toward more sophisticated modelling strategies. By 2022–2023, natural language processing and hate speech have increased in prominence, with larger bubble sizes reflecting intensified research activity. This phase suggests a movement from exploratory experimentation toward methodological consolidation. Computational approaches, particularly those centered on language-based detection systems, appear to have matured into standardized analytical frameworks, reinforcing the structural centrality of NLP-driven architectures within the domain.

The most recent period, extending into 2024–2025, introduces keywords such as large language models, low-resource language, and bidirectional encoder representations from transformers. Their emergence signals a new stage characterized by technological refinement and conceptual expansion. The emergence of large language models signals engagement with generative, transformer-based architectures operating at scale, while attention to low-resource languages reflects growing concern for linguistic inclusivity and cross-lingual adaptation. This development suggests that the field is gradually extending beyond high-resource, predominantly Anglophone datasets toward more globally representative linguistic contexts.

Throughout the period, artificial intelligence and hate speech detection remain visible, underscoring their structural embeddedness within the research trajectory. Despite evolving methodologies, the core objective remains stable: the computational identification, classification, and mitigation of harmful linguistic content in digital communication environments.

Overall, the temporal configuration reveals three analytically distinct yet sequential phases: an initial infrastructure-building stage focused on annotation and feature extraction; a consolidation phase centered on neural and NLP-based modelling; and a recent expansion toward transformer architectures and multilingual challenges. This progression reflects not only increasing technical sophistication but also a widening epistemic and linguistic scope in the automated analysis of hate speech (see Figure 4).

4.5. Thematic Map

The thematic map organizes keyword clusters along two analytical dimensions: centrality, which represents the degree of interaction with other themes in the network, and density, which indicates the internal development and conceptual cohesion of a cluster. The four resulting quadrants distinguish motor themes (high centrality, high density), basic themes (high centrality, low density), niche themes (low centrality, high density), and emerging or declining themes (low centrality, low density).

The cluster comprising hate speech, machine learning, and deep learning is located within the basic themes quadrant. Its high centrality demonstrates its role as a structural core of the dataset, maintaining extensive connections with other thematic domains. However, its relatively low density indicates that the cluster remains conceptually broad and internally diverse. The combination of the social phenomenon of hate speech with broad methodological approaches such as machine learning and deep learning suggests that the field is primarily structured around applying computational techniques to a specific societal issue. Rather than constituting a highly specialized subdomain, this cluster serves as a foundational axis around which more technically differentiated themes are organized.

In contrast, the cluster comprising natural language processing, hate speech detection, and text classification falls within the motor themes quadrant. Its combination of high centrality and high density indicates both strong integration within the broader the-

matic network and significant internal consolidation. This suggests that language-centered computational detection has reached methodological maturity, as evidenced by stable terminology and frequent use. The prominence of text classification in this cluster highlights the operational focus on automated categorization of textual data, confirming that computational language analysis is the field's primary technical trajectory.

The upper-left quadrant contains the cluster of support vector machine, random forest, and logistic regression, which are categorized as niche themes. Although these approaches are well-developed internally, their relatively low centrality suggests limited interaction with dominant thematic areas. This suggests that classical machine learning methods persist in specific research strands but no longer define the field's main direction. Their continued presence reflects methodological continuity rather than current centrality. Finally, the cluster comprising BERT, a convolutional neural network, and an LSTM occupies the emerging or declining quadrant. Their low density and limited centrality indicate that these architectures, while technically significant, have not yet formed a cohesive, widely integrated thematic core within the dataset. This may reflect transitional dynamics in model architectures or fragmentation across specific applications.

Collectively, the map illustrates a field anchored in computational detection methodologies, with increasing consolidation around natural language processing. The thematic structure emphasizes the dominance of technical approaches to hate speech analysis and delineates a clear empirical context for situating broader governance and normative questions (see Figure 5).

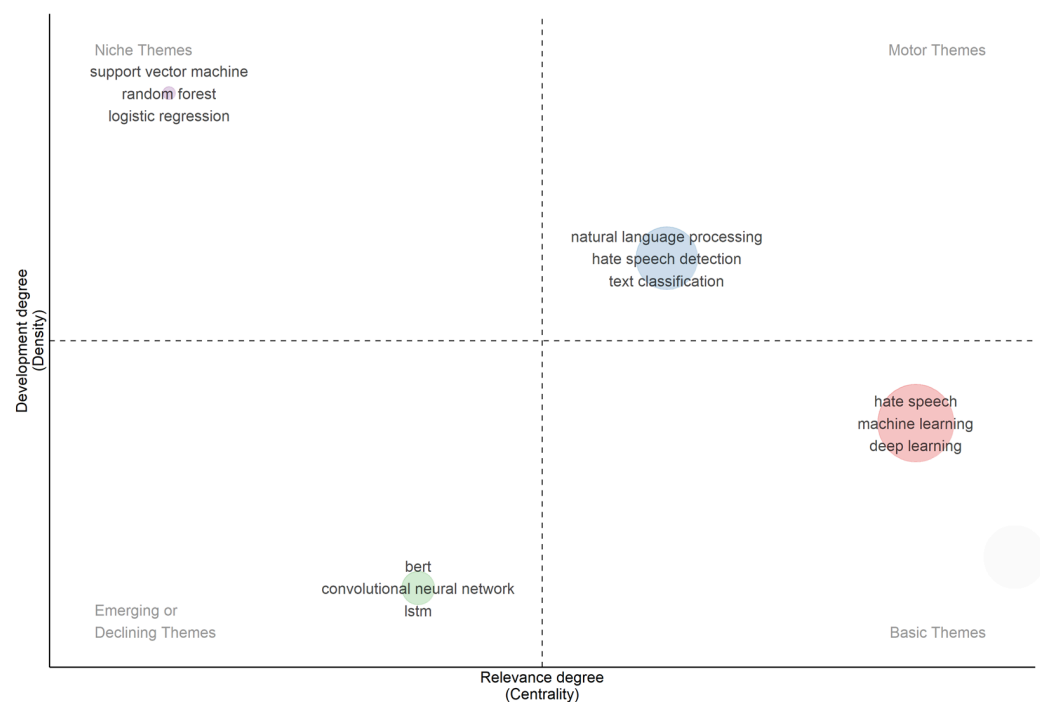


Figure 5. Thematic Map.

4.6. Top Ten Most Cited Papers

In Table 2, the top ten most cited papers are reported along with their total citation counts. Survey-oriented and framing contributions occupy a foundational position. Most prominently, Schmidt and Wiegand (2017), with 1136 citations, consolidate hate speech detection as a task within natural language processing while simultaneously foregrounding definitional ambiguity, contextual dependence, and the centrality of annotation practices. Their review systematizes modelling strategies and identifies persistent constraints, including dataset and domain dependence, limited cross-dataset comparability, and substantial

annotation variability. In parallel, [Gorwa et al. \(2020\)](#), cited 630 times, reconceptualize algorithmic content moderation as a socio-technical governance infrastructure rather than a purely technical classification problem. By emphasizing scale, institutional embedding, and the political consequences of classification choices, the analysis demonstrates that improvements in predictive accuracy alone cannot resolve concerns related to opacity, accountability, and legitimacy. Taken together, these two publications delineate the conceptual perimeter of the domain: one consolidates its technical foundations, while the other situates detection systems within regulatory and institutional architectures.

Table 2. The Top Ten Papers With the Most Citations.

Paper	Total Citations
(Schmidt & Wiegand, 2017) Socialnlp—Int Workshop Nat Lang Process Soc Media, Proc Workshop Afnlp Sig Socialnlp	1136
(Badjatiya et al., 2017) Int World Wide Web Conf, WWW Companion	1013
(Nobata et al., 2016) Int World Wide Web Conf, WWW	985
(Gorwa et al., 2020) Big Data Soc	630
(Zampieri et al., 2019) NaacL Hlt—Conf N Am Chapter Assoc Comput Linguistics: Hum Lang Technol—Proc Conf	616
(Burnap & Williams, 2015) Policy Internet	511
(Gambäck & Sikdar, 2017) Proc Annu Meet Assoc Comput Linguist	462
(Mathew et al., 2021) AAAI Conf Artif Intell, AAAI	452
(Watanabe et al., 2018) IEEE Access	369
(Kwok & Wang, 2013) Proc AAAI Conf Artif Intell, AAAI	345

Methodological consolidation and acceleration are strongly reflected in highly cited deep learning contributions. [Badjatiya et al. \(2017\)](#), with 1013 citations, report neural network experiments for tweet-level hate speech detection, demonstrating performance gains over earlier feature-engineered baselines on benchmark datasets. The study exemplifies the transition from manual feature construction to distributed representations and end-to-end learning architectures. Complementing this trajectory, [Gambäck and Sikdar \(2017\)](#), cited 462 times, introduce a convolutional neural network model for classifying short Twitter texts into racism and sexism categories, thereby reinforcing the applicability of neural approaches in noisy, user-generated language environments. A related operational orientation appears in [Watanabe et al. \(2018\)](#), with 369 citations, where a lexicon-driven data collection strategy is combined with supervised classification to distinguish hateful, offensive, and neutral content. Collectively, these works entrench the treatment of hate-related content as a detection task centered on extracting linguistic signals, with deep learning architectures emerging as the methodological standard.

A further cluster of highly cited publications provides the empirical and taxonomic infrastructures that enable cumulative research. [Nobata et al. \(2016\)](#), cited 985 times, develop large-scale abusive language detection across multiple topical domains using annotated user comments. The study explicitly demonstrates the limitations of keyword-based filtering and highlights cross-domain variability, thereby underscoring the contextual and distributional sensitivity of abusive language classification. [Zampieri et al. \(2019\)](#), with 616 citations, advances taxonomic refinement through a hierarchical annotation framework for offensive-language identification, distinguishing among offensive and non-offensive content, targeted versus untargeted offense, and the nature of the target. This layered structure extends analytical granularity beyond binary hate detection, enabling more differ-

entiated evaluation protocols. Earlier dataset-driven modelling is exemplified by Kwok and Wang (2013), cited 345 times, which formulates anti-Black racist content detection on Twitter as a supervised classification task while foregrounding the methodological difficulty of identifying low-base-rate harmful content within high-volume communication streams.

Another influential strand connects detection to governance-relevant contexts, interpretability, and decision support. Burnap and Williams (2015), with 511 citations, utilize human-annotated Twitter data following a specific event to train classifiers and subsequently apply statistical modelling to analyze the temporal diffusion of cyber hate. Here, detection is embedded within broader analytical objectives, including forecasting and situational awareness, rather than treated as an isolated optimization exercise. More recently, Mathew et al. (2021), cited 452 times, introduced the HateXplain benchmark dataset, incorporating not only hate speech labels but also identified target communities and human-provided rationales. By facilitating evaluation of model explanations and bias alignment, this framework reorients evaluative emphasis from predictive performance alone toward transparency, interpretability, and accountability.

To conclude, across the table, citation prominence converges around three interrelated functions: defining the conceptual and methodological boundaries of hate speech detection; consolidating technical practice through reusable neural baselines; and constructing datasets, taxonomies, and analytical infrastructures that enable governance-aware assessment. For scholarship on international policy and legislative frameworks for platform moderation, the most salient conceptual bridge emerges when detection systems are held accountable through structured taxonomies, transparent annotation logic, and interpretable outputs. These elements are central to evaluating automated moderation for fairness, proportionality, and regulatory oversight, thereby linking computational detection practices to normative and institutional considerations (see Table 2).

5. Discussion

The bibliometric evidence indicates that research on automated hate speech analysis has evolved into a distinctly technology-centered domain, with computational language modelling as the primary analytical orientation. The thematic structure reveals that the conjunction of hate speech with machine learning and deep learning functions as the principal organizing axis of the literature. Its position as a highly connected yet internally heterogeneous cluster suggests that a single dominant modelling paradigm does not define the field; rather, it is unified by the shared deployment of algorithmic approaches to a common social problem. By contrast, the cluster linking natural language processing with detection and classification appears both internally cohesive and structurally influential, indicating that language-driven automated categorization has consolidated as the most stable methodological trajectory. This configuration implies that the empirical investigation of hate speech has largely crystallized around text-processing pipelines, while alternative methodological strands remain comparatively marginal.

Patterns of keyword frequency reinforce this interpretation by disclosing an unusually concentrated thematic core. The predominance of hate speech as the principal descriptor, combined with the recurrent presence of central computational methodologies, confirms that the literature is firmly anchored in the algorithmic operationalization of a clearly delimited social phenomenon. Secondary terms associated with detection tasks and social media environments underscore the applied orientation of the research, signaling that platform-mediated communication spaces function as the primary empirical laboratory for model development and evaluation. The relatively limited salience of specific model architectures suggests that scholarly attention is directed less toward discrete technical implementations and more toward scalable methodological frameworks. Taken together,

the keyword configuration portrays a research landscape structured predominantly by methodological functionality rather than by competing theoretical traditions.

Temporal topic evolution further elucidates the emergence of this configuration. Early scholarly activity centered on dataset construction, annotation protocols, and conventional text-representation techniques, reflecting the infrastructural prerequisites of supervised learning paradigms. A subsequent phase incorporated neural modelling strategies and placed greater emphasis on computational language processing, marking the transition from exploratory experimentation to methodological consolidation. More recent developments indicate the integration of transformer-based architectures alongside growing concern for multilingual and low-resource contexts. This shift signals a movement beyond performance-driven optimization toward broader considerations of linguistic inclusivity and global applicability. The appearance of large-scale language models suggests that the research frontier is now oriented toward refinement, scalability, and cross-contextual adaptation rather than toward demonstrating the fundamental feasibility of automated detection.

Descriptive indicators in the dataset closely align with this developmental trajectory. The rapid proliferation of publications, coupled with a relatively young average document age, confirms that automated hate speech research is a dynamic, still-expanding scholarly domain. High levels of collaboration and substantial international co-authorship reflect the interdisciplinary character of the field, which necessitates integrating computational expertise, linguistic analysis, and socio-cultural interpretation. The predominance of conference proceedings over journal articles further signals a research culture shaped by accelerated technological iteration and compressed innovation cycles, features commonly associated with artificial intelligence research ecosystems. Collectively, these structural characteristics point to a domain that prioritizes methodological experimentation and technical advancement, while theoretical consolidation remains comparatively secondary.

Citation patterns among the most influential publications provide further insight into the intellectual structure of the field. Highly cited contributions tend to cluster around three interconnected domains: the clarification of definitional and annotation challenges, the demonstration of the advantages of neural architectures over earlier computational approaches, and the development of reusable datasets and taxonomic frameworks that support cumulative research. This distribution suggests that scholarly influence is derived not only from improvements in algorithmic performance but also from infrastructural contributions that standardize research tasks, benchmarking procedures, and evaluation criteria. The prominence of studies addressing annotation practices, dataset construction, and model interpretability further indicates a growing awareness that detection systems should be understood not merely as technical classifiers, but as socio-technical instruments situated within broader analytical, institutional, and regulatory environments. Nevertheless, the citation structure also reveals that scholarly attention remains concentrated on the design, assessment, and refinement of detection methodologies rather than on examining the prevalence, drivers, or societal consequences of hate speech itself. Consequently, much of the field's intellectual impact continues to stem from advances in classification techniques, data infrastructures, and evaluation frameworks, rather than from empirical investigations of hate speech as a social phenomenon.

Finally, longitudinal publication trends illustrate the transformation of automated hate speech analysis from a peripheral research endeavor into a large-scale and institutionalized scholarly agenda. Limited output before the late 2010s gave way to steady growth and, subsequently, rapid expansion, coinciding with the widespread adoption of deep learning techniques in natural language processing. The sustained high volume of publications in subsequent years indicates that computational hate speech analysis has become firmly established as a central research priority. This expansion reflects both technological matura-

tion and intensifying societal concern regarding digital communication risks, suggesting that methodological innovation and external institutional pressures have jointly contributed to the field's consolidation.

Overall, the cumulative bibliometric evidence portrays a research domain whose empirical core remains firmly anchored in automated language processing and scalable detection systems. The predominance of computational methodologies, the concentration of empirical inquiry on social media platforms, and the sustained growth in publication output collectively suggest that scholarly attention has been directed primarily toward enhancing detection capabilities. In contrast, the comparatively limited visibility of governance-related themes indicates that normative and regulatory considerations have received considerably less attention than methodological advancement. This imbalance is particularly significant because automated hate speech detection systems do more than identify harmful content; they influence the conditions under which expression, participation, and protection are negotiated within digital public spheres. Although such technologies may contribute to safeguarding targeted communities and reducing exposure to harmful communicative environments, they may also restrict lawful expression or reproduce existing biases affecting minority groups when implemented without adequate safeguards for transparency, accountability, and proportionality. Taken together, these findings point to a persistent gap between progress in computational detection and the governance mechanisms necessary to ensure that digital content moderation remains both rights-sensitive and publicly accountable.

6. Conclusions

This study demonstrates that research on computational hate speech detection has consolidated around language-processing methodologies that serve as the field's dominant analytical infrastructure. Bibliometric evidence indicates that natural language processing, machine learning, and deep learning collectively constitute the structural core through which hate-related communication is operationalized, categorized, and empirically interrogated. Thematic mapping reveals that broad computational paradigms provide the foundational axis of the research landscape, while language-centered detection systems exhibit the highest degree of methodological consolidation and internal coherence. At the same time, classical machine-learning approaches persist within specialized research strands, and emergent neural architectures continue to surface without yet crystallizing into fully integrated thematic cores. These patterns point to a domain characterized by simultaneous methodological stabilization and ongoing architectural experimentation.

Temporal analysis further delineates a sequential developmental trajectory. Initial scholarly activity centered on the construction of annotated datasets, the refinement of labelling protocols, and the application of conventional feature-engineering techniques required for supervised classification. This infrastructure-building phase was succeeded by a period of consolidation marked by the widespread adoption of neural modelling strategies and the normalization of language-processing pipelines as the default analytical framework. More recent developments signal a turn toward transformer-based architectures and an expanding focus on multilingual and low-resource contexts. This progression reflects not only increasing technical sophistication but also a growing recognition that computational detection systems must operate across heterogeneous linguistic environments and globally distributed communication platforms.

Descriptive indicators reinforce the interpretation of a rapidly expanding and technologically dynamic research ecosystem. A pronounced annual growth rate, the concentration of publications within recent years, and elevated levels of collaborative authorship collectively demonstrate that AI-driven hate speech detection has matured into a large-scale interdisciplinary research priority. The predominance of conference proceedings over

journal articles further suggests that innovation cycles remain closely aligned with the accelerated rhythms characteristic of contemporary artificial intelligence research cultures. Citation patterns among the most influential studies indicate that scholarly impact is concentrated in contributions that clarify annotation practices, develop reusable and benchmark datasets, and establish methodological baselines for neural classification. This distribution underscores the centrality of infrastructural and evaluative frameworks alongside incremental improvements in predictive performance.

Collectively, these findings suggest that the scholarly literature is primarily oriented toward the development of scalable algorithmic detection systems, with social media platforms constituting the dominant empirical setting for their design, testing, and implementation. This technical focus carries important governance implications. First, the operational categories employed in computational detection may blur the distinction between hate speech as a legal and human rights concept and broader forms of harmful online communication, such as offensive language, toxicity, and abuse. Second, automated detection systems may have significant implications for freedom of expression when classification errors result in the removal, demotion, or indirect suppression of lawful content. Third, the limited prominence of governance-related concepts within the keyword structure indicates that issues such as fairness, accountability, and transparency remain secondary to the development of detection capabilities. Taken together, these challenges underscore the need for transparent, accountable, and proportionate governance frameworks that extend beyond measures of classification performance and address the broader social and normative consequences of automated content moderation.

Future research should move toward more fully operationalized forms of interdisciplinary integration that connect computational modelling practices with governance-oriented evaluative criteria. Particularly, funding agencies and research councils could incorporate multilingual benchmark evaluation and subgroup-sensitive performance assessment as explicit project requirements. Such measures would strengthen the structural alignment between technical development and normative commitments to equity, ensuring that questions of fairness are not treated as peripheral add-ons but as constitutive dimensions of system design. In parallel, leading conference venues in computational linguistics and AI might establish dedicated tracks or workshops that explicitly link modelling performance to fairness auditing and regulatory compliance frameworks, thereby institutionalizing dialogue between technical innovation and public accountability.

Greater attention to interpretable decision-making workflows can be advanced by systematically integrating established explanation techniques, such as LIME, SHAP, and counterfactual explanation methods, alongside structured documentation instruments, including model cards and dataset datasheets. Embedding these tools within content moderation pipelines would facilitate more transparent reporting of system limitations, epistemic uncertainties, and decision rationales, particularly in high-impact deployment contexts where automated judgments may carry significant social consequences. Such integration would also promote more reflexive forms of model evaluation, encouraging practitioners to articulate the normative assumptions embedded within algorithmic architectures.

Expanding empirical coverage beyond a narrow set of dominant platforms and frequently reused datasets is equally imperative. A broader evidentiary base would contribute to a more representative and contextually grounded understanding of online communicative dynamics, especially across diverse linguistic, cultural, and regulatory environments. Structured collaboration mechanisms, such as joint funding calls between computer science and law faculties, interdisciplinary doctoral training networks, and co-organized academic-policy workshops, offer promising institutional pathways for sustained cross-disciplinary engagement. More fundamentally, durable collaboration among computational researchers,

social scientists, and policy specialists is most likely to prove effective when embedded at the design stage of research projects rather than appended retrospectively. Such early integration ensures that AI-based hate speech detection systems evolve not merely as technically efficient classifiers, but as accountable socio-technical instruments capable of supporting equitable, transparent, and context-sensitive moderation practices within increasingly complex digital communication ecosystems.

Author Contributions: Conceptualization: N.K. and M.N.; methodology: N.K. and M.N.; software: M.N.; validation: N.K.; formal analysis: M.N.; investigation: N.K. and M.N.; resources: N.K. and M.N.; data curation: N.K. and M.N.; writing—original draft preparation: N.K. and M.N.; writing—review and editing: N.K.; visualization: M.N.; supervision: N.K.; project administration: N.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Alkiviadou, N. (2018). The legal regulation of hate speech: The international and European frameworks. *Politička Misao*, 55(4), 203–229. [\[CrossRef\]](#)
- Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. [\[CrossRef\]](#)
- Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105, 101584. [\[CrossRef\]](#)
- Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975. [\[CrossRef\]](#)
- Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1), 377–386. [\[CrossRef\]](#)
- Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April 3–7). *Deep learning for hate speech detection in tweets*. Companion: Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759–760), Perth, Australia. [\[CrossRef\]](#)
- Blodgett, S. L., Green, L., & O'Connor, B. (2016, November 1–4). *Demographic dialectal variation in social media: A case study of African-American English*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1119–1130), Austin, TX, USA. [\[CrossRef\]](#)
- Brown, A. (2015). *Hate speech law: A philosophical examination*. Routledge.
- Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2), 223–242. [\[CrossRef\]](#)
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512. [\[CrossRef\]](#)
- Buyse, A. (2014). Dangerous expressions: The echr, violence and free speech. *International and Comparative Law Quarterly*, 63(2), 491–503. [\[CrossRef\]](#)
- Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the ‘good society’: The US, EU, and UK approach. *Science and Engineering Ethics*, 24(2), 505–528. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can’t stay here. *Proceedings of the ACM on Human-Computer Interaction*, 1, 1–22. [\[CrossRef\]](#)
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), e2023301118. [\[CrossRef\]](#) [\[PubMed\]](#)
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512–515. [\[CrossRef\]](#)
- DeNardis, L., & Hackl, A. (2015). Internet governance by social media platforms. *Telecommunications Policy*, 39(9), 761–770. [\[CrossRef\]](#)
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the Association for Computational Linguistics: Human language technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. [\[CrossRef\]](#)

- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. [\[CrossRef\]](#) [\[PubMed\]](#)
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4), 1–30. [\[CrossRef\]](#)
- Frosio, G., & Geiger, C. (2023). Taking fundamental rights seriously in the digital services act's platform liability regime. *European Law Journal*, 29(1–2), 31–77. [\[CrossRef\]](#)
- Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. UNESCO Publishing.
- Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online, Vancouver, BC, Canada, August 4* (pp. 85–90). Association for Computational Linguistics. [\[CrossRef\]](#)
- Gillespie, T. (2018). Custodians of the internet. In *Yale university press eBooks*. Yale University Press. [\[CrossRef\]](#)
- Gorwa, R. (2019). What is platform governance? *Information Communication & Society*, 22(6), 854–871. [\[CrossRef\]](#)
- Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 205395171989794. [\[CrossRef\]](#)
- Grizzle, A., Moore, P., Dezuanni, M., Asthana, S., Wilson, C., Banda, F., & Onumah, C. (2014). *Media and information literacy: Policy and strategy guidelines*. UNESCO.
- Heinze, E. (2016). *Hate speech and democratic citizenship*. Oxford University Press.
- Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation. *ACM Transactions on Computer-Human Interaction*, 25(2), 1–33. [\[CrossRef\]](#)
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics, Online, July 5–10* (pp. 6282–6293). Association for Computational Linguistics. [\[CrossRef\]](#)
- Kapelańska-Pręgowska, J., & Pucelj, M. (2023). Freedom of expression and hate speech: Human rights standards and their application in Poland and Slovenia. *Laws*, 12(4), 64. [\[CrossRef\]](#)
- Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1), 1621–1622. [\[CrossRef\]](#)
- Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 16(10), 31–43. [\[CrossRef\]](#)
- Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information Communication & Society*, 20(6), 930–946. [\[CrossRef\]](#)
- Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 14867–14875. [\[CrossRef\]](#)
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. [\[CrossRef\]](#)
- Mihailidis, P., & Viotty, S. (2017). Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in “Post-Fact” society. *American Behavioral Scientist*, 61(4), 441–454. [\[CrossRef\]](#)
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency, Atlanta, GA, USA, January 29–31* (pp. 220–229). Association for Computing Machinery. [\[CrossRef\]](#)
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web, Montreal, QC, Canada, April 11–15* (pp. 145–153). International World Wide Web Conferences Steering Committee. [\[CrossRef\]](#)
- 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*, 372, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
- Peitz, M. (2025). Governance and regulation of platforms. In *Handbook of new institutional economics* (pp. 565–593). Springer. [\[CrossRef\]](#)
- Penney, J. W. (2016). Chilling effects: Online surveillance and Wikipedia use. *eYLS (Yale Law School)*, 31(1), 117. [\[CrossRef\]](#)
- Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5–14. [\[CrossRef\]](#)
- Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. a. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency, Barcelona, Spain, January 27–30* (pp. 131–141). Association for Computing Machinery. [\[CrossRef\]](#)
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should i trust you?”. In *The 22nd ACM SIGKDD international conference, San Francisco, CA, USA, August 13–17* (pp. 1135–1144). Association for Computing Machinery. [\[CrossRef\]](#)

- Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing, Online, August 1–6* (pp. 41–58). Association for Computational Linguistics. [\[CrossRef\]](#)
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics, Florence, Italy, July 28–August 2* (pp. 1668–1678). Association for Computational Linguistics. [\[CrossRef\]](#)
- Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media, Valencia, Spain, April 3* (pp. 1–10). Association for Computational Linguistics. [\[CrossRef\]](#)
- Schotten, M., Aisati, M. E., Meester, W. J. N., Steinginga, S., & Ross, C. A. (2017). A brief history of Scopus: The world's largest abstract and citation database of scientific literature. In *Research analytics boosting university productivity and competitiveness through scientometrics* (pp. 31–58). Auerbach Publications. [\[CrossRef\]](#)
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency, Atlanta, GA, USA, January 29–31* (pp. 59–68). Association for Computing Machinery. [\[CrossRef\]](#)
- Suzor, N., Van Geelen, T., & West, S. M. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette*, 80(4), 385–400. [\[CrossRef\]](#)
- Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press.
- Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature*, 556(7701), 296–298. [\[CrossRef\]](#) [\[PubMed\]](#)
- Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, 15(12), e0243300. [\[CrossRef\]](#) [\[PubMed\]](#)
- Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science, Austin, TX, USA, November 5* (pp. 138–142). Association for Computational Linguistics. [\[CrossRef\]](#)
- Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop, San Diego, CA, USA, June 12–17* (pp. 88–93). Association for Computational Linguistics. [\[CrossRef\]](#)
- Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6, 13825–13835. [\[CrossRef\]](#)
- York, J. C., & Zuckerman, E. (2019). Moderating the public sphere. In *Human rights in the age of platforms* (pp. 137–162). The MIT Press. [\[CrossRef\]](#)
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Minneapolis, MN, USA, June 2–7* (pp. 1415–1420). Association for Computational Linguistics. [\[CrossRef\]](#)
- Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Lecture notes in computer science* (pp. 745–760). Springer. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.