Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery

WILEY

WIREs DATA MINING AND KNOWLEDGE DISCOVERY

ADVANCED REVIEW OPEN ACCESS

A Literature Review of Textual Cyber Abuse Detection Using Cutting-Edge Natural Language Processing Techniques: Language Models and Large Language Models

J. Angel Diaz-Garcia[1] | Joao Paulo Carvalho[2]

[1]Department of Computer Science and A.I, University of Granada, Granada, Spain | [2]INESC-ID, Instituto Superior Tecnico, Universidade de Lisboa, Lisboa, Spain

Correspondence: J. Angel Diaz-Garcia (joseangeldiazg@ugr.es)

Received: 22 November 2024 | **Revised:** 6 June 2025 | **Accepted:** 7 June 2025

Editor-in-Chief: Witold Pedrycz

Funding: Funding for open access charge: Universidad de Granada/CBUA. This research was funded by the European Union, the Spanish Ministry of Science, Innovation, and Universities, and the University of Granada. European Commission (NextGenerationEU/PRTR), and the DesinfoScan project: Grant TED2021-129402B-C21 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and FederaMed project: Grant PID2021-123960OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. Finally, the research reported in this paper is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309). This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project (UIDB/50021/2020), and by the Recovery and Resilience Plan (RRP) and Next Generation EU European Funds through project C644865762-00000008 Accelerat.AI.

Keywords: cyberabuse | generative AI | literature review | LLMs | NLP | social media analysis

ABSTRACT

The success of social media platforms has facilitated the emergence of various forms of online abuse within digital communities. This abuse manifests in multiple ways, including hate speech, cyberbullying, emotional abuse, grooming, and shame sexting or sextortion. In this paper, we present a comprehensive analysis of the different forms of abuse prevalent in social media, with a particular focus on how emerging technologies, such as Language Models (LMs) and Large Language Models (LLMs), are reshaping both the detection and generation of abusive content within these networks. We delve into the mechanisms through which social media abuse is perpetuated, exploring the psychological and social impact. To achieve this, we conducted a literature review based on PRISMA methodology, deriving key insights in the field of cyber abuse detection. Additionally, we examine the dual role of advanced language models—highlighting their potential to enhance automated detection systems for abusive behavior while also acknowledging their capacity to generate harmful content. This paper contributes to the ongoing discourse on online safety and ethics by offering both theoretical and practical insights into the evolving landscape of cyber abuse, as well as the technological innovations that simultaneously mitigate and exacerbate it. The findings support platform administrators and policymakers in developing more effective moderation strategies, conducting comprehensive risk assessments, and integrating AI responsibly to create safer digital environments.

This article is categorized under:

Algorithmic Development > Web Mining

Technologies > Classification

JEL Classification: Web Mining, Classification

1 | Introduction

In the digital age, social networks and online platforms have become integral to everyday life, facilitating communication, information sharing, and social interaction. However, this digital connectivity has also given rise to various forms of cyber abuse, affecting millions of individuals globally (Altobelli 2010). The pervasive nature of cyber abuse, including hate speech, cyberbullying, emotional abuse, doxing, trolling, impersonation, and shaming, poses significant risks to mental health and well-being, causing psychosocial problems (Kwan et al. 2020) as anxiety or depression (Fisher et al. 2016).

Traditional methods of addressing cyber abuse have often relied on manual reporting systems and heuristic-based detection (Vandebosch and Van Cleemput 2009). However, recent advancements in artificial intelligence (AI), particularly through LLMs and LMS, offer new avenues for understanding and combating these issues (Rajeshwari and Divya 2025; Albladi et al. 2025). Language Models (LMs) and Large Language Models (LLMs) process sequences of tokens (words), denoted as $w = \{w_1, w_2, ..., w_N\}$, by assigning a probability $p(w)$ to the sequence. These models incorporate information from previous tokens to predict the next token in the sequence (Li, Huang, et al. 2024). Through training, LMs learn relationships between words, providing tools for tasks such as question answering, relation extraction, or natural language translation.

At this point, it is important to distinguish between LMs and LLMs. The primary difference lies in their size. LMs are typically smaller in terms of the number of parameters, while LLMs are significantly larger, with billions or even trillions of parameters. Due to this, LLMs have far greater computational capacity (Kumar 2024). Traditional LMs have limited capacity to handle complex tasks, smaller datasets, or nuanced language patterns. In contrast, LLMs have the vast capacity to understand and generate complex language, as they can process extensive datasets that are closely tied to context. For example, while a typical language model may have a context window of 1024 tokens, models like GPT-3 can handle up to 8192 tokens, and GPT-4 even supports much larger windows.

Furthermore, LMs and LLMs differ not only in terms of size but also in their training processes and practical applications. LLMs require more extensive computational resources for training and can be used for a wider variety of NLP tasks, including those that involve deeper contextual understanding and longer-term dependencies. LLMs, such as GPT (Radford et al. 2018), Mistral (Jiang et al. 2023), or T5 (Raffel et al. 2020) and other state-of-the-art LMs, such as BERT (Devlin et al. 2019) or RoBERTa (Liu et al. 2019), have demonstrated remarkable capabilities in natural language processing, providing powerful tools for analyzing and detecting harmful content.

This paper presents a literature review on the current use of advanced LMs and LLMs in detecting different forms of cyber abuse. Our goal is to bring together and analyze recent research to better understand how these technologies are being used to identify, classify, and respond to online abuse. This review is important because cyber abuse continues to cause serious harm to individuals and communities, including emotional distress and, in extreme cases, self-harm or suicidal thoughts (Hamm et al. 2015; Kim et al. 2019; Kowalski et al. 2014). Although there has been considerable work on well-known issues like hate speech and cyberbullying, other types of abuse—such as doxing, impersonation, and trolling—are still not well covered in existing reviews. Our work addresses this gap and helps guide future research toward more complete and inclusive solutions.

In this paper, we adopt the PRISMA methodology (Moher et al. 2010) to conduct a structured and systematic literature review aligned with our research objectives. Our review offers a comprehensive overview of how language models, including both traditional and large-scale architectures, are being applied to detect and address various forms of cyber abuse—ranging from well-established categories to those that remain underexplored. Through this analysis, we underscore the growing potential of AI-driven approaches to provide more nuanced, scalable, and effective solutions for promoting safer and more inclusive online environments.

In recent years, several surveys have explored hate speech detection (Fortuna and Nunes 2018); however, there is a noticeable gap in covering the latest advancements, particularly the emergence of cutting-edge models like LLMs that have been released in the past year. Therefore, our primary goal is to update the existing literature with these recent developments. Additionally, since most studies in this domain approach the problem as a classification task, we aim to conduct an in-depth analysis of the evaluation processes, especially considering the challenges posed by the unbalanced nature of cyber abuse detection. Moreover, existing reviews predominantly concentrate on cyberbullying (Salawu et al. 2017) or hate speech (Paz et al. 2020; Schmidt and Wiegand 2017), leaving other forms of cyber abuse significantly underrepresented. With this in mind, the main contributions of our paper are as follows:

- We present the most up-to-date analysis of textual cyber abuse, with a particular emphasis on cutting-edge techniques that leverage LMs and LLMs, identifying current gaps and suggesting avenues for future research and improvement.
- To the best of our knowledge, we conduct the first literature review that addresses a broader spectrum of cyber abuse forms, including doxing, trolling, and impersonation, which have been largely overlooked in previous reviews.
- We perform an in-depth analysis of the evaluation metrics used in cyber abuse detection, paying special attention to the challenges posed by class imbalance, a common issue in this domain.

Our paper aligns these contributions with a set of five research questions, each designed to provide specific and informative insights into key challenges and developments emerging from the rise of LMs and LLMs in the context of cyber abuse detection. To this end, we focus on the following research questions:

RQ1. *How representative are LMs and LLMS in detecting textual cyber abuse compared to traditional methods?*

RQ2. *Considering that cyber abuse frequently involves imbalanced classification challenges, are researchers adequately addressing this issue in their evaluation metrics and pre-processing methods?*

RQ3. *Based on the reviewed literature, what emerging trends and future directions are shaping the development of cyber abuse detection?*

RQ4. *Which types of cyber abuse detection are most significantly influenced by LLMs?*

RQ5. *Are there specific forms of cyber abuse that remain underexplored and lack adequate detection solutions?*

To address these questions, the paper is organized as follows: Section 2 examines the various forms of textual cyber abuse encountered in online social interactions. Section 4 analyzes the impact of LMs and LLMs in two main aspects: their role in detecting cyber abuse and hate speech, as well as their potential in generating automated instances of such content. Section 5 delves into how artificial intelligence is advancing the detection and analysis of other, less-explored forms of cyber abuse. Finally, we present a comprehensive discussion in Section 7 and draw our conclusions in Section 8.

2 | Forms of Textual Cyberabuse

In this section, we explore the various forms of textual cyber abuse prevalent in social networks. Each subsection delves into the definition, characteristics, and impact of a specific type of abuse.

In (van Huijstee et al. 2022), the authors identify a range of harmful and immoral uses of social media, categorizing these into six primary types, which can be further analyzed into 22 distinct tasks. The full categorization is depicted in Figure 1, reproduced from (van Huijstee et al. 2022).

It is important to note that the advent of generative AI has boosted all types of cyber abuse, including the production of fake sexual content (Arteaga 2024), the generation of deceptive

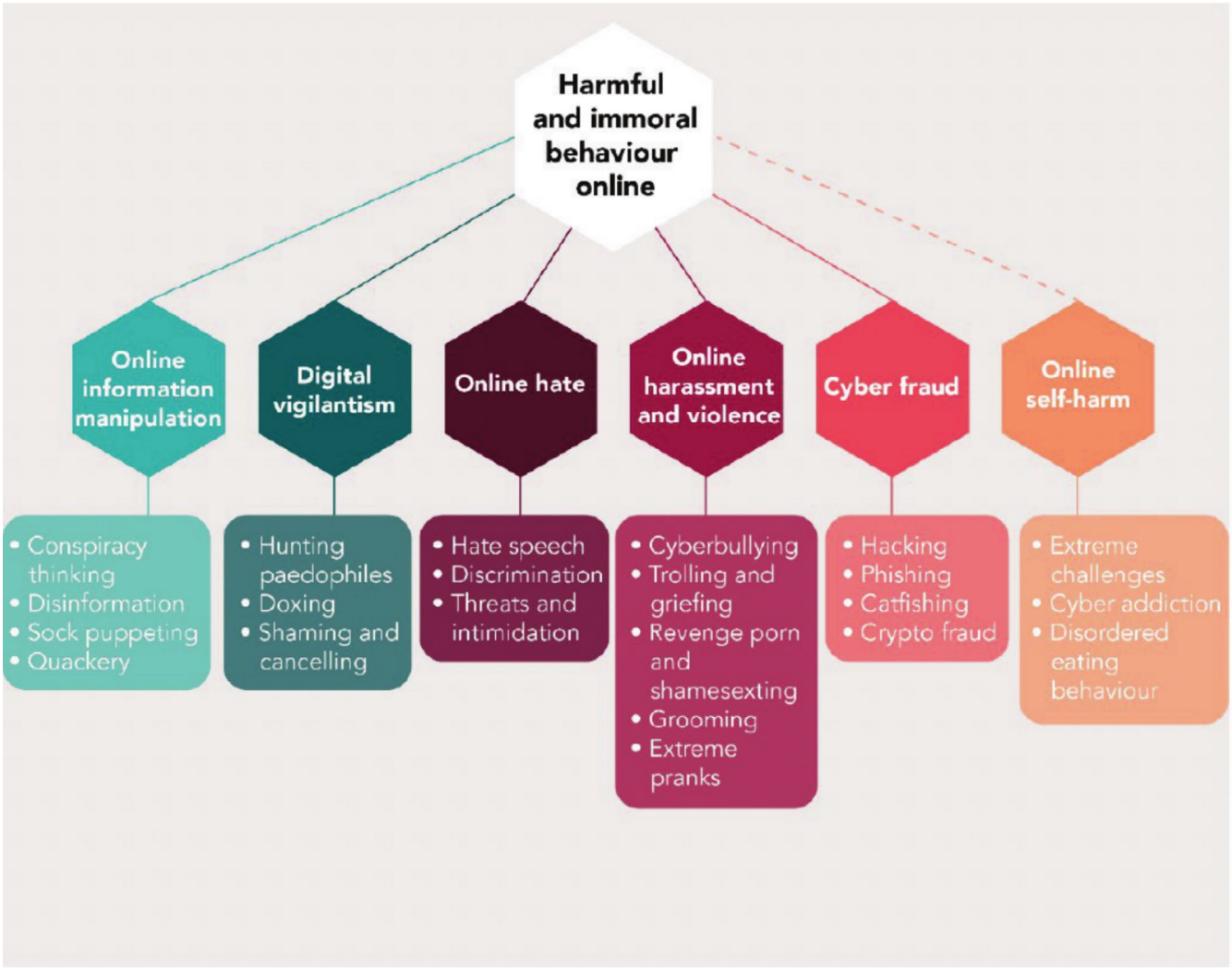

FIGURE 1 | Harmful and immoral behavior online (Hossain et al. 2022, 6).

images (Appel and Prietzel 2022), and the mass distribution of both real and created content on social media platforms (Franks and Waldman 2018). These types of abuse are beyond the scope of this work because we focus exclusively on text content. Therefore, our analysis is centered on various types of text-based cyber abuse.

To align with existing research in the field, we have selected the following sub-genres of cyber abuse in social media for our analysis: hate speech, cyberbullying, doxing, shaming, cancel culture, and trolling as the primary components. Additionally, we include other forms of cyber abuse such as impersonation (Gharawi et al. 2021) and emotional abuse (Stephenson et al. 2018). In the following subsections, we introduce each of these subgenres in detail.

2.1 | Hate Speech

Hate speech includes language that belittles, threatens, or encourages violence against individuals or groups based on characteristics like religion (Albadi et al. 2018), ethnicity, gender, sexual orientation, or other defining traits (Sugarman et al. 2018). It can seriously undermine personal dignity, democratic values, social stability, and peace. Yet, despite its far-reaching consequences, many people remain unaware of just how deeply hate speech can affect different aspects of life (Ștefăniță and Buf 2021).

Hate speech on social media is a pressing social issue. Research indicates that exposure to such speech can lead to a range of mental health problems, particularly among young adults, as well as psychosocial issues (Kwan et al. 2020). Consequences may include anxiety (Fisher et al. 2016), depression (Fisher et al. 2016; Hamm et al. 2015; Wachs et al. 2022), and post-traumatic stress disorder (Cassiani-Miranda et al. 2022; Liu et al. 2020; Wypych and Bilewicz 2022).

Detecting hate speech requires analyzing text for derogatory terms, slurs, and contextual indicators of malicious intent. NLP techniques, and more recently LMs and LLMs, are crucial tools in this effort. These technologies can automatically capture the specific nuances of hate speech in social networks with high accuracy. This survey focuses on recent advancements in hate speech detection, particularly the integration of new language models to enhance detection capabilities.

2.2 | Cyberbullying

Cyberbullying affects both children and adolescents and is widely recognized as intentional, violent, cruel, and repetitive behavior directed at peers. This pervasive issue is identified by the US Centers for Disease Control and Prevention (CDC) as a significant public health threat (Centers for Disease Control and Prevention 2009). The emotional and psychological damage caused by cyberbullying can be severe, with extreme cases potentially leading to suicide (Bauman et al. 2013).

According to (McLoughlin and Burgess 2009), cyberbullying can be categorized into several types, including flaming, harassment, denigration, impersonation, outing, boycott, and cyberstalking. Particularly in more severe forms such as flaming and harassment, cyberbullying often manifests through texts, comments, or messages exchanged between the bully and the victim. These communications typically involve insults and attacks related to race, sexuality, ethnicity, or social status. Such messages represent user-generated content in an unstructured format, which can be analyzed automatically using AI systems based on NLP and LLMs to identify and address hate speech.

In the current digital era, cyberbullying has evolved to encompass a wide range of multimodal categories, including the creation of sexual deep fakes targeting individuals (Rini and Cohen 2022; Rousay 2023; Dunn 2024). While this is a significant area of concern, our focus remains on the textual aspects of cyberbullying, specifically analyzing comments and interactions on social networks and online platforms.

2.3 | Emotional and Psychological Abuse

This form of abuse involves manipulative or controlling language aimed at undermining an individual's emotional well-being (Akar 2025). Examples include gaslighting (Divon et al. 2025; Thulin and Heinze 2025), persistent criticism (Aitken et al. 2025), and verbal assaults (Putri et al. 2025). While emotional and psychological abuse often overlaps with cyberbullying, especially since all forms of cyberbullying inherently involve emotional harm, it is important to distinguish between the two. Cyberbullying is more commonly associated with early life stages, such as during school or adolescence, whereas psychological abuse can occur at any stage of life, including in relationships with partners in adulthood (de la Villa Moral and Prieto 2022). So in our survey, we aim to establish a more fine-grained distinction by categorizing and summarizing various forms of cyberbullying and their intersections with artificial intelligence, with a particular focus on LLMs. Regarding psychological abuse, our inclusion criteria specifically exclude cases of psychological abuse occurring during childhood or other life stages unrelated to online interactions (Iyer et al. 2022), as our focus is solely on cyber abuse, and the realm of these kinds of applications is more related to the medicine area (Luo et al. 2022) or psychology (Ronneberg et al. 2024).

This form of abuse involves the use of manipulative or controlling language intended to diminish an individual's emotional stability and psychological well-being (Akar 2025). Common examples include gaslighting (Divon et al. 2025; Thulin and Heinze 2025), persistent criticism (Aitken et al. 2025), and direct verbal aggression (Putri et al. 2025). Although emotional and psychological abuse frequently overlaps with cyberbullying, given that all forms of cyberbullying inherently inflict emotional distress, it is important to differentiate between the two. Cyberbullying typically emerges during earlier life stages, such as adolescence or school years, whereas psychological abuse may occur across the lifespan, including in adult intimate relationships (de la Villa Moral and Prieto 2022). In this review, we aim to draw a more precise distinction by organizing and summarizing the various forms of cyberbullying and their intersections with artificial

intelligence, with a particular emphasis on the role of LLMs. Concerning psychological abuse, we have specifically excluded instances that occur outside the context of online environments—such as those related to childhood trauma or non-digital interpersonal relationships—from our analysis (Iyer et al. 2022), as such cases fall more directly within the domains of clinical medicine (Luo et al. 2022) or traditional psychology research (Ronneberg et al. 2024).

2.4 | Doxing

Doxing refers to the unauthorized publication of an individual's private or personally identifiable information without their consent (Karimi et al. 2022). This form of cyber abuse can take various forms, such as posts or messages that expose sensitive details like home addresses, phone numbers, or other confidential data. Doxing is often comparable to a targeted data breach and is usually carried out with malicious intent—intended to threaten, harass, or cause reputational or emotional harm to the victim. On social media platforms, the consequences of doxing can be severe and may include:

- **Privacy Violations**: Doxing constitutes a serious breach of personal privacy by making sensitive information publicly accessible without consent. This exposure often results in unwanted scrutiny and targeted harassment (Bottlapally and Zhou 2025).
- **Emotional and Psychological Impact**: Individuals targeted by doxing may suffer significant psychological consequences, such as increased anxiety or fear stemming from the loss of control over their personal data (Shan et al. 2024).
- **Professional Consequences**: Doxing can have detrimental effects on a person's career, damaging professional relationships, leading to job loss, and causing reputational harm (Solo 2025).

In the literature, it is essential to distinguish between doxing experienced by individuals and that which affects organizations (Herrera Montano et al. 2022). This paper focuses exclusively on the impact of doxing on individuals. Finally, it is necessary to mention that data leakage in the realm of machine learning can be related to the unintentional inclusion of information from the test set in the training process, this topic falls out of our analysis.

2.5 | Trolling

According to Bishop (Bishop 2013), trolling is defined as "the activity of posting messages via a communications network that are intended to be provocative, offensive, or menacing." This definition highlights that trolling involves posting inflammatory or irrelevant messages with the intent to provoke or disrupt online conversations.

In the literature, various approaches categorize trolling not only by the content of the messages but also by the underlying intentions. For example, De et al. (de la Vega and Ng 2018) identify four key aspects of trolling:

- **Intention (I):** The author's underlying purpose or motive behind the trolling activity.
- **Intention Disclosure (D):** This aspect evaluates whether the author is concealing their true (malicious) intentions from the audience.
- **Intention Interpretation (R):** This reflects how responders perceive and interpret the troll's intentions.
- **Response Strategy (B):** This describes the responder's reaction to the trolling attempt, which may also be a form of trolling.

Trolling is a significant issue that requires attention, as highlighted by De et al. (de la Vega and Ng 2018). Their study found that in 26.9% of trolling incidents, the victims engage with the trolling attempt, often experiencing some form of emotional distress.

In our survey, we concentrate on the core aspects of trolling within social network conversations, specifically examining how trolling manifests through interactions in textual dialogues. We exclude from our analysis newer forms of trolling, such as those involving memes, as they represent a different avenue of exploration that falls outside the direct scope of conversational interactions (Hossain et al. 2022; Suryawanshi et al. 2023).

2.6 | Impersonation

Impersonation involves creating fake profiles or hacking into existing accounts to deceive others. Typically, impersonators target brands or well-known individuals, aiming to extract some form of value through their deception (Zarei et al. 2020).

Numerous studies have explored the issue of impersonation, with a particular focus on understanding its impact and vulnerabilities. The findings from Yu et al. (2023) on the susceptibility of older adults to government impersonation scams highlight alarming trends that call for urgent, targeted interventions. Approximately 16% of participants engaged in conversations with an agent impersonating a government representative without displaying skepticism, indicating a high risk of victimization within this demographic. Even more concerning, 12% of participants willingly shared personal information, and nearly 5% provided the last four digits of their Social Security number. While impersonation has been widely recognized and studied as a growing problem, efforts to detect and prevent these scams remain inadequate and underdeveloped, leaving vulnerable populations at continued risk.

Impersonation often relies on the use of altered images, deepfakes, or even biometric information of the scammed individuals (Liu et al. 2024). Also, we found a stream regarding the impersonations and infiltration between networks, as a kind of network attack (Shukla 2023). In this survey, we have excluded these types of impersonation detection methods, as they fall within the domain of image analysis. Our focus is on text-based detection based on textual indicators or features, such as sudden changes in language style or tone, and inconsistencies in personal information.

2.7 | Shaming and Cancel Culture

Shaming and cancel culture involve public criticism or ostracism of individuals based on their actions or statements, sometimes leading to online harassment. Public Shaming refers to criticizing or condemning individuals or groups in a public forum, often through social media posts, comments, or viral campaigns. This practice aims to invoke feelings of guilt or shame and can escalate quickly, resulting in widespread condemnation from the online community (Basak et al. 2019). Cancel Culture involves withdrawing support from individuals, organizations, or brands after they have made statements or engaged in actions deemed objectionable or offensive. Harmful Implications of Shaming and Cancel Culture:

- **Emotional and psychological distress**: Victims may experience significant emotional and psychological impacts, such as anxiety, depression, and a sense of isolation (Muir et al. 2023).
- **Effect on free expression**: The fear of public shaming or being "canceled" can discourage individuals from expressing their opinions or participating in discussions on important issues (Koivukari and Korpisaari 2021).
- **Normalization of aggressive behavior**: These practices can foster a hostile online environment, where aggressive behavior becomes normalized and further encouraged (Beres et al. 2021).

Although there is significant research addressing issues related to social networks such as Twitter, our survey specifically focuses on recent advancements involving language models or neural models published in academic journals. Consequently, some notable studies, such as those by Basak et al. (2019), Surani and Mangrulkar (2021), and Nalawade and Kulkarni (2021), are not included in our survey.

3 | Methodology

As introduced in the introduction section, given the rapid development of advanced language models and their growing impact on cyber abuse detection, a comprehensive review of the current literature is essential. While previous reviews have focused primarily on well-established forms of cyber abuse, such as hate speech, there is a notable gap in the analysis of newer and underexplored forms, including doxing, trolling, and impersonation. This review aims to fill that gap by providing an in-depth, advanced analysis of how LMs and LLMs are being applied to these emerging issues. Unlike earlier reviews that predominantly concentrate on widespread techniques, our review takes a broader perspective, incorporating a wider range of cyber abuse types and evaluating the influence of cutting-edge AI models on the field. In this paper, we applied the PRISMA methodology (Moher et al. 2010) to conduct a thorough and systematic review aligned with our research objectives. The PRISMA methodology is a standardized framework designed to enhance the transparency and rigor of systematic literature reviews. It provides structured guidelines for identifying, screening, and selecting relevant studies, ensuring that the review process is comprehensive and reproducible. The PRISMA framework comprises four key stages: formulating research questions, developing a search strategy, defining eligibility criteria, and selecting relevant studies. We have dedicated a section to each step of the PRISMA methodology, consolidating the eligibility criteria and relevant studies into a single section. It is important to note that the research questions are presented in the Introduction section. Therefore, this element is not included in the PRISMA methodology section, as we considered it essential to introduce this key information early in the paper.

3.1 | Search Strategy

The vast majority of research in the field of textual cyber abuse detection focuses on hate speech and cyberbullying. To capture the most recent advancements, our search strategy is twofold. For hate speech, trolling, and cyberbullying, we have narrowed the results to papers that address these issues using cutting-edge techniques based on LLMs. In contrast, for other forms of abuse—where research is less prominent—we have broadened our queries to include additional analysis-related terms. To ensure robust and comprehensive results, we focused our analysis on the Web of Science (WOS) database. This database exclusively indexes peer-reviewed papers and encompasses other databases, such as IEEE Xplore and SpringerLink, allowing us to cover multiple indexed sources while benefiting from WOS's advanced search capabilities. The choice of WoS over other databases is supported by previous studies that have also employed the PRISMA methodology, highlighting its reliability and relevance for systematic reviews (Florido-Benítez 2025; Bathaei and Štreimikienė 2023). For each topic, we used the following research queries.

- **Hate Speech**

  TI = ("hate speech" OR "hate content" OR "hate messages") AND AB = ("transformers" OR "GPT" OR "BERT" OR "T5" OR "LLMs" OR MISTRAL or Gemini OR LLAMA) AND TS = ("detection" OR "classification" OR "analysis").

- **Cyberbullying**

  TI = ("cyberbullying" OR "online harassment" OR "cyber harassment") AND AB = ("transformers" OR "GPT" OR "BERT" OR "T5" OR "LLMs" OR MISTRAL or Gemini OR LLAMA) AND TS = ("detection" OR "classification" OR "prediction" OR "analysis").

- **Emotional and Psychological Abuse**

  (TI = ("emotional abuse" OR "psychological abuse" OR "verbal abuse" OR "emotional manipulation" OR "mental abuse" OR "emotional distress" OR "psychological distress" OR "emotional harm" OR "psychological harm" OR "emotional violence" OR "psychological violence" OR "verbal harassment")) AND AB = ("artificial intelligence" OR "natural language processing" OR "deep learning" OR "transformers" OR "GPT" OR "BERT" OR "T5" OR "LLMs" OR MISTRAL or Gemini OR LLAMA)

- **Doxing**

  (TI = (("doxing" OR "doxxing" OR "doxing attack" OR "doxxing attack" OR "personal information exposure" OR

"personal data exposure" OR "data leakage" OR "identity exposure" OR "identity disclosure" OR "public exposure" OR "private information leak" OR "information leak" OR "confidential information disclosure" OR "online doxing" OR "online doxxing" OR "Privacy-Disclosure" OR "PrivacyLeak" OR "private information")) OR AB = (doxing)) AND TS = ("artificial intelligence" OR "machine learning" OR "natural language processing" OR "data analysis" OR "deep learning" OR "transformers" OR "GPT" OR "BERT" OR "T5" OR "LLMs" OR MISTRAL or Gemini OR LLAMA) AND TS=("detection" OR "prevention" OR "analysis").

- **Trolling**

  (TI=(trolling OR "online trolling" OR "internet trolling")) AND TS=("artificial intelligence" OR "machine learning" OR "natural language processing" OR "data analysis" OR "deep learning" OR "transformers" OR "GPT" OR "BERT" OR "T5" OR "LLMs" OR MISTRAL or Gemini OR LLAMA OR "neural networks")

- **Impersonation**

  (TI = ("impersonation" OR "identity theft" OR "fake profiles" OR "catfishing" OR "fake social media accounts ")) AND TS = ("artificial intelligence" OR "machine learning" OR "natural language processing" OR "data analysis" OR "deep learning" OR "transformers" OR "GPT" OR "BERT" OR "T5" OR "LLMs" OR MISTRAL or Gemini OR LLAMA OR "neural networks")

- **Shaming and Cancel Culture**

  (TI = ("shaming" OR "cancel culture" OR "public shaming" OR "online shaming" OR "social shaming" OR "digital shaming" OR "shaming campaigns" OR "social ostracism" OR "public humiliation" OR "online humiliation" OR "reputation damage" OR "character assassination" OR "social backlash" OR "public backlash" OR "mob justice" OR "cyber shaming" OR "call-out culture" OR "name and shame" OR "internet shaming" OR "public reprimand")) AND AB=("artificial intelligence" OR "machine learning" OR "natural language processing" OR "data analysis" OR "deep learning" OR "transformers" OR "GPT" OR "BERT" OR "T5" OR "LLMs" OR MISTRAL or Gemini OR LLAMA OR "neural networks") AND TS = ("artificial intelligence" OR "machine learning" OR "natural language processing" OR "data analysis" OR "deep learning") AND TS = ("detection" OR "prevention" OR "analysis")

3.2 | Eligibility Criteria and Relevant Study Selection

Regarding the eligibility criteria, we have focused on journal papers published in the last 3 years (January 2022–December 2024), as this period aligns with the significant advancements and widespread adoption of LLMs (Schillaci 2024). We selected papers that primarily involve text-based datasets, ensuring relevance to our focus on LLMs. While some papers may also address social network analysis through graph theory or image detection, they were only included if they incorporated a textual analysis component. For instance, a paper primarily centered on graph theory but also analyzing tweet content was considered. Additionally, we excluded survey-based studies focusing on psychological assessments through participant responses, as our interest lies in the direct analysis of textual data rather than self-reported experiences. Table 1 provides a summary of the inclusion criteria and the final outcomes of the selected papers.

Finally, for relevant study selection, it is important to clarify that while some studies from 2019 to 2021 may appear in the survey, they are not part of the papers considered for our review. These references are included to provide background and context for certain theoretical points or foundational methods used in the surveyed period. We found it necessary to include this research in order to provide valuable insights that enhance the understanding of our survey. The entire data flow for our methodology is (Table 1) depicted in Figure 2, which illustrates how we discarded and selected the final 87 papers from the WOS database based on our eligibility and inclusion criteria for relevant studies (Tables 2, 3, 4, 6).

4 | Impact of Language Models on Hate Speech and Cyberbullying Detection

In this section, we conduct an analysis focusing on the adoption of different LMs and LLMs for hate speech and cyberbullying detection. To enhance clarity, the section is divided into two parts: hate speech detection (Section 4.1) and cyberbullying detection (Section 4.2). Acknowledging the rise of generative AI being used for harmful purposes, we have also included a section that explores the generation of abusive content (Section 4.4).

4.1 | Hate Speech Detection

The complexities of hate speech often necessitate the use of a combination of techniques and models to effectively address each specific task. In this context, Khan et al. (2022) proposed a system that utilizes BERT for embeddings, a CNN to capture spatial features of the text, and a BiLSTM to model sequential dependencies. The model also incorporates hierarchical attention, which assigns different weights to features extracted from both the CNN and BiLSTM outputs. This hierarchical attention mechanism is a key contribution, featuring two levels of attention: a high-level focus on the CNN outputs and a low-level focus on the BiLSTM outputs.

Gupta et al. (2021) highlighted that the main contribution of their paper was the application of BERT to the problem of hate speech detection. However, we found the most compelling aspect to be their comprehensive comparative analysis across four different datasets, with particular attention to fine-tuning and parameter adjustments. The authors achieved good results and provided insightful interpretations for each dataset.

Hate speech frequently targets specific groups or individuals based on characteristics such as race, gender, or sexual orientation. To address this issue, some studies have adopted multi-task, multi-aspect, or dual-objective approaches, with ensemble methods often outperforming state-of-the-art models in this task. Mazari et al. (2024) introduced novel architectures that combine the pre-trained BERT model with deep learning techniques, specifically leveraging ensemble learning strategies that incorporate Bi-LSTMs

TABLE 1 | Inclusion criteria and results summary.

Category	Inclusion years	Document type	Raw results	Inclusion criteria	Final results
Hate speech	2022–2024	Article	52	Text and online related, accessible	48
Cyberbullying	2022–2024	Article	22	Text and online related, accessible	21
Shaming and cancel culture	2022–2024	Article	1	Text based and online related, accessible	1
Doxing	2022–2024	Article	12	Text-based and individuals' information related, private information related, accessible	3
Trolling	2022–2024	Article	16	Text based in online environments. Not meme-based, accessible	10
Impersonation	2022–2024	Article	6	Text based and online related, accessible	1
Emotional and psychological abuse	2022–2024	Article	12	Text based with possible applications in cyber abuse detection, accessible	3

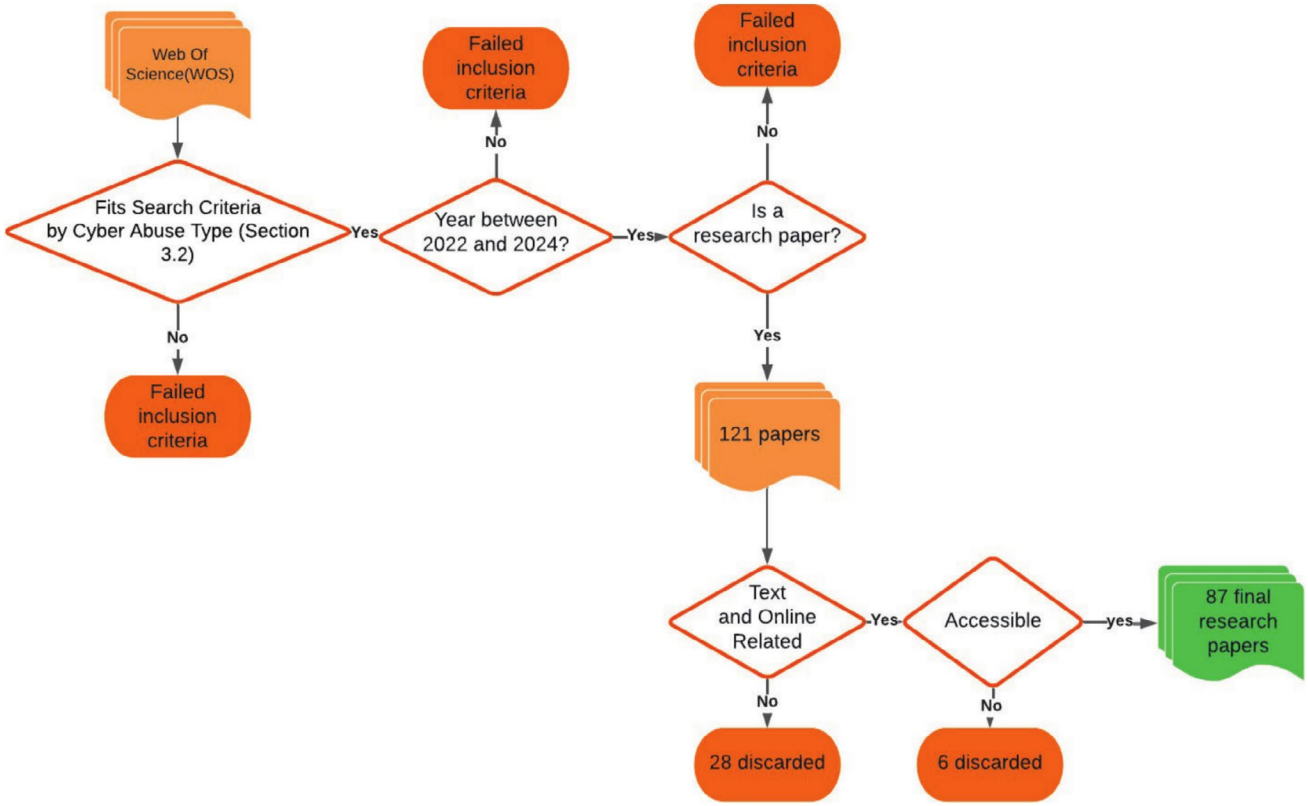

FIGURE 2 | Literature review—Data identification and collection flowchart.

and Bi-GRUs. Their approach effectively tackles multi-aspect hate speech, covering various forms such as identity-based hate and threats. Yun et al. (2023) tackle both hate speech detection and gender bias, which refers to the unfair or prejudiced treatment of individuals based on their gender. The authors proposed a BERT-based ensemble model that integrates various pre-trained models, including Soongsil-BERT, KcELECTRA, BERT, and RoBERTa, to enhance detection accuracy.

The research emphasizes the effectiveness of combining multiple pre-trained models using a weighting scheme that accounts for label distribution, improving performance on imbalanced datasets. To further address class imbalance, they applied data oversampling techniques, allowing the model to learn more effectively from underrepresented classes. Karayi˘git et al. (2022) made a similar contribution by focusing on homophobic and hate speech detection in the Turkish language. Their work stands out as one of the first contributions in this area, utilizing an M-BERT model for this task. Malik et al. (2024), the author also delves into the multifaceted nature of hate speech by not only developing a language model-based classifier for hate speech detection but also conducting community detection (Bedi and Sharma 2016) to gain deeper insights into the communities most affected by hate speech. The study particularly focuses on

TABLE 2 | Summary of hate speech detection studies using LLMs.

CITE (author, year)	LM or LLM used?	Traditional techniques used?	Unbalanced data handling
(Gupta et al. 2021)	BERT	None	Partially Handled (acknowledged issue of imbalance)
(Alkomah et al. 2022)	BERT, bi-LSTMS	Random Forest, Naive Bayes, KNN, XG-Boost, Word2Vec, GLOVE, TF-IDF	Correctly handled with weighting and oversampling
(Mehta and Passi 2022)	BERT + ANN, BERT + MLP, LSTM	Decision Trees, Random Forest, Logistic Regression, Naïve Bayes	Handled using weight optimization
(Arcila-Calderón et al. 2022)	BERT	NB, Logistic regression, SVM, RNN	Handled via instance weighting and macro-F1/AUC
(Ali et al. 2022)	BERT, DistilBERT	FastText + BiGRU	Not handled, evaluated with macro-F1
(Toliyat et al. 2022)	BERT, bi-LSTMs	NB, KKN, XGBoost, Random Forest	Correctly handled and evaluated
(Khan et al. 2022)	BERT, bi-LSTMs	CNN	Correctly evaluated on balanced/unbalanced configurations
(Althobaiti 2022)	BERT	SVM, Logistic Regression	Partially handled with synthetic datasets
(Karayi˘git et al. 2022)	M-BERT	None	Correctly handled using over-sampling/ undersampling
(Wullach et al. 2022)	BiGRU, GPT-2	None	Handled with data augmentation
(Husain and Uzuner 2022)	AraBERT, ArabicBert, LSTM	SVM, Bagging, RF, Logistic Regression, RNN	Correctly evaluated with macro-averaged measures
(Saleh et al. 2023)	BERT, bi-LSTMS	Word2Vec, Glove	Not handled
(Singh et al. 2023)	BERT, BiGRU	None	Not handled
(García-Díaz, Jiménez-Zafra, et al. 2023)	BERT-based: BERTIN, BETO, BNE, M-BERT	None	Discussed but not implemented
(Zhou et al. 2023)	BERT	SVM, CNN-GRU	Not handled
(García-Díaz, Pan, et al. 2023)	FLAN-T5, LLAMA-2, mT0, Flan-Alpaca, ALBETO, BETO, DistilBETO, mBERT, MarIA, mDeBERTa, twHIN	None	Not handled but evaluated from a binary perspective
(Montesinos-Cánovas et al. 2023)	BETO, MarIA, BERTIN, Distil-BETO, AlBETO, Multilingual BERT, mBERT, XLM, Multilingual DeBERTa, mDeBERTa, TwHIN-BERT	None	Partially handled using weighted F1-score
(Su et al. 2023)	RoBERTA	GAN	Not handled
(del Valle-Cano et al. 2023)	BERT, BETO	None	Not handled but well-evaluated

(Continues)

TABLE 2 | (Continued)

CITE (author, year)	LM or LLM used?	Traditional techniques used?	Unbalanced data handling
(Yun et al. 2023)	Soongsil-BERT, Kc-ELECTRA, BERT, RoBERTa	XGBoost	Correctly handled with weighting and oversampling
(Saleh et al. 2023)	BERT, bi-LSTMs	Word2Vec	Well evaluated with weighted metrics
(Ollagnier et al. 2023)	mBERT, mUSE	Multi-view Spectral Clustering	Not handled, dataset adjusted by removing labels
(Arshad et al. 2023)	RoBERTa	FastText	Not handled
(Bilal et al. 2023)	BiLSTM, BERT	Logistic Regression, XGBoost, Random Forest, CNN	Not handled
(Keya et al. 2023)	BERT, GRU	None	Not handled
(Almaliki et al. 2023)	BERT	None	Not handled but evaluated by class F1
(Faria et al. 2024)	BERT, BanglaBERT, mBERT, Gemini, GPT-3.5	None	Handled using weight optimization in the loss function
(Moreno-Sandoval et al. 2024)	BERT, RoBERTa, BETO, RuPERTa, GPT-3.5	None	Not handled
(Ghosh and Senapati 2024)	BERT, ALBERT, RoBERTa, and Distil-BERT, mBERT	None	Well evaluated with weighted metrics
(Siddiqui et al. 2024)	Distil-RoBERTa, mBERT, XLM-RoBERTa	None	Not handled
(Garg et al. 2024)	BERT, bi-LSTMS, Bi-GRU	SVM, LR, ANN, RNN, GloVe	Not handled
(Madhavi et al. 2024)	BERT, BiGRU	CNN	Not handled
(Yadav and Singh 2024)	BERT, DistilBERT	None	Well evaluated with weighted metrics
(Badri et al. 2024)	BERT	FastText	Data augmentation using random word insertion
(Chhikara et al. 2024)	BERT, GPT-2	None	Handled with data augmentation
(Ansari et al. 2024)	BERT, bi-LSTMs	CNN	Addressed with data augmentation techniques
(Sreelakshmi et al. 2024)	BERT, DistilBERT, LaBSE, MuRIL, XLM, IndicBERT, FNET	SVM, Random Forest, Naive Bayes, KNN, Logistic Regression	Handled using cost-sensitive learning
(Pookpanich and Siriborvornratanakul 2024)	BERT, XLM-RoBERTa, Distil-BERT, Wangchan-BERTa, TwHIN-BERT	None	Correctly handled using over-sampling
(Maity et al. 2024)	BERT, bi-LSTMs	FastText	Not handled
(Anjum and Katarya 2024)	BERT	MLP, Logistic Regression, BoW	Not handled
(Nandi et al. 2024)	BERT (mBERT) and Indic BERT	None	Not handled, evaluated with multiple F1 measures
(Putra and Wang 2024)	BERT	CNN	Correctly handled using class importance

(Continues)

TABLE 2 | (Continued)

CITE (author, year)	LM or LLM used?	Traditional techniques used?	Unbalanced data handling
(Ramos et al. 2024)	BERTimbau, BERTimbau-hatebr, mDeBERTa-hatebr, HateBERTimbau, GPT-3.5, Gemini-Pro, Mistral-7B-Instructv0.3, CNN, LSTM	None	Correctly evaluated with emphasis on positive F-score
(Malik et al. 2024)	Urdu-RoBERTa, Urdu-DistilBERT	None	Not handled, only weighted F1 used
(Pan et al. 2024)	Mistral-7B-Instruct, Zephyr-7B, StableBeluga-7B, Tulu-2, LlaMa-2, Mistral-7B-v0.1, BERT, RoBERTa, DeBERTa, mBERT, AL-BERT, DistilBERT, XLM-RoBERTa	None	Evaluated with macro and weighted F1, not handled
(Firmino et al. 2024)	BERTimbau, Italian BERT, BERT, XLM-R	None	Correctly handled with extensive evaluation
(Altinel et al. 2024)	BERT	TextGCN, Kmeans, Word2Vec, Doc2Vec, Glove, TF	Partially Handled (acknowledged issue of imbalance)
(Mazari et al. 2024)	BERT, bi-LSTMS, Bi-GRU	MLP	Correctly evaluated with macro-averaged measures

religious, political, and ethnic communities, offering a comprehensive approach to understanding both the prevalence of hate speech and the groups that are disproportionately targeted. An important contribution is made by Alkomah et al. (2022), where the authors propose a combination of text features and psychological features for hate speech detection in a multiclass classification setting. Their study demonstrates that integrating these features enhances the performance of both BERT and Random Forest, with the latter proving to be a highly accurate model. Additionally, they propose the use of SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al. 2002; Suhas Bharadwaj et al. 2022) address the issue of imbalanced datasets, improving the robustness of their approach.

Su et al. (2023) proposed a semi-supervised system for hate speech detection that learns from a limited set of annotated examples while leveraging a larger pool of unannotated data. This approach is particularly beneficial in scenarios where labeled data are scarce. The model is built using a Generative Adversarial Network (GAN) in conjunction with RoBERTa as the encoder. The GAN consists of two primary components: a generator and a discriminator. The generator creates synthetic data that closely mimics the real data, while the discriminator evaluates the authenticity of the generated data in comparison to real data. This architecture enables the model to improve its ability to detect hate speech. The results show that the proposed model achieves significant performance gains over the baseline RoBERTa model, particularly in terms of accuracy and Macro-F1, highlighting its effectiveness in semi-supervised learning scenarios. In a similar study, (Putra and Wang 2024) tackled hate speech detection using a semi-supervised approach by employing BERT as an encoder alongside a CNN architecture composed of three parallel branches to process the input embeddings. The authors introduced a method of generating pseudo-labels for unlabeled data, where a model trained on the labeled data predicts labels for unlabeled instances, thereby expanding the training set. A notable enhancement in their model is the inclusion of a shared BERT layer that enables the model to learn from multiple datasets concurrently. This shared layer works alongside private layers dedicated to individual tasks; the shared layer captures generalized features relevant across tasks, while the private layers focus on task-specific nuances. The datasets used for training included Hatebase, Supremacist, Cybertroll, TRAC, and TRAC 2020. The BERT-3CNN model demonstrated substantial performance improvements, achieving an F1 score enhancement of 18% over state-of-the-art models.

One significant challenge in hate speech and other forms of cyber abuse detection is the lack of sufficient labeled examples to train robust models. Addressing this issue, while also aiming to improve the explainability of typically opaque models, Ansari et al. (2024) investigate the impact of three distinct data augmentation techniques on two publicly available datasets: one containing comments from Facebook and YouTube, and another composed of tweets from Twitter, all labeled for hatefulness or neutrality. The augmentation techniques applied include Easy Data Augmentation, BERT-based paraphrasing, and Back Translation. EDA involves simple text manipulation methods to create additional training samples. BERT-based augmentation leverages the model to generate paraphrases by predicting masked words or producing semantically equivalent sentences. Back Translation translates text into another language and back into the original to produce syntactically varied but semantically similar sentences. For classification, the authors utilized

TABLE 3 | Overview of cyberbullying detection research (sorted by year).

CITE (author, year)	LM or LLM used?	Traditional techniques used?	Unbalanced data correctly processed?
(Ahmed et al. 2022)	BERT, GPT-2, XLNet, RoBERTa, DistilBERT	SVM, XGBoost	Handled effectively (dynamic query expansion)
(Maity et al. 2022)	BERT, bi-GRU	VecMap	Not handled
(Abarna et al. 2023)	BERT	ELMO, XGBoost, FP-Growth, LSA, K-Means	Partially handled (evaluated with weighted metrics)
(Muneer et al. 2023)	BERT, bi-LSTMs	CBOW, Word2Vec	Partially handled (acknowledged issue of imbalance)
(Paul et al. 2023)	BERT, bi-LSTMs	SVM, logistic regression, CBOW, FastText, TF-IDF	Not handled
(Samee et al. 2023)	BERT	None	Handled effectively (under-sampling applied)
(Saini et al. 2023)	BERT	SVM + CNN Ensemble	Partially handled (addressed via ROC curves)
(Razi and Ejaz 2024)	m-BERT, muRIL	None	Not handled
(Kumar et al. 2024)	DeBERTa, Longformer, BigBird, HateBERT, MobileBERT, DistilBERT, BERT, RoBERTa, ELECTRA, XLNet, GPT-4, Gemini	None	Handled with data augmentation techniques
(Khafajeh 2024)	BERT, bi-LSTMs	CNN, TF-IDF	Handled effectively (under-sampling applied)
(Tabassum and Nunavath 2024)	BERT, DistilBERT, RoBERTa, GRU, LSTMs	CNNs	Not handled
(Dobrojevic et al. 2024)	BERT	TF-IDF, XGBoost	Partially handled (evaluated with weighted metrics)
(Umer et al. 2024)	RoBERTa, BERT, bi-LSTMs	GLOVE, Word2Vec, RF, SVM, KNN, NB, CNN	Not handled
(Ferrer et al. 2024)	T5-small, MobileBERT	None	Handled with data augmentation techniques
(Chen et al. 2024)	XLNet, bi-LSTM	None	Handled effectively (balanced dataset via over/under-sampling)
(Mahmud et al. 2024)	BERT, XLM-RoBERTa, BiLSTM, CNN, GRU	SVM, logistic regression, RF, KNN, NB	Not handled
(Li, Zeng, et al. 2024)	BERT, Graph Isomorphism Networks	XGB, RF, LR, KNN, ML, LGBM	Handled (SMOTE applied)
(Alfurayj et al. 2024)	BERT, bi-LSTMs	None	Partially handled (evaluated with weighted metrics)
(Al-Harigy et al. 2024)	BERT, bi-LSTMs, Sim-CSE	None	Handled with data augmentation techniques
(Mahdi et al. 2024)	BERT	None	Not handled
(Alrowais et al. 2024)	RoBERTa	GloVe	Not handled

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. This approach not only addresses the data scarcity issue but also reduces the risk of overfitting, enhancing the models' generalization capabilities and improving their performance in real-world hate speech detection scenarios. Wullach et al. (2022) proposed an intriguing approach that leverages HyperNetworks, a type of neural network architecture designed to generate weights for another neural network,

TABLE 4 | Overview of other ways of textual cyber abuse detection.

CITE (author, year)	Cyber abuse type	Advanced NN model/LLM used?	Traditional techniques used?	Unbalanced data processed?
(Aarthi and Chelliah 2022)	Shaming	DRNN	SVM	Not handled
(Li et al. 2022)	Doxing	DNN	Regularization	Handled effectively including regularization terms to the training loss
(Aarthi and Chelliah 2023)	Doxing	biLSTM, biGRU	SVM	Partially handled
(Miao et al. 2024)	Doxing	deBERTa, BiLSTM	None	Not handled
(Jang et al. 2022)	Emotional abuse	BERT	LDA	Not handled
(Chen et al. 2022)	Emotional abuse	None	Logistic regression, RF, ANN	Not handled
(Alemu et al. 2023)	Emotional abuse	LSTM, GMM	None	Partially handled, taken into account in the experimentation design
(Uyheng et al. 2022)	Trolling	BERT, LSTM	Logistic Regression, SVM, RF	Not handled
(MacDermott et al. 2022)	Trolling	biLSTM, biGRU	None	Handled effectively by means of under-sampling
(Bakar et al. 2023)	Trolling	None	SVM	Handled effectively with class weighting and RBF kernel in SVM
(Nepomuceno et al. 2023)	Trolling	None	LDA, MLP	Not handled
(Ezzeddine et al. 2023)	Trolling	LSTM	None	Handled effectively by means of under-sampling
(Cranmer et al. 2024)	Trolling	BERT	None	Not handled
(Asif et al. 2024)	Trolling	GCN, GloVe	None	Not handled
(Sultan et al. 2024)	Trolling	DeBERTa, GPT 2, GPT 4, BERT, XLnet	None	Handled effectively by means of over-sampling
(Been Choi et al. 2024)	Trolling	BiLSTM	CNN, XGboost, RF	Handled effectively by means of over-sampling
(Swed et al. 2024)	Trolling	None	Random forest	Handled effectively compensating in the algorithm the unequal distribution
(Huang et al. 2021)	Impersonation	None	Gaussian Dist., Euclidean Dist.	Not handled
(Mewada and Dewang 2024)	Impersonation	CNN	None	Not handled

enabling efficient weight sharing across layers and operating at the character level. This character-level focus allows the model to effectively manage morphological variations and the noisy text commonly found in social media. Additionally, using GPT-2, the authors generated a large-scale corpus of synthetic hate and non-hate sequences to augment existing datasets, effectively addressing challenges related to limited labeled data and class imbalance. Badri et al. (2024) demonstrate how data augmentation enhances the accuracy of a BERT-based model for Arabic text. The authors propose using random word insertion as an augmentation technique to generate new examples, thereby improving model performance and overall data quality.

One of the most comprehensive datasets for hate speech detection is presented by Toliyat et al. (2022), comprising over 10 million tweets related to COVID-19 and hate speech targeting

Asian communities. The authors collected this dataset by crawling Twitter using specific keywords. To establish a reliable ground truth, they manually annotated 3000 hate speech examples from the dataset proposed by He et al. (2021). This manually labeled subset was then leveraged to label the entire 10 million tweets. The study also conducts an extensive evaluation of various machine learning approaches, including deep learning, transformer models, and traditional machine learning techniques, across different configurations of balanced and imbalanced datasets.

To address the scarcity of hate speech research in languages other than English, which has over 60% of efforts in this field (Pikuliak et al. 2021), Firmino et al. (2024) proposed the use of cross-lingual methodologies. The authors first gathered data from English and Italian, leveraging these larger datasets to train language models such as BERT. In a subsequent step, they fine-tuned the model on the target language, Portuguese, using smaller datasets. The results demonstrated the effectiveness of this cross-lingual approach, achieving an F1-measure of 92% on the OffComBr-2 corpus. The study's findings highlight that incorporating Latin-based languages, like Italian, significantly improves hate speech detection in Portuguese, showcasing the potential of transfer learning in enhancing performance for low-resource languages.

Zhou et al. (2023) introduced a manually annotated dataset comprising posts from HackForums, Stormfront, and Incels.co, focusing on the prevalence and nature of hate speech across these platforms. The study revealed that while underground forums like HackForums exhibited lower levels of hate speech, extremist platforms had significantly higher volumes, driven by specific ideologies. To address the challenge of hate speech detection, the authors employed BERT and explored two methods: span prediction and sequence labeling. Additionally, the paper evaluated several models beyond BERT, including a Convolutional Neural Network with Gated Recurrent Units (CNN-GRU) and a SVM. The SVM, trained with multi-platform data, achieved the highest F1-score on the HatEval test set, indicating its strength in cross-platform detection. However, BERT outperformed the other models in detecting hate speech on the Extremist forums and HackForums test sets, demonstrating its superior ability to capture the linguistic nuances of hate speech within these contexts. A hybrid model combining CNN and GRU with BERT-based encoding is proposed by Madhavi et al. (2024), reinforcing ideas from prior research and demonstrating the necessity of specialized models to capture the nuances of hate speech.

Hate speech is a global issue that remains unexplored in many languages. del Valle-Cano et al. (2023) introduce SocialHaterBERT, a model designed for hate speech detection that is trained in both English and Spanish. This research builds upon the HaterNet algorithm by leveraging the capabilities of BERT. One of the most significant contributions of this study to the state of the art is the incorporation of social network features into the classification model. By analyzing statistics and characteristics derived from the interaction graph of Twitter users, the model enhances its ability to detect hate speech by considering not only the content of the messages but also the social context in which they are shared. The results demonstrate that SocialHaterBERT outperforms HaterBERT by 4% and shows a 19% improvement over the original HaterNet algorithm. These findings underscore the importance of integrating user characteristics into hate speech detection systems, as this approach allows for a more nuanced and contextual understanding of the messages being analyzed. The study concludes that the fusion of user profile data with textual analysis not only enhances the detection capabilities of hate speech models but also paves the way for new research opportunities in the field of cyber abuse.

Ramos et al. (2024) conducted an exhaustive review that systematically evaluates various transfer learning models for hate speech detection in European Portuguese, focusing particularly on BERT-based models and generative models. The study utilized two distinct annotated corpora consisting of YouTube comments and Twitter tweets to assess model performance. The findings reveal that the HateBERTimbau model achieved the highest F-score of 87.1% on the YouTube corpus, while GPT-3.5 performed best on the Twitter corpus with an F-score of 50.2%. The evaluation was centered on the positive classes, ensuring robustness in the results despite the absence of oversampling or undersampling techniques, demonstrating the models' efficacy in handling class imbalances within the datasets.

García-Díaz, Jiménez-Zafra, et al. (2023), compiled existing Spanish-language datasets, such as the MisoCorpus 2020, AMI 2018, HaterNet, and HatEval 2019, creating a valuable resource for analyzing hate speech across different contexts. The study focuses on knowledge integration strategies, which combine diverse feature sets to build more robust detection systems. Specifically, the authors proposed a method that fuses linguistic features with transformer-based models. These linguistic features include morphosyntactic, semantic, stylometric, and negation-related elements. For the detection process, they utilized Spanish BERT variants like BETO and BERTIN to improve accuracy. Additionally, the authors introduced a model resolver that integrates various feature sets, employing ensemble learning techniques to further enhance the system's ability to detect hate speech in Spanish-language texts. This approach demonstrates the value of combining linguistic insights with advanced machine learning techniques for more nuanced hate speech detection. García-Díaz, Pan, et al. (2023), investigate the effectiveness of Zero-Shot Learning (ZSL) and Few-Shot Learning (FSL) strategies for hate speech detection in low-resource languages, specifically focusing on Spanish applications. The study evaluates the performance of various generative models, including T5, BLOOMZ, and Llama-2, in detecting hate speech. The research explores the cross-lingual capabilities of these models, assessing their effectiveness in both English and Spanish contexts. Under ZSL conditions, the models classify hate speech without prior examples, relying on descriptive prompts of hate speech categories to infer classification based on their contextual understanding. In FSL, the models are provided with a small number of labeled examples (typically five) per hate speech category, which are included in the prompts to help the models grasp the nuances of the task. Llama-2 consistently outperformed other models, especially in English datasets, demonstrating that generative models can effectively bridge the gap between data scarcity and accurate hate speech detection. Similarly, Pan et al. (2024) examine the effectiveness of various models in detecting hate speech and cyber abuse, with a particular emphasis on sexism. The study evaluates several models, including BERT, RoBERTa, DeBERTa,

mBERT, ALBERT, DistilBERT, and XLM-RoBERTa, using both fine-tuning and contextual learning approaches, such as zero-shot and few-shot learning. An in-depth error analysis is also conducted, revealing challenges like difficulties in differentiating between hate speech categories and handling figurative language. Siddiqui et al. (2024) explore various types of hate speech, including discrimination based on disability, gender, nationality, race, and religion, within the domain of multilingual text, particularly in English, Urdu, and Sindhi. The study highlights the effectiveness of XLM-RoBERTa in multilingual contexts while also emphasizing the need for improved data quality and language-specific models to further enhance hate speech detection in low-resource languages. Moreno-Sandoval et al. (2024) also explore other forms of cyber abuse related to hate, particularly in the context of sexuality and misogyny. The authors propose a pipeline that leverages various text representations using different BERT-based techniques, some of which are trained on Spanish content. In the final classification layer, they also incorporate GPT. Their results indicate that the BERT model achieves an accuracy of 83.6% for hate speech detection, while the GPT model outperforms it, reaching 90.8% accuracy for detecting racist speech and 90.4% for misogyny detection. These findings highlight the promising capabilities of LLMs in capturing specific nuances of cyber abuse. Additionally, Faria et al. (2024) demonstrates that GPT excels in a few-shot configuration for Bengali text. The authors employ novel prompting strategies using Gemini and GPT, along with class-balanced weighting, to enhance model training and improve performance across various hate speech categories. Their results reveal that LLMs not only achieve higher detection accuracy but also exhibit greater adaptability in real-time applications. This suggests that LLMs could revolutionize the fight against online hate speech in Bengali-speaking communities, addressing a critical gap in existing methodologies.

Another significant challenge in low-resource languages is the lack of available datasets, obstructing the development of effective models for tasks like hate speech and cyberbullying detection. Recently, there has been a growing trend of research papers proposing their own datasets for these underrepresented languages, which is highly valuable for advancing research beyond English. Expanding the review of the influence of transformer-based models in low-resource languages and the multilingual domain, (Ghosh and Senapati 2024) introduces two new hate speech datasets for binary classification tasks across Hindi, Marathi, Bangla, Assamese, and Bodo. The study evaluates various transformer-based models, comparing the performance of both monolingual and multilingual approaches. Montesinos-Cánovas et al. (2023), the authors introduced a novel multi-label corpus of 7483 manually annotated tweets, categorized into four classes: safe, aggressive, misogynistic, and racist, specifically within the context of Spanish football matches. They employed state-of-the-art pre-trained LLMs like BETO, MarIA, and multilingual BERT to develop an ensemble classifier. The model achieved an impressive macro-weighted F1-score of 88.71%, demonstrating its effectiveness in identifying and classifying various forms of abusive content, including aggression, misogyny, and racism. The study highlighted that while individual LLMs performed well, their combination with linguistic features—such as proper nouns, contextual keywords, and psycho-linguistic elements—led to a significant improvement in classification accuracy. The authors noted that the models showed strong results across different hate speech categories, though they also identified opportunities for further enhancement by expanding the dataset and refining the classifiers to better capture a broader spectrum of cyber abuse expressions in sports-related social media content.

Sreelakshmi et al. (2024) developed an annotated CodeMix Malayalam-English hate speech corpus and conducted an in-depth study on detecting hate speech and offensive language in CodeMix texts from Dravidian languages, specifically Kannada, Malayalam, and Tamil. Their key contributions include the introduction of a cost-sensitive learning approach to address class imbalance in the datasets and a thorough evaluation of various multilingual transformer models such as BERT, DistilBERT, LaBSE, and MuRIL, alongside traditional machine learning techniques like SVM. The results demonstrated that models such as MuRIL and DistilBERT outperformed others, achieving notable accuracy and F1-scores, highlighting their effectiveness in hate speech detection in CodeMix text. Similarly to previously reviewed works, Almaliki et al. (2023) introduced a new dataset comprising ~10,000 annotated tweets specifically for hate speech detection in the Arabic language. To evaluate the dataset's effectiveness, they conducted experiments using BERT models. Urdu has garnered significant attention in recent years, with researchers increasingly focusing on the language's unique nuances. Ali et al. (2022) contributed to this growing body of work by compiling a dataset of around 20,000 tweets using the open-source tool Twint, specifically targeting hate speech-related keywords. The study explores the effectiveness of transfer learning techniques, utilizing pre-trained transformer models like BERT to tackle the challenges of hate speech detection in Urdu. In a similar approach, Arshad et al. (2023) present their work on hate speech detection. This study leverages transfer learning techniques with RoBERTa, specifically tailored for Urdu, to address the challenges associated with detecting hate speech in this underrepresented language. Also in Urdu, Bilal et al. (2023) introduced a new dataset comprising 173,714 Roman Urdu hate speech examples and conducted a comparative analysis between deep learning and transformer models on this dataset.

Nandi et al. also highlight the potential of using pre-trained language models to construct an ensemble method for hate speech classification, particularly focusing on low-resource Indian languages (Nandi et al. 2024). The study introduces a feature fusion-based ensemble model designed to detect hate speech in Bengali, Marathi, and Hindi. The approach leverages two pre-trained language models: multilingual BERT (mBERT) and IndicBERT. The features extracted from both models are concatenated, and a traditional softmax layer is employed to perform the final classification. This fusion of feature representation proved effective, with the model achieving weighted F1 scores of 0.923 for Bengali, 0.815 for Hindi, and 0.924 for Marathi, significantly outperforming existing benchmarks for the Bengali and Marathi datasets. In the multimodal domain of English, Hindi, and Hindi-English, (Singh et al. 2023) proposed an ensemble model combining mBERT and GRU, a type of recurrent neural network that is particularly well-suited for handling sequential data such as text, further advancing performance in multilingual hate speech detection. For hate speech detection in Bengali,

Keya et al. (2023) proposed the same combination of BERT architecture combined with GRU.

Arcila-Calderón et al. (2022), introduced an ad hoc dataset manually annotated for the detection of hate speech and racist messages, specifically focusing on topics related to migration and refugees. The dataset primarily comprises Spanish-language messages but has been extended to include messages in Greek and Italian as well, making it more comprehensive in its coverage of multilingual hate speech. The authors conducted a comparative analysis of traditional data mining techniques, deep learning architectures, and transformer-based models for binary classification of hate speech. The results indicated that deep learning and transformer models significantly outperformed traditional algorithms, demonstrating superior performance metrics in detecting hate speech across different languages. The study emphasizes the importance of developing high-quality, validated datasets as a fundamental step for training effective and reliable predictive models in the domain of hate speech detection. This work underscores the critical role that diverse, annotated datasets play in enhancing the accuracy and generalizability of machine learning models across multilingual contexts.

Maity et al. (2024) proposed a novel tagged dataset for hate speech detection in the Thai language. A key contribution of this work is the incorporation of sentiment analysis through a systematic annotation process, where each social media post is labeled with both hate speech and sentiment information. To classify hate speech, the study employs a dual-channel multitask model called ThaiCaps, which integrates both BERT and FastText embeddings. BERT effectively captures contextual information and semantic nuances, while FastText offers character-level features, enhancing the model's robustness against the noisy nature of social media data. This multitask approach enables the model to simultaneously tackle hate speech detection and sentiment analysis, leveraging the interrelated nature of these tasks to improve overall performance. The results from the ThaiCaps model are impressive, achieving an overall accuracy of 89.67% and a macro F1 score of 89.79% for hate speech detection, alongside accuracy and F1 scores of 80.92% and 80.97% for sentiment detection, respectively. This highlights the model's effectiveness in addressing both hate speech and sentiment analysis within the Thai language context.

In the context of Arabic language hate speech detection, Althobaiti (2022) and Husain and Uzuner (2022) explored the use of BERT architectures to enhance performance in this area. Husain and Uzuner (2022) conducted a comparative analysis between traditional models and BERT, emphasizing the importance and different approaches to preprocessing for improving hate speech detection. Their study revealed significant variations in classifier performance based on the preprocessing techniques employed. While traditional classifiers like SVM demonstrated better results with fully processed text, BERT-based models excelled with targeted preprocessing methods, such as emoji conversion and dialect normalization. Althobaiti (2022) proposed a BERT-based classification system that integrates sentiment analysis features and emoji transcriptions into the model's feature set. The incorporation of sentiment features led to a modest improvement in the model's overall performance. Additionally, the impact of emoji transcriptions on the model's effectiveness varied depending on the size of the annotated dataset and the uniqueness of the emojis as distinguishing features. The potential of incorporating emotional features and tone analysis is further explored in (Anjum and Katarya 2024), where the authors propose a multilingual hate speech detection model that combines BERT with a Multilayer Perceptron (MLP) and a Logistic Regression classifier. This approach aims to effectively analyze hate speech across multiple languages. A key feature of their system is the incorporation of a profanity detection technique, which leverages a curated list of profane words to assess the emotional tone of tweets, thereby enhancing the model's ability to accurately identify hate speech.

Pookpanich and Siriborvornratanakul (2024), the authors investigated the impact of hate speech in live-streaming messages during football matches, focusing specifically on detecting offensive language in Thai YouTube live chat. The study utilized five transformer-based language models: BERT, XLM-RoBERTa, DistilBERT, WangchanBERTa, and TwHIN-BERT. They carried out an implementation of a two-stage labeling process, combining automated and manual labeling to enhance the quality of the dataset, which consists of over 2 million messages categorized as neutral, positive, negative, or hate speech. The authors placed special emphasis on identifying negative messages with offensive content, paying attention to the intricacies of Thai slang and colloquial language, which are often indicators of hate speech. To address the challenge of class imbalance, they introduced "fake hate speech" messages—containing hate speech keywords without necessarily being abusive—helping to balance the dataset and improve model training. Results showed that WangchanBERTa, in particular, achieved the best performance, with high accuracy, precision, recall, and F1-scores, highlighting its effectiveness in detecting nuanced hate speech in the Thai language.

Most studies on hate speech detection rely on deep learning-based classification models, which often function as "black boxes" with limited interpretability. Mehta and Passi (2022), the authors take a different approach by focusing on model explainability. They apply interpretability techniques such as SHAP (Shapley Additive Explanations) (Lundberg and Lee 2017) values and LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al. 2016) to shed light on the decision-making processes of various models. LIME, in particular, is applied to predictions made by both traditional classifiers and deep learning models like BERT variants. It generates local explanations by assigning weights to individual words, revealing how much each word contributes to the model's decision on whether a comment contains hate speech. The study evaluates a combination of traditional machine learning techniques alongside more advanced deep learning architectures, such as BERT integrated with different kinds of traditional neural network architectures. The results show that BERT-based models significantly outperform traditional classifiers, achieving accuracies of up to 93.67%. By employing explainability techniques like LIME, the authors not only enhance the performance of hate speech detection models but also make them more transparent, offering valuable insights into how specific words or phrases influence the predictions. Similarly, (Garg et al. 2024) utilized SHAP values and

LIME to gain interpretability insights into hate speech detection, conducting a robust study on which words have the most influence in toxic messages. The integration of XAI techniques provided transparent insights into model predictions, with a BERT and Bi-GRU fusion under an ensemble model achieving the best results. Their findings demonstrate that ensemble methods outperform single-model systems by 2%, highlighting their effectiveness. LIME interpretations are also used by Saleh et al. (2023), where BERT outputs are analyzed through this technique. In this study, the authors propose the use of transfer learning with BERT and a Bi-LSTM classifier, leveraging domain-specific word embeddings alongside Word2Vec embeddings. Their results demonstrate that BERT achieves superior performance compared to other approaches. Building on this, (Yadav and Singh 2024) take a step further by introducing HateFusion, an interactive tool designed for both explicit and implicit hate speech detection using XAI techniques. The study employs DistilBERT and BERT as language models, with BERT proving to be the most accurate. Additionally, SHAP, and LIME are utilized for explainability, and BertViz is incorporated to enhance model interpretability. These approaches bridge the gap between model accuracy and interpretability, making the models more trustworthy and actionable.

Saleh et al. (2023) approached hate speech detection from two distinct approaches: one utilizing BERT and the other focusing on constructing an extensive hate speech-specific word embedding to serve as features for a biLSTM classifier. The primary contribution of the paper is the creation of this comprehensive word embedding, which was trained on a curated dataset comprising ~1 million sentences sourced from hate speech domains on Twitter. By building the embedding in an unsupervised manner, the authors provide a valuable resource that captures the nuances and context of hate speech language. The results demonstrated that this method, which combines unsupervised word embedding with a biLSTM classifier, is highly effective, achieving state-of-the-art performance. The study underscores the potential of leveraging unsupervised techniques in the hate speech detection domain, suggesting that they can be competitive.

The effectiveness and necessity of unsupervised techniques in understanding hate speech are explored in Ollagnier et al. (2023), where the authors conduct a fine-grained analysis of hate speech through multiple perspectives, including the type of hate, its directness, and target. They undertake community detection and clustering experiments using a novel approach that models cyberbullying as a graph structure. The process begins with generating embeddings using mBER and multilingual Universal Sentence Encoder (mUSE). They then compute similarities between the embeddings to construct a similarity matrix, from which a graph is built with tweets as nodes and edges representing the similarity scores. This graph serves as the basis for community analysis and clustering, employing Multi-view Spectral Clustering. The resulting clusters provide valuable insights into the different forms of hate speech, their interactions, and how they manifest across various contexts. The importance and effectiveness of hybrid systems that combine unsupervised and supervised learning methods are also emphasized by Altinel et al. (2024). In their work, the authors developed a system that integrates K-means clustering, BERT, and a textGCN classification layer for hate speech detection. In this hybrid approach, the test set is first partitioned into k subclusters using the K-means algorithm. Subsequently, for each distinct subset identified by the features resulting from the K-means division, the textGCN technique is applied. Additionally, the authors introduced a manually labeled dataset for hate speech in the Turkish language, further contributing to research in this domain.

Finally, Chhikara et al. (2024), delve into the issue of automated accounts spreading hate speech, focusing on identifying and differentiating AI-generated hate speech from genuine human-generated content. They introduce a Giant Language Model Test Room (GLTR), a tool designed to analyze and detect the authenticity of text produced by large language models. The tool combines BERT and GPT-2 to effectively tackle this challenge. BERT is fine-tuned on labeled datasets to classify individual words and phrases, enabling it to detect instances of hate speech based on contextual meanings. Meanwhile, GPT-2 is employed to generate synthetic examples of hate speech, which are then used to augment the training data, enhancing the model's ability to recognize various patterns. The outputs of BERT and GPT-2 are combined to evaluate how effectively generative hate speech examples can be detected. This innovative approach provides valuable insights into the challenge of identifying AI-generated hate speech, a topic that will be further explored in Section 4.4.

4.1.1 | Quantitative Analysis of Hate Speech Detection Leveraging LMs and LLMs

In this quantitative analysis (Table 2), we examine the prevalence of various language models used in the reviewed papers, as well as their approaches to handling unbalanced data.

BERT and its variants—such as BETO, DistilBERT, mBERT, and BERTIN—are the most frequently employed models, appearing in nearly 95% of the studies. This underscores BERT's dominance in natural language processing tasks, particularly in hate speech detection. Specifically, the base BERT model alone is present in 75% of the reviewed papers. A notable trend is the increasing use of multilingual and language-specific BERT variants, reflecting a growing focus on hate speech detection across different languages.

Beyond BERT, RoBERTa and its derivatives appear in 20% of the studies. Large language models (LLMs) such as Mistral, T5, LLaMA, and GPT have a more limited presence, collectively appearing in 6.25% and 10.42% of the studies, respectively. Additionally, the emerging Gemini model is found in 4.17% of the papers. These findings suggest that while newer LLMs are gaining traction, BERT remains the dominant choice, particularly excelling in hate speech detection. Notably, BERT is frequently combined with LSTM models for classification tasks. Its widespread adoption is further driven by the nature of recent publications, which primarily focus on fine-tuning BERT for various languages, making it highly suitable for transfer learning applications.

In addition to transformer-based architectures, several classical and deep learning models are still in use. GRUs (Gated Recurrent Units) are employed in 16.67% of the studies, while

CNNs appear in 14.58%. Traditional machine learning models, such as XGBoost (8.33%), Random Forest (12.5%), and SVMs (12.5%), continue to play a role, particularly in feature-based text classification. Vector-based embedding techniques, including Word2Vec, GloVe, and FastText, are used in 14.58% of the studies.

Regarding the handling of unbalanced data, approaches vary significantly. A total of 27.08% of studies employ techniques such as data augmentation, oversampling, or undersampling to address class imbalance. However, 41.67% of papers do not adequately tackle this issue, potentially leading to biased model performance. Meanwhile, 31.25% of studies implement partial mitigation strategies, often relying on robust evaluation metrics such as weighted F1 scores to assess model effectiveness. Weighted metrics emerge as the most frequently used approach for handling unbalanced datasets.

Overall, these findings highlight the continued preference for BERT-based models in hate speech detection, emphasizing their adaptability across languages and fine-tuning capabilities. While newer models like Mistral and T5 are being introduced, their adoption remains relatively low compared to BERT and RoBERTa. Additionally, the inconsistent use of data imbalance mitigation techniques suggests that more comprehensive strategies are needed to ensure fair and unbiased model performance.

4.2 | Cyberbullying Detection

With the exponential growth of internet users and the increasing prevalence of social media, harmful behaviors like cyberbullying have become widespread. Detecting cyberbullying, however, remains a challenging task due to the linguistic subtleties and contextual nuances often embedded in offensive language. Traditional approaches rely heavily on extensive human-annotated datasets, which are both time-consuming and costly to produce. In an effort to overcome these limitations, Al-Harigy et al. (2024) propose a novel approach using Deep Contrastive Self-Supervised Learning (DCSSL). This method automates data augmentation and labeling, enabling the training of models on larger datasets without the need for extensive manual intervention. By starting with a small, labeled dataset, the model identifies similar examples in unlabeled data, effectively generating new instances that capture the complex nature of cyberbullying language. To automate the labeling process, a fine-tuned sentence encoder is used to compute sentence similarities, assigning pseudo-labels to unlabeled data based on their resemblance to labeled examples. The model combines BERT with a Bi-LSTM network to leverage the strengths of both architectures, resulting in richer sentence representations. The DCSSL model achieves a macro-averaged F1 score of 0.9231, significantly outperforming baseline models in detecting cyberbullying. In a similar way, Khafajeh (2024) proposed a hierarchical architecture by comparing CNN and LSTM models for cyberbullying detection. They also utilize BERT for a traditional cyberbullying classification task, achieving the best results with this approach.

Has we shown in Section 2 cyberbullying is closely linked to significant psychological and emotional distress. In this context, Maity et al. (2022) approach the issue of cyberbullying from the perspective of sentiment and emotional analysis. They point out a notable gap in existing research, which often overlooks the integration of sentiment and emotion data for detecting cyberbullying—especially in code-mixed languages like Hindi and English. To address this, the authors developed a multitask, multimodal framework that incorporates both sentiment and emotion analysis to enhance the detection of cyberbullying. Their contribution includes the creation of a novel dataset, BullySentEmo, annotated with labels for cyberbullying, sentiment, and emotion. The proposed framework utilizes BERT and VecMap embeddings along with a bi-GRU model to analyze both text and emoji modalities. The results are impressive, with accuracy rates of 82.05% for cyberbullying detection, 77.87% for sentiment analysis, and 58.05% for emotion recognition. The study underscores the crucial role of sentiment and emotional cues in cyberbullying detection, revealing that posts labeled as bullying tend to exhibit negative sentiments and emotions such as anger and disgust. Also focusing on emotional features within a privacy-preserving framework, Samee et al. (2023) introduce an innovative federated architecture aimed at enhancing both the security and accuracy of cyberbullying detection models. A key contribution of the study is the introduction of eight novel emotional features extracted from textual tweets, which provide deeper insights into the emotional undertones of messages, thereby improving the detection of cyberbullying instances. The authors address the challenge of imbalanced data by employing data undersampling techniques before distributing the data to each federated node for analysis. The results demonstrate that the proposed framework, particularly through the integration of the BERT model, not only improves detection accuracy but also ensures user privacy via federated learning, outperforming traditional centralized methods in cyberbullying identification.

Cyberbullying can manifest in various forms, particularly on online platforms like Twitter, where the role of bystanders can significantly influence the nature of the bullying. Alfurayj et al. (2024), this complexity is explored by introducing a dataset that incorporates bystander roles as a feature in cyberbullying detection, enabling a more nuanced analysis of interactions. Bystanders are categorized into defenders, instigators, neutrals, and others. The proposed model utilizes a novel chained architecture consisting of two sequential classification layers. The first layer classifies bystander roles, while the second layer detects cyberbullying, allowing the model to capture the interdependencies between bystander behavior and bullying. The authors fine-tune a pre-trained BERT model, enhanced with LSTM networks, to extract textual features. The model achieves a significant F1 score of 89% in classifying bystander roles, marking a 25.35% performance improvement over traditional methods. While the paper acknowledges the challenge of unbalanced datasets, no specific solutions for this issue are provided. However, they propose the use of weighted metrics to analyze the results, yielding more robust evaluations in unbalanced scenarios.

Recent trends in social media interactions increasingly involve the use of text overlaid on images, creating new opportunities for cyberbullying to emerge in the multimodal domain of social networks. In their work, Li, Zeng, et al. (2024) address

this issue by developing a multimodal system for cyberbullying detection, leveraging a dataset consisting of 5211 irony-aware cyberbullying text-image pairs and 6598 non-cyberbullying pairs. This dataset enables a detailed analysis of the subtleties in cyberbullying expression. The proposed model integrates cutting-edge techniques such as Graph Isomorphism Networks (GIN) for feature transformation, BERT for text processing, and Vision Transformer (ViT) for image analysis. By combining these methods, the multimodal approach enhances the structural information embedded in the data, allowing for a more nuanced understanding of cyber abuse behaviors. The results demonstrate that the framework significantly outperforms traditional single-modality approaches, whether based solely on text classification or image recognition, particularly in detecting irony-aware cyberbullying. Additionally, the model reduces decision-making bias through a soft voting mechanism that aggregates predictions from various feature combinations, improving accuracy and robustness. The study also effectively addresses the challenges posed by the imbalanced dataset using the SMOTE (Chawla et al. 2002; Suhas Bharadwaj et al. 2022). Performance was evaluated through metrics such as weighted F1 and AUC scores. The effectiveness of ViT combined with BERT-based techniques is further highlighted in Tabassum and Nunavath (2024), where the authors address the multimodal classification of memes and text. They compare 15 different configurations of BERT-based models (e.g., BERT, DistilBERT, RoBERTa) with computer vision techniques such as CNNs and ResNet. Their findings conclude that the best performance is achieved using RoBERTa + ViT. To address class imbalance, the authors apply data augmentation, ultimately achieving an 0.992 accuracy on a public dataset, which combines data from four widely used datasets. RoBERTa also achieves the best results (Umer et al. 2024), where the authors propose an innovative model that leverages GloVe embeddings and applies PCA to select the most relevant vector components for processing by RoBERTa. The study compares their approach against various machine learning, deep learning, and transformer-based techniques, demonstrating its high accuracy on a balanced dataset. However, the influence of the model on imbalanced scenarios remains unexplored, leaving room for further research in real-world applications.

In the multilingual domain, (Razi and Ejaz 2024) explores linguistic variations in text messages written in Urdu, Roman Urdu, and English. The research focuses on developing a comprehensive dataset that captures these variations, with a particular emphasis on cyberbullying instances characterized by aggression, repetition, and intent to harm, key elements for a deeper understanding of cyber abuse. The authors propose a novel framework for classifying text messages as aggressive or non-aggressive, leveraging fine-tuned models like m-BERT and MuRIL to improve detection accuracy. The approach demonstrates strong performance, achieving a precision of 0.93, recall of 0.92, and an F1-score of 0.92, showcasing its effectiveness in identifying cyberbullying. However, while the results are promising, the dataset is somewhat imbalanced, and no techniques like undersampling or oversampling were applied to address this issue. The model's strength lies in its ability to capture the core aspects of cyber abuse, namely aggression, repetition, and intent to harm, allowing for a more comprehensive classification of cyberbullying incidents.

Mahmud et al. (2024) presents a dataset of over 5000 manually curated text samples from social media platforms, focusing on the Bangla and Chittagonian languages. The study offers a comprehensive comparison of popular deep-learning models, transformer-based architectures, and traditional machine learning techniques for text classification. The results highlight that deep learning models, especially CNNs and transformer-based architectures, significantly outperform traditional machine learning approaches. These advanced models achieved accuracies ranging from 0.69 to 0.811, showcasing their superior ability to handle the complexities of multilingual and context-specific cyberbullying detection in these languages. Similarly, Saini et al. (2023) present findings where the authors compare an ensemble method, combining a Convolutional Neural Network (CNN) with a Support Vector Machine (SVM), against a BERT-based model for cyberbullying detection. Their results highlight that while the CNN–SVM ensemble achieved a strong accuracy of 96.88%, the BERT model outperformed it, achieving a higher accuracy of 97.34%, showcasing the superior capability of transformer-based models in handling complex language tasks. The effectiveness of transformers is further highlighted in (Mahdi et al. 2024), where the authors demonstrate the significant impact of fine-tuning a BERT model specifically for Arabic languages in the context of cyberbullying classification. The study shows how this targeted fine-tuning leads to remarkable improvements over baseline models, showcasing the model's ability to capture the unique linguistic characteristics of Arabic and significantly enhance classification performance.

In multilingual societies like India, code-switching—mixing multiple languages within the same conversation—is prevalent, and Paul et al. (2023) address this by curating a new annotated dataset of over 22,000 tweets that blend English and Hindi, specifically focusing on instances of cyberbullying. Sayanta et al. proposed a powerful ensemble model that integrates both machine learning and deep learning techniques to improve detection accuracy. This model combines traditional classifiers, such as SVM and Logistic Regression, with advanced deep learning architectures like Multilayer Perceptron (MLP), CNN, BiLSTM, and BERT. For embeddings, the study utilized methods like FastText, CBOW, and TF-IDF vectorization to build domain-specific representations. The proposed deep ensemble model achieved a macro-averaged F1 score of 0.93, demonstrating state-of-the-art performance in identifying cyberbullying in the code-switched dataset. The authors emphasized the critical role of language-specific word embeddings and their concatenation in effectively processing code-switched text, enabling the model to better capture the linguistic nuances present in cyberbullying. However, the study did not account for unbalanced data scenarios, and while the performance is strong, certain cyberbullying techniques are underrepresented, potentially impacting the model's generalization to more complex forms of abuse.

On Instagram, one of the most widely used social networks among teenagers, (Abarna et al. 2023) proposes a novel approach that combines traditional machine learning techniques, such as pattern mining and clustering, with advanced models like BERT and K-BERT. Their method builds knowledge graphs of words associated with different types of cyber abuse, which are then leveraged in an ensemble model using XGBoost to classify not only the presence of cyber abuse but also its intention

and specific type. The results and provided datasets are highly promising, demonstrating the value of integrating traditional machine learning techniques for feature extraction alongside LMs and LLMs. This study underscores the potential of hybrid approaches in improving cyber abuse detection.

Continuing with ensemble techniques Muneer et al. (2023) introduce an approach to detecting cyberbullying on social media platforms, specifically targeting Twitter (now X) and Facebook. The study's key contribution is the development of a stacking ensemble learning model that integrates several deep learning techniques, including a fine-tuned BERT model adapted for cyberbullying data, along with biLSTMs and CNNs. This combination enhances the accuracy and robustness of cyberbullying detection. The research utilizes labeled text datasets from both Twitter and Facebook, focusing on offensive and non-offensive comments to train the models in distinguishing cyberbullying from benign interactions. The stacking ensemble model demonstrates strong performance, achieving an accuracy of 97.4% on the Twitter dataset and 90.97% on a combined dataset of Twitter and Facebook. Furthermore, the study highlights notable improvements in processing efficiency, with the proposed model requiring only 2 min and 45 s for data processing, significantly faster than baseline models. The authors address the challenge of imbalanced data, which is prevalent in cyberbullying detection tasks, by employing robust evaluation metrics such as F1-score and AUC, well-suited for such scenarios. Notably, despite being older techniques, the paper effectively leverages CBOW and Word2Vec embeddings, achieving impressive results in the context of cyberbullying detection. This finding is further supported by Dobrojevic et al. (2024), where authors propose a metaheuristic algorithm to optimize the parameter selection for an XGBoost classifier. They encode text related to sexist cyberbullying using both TF-IDF and BERT, with results showing that TF-IDF slightly outperforms BERT. This leads the authors to argue that while harmful content of this nature is inherently difficult to distinguish, in cases where it is absent, negative classifications tend to be highly accurate. The study is conducted on a highly balanced dataset, yet the authors introduce a robust evaluation method designed to assess performance in imbalanced scenarios.

Ahmed et al. (2022) presented a comprehensive analysis of transformer-based architectures for detecting trait-based cyberbullying in social media comments, contributing to the development of automated detection systems. The authors evaluate several well-known models, including BERT, DistilBERT, RoBERTa, XLNet, and GPT-2, highlighting their effectiveness in capturing the nuances of text data. The study utilizes two publicly available datasets: the Fine-Grained Cyberbullying Dataset (FGCD), which is balanced and categorizes comments into six distinct classes, and the Twitter-parsed Cyberbullying Dataset, which is imbalanced and includes three classes. To address the challenges posed by imbalanced data, the authors implement a semi-supervised learning technique called Dynamic Query Expansion to enhance the representation of underrepresented classes.

In an effort to maximize the benefits of combining multiple models, Chen et al. (2024) integrate two distinct architectures: XLNet and Deep BiLSTM. This approach exemplifies model fusion, where the strengths of each architecture are leveraged to enhance the performance of cyberbullying detection. The study utilizes the COLDATASET, which was meticulously constructed by augmenting an existing Chinese offensive language dataset with 1.66 k offensive remarks from real cyberbullying incidents and one-star reviews sourced from the Douban platform. This enriched dataset captures both explicit and implicit forms of cyberbullying, which are essential for effective detection. The results show that the proposed hybrid model outperforms traditional machine learning and other deep learning approaches in detecting cyberbullying. The use of pre-trained LLMs like XLNet for representation learning proves especially beneficial. In the classification phase, the model addresses data imbalance through oversampling techniques, allowing it to learn implicit and explicit features of cyberbullying more effectively. While the dataset expansion improves balance, it is worth noting that metrics such as AUC, crucial for evaluating performance in imbalanced scenarios, are missing.

Ferrer et al. (2024) contribute to the field of cyberbullying detection by developing an AI-based virtual companion aimed at helping adolescents, particularly those with autism, recognize and respond to instances of cyberbullying. The research leverages a combination of datasets, including the limited KIDS-cyberbullying dataset and a larger Kaggle cyberbullying dataset, which consolidates diverse examples of online harassment from various social media platforms. To address the imbalanced nature of the datasets, data augmentation techniques such as paraphrasing and back translation were employed. The results demonstrate that lightweight models, specifically T5-small and MobileBERT, are capable of effectively detecting cyberbullying in real time within a web-based system. The study emphasizes the importance of considering context in cyberbullying detection, as traditional approaches that focus solely on profanity often miss more subtle forms of abuse. The paper concludes by highlighting the potential of AI-based virtual companions in educating adolescents about cyberbullying and improving their ability to identify and cope with such harmful behaviors.

Finally, Kumar et al. (2024) present an intriguing approach by combining LLMs and LMs for cyberbullying detection. Their method leverages LLMs to generate synthetic data, which is then used to enhance LMs, making them more suitable for classification tasks. The study conducts extensive experiments using GPT-4 and Gemini for synthetic data generation, while various models—including De-BERTa, Longformer, BigBird, HateBERT, MobileBERT, DistilBERT, BERT, RoBERTa, ELECTRA, and XLNet—are evaluated for classification. The best performance is achieved using BERT and Distil-BERT. Additionally, the authors provide a practical tool for cyberbullying detection utilizing OpenAI's ChatGPT. To address the class imbalance problem, they employ data augmentation techniques with LLMs to create a more balanced dataset, further improving classification accuracy.

4.2.1 | Quantitative Analysis of Cyberbullying Detection Leveraging LMs and LLMs

In Table 3, we present a quantitative analysis of cyberbullying detection models, paralleling the insights obtained from hate

speech detection studies. This section explores the prevalence of different models, compares their adoption rates, and examines strategies used to handle data imbalance in the reviewed research.

The results indicate that transformer-based architectures continue to dominate cyberbullying detection tasks. BERT and its variants remain the most widely used language models, appearing in 80.95% of the reviewed studies. This reinforces its reputation as the leading NLP model for cyber abuse detection, consistent with its prominence in hate speech research. However, unlike hate speech detection, where multilingual BERT (mBERT) showed some adoption, it is entirely absent in the analyzed cyberbullying studies, suggesting a lack of focus on multilingual cyber abuse detection.

RoBERTa appears in 23.81% of the studies, demonstrating a moderate level of adoption, but still significantly lower than BERT. Other transformer-based models, such as Mistral, T5, and Gemini, each appear in only 4.76% of the studies, highlighting their limited impact on cyberbullying detection thus far. Similarly, GPT models are used in just 9.52% of the papers, indicating that large-scale generative models have not yet gained widespread traction in this domain.

Beyond transformer-based models, recurrent and convolutional neural networks play a crucial role in cyberbullying detection. BiLSTM models show significant adoption, appearing in 38.10% of the studies, making them the second most frequently used architecture after BERT. GRUs and CNNs also exhibit notable presence, with adoption rates of 14.29% and 19.05%, respectively. These findings suggest that sequence-based and feature-extraction models still offer competitive performance, even in the era of transformers.

Traditional machine-learning approaches also remain relevant. SVMs and XGBoost each appear in 19.05% of the studies, while Random Forest is present in 14.29%. Additionally, vector-based word embeddings such as Word2Vec, GloVe, and FastText are applied in 23.81% of the studies, showing that despite the rise of contextual embeddings, traditional word vectorization methods still play a role in cyberbullying detection.

Data imbalance is a persistent issue in cyberbullying detection, as harmful content is often underrepresented in datasets. Our analysis reveals that 38.10% of the studies implemented explicit strategies to address data imbalance, including data augmentation, oversampling, and undersampling techniques. This is an improvement compared to hate speech studies, where only 31.4% of papers successfully addressed imbalance. The most commonly applied technique in cyberbullying detection is data augmentation, used in three studies, while oversampling and undersampling methods are each employed in three papers as well.

A significant portion of studies (23.81%) did not implement any measures to address data imbalance, potentially leading to biased model performance and suboptimal classification results. However, 38.10% of the papers employed partial mitigation strategies by incorporating robust evaluation metrics such as weighted F1 scores. This suggests that while data balancing techniques are becoming more common, many studies still rely on post hoc evaluation methods rather than proactive dataset adjustments.

4.3 | Comparative Observations and Insights

The analysis highlights key similarities and differences between cyberbullying and hate speech detection models. BERT remains the dominant architecture in both domains, but BiLSTM models exhibit even higher adoption in cyberbullying detection compared to hate speech research. This suggests that recurrent architectures remain particularly valuable for sequential text analysis in cyberbullying contexts.

Additionally, the absence of multilingual BERT in cyberbullying studies indicates a research gap in addressing cyber abuse across different languages. Given the increasing global prevalence of online harassment, future work should explore multilingual and cross-lingual approaches to improve detection in diverse linguistic settings.

The handling of data imbalance has improved in cyberbullying studies compared to hate speech detection research. However, a considerable number of papers still fail to address this issue comprehensively. The reliance on weighted evaluation metrics rather than direct dataset balancing strategies suggests that further methodological advancements are needed to ensure unbiased model performance.

Overall, while transformer-based models like BERT continue to lead cyberbullying detection, traditional machine learning methods, and recurrent architectures maintain a significant presence. Future research should focus on enhancing multilingual capabilities, improving dataset balance methodologies, and further evaluating the role of emerging large language models in cyber abuse detection.

4.4 | Generation of Abusive Content

Before concluding, it is important to acknowledge that LLMs are also being misused for malicious purposes. While this aspect of artificial intelligence is less visible, a review of the literature reveals several papers discussing how LLMs are being exploited to facilitate cyber abuse and other harmful activities on social media. It is crucial to note that this section offers a perspective on the broader implications of LLMs and falls outside the scope of our primary survey.

Divya and Mirza (2024), the authors acknowledge the challenges associated with the misuse of generative AI, particularly in harmful contexts such as the creation of fake news and disinformation. The realistic content generated by AI poses significant risks, as it can be exploited for social engineering and manipulation. Moreover, generative AI's capacity to produce highly realistic images and videos, such as deepfakes, raises additional concerns around identity theft and privacy invasion. Similarly, Wach et al. (2023) discuss several controversies, threats, and drawbacks associated with generative AI, including issues such as algorithmic bias and violations of personal data privacy. The authors explicitly note that "Generative AI can be used for cyberbullying or

other forms of online harassment." They explain that by employing AI to create highly personalized attacks, individuals or groups can leverage social surveillance to identify vulnerable targets and initiate precisely targeted campaigns of harassment or abuse.

In the realm of deepfakes—images and videos generated by generative AI that often closely resemble reality— (Romero Moreno 2024) examines their implications in relation to the EU's Artificial Intelligence Act. The article underscores the necessity of a balanced approach that protects fundamental rights while addressing the risks associated with deepfakes, including misinformation (Barari et al. 2021), extortion (Blancaflor et al. 2024), and privacy violations (Dobrobaba 2022).

In the realm of text generation, (Bullock and Luengo-Oroz 2019) conducts an insightful review of the risks associated with using generative AI for political content, particularly its potential for societal disinformation. The authors identify several risks, including the automated generation of disinformation, fake news dissemination, hate speech, and impersonation. To explore these risks, they conduct a proof-of-concept experiment that illustrates how easily AI can generate politically sensitive text, emphasizing the implications for political stability, peace, and security. The study uses English-language transcripts of speeches delivered by political leaders at the United Nations General Assembly (UNGA) from 1970 to 2015 as training data. Leveraging the AWD-LSTM model, they generated text by "seeding" it with the start of a sentence or paragraph. The quality of the generated text was evaluated based on its coherence and similarity to human-produced content. The results were notable: high-quality text for benign topics was generated around 90% of the time, while generating more inflammatory or sensitive speech required multiple attempts, with approximately 60% success in producing acceptable outputs. The findings highlight both the potential and the challenges of using AI to generate politically charged content.

One of the most interesting studies in this field is Marcondes et al. (2024). This paper reveals that troll factories likely utilize a hybrid approach, integrating both human oversight and automated tools, rather than relying solely on automation. The study investigates the application of LLMs for generating persuasive content on social media, demonstrating their effectiveness in producing contextually relevant tweets. Additionally, the research introduced a General Twitter Corpus to improve LLM performance. The findings suggest that while LLMs can significantly contribute to the dissemination of misinformation and trolling behaviors, human involvement remains crucial for the effective management and coordination of troll operations.

Conversely, we identified a paper that highlights the potential of generative models for creating counter-speech, which serves to respond to hate speech and mitigate the spread of harmful content. Specifically, (Pranesh et al. 2021) propose a data-driven natural language generation task aimed at automatically generating counter-narratives in response to online hate speech. This research utilizes three state-of-the-art pre-trained language models—BART, DialoGPT, and BERT. The proposed method is noted for its scalability and flexibility, enabling it to address hate speech across multiple languages and contexts.

5 | Impact of Artificial Intelligence on Textual Cyber Abuse Detection

After reviewing the most prominent papers on LLMs for cyberbullying or hate speech detection, we extended our analysis to explore the impact of these models, as well as traditional machine learning and deep learning techniques, on other types of textual cyber abuse. In this part of the survey, we broadened our scope beyond LLMs to include other artificial intelligence techniques, given that forms of cyber abuse like doxing and shaming are less frequently studied. This broader approach allowed us to identify a wider range of papers and derive more comprehensive insights. Despite this more general analysis, we remained focused on the same key questions as in the previous section, aiming to understand the influence of models like BERT in the context of textual cyber abuse.

Table 4 summarizes our findings regarding model usage in textual cyber abuse detection. The results indicate that transformer-based architectures, particularly BERT and its variants, are among the most frequently employed approaches. Specifically, BERT is used in 21.05% of the reviewed studies, highlighting its continued dominance in NLP applications related to cyber abuse. However, notable gaps exist in the adoption of multilingual BERT (mBERT) and other large-scale generative models such as Mistral, T5, and Gemini, all of which show no presence in the analyzed papers. GPT-based models appear in only 5.26% of the studies, suggesting that their role in this domain remains limited compared to classification-focused architectures.

Among transformer alternatives, DeBERTa is found in 10.53% of the papers, showing moderate adoption but still lagging behind the original BERT model. Similarly, language-specific BERT models do not appear in the reviewed studies, which may indicate that research on regional or domain-specific adaptations remains underexplored in the context of textual cyber abuse.

In contrast, BiLSTM models exhibit the highest adoption rate at 31.58%, surpassing even BERT-based architectures. This suggests that despite the rise of transformer-based models, recurrent neural networks (RNNs) like BiLSTMs continue to play a crucial role in sequence modeling for text classification tasks, especially in kinds of cyber abuse that remain underexplored and outside the focus of cutting-edge techniques and advancements. Other deep learning approaches, such as GRUs and CNNs, are employed in 10.53% of the studies, demonstrating that alternative architectures remain relevant for feature extraction and pattern recognition.

Traditional machine learning models also show significant presence, with SVMs and Random Forest each appearing in 21.05% of the reviewed papers. XGBoost, another widely used gradient boosting algorithm, is present in 10.53% of the studies, reinforcing the role of ensemble learning methods in cyber abuse detection. Additionally, vector-based word embeddings such as Word2Vec, GloVe, and FastText appear in 6% of the papers, indicating their continued application despite the growing preference for contextualized embeddings from transformers.

Data imbalance remains a critical challenge in cyber abuse detection, as class distributions are often skewed, leading to biased

model performance. Our analysis reveals that only 31.58% of studies actively address this issue through data augmentation, oversampling, or undersampling techniques. In contrast, a significant portion (57.89%) of studies do not implement any explicit strategy for handling imbalanced datasets, potentially leading to suboptimal results and reduced generalizability. A smaller subset (10.53%) employs partial mitigation strategies by incorporating robust evaluation metrics such as weighted F1 scores, ensuring that performance assessments remain reliable despite class imbalances.

Our findings indicate that while transformers like BERT continue to be widely used in textual cyber abuse detection, BiLSTM models remain the most frequently adopted technique. This suggests that recurrent networks still offer competitive performance, particularly for tasks requiring sequential dependencies. The absence of multilingual models, such as mBERT and language-specific BERTs, highlights a gap in research that could be explored further to improve cyber abuse detection across diverse linguistic contexts.

Furthermore, traditional machine learning models such as SVMs and Random Forest continue to hold relevance, particularly in cases where feature-based approaches remain effective. However, their reliance on handcrafted features may limit scalability compared to deep learning methods. The lack of widespread adoption of large-scale generative models like GPT, Mistral, and Gemini suggests that their utility in cyber abuse detection is not yet fully realized.

Finally, the inconsistent handling of data imbalance underscores the need for standardized methodologies to ensure fair and unbiased model performance. Future research should focus on integrating more robust data balancing techniques, particularly in studies employing deep learning approaches, to enhance model effectiveness and reliability in detecting various forms of textual cyber abuse.

Finally, to gain a comprehensive understanding of the proposed papers, we conducted an analysis of their key findings, models, advancements, results, and the handling of unbalanced data. This analysis focused on the five subcategories of cyber abuse identified in our Web of Science queries.

5.1 | Shaming and Cancel Culture

Regarding the detection of shaming and cancel culture in text using LMs or neural networks, only one comprehensive study was retrieved (Aarthi and Chelliah 2022). Aarthi and Balika present a novel approach to detecting and classifying online shaming comments on social media, specifically Twitter, by utilizing a deep recurrent neural network (DRNN) optimized with the Aquila optimization algorithm (Abualigah et al. 2021), a population-based optimization technique inspired by the hunting behavior of Aquila birds. The study employs a substantial dataset comprising ~1,600,034 tweets collected through various public APIs. The results demonstrate that the proposed DRNN model, enhanced by the Aquila optimization algorithm, significantly improves classification accuracy in identifying shaming comments compared to traditional methods. By focusing on the syntactic, semantic, and contextual features extracted from the tweets, the model effectively addresses the complexities of cyber abuse, such as sarcasm and context-dependent meanings. This research contributes to the field by providing a robust framework for detecting online shaming.

5.2 | Doxing

Aarthi and Chelliah (2023) introduce a novel approach for detecting and classifying offensive comments on social media, with a specific focus on doxing. The model employs advanced techniques such as CNN for feature extraction and Bidirectional Gated Recurrent Units (BiGRU) for sequence modeling, allowing it to capture both local patterns and contextual information in the text. The study utilizes various datasets, including a labeled hate speech detection dataset with 3000 comments from platforms like Twitter and Reddit, and the Tweepy dataset, which contains over a million tweets. The results demonstrate the model's strong performance, achieving an F1-score of 0.808 for English subtask A, indicating effective classification capabilities. Specifically, the model incorporates intrinsic methods to identify sensitive keywords and phrases associated with doxing, enhancing its ability to detect harmful content.

More recently, Miao et al. (2024) deep learning-based method for preventing data leakage in the Electric Power Industrial Internet of Things specifically addresses the challenges of safeguarding sensitive personal data during interactions between the State Grid business platform and third-party platforms. The main contributions include the development of a robust framework that combines the DeBERTa (He et al. 2020) model with a BiLSTM-CRF (Huang 2015) architecture to enhance the identification of privacy-sensitive information, such as identity cards, phone numbers, and email addresses, within both structured and unstructured text data. The experimental results demonstrate an F1 score of 81.26% on the CLUENER 2020 dataset (Xu et al. 2020), indicating the method's effectiveness in accurately detecting sensitive data entities and mitigating the risk of data leakage. In terms of data handling, the approach employs a combination of regular expressions and keyword detection to construct a sensitive data distribution library, facilitating the identification and filtering of personal information. The model processes text data by extracting semantic features and contextual information, allowing for precise matching of sensitive content. This dual strategy not only enhances the recognition of personal data but also ensures that privacy-sensitive information is desensitized before being shared, thereby providing a reliable solution for data security in the power industry context.

Finally, Li et al. (2022) presents the covert task embedding (CTE) attack, a novel method that exploits deep neural networks to extract sensitive personal information while performing a legitimate task, such as age estimation. The main contribution lies in demonstrating how a deep neural networks can be trained to simultaneously execute a primary task and a covert task (e.g., detecting gender or ethnicity) without compromising the accuracy of the primary function. The experiments utilized face-based images as input data, showcasing the feasibility of the CTE

attack across different neural network architectures, including multiclass classification and soft-ranking networks. The results indicated that the hidden information could be reliably extracted without impairing the performance on the primary task, highlighting significant privacy risks. In terms of data handling, the study emphasizes the importance of safeguarding personal information by illustrating how covert tasks can be embedded within seemingly innocuous models. The authors propose a key-protected CTE approach, where the extraction of sensitive information requires a secret key, thereby adding a layer of complexity to the data leakage process. This method underscores the need for robust privacy measures in neural network applications in contexts where personal data is processed.

5.3 | Emotional and Psychological Abuse

Although the subsequent papers do not specifically target cyber abuse, they are relevant to our study due to their focus on text-based data and advanced analytical techniques. Although these studies do not directly address cyber abuse, their methodologies and findings provide a foundation that can be leveraged to enhance the detection and understanding of emotional abuse in digital environments. Thus, while not directly applicable, these papers contribute to the broader context and future development of cyber abuse detection tools.

The study (Jang et al. 2022) investigates the use of NLP techniques to analyze text data for predicting personality traits and psychological distress. It combines qualitative data from written responses and transcribed speech, employing models such as Latent Dirichlet Allocation (LDA) and BERT (Devlin et al. 2019) for data processing and analysis. Although the study does not directly address cyber harassment, its methodologies could be adapted to study emotional abuse in online contexts by analyzing language patterns indicative of distress or victimization. The research highlights the importance of developing accurate algorithms for predicting psychological outcomes from various linguistic inputs, but it does not address the issue of imbalanced data. The findings could be used to inform the development of tools for detecting emotional abuse by identifying specific linguistic markers associated with distress in online communications.

The paper (Chen et al. 2022) explores psychological distress by applying four machine learning algorithms—logistic regression, random forests, artificial neural networks, and gradient boosting—to data from the Health Information National Trends Survey (HINTS), which includes 5484 respondents. The study identifies 20 significant predictors of psychological distress, including sociodemographic and lifestyle-related factors. The dataset was a bit unbalance, with 2874 respondents reporting psychological distress and 2610 without, but this scenario was correctly addressed using techniques such as 10-fold cross-validation to improve model robustness, and they provided metrics such as AUC and F1 score to evaluate model performance.

Although not focused solely on text, a study by (Alemu et al. 2023) examines speech features related to emotional distress. The research aims to develop an automated speech analysis algorithm capable of detecting clinically relevant emotional distress and functional impairments in children and adolescents with mental health disorders. The study's key contribution lies in its integration of speech analysis with validated psychological assessments, providing a comprehensive understanding of emotional states. Preliminary results are promising, with the algorithm achieving up to 85.35% accuracy in emotion classification using a Gaussian mixture model (GMM). The GMM effectively handles unbalanced data by emphasizing feature extraction and frame selection techniques. While the study primarily focuses on detecting emotional distress, it also has implications for identifying issues related to cyber harassment and emotional abuse by recognizing emotional signals in speech that may originate from such experiences.

5.4 | Trolling

Bakar et al. (2023) introduce a Support Vector Machine (SVM) classification model aimed at detecting troll threat sentences in the Malay language, contributing significantly to the field of text simplification and addressing the challenges of cyber abuse associated with trolling. The study utilized the Malay Text Simplification Dataset, comprising 6836 instances categorized as complex or non-complex, and implemented class weighting in the SVM classifier to effectively tackle the issue of class imbalance within the dataset. This approach ensures that both complex and non-complex sentences are accurately identified, enhancing the model's performance. Results indicated that the SVM classifier achieved an impressive average accuracy of 92.37% when incorporating semantic features. The model's focus on semantic and lexical features allows for a nuanced understanding of trolling behavior, which often involves complex language and hidden meanings.

In their study, Nepomuceno et al. (2023) examine the impact of marketer-generated content (MGC)—marketing communications initiated by companies on their official social media pages—on user interactions in online communities, with a particular focus on trolling. The authors analyze a large dataset of posts from platforms such as Facebook, Instagram, and Twitter, employing a mix of unsupervised and supervised machine learning techniques, including Latent Dirichlet Allocation (LDA) (Blei et al. 2003) and vector autoregression (VAR) to classify content and assess its toxicity. Their findings reveal a positive correlation between the average toxicity of comments and user engagement, suggesting that toxic interactions may inadvertently boost product usage. To address the challenge of toxicity classification, the study employs a time-series model (panel-data vector autoregression) to effectively capture the dynamics of user interactions over time. However, despite the unbalanced nature of the problem and the dataset, the study does not provide evaluations related to this imbalance. In the context of cyber abuse, the research underscores the psychological motivations behind trolling behavior, which often involve traits like narcissism and sadism. Trolling is identified as the intentional disruption of online communities through toxic comments. The authors stress the importance of understanding the elements of MGC that may contribute to such negative interactions and advocate for strategies to mitigate trolling behavior.

Cranmer et al. (2024) contribute to the understanding of state-sponsored disinformation and the role of social media trolls in shaping public discourse. The study analyzes a dataset of 163,549 tweets related to Daryl Morey's controversial tweet supporting pro-democracy protests in Hong Kong. The research categorizes these tweets into various themes, revealing that 68.81% of the responses were focused on promoting the image of the People's Republic of China (PRC) and attacking Morey. This highlights the significant prevalence of troll activity and its impact on public perception. To address the challenge of unbalanced data, the authors employed a natural language classifier, achieving an overall accuracy of 78% in categorizing tweets, comparable to the performance of the BERTweet model (Nguyen et al. 2020). The dataset was randomly sampled and categorized into five distinct themes, some of which had low representation. However, the paper does not provide detailed explanations regarding the handling of data imbalance. The study emphasizes the need for advanced analytical frameworks to effectively combat disinformation and online abuse. By leveraging machine learning and NLP techniques, the authors successfully identified and categorized troll activity, offering valuable insights into how these tactics distort online discourse.

In line with precedent work examining how trolls are utilized by foreign entities, (Swed et al. 2024) advances our understanding of information operations by highlighting the distinct social footprint of Russian troll accounts, shaped by specific social norms and expectations. The study leverages extensive datasets, including over 2.9 million English tweets and nearly 4.9 million Russian tweets from the Internet Research Agency (IRA), focusing on text data to analyze behavioral patterns. The research categorizes accounts into four types: Fake News, Organizations, Political Affiliates, and Default Individuals. The model is built on a random forest classifier (RFC), utilizing carefully selected features to minimize multicollinearity, thereby improving the model's learning efficiency. To address the challenge of imbalanced data, the RFC algorithm compensates for the unequal distribution of categories within the dataset. Additionally, least absolute deviation (L1) normalization is applied to balance the weights of each variable, preserving the dataset's structure while enhancing the model's predictive accuracy. The model achieved high accuracy rates, up to 90.7% for certain categories, by leveraging features such as tweet frequency, timing, and linguistic characteristics to identify and predict the behavior of troll accounts. This approach is particularly relevant to understanding cyber abuse, as it provides a robust framework for detecting and analyzing trolling behaviors that frequently manifest as targeted harassment on social media platforms. Ezzeddine et al. (2023) introduce a novel behavioral-based approach for detecting state-sponsored troll activity on social media, emphasizing the analysis of online activity sequences rather than just textual content. The key contributions of the study include the development of a "Troll Score," a metric designed to quantify troll-like behavior, and the implementation of an LSTM model for classifying user trajectories with high accuracy. The study utilizes a dataset from Twitter, examining both active and passive online activities to distinguish the behavioral patterns of trolls from those of organic users. These sequences, termed "trajectories," capture the chronological order of actions such as posting, retweeting, and interacting with others. The results highlight the model's effectiveness, achieving an AUC of ~91% in differentiating trolls from legitimate users, even in the presence of unbalanced data—demonstrating the robustness of the model's performance. The study addresses the challenges of cyber abuse related to trolling by focusing on behavioral cues that are more difficult to mimic than textual content, thereby increasing the resilience of their detection method against the sophisticated tactics employed by LLMs in influence operations. In a similar vein, using LSTMs, (Been Choi et al. 2024) introduce the term "agro" to describe a broader category of attention-seeking behaviors in online discourse, such as clickbait and trolling. The term "agro" originates from the Korean word for "aggravation" and refers to a specific type of provocative content designed to capture attention in online environments. It typically involves eye-catching language or imagery aimed at engaging users on platforms like YouTube and social media, often relying on intense emotional, sexual, or sensational appeals. The study delves into the psychological mechanisms behind such attention-seeking behaviors. The authors created and rigorously labeled a dataset focused on South Korean content, consisting of 1881 headlines and thumbnails. Due to class imbalance, they applied SMOTE to balance the dataset. Their findings highlight the effectiveness of machine learning models, particularly Bi-LSTM and CNN, in accurately detecting agro versus non-agro content.

Asif et al. (2024) introduce a novel machine learning framework designed to detect trolls and toxic content on social media by leveraging deep learning techniques, specifically Graph Convolutional Networks (GCNs) and GloVe word embeddings. A key innovation of this work is its ability to analyze text embedded within images, effectively addressing a significant limitation in existing content moderation methods. The study employs diverse datasets encompassing both textual and visual elements, acknowledging the challenges posed by imbalanced data, where non-toxic comments vastly outnumber toxic ones. To counter this imbalance, the authors apply data augmentation techniques, which significantly enhance the model's ability to identify trolling behaviors. The results demonstrate that the GCN-based model achieves a testing accuracy of 0.92 and an F1-score of ~0.4, indicating its effectiveness in detecting toxic content, though with noted areas for improvement in precision and recall. The framework underscores the importance of contextual analysis in text data, utilizing natural language processing techniques to capture the subtleties of online interactions.

Uyheng et al. (2022) advance our understanding of online trolling by employing a psycholinguistic framework to differentiate between trolls and bots, emphasizing that not all trolls are automated accounts and vice versa. The study uses a dataset of 4917 labeled instances of trolling and non-trolling interactions, focusing on psycholinguistic features such as toxicity, cognitive complexity, and direct targeting. Leveraging these features, the researchers propose a classification system that utilizes both traditional machine learning models and advanced deep learning techniques, including BERT. The study provides insights into the nature of cyber abuse by analyzing the linguistic properties of trolling behavior. By examining the language used in trolling, the research lays the groundwork for developing tools to effectively identify and mitigate instances of cyber abuse. The authors found that an LSTM model achieved the best results. However, the classes in the dataset were manually balanced,

which limits the generalizability of the system. Despite this limitation, the study underscores the importance of understanding the underlying psycholinguistic characteristics of malicious online interactions to create more effective detection and prevention strategies.

In the multimodal domain, MacDermott et al. (2022), the authors present a framework for detecting trolls and toxic content on social media by integrating deep learning techniques for both text and image analysis. The framework utilizes GloVe word embeddings and recurrent neural networks (RNNs), including Bidirectional LSTM and GRU, to analyze textual content. It employs a dataset of 223,549 comments from the Jigsaw dataset, which is characterized by a significant class imbalance, with less than 7% of comments labeled as toxic. To address this imbalance, the authors apply undersampling techniques. The results indicate that the proposed models achieved high detection rates for toxic comments. However, accuracy decreased when analyzing text extracted from images due to the challenges associated with this task. The framework's ability to extract and classify text from images is particularly important for addressing cyber abuse, as it facilitates the analysis of content in non-traditional formats such as memes or screenshots. The highest F-measure achieved was 0.88, using an ensemble model, supported by undersampling techniques. Finally, in the multimodal domain, (Sultan et al. 2024) propose combining image transformers and image classification techniques with various LMs and LLMs for the multimodal classification of troll memes. Their experiments demonstrate the practical applicability of these methods, particularly in zero-shot and few-shot learning scenarios using models such as GPT-2 and GPT-4. The results highlight the effectiveness of LLMs in environments with limited labeled data. Among the tested models, DeBERTa, a variation of BERT, achieves the highest accuracy. This model, along with several other LMs and LLMs, is extensively evaluated throughout the study, showcasing its superiority in multimodal meme classification.

5.5 | Impersonation

In the context of impersonation, Mewada and Dewang (2024) address the issue from a dual perspective: analyzing both the content generated by accounts and their interactions within networks. The study leverages techniques from text analysis and network analysis, particularly focusing on homophily—the tendency of individuals to associate with others who are similar in interests, demographics, or behaviors. This approach is used to propose a novel method for identifying fake profiles on social media, specifically targeting cyber abuse through impersonation and other malicious activities. The primary contribution of the paper is the integration of linguistic features, profile-centric features, and group-centric features to improve the detection of sockpuppets and crowdturfing communities. The study utilizes a dataset of user-generated content from various social media platforms, employing a Convolutional Neural Network (CNN) to classify users as either malicious or genuine. The results demonstrate the framework's effectiveness, with the CIPF framework achieving high precision, recall, and F1-scores, significantly outperforming state-of-the-art techniques in detecting fake profiles. A key aspect of the framework is its incorporation of linguistic features, including sentiment analysis, grammatical quality, and emotional context, which provide insights into the language patterns used by individuals. These linguistic features are critical for detecting impersonation, as they help identify anomalies in user-generated content that may indicate deceptive practices. This enhances the framework's ability to distinguish between genuine users and malicious actors, making it a robust tool for combating cyber abuse through impersonation.

We identified a borderline paper by Huang et al. (2021), where the authors proposed an automatic system to detect deceptive voices used in impersonation attempts. The reason for selecting this study is its twofold approach, which analyzes both text-independent and text-dependent features. The latter involves detecting when an impersonator mimics specific phrases or commands typically used by the genuine speaker, thus linking the process to text analysis. The paper introduces a quasiGaussian distribution (QGD) model that effectively addresses both types of impersonation attacks on smart devices. The authors collected a unique dataset from a reality TV show, which enhances the robustness of their findings and enables a comprehensive evaluation of voice and text features. The results demonstrate that the QGD model can accurately distinguish between genuine and impersonated voices by analyzing the Euclidean distance within the voice (and text) feature space. The study reveals that voice features from the same individual are tightly clustered, while those from different individuals are more dispersed, making it a promising approach for detecting impersonation.

6 | ACL Comparison

Since ACL conferences are widely regarded as a leading venue for state-of-the-art NLP research, we found it essential to include a new section that compares our survey results with the most recent ACL conference proceedings. To do this, we conducted a search using the same topic terms as in our study, specifically focusing on papers with “hate speech,” “toxic,” “doxing,” “impersonation,” “trolling,” “shaming” or “emotional abuse” in their titles or abstracts. This search yielded 10 papers from the past year, of which 9 were relevant to cyber abuse detection, and all of them were based on hate speech. It is notable that other forms of cyber abuse, such as cyberbullying, were conspicuously absent, highlighting a gap in research within this domain.

Our analysis reveals that BERT-based models continue to play a central role in hate speech detection. Despite the rise of LLMs such as GPT-3.5, LLaMA, and Mistral, BERT and its variants (RoBERTa, BERTweet, DeBERTa, HateBERT, etc.) remain widely utilized for a variety of tasks. In most studies, BERT-based models remain competitive and often outperform LLMs, particularly in traditional classification tasks. For instance, in the NaijaHate study (Tonneau et al. 2024), BERTweet and DeBERTa achieved top results alongside XLM on a specific Nigerian dataset.

Our analysis also highlights the growing use of new techniques such as contrastive learning and models like T5, which have shown promising results. However, BERT remains instrumental

in these advancements. In VIHATET5 (Thanh Nguyen 2024), BERT played a crucial role in dataset annotation. In the realm of contrastive learning, models like SharedCon (Ahn et al. 2024) and LAHN (Kim et al. 2024) have demonstrated significant improvements in implicit hate speech detection. Both methods leverage BERT as an encoder, further reinforcing the model's importance in enhancing classification performance.

While LLMs are explored in some studies, they do not consistently outperform fine-tuned BERT-based models. For instance, Tox-BART (Yadav et al. 2024) found BART to be the best-performing model but still used BERT as a regressor for in-domain attributes. Similarly, Don't Go to Extremes (Zhang et al. 2024) found GPT-3 to be the best model in their study, but the paper also highlights limitations of LLMs, such as oversensitivity and calibration issues in hate speech detection.

In the multilingual context, many studies still rely on BERT-based models for low-resource and multilingual language tasks. Both NaijaHate (Tonneau et al. 2024) and VIHATET5 (Thanh Nguyen 2024) demonstrate that fine-tuned BERT variants remain highly effective in these contexts, often outperforming LLMs.

Finally, the growing interest in explainability and text generation to enhance moderator efficiency in social spaces is noteworthy. For instance, Explainability and Hate Speech (Calabrese et al. 2024) explores how structured explanations can help moderators work more efficiently, though not necessarily improving model performance alone.

In conclusion, while LLMs are increasingly tested for hate speech detection, they have not yet replaced BERT-based models as the dominant approach. Research is evolving toward explainability (XAI), contrastive learning, and domain-specific optimizations, which enhance existing transformer-based architectures rather than replacing them. This suggests that BERT-based models remain a cornerstone of hate speech detection research, with LLMs playing a complementary role rather than serving as the default solution.

7 | Discussion

In this section, we discuss the results of our survey, focusing specifically on the research questions guiding this study. To enhance clarity and the practical utility of our findings, we have structured the discussion by addressing each research question individually. This approach ensures a comprehensive and organized analysis of the key insights drawn from the reviewed literature.

7.1 | RQ1: How Representative Are LMs and LLMS in Detecting Textual Cyber Abuse Compared to Traditional Methods?

To address this question, we performed an occurrence analysis based on the models used across the reviewed literature, focusing on those referenced in at least three papers. Table 6 presents the results for LMs and LLMs, while Table 5 shows the results for traditional machine learning techniques, being traditional methods algorithms and techniques that have been foundational to the field of machine learning prior to the widespread adoption of deep learning, including early deep learning methods (as CNNs) and the first static embedding models. It is worth mentioning that, for our purposes, "representative" is defined by how frequently a specific LLM or LM is mentioned or used in articles. In other words, the more frequently a model is used, the more representative it is considered. As shown in the tables, BERT is the most widely used model, significantly outperforming both traditional machine learning techniques and other language models. Among LLMs, GPT and Gemini stand out as the only models with multiple occurrences, while models like T5, LLAMA, and Mistral appear only twice.

When comparing the adoption of LMs to LLMs, we can observe that LMs remain the dominant tool in cyber abuse detection. BERT-based models are particularly widespread due to their flexibility in training and fine-tuning, making them highly effective. While LLMs are underrepresented, they present promising opportunities, especially for few-shot learning and data augmentation. Traditional machine learning techniques, though less prevalent in hate speech or cyberbullying detection, continue to play a role in other forms of cyber abuse. Notably, SVM and random forests remain useful due to their interpretability and ability to explain results. Interestingly, even older embedding techniques like FastText and Word2Vec still appear frequently in recent literature, highlighting their enduring relevance in specific contexts.

Analyzing the performance of hate speech detection and other cyber abuse detection techniques presents challenges due to the lack of a standardized benchmark across studies. Many papers introduce new techniques and validate them on custom datasets or adapt them for specific languages, making direct comparisons difficult. This fragmentation in datasets and evaluation methods complicates the ability to gauge overall progress in the field.

However, from a broader perspective, the use of transformer-based architectures consistently outperforms other approaches. This trend is evident in widely used datasets like Ethos Binary (Mollas et al. 2022), HateXplain (Mathew et al. 2021), and OffensEval 2019 (Zampieri et al. 2019), where BERT-based models consistently top the leaderboards (Rajput et al. 2021; Caselli et al. 2021; Kim et al. 2022), reaffirming their superior performance in cyber abuse detection tasks. This trend and these findings are also corroborated by the latest edition of OffensEval in 2023, where the baselines were established using BERT and ELMo, achieving impressive results. Extensive research has been conducted on other types of LLMs, such as (Zampieri et al. 2023), where the authors reviewed recent results from the competition. They concluded that LLMs like Falcon-7B-Instruct (Penedo et al. 2023), RedPajama-INCITE-7B-Instruct (Weber et al. 2024), MPT-7B-Instruct (Team 2023), Llama-2-7B-Chat (Touvron et al. 2023), T0-3B (Sanh et al. 2021), and Flan-T5-large (Chung et al. 2024) underperformed compared to the top three systems in these competitions, which were based on more established transformer-based models such as BERT. In contrast, other papers as (Faria et al. 2024) found that models as GPT-3.5 and

TABLE 5 | Most used traditional machine learning methods with more than three occurrences.

Machine learning or traditional embeddings	Occurrences	Citations
SVM	12	(Paul et al. 2023), (Mahmud et al. 2024), (Ahmed et al. 2022), (Saini et al. 2023), (Zhou et al. 2023), (Arcila-Calderón et al. 2022), (Sreelakshmi et al. 2024), (Althobaiti 2022), (Aarthi and Chelliah 2022), (Aarthi and Chelliah 2023), (Umer et al. 2024), (Garg et al. 2024)
CNN	12	(Ansari et al. 2024), (Putra and Wang 2024), (Bilal et al. 2023), (Khan et al. 2022), (Mewada and Dewang 2024), (Paul et al. 2023), (Saini et al. 2023), (Been Choi et al. 2024), (Khafajeh 2024), (Tabassum and Nunavath 2024), (Umer et al. 2024), (Madhavi et al. 2024)
Random Forest (RF)	11	(Mahmud et al. 2024), (Li, Zeng, et al. 2024), (Sreelakshmi et al. 2024), (Mehta and Passi 2022), (Bilal et al. 2023), (Toliyat et al. 2022), (Husain and Uzuner 2022), (Swed et al. 2024), (Been Choi et al. 2024), (Umer et al. 2024), (Alkomah et al. 2022)
Logistic regression (LR)	10	(Paul et al. 2023), (Mahmud et al. 2024), (Li, Zeng, et al. 2024), (Mehta and Passi 2022), (Arcila-Calderón et al. 2022), (Sreelakshmi et al. 2024), (Anjum and Katarya 2024), (Bilal et al. 2023), (Uyheng et al. 2022), (Garg et al. 2024)
XGBoost	9	(Ahmed et al. 2022), (Yun et al. 2023), (Bilal et al. 2023), (Toliyat et al. 2022), (Li, Zeng, et al. 2024), (Been Choi et al. 2024), (Abarna et al. 2023), (Dobrojevic et al. 2024), (Alkomah et al. 2022)
Naive Bayes (NB)	7	(Mahmud et al. 2024), (Sreelakshmi et al. 2024), (Arcila-Calderón et al. 2022), (Mehta and Passi 2022), (Toliyat et al. 2022), (Umer et al. 2024), (Alkomah et al. 2022)
GLOVE	7	(Asif et al. 2024), (Altinel et al. 2024), (Alrowais et al. 2024), (Umer et al. 2024), (Garg et al. 2024), (Saleh et al. 2023), (Alkomah et al. 2022)
Word2Vec	6	(Muneer et al. 2023), (Saleh et al. 2023), (Altinel et al. 2024), (Umer et al. 2024), (Saleh et al. 2023), (Alkomah et al. 2022)
KNN	5	(Umer et al. 2024), (Mahmud et al. 2024), (Li, Zeng, et al. 2024), (Alkomah et al. 2022), (Sreelakshmi et al. 2024)
TF-IDF	5	(Paul et al. 2023), (Altinel et al. 2024), (Khafajeh 2024), (Dobrojevic et al. 2024), (Alkomah et al. 2022)
FastText	5	(Badri et al. 2024), (Maity et al. 2024), (Arshad et al. 2023), (Paul et al. 2023), (Ali et al. 2022)
FastText	3	(Paul et al. 2023), (Maity et al. 2024), (Arshad et al. 2023)
RNN	3	(Arcila-Calderón et al. 2022), (Garg et al. 2024), (Husain and Uzuner 2022)

Gemini performs better than BERT under certain configurations, which opens new paths of research.

7.2 | RQ2: Considering That Cyber Abuse Frequently Involves Imbalanced Classification Challenges, Are Researchers Adequately Addressing This Issue in Their Evaluation Metrics and Pre-Processing Methods?

Based on the reviewed literature, it is clear that researchers in the field of cyber abuse detection are increasingly recognizing the challenges posed by imbalanced classification. Many studies now utilize a range of evaluation metrics—such as precision, recall, and F1-score—which provide a more nuanced understanding of model performance than accuracy alone, particularly in contexts where class distribution is skewed. However, some research still relies on overall accuracy or non-weighted metrics, potentially obscuring the true effectiveness of models in detecting minority classes.

A notable trend is the adoption of various pre-processing methods to address class imbalance, including oversampling techniques like SMOTE (Li, Zeng, et al. 2024) and undersampling methods. These strategies aim to create a more balanced training dataset, ultimately improving model performance on underrepresented classes. Additionally, some studies are incorporating cost-sensitive learning, which assigns varying penalties for misclassifications based on the importance of each class. The introduction of generative models, such as GPT, for data augmentation also represents a promising avenue for enhancing the minority class representation,

TABLE 6 | Most used models LMs and LLMs models with more than three occurrences.

Model	Occurrences	Citations
BERT	50	(Muneer et al. 2023), (Paul et al. 2023), (Mahmud et al. 2024), (Li, Zeng, et al. 2024), (Alfurayj et al. 2024), (Al-Harigy et al. 2024), (Maity et al. 2022), (Mahdi et al. 2024), (Samee et al. 2023), (Saini et al. 2023), (Ahmed et al. 2022), (Ansari et al. 2024), (Zhou et al. 2023), (García-Díaz, Jiménez-Zafra, et al. 2023), (García-Díaz, Pan, et al. 2023), (del Valle-Cano et al. 2023), (Pookpanich and Siriborvornratanakul 2024), (Arcila-Calderón et al. 2022), (Maity et al. 2024), (Anjum and Katarya 2024), (Nandi et al. 2024), (Putra and Wang 2024), (Yun et al. 2023), (Chhikara et al. 2024), (Saleh et al. 2023), (Almaliki et al. 2023), (Ramos et al. 2024), (Pan et al. 2024), (Bilal et al. 2023), (Ali et al. 2022), (Firmino et al. 2024), (Toliyat et al. 2022), (Khan et al. 2022), (Althobaiti 2022), (Gupta et al. 2021), (Sultan et al. 2024), (Kumar et al. 2024), (Abarna et al. 2023), (Khafajeh 2024), (Tabassum and Nunavath 2024), (Dobrojevic et al. 2024), (Umer et al. 2024), (Faria et al. 2024), (Moreno-Sandoval et al. 2024), (Ghosh and Senapati 2024), (Garg et al. 2024), (Madhavi et al. 2024), (Saleh et al. 2023), (Yadav and Singh 2024), (Alkomah et al. 2022), (Badri et al. 2024)
bi-LSTM	20	(Muneer et al. 2023), (Chen et al. 2024), (Paul et al. 2023), (Mahmud et al. 2024), (Alfurayj et al. 2024), (Al-Harigy et al. 2024), (Maity et al. 2022), (Ansari et al. 2024), (Maity et al. 2024), (Saleh et al. 2023), (Bilal et al. 2023), (Mazari et al. 2024), (Miao et al. 2024), (Been Choi et al. 2024), (Khafajeh 2024), (Tabassum and Nunavath 2024), (Umer et al. 2024), (Garg et al. 2024), (Saleh et al. 2023), (Alkomah et al. 2022)
RoBERTa	14	(Alrowais et al. 2024), (Ahmed et al. 2022), (Su et al. 2023), (Yun et al. 2023), (Pan et al. 2024), (Arshad et al. 2023), (Ali et al. 2022), (Gupta et al. 2021), (Pookpanich and Siriborvornratanakul 2024), (Kumar et al. 2024), (Tabassum and Nunavath 2024), (Umer et al. 2024), (Moreno-Sandoval et al. 2024), (Ghosh and Senapati 2024)
DistilBERT	12	(Ahmed et al. 2022), (Pookpanich and Siriborvornratanakul 2024), (Montesinos-Cánovas et al. 2023), (Sreelakshmi et al. 2024), (Ali et al. 2022), (Pan et al. 2024), (Malik et al. 2024), (Firmino et al. 2024), (Kumar et al. 2024), (Tabassum and Nunavath 2024), (Ghosh and Senapati 2024), (Yadav and Singh 2024)
BiGRU and GRU	11	(Maity et al. 2022), (Wullach et al. 2022), (Mazari et al. 2024), (Tabassum and Nunavath 2024), (Garg et al. 2024), (Madhavi et al. 2024), (Singh et al. 2023), (Mahmud et al. 2024), (Keya et al. 2023), (Aarthi and Chelliah 2023), (Mazari et al. 2024)
m-BERT	10	(Razi and Ejaz 2024), (García-Díaz, Pan, et al. 2023), (Montesinos-Cánovas et al. 2023), (Karayi˘git et al. 2022), (Ollagnier et al. 2023), (Kumar et al. 2024), (Faria et al. 2024), (Ghosh and Senapati 2024), (Siddiqui et al. 2024), (Singh et al. 2023)
GPT (including GPT-2, GPT-3.5)	7	(Ahmed et al. 2022), (Chhikara et al. 2024), (Ramos et al. 2024), (Sultan et al. 2024), (Kumar et al. 2024), (Faria et al. 2024), (Moreno-Sandoval et al. 2024)
BETO	5	(Moreno-Sandoval et al. 2024), (García-Díaz, Jiménez-Zafra, et al. 2023), (García-Díaz, Pan, et al. 2023), (Montesinos-Cánovas et al. 2023), (del Valle-Cano et al. 2023)
XLM-RoBERTa	5	(Mahmud et al. 2024), (Pookpanich and Siriborvornratanakul 2024), (Pan et al. 2024), (Ali et al. 2022), (Siddiqui et al. 2024)
TwHIN-BERT	3	(García-Díaz, Pan, et al. 2023), (Pookpanich and Siriborvornratanakul 2024), (Montesinos-Cánovas et al. 2023)
Gemini	3	(Faria et al. 2024), (Kumar et al. 2024), (Ramos et al. 2024)

functioning similarly to oversampling techniques (Wullach et al. 2022). Despite these advancements, a lack of standardization in how imbalanced classification is addressed remains a significant issue across studies. This inconsistency complicates efforts to compare results and identify best practices. While there is growing awareness of the need to tackle class imbalance in cyber abuse detection, substantial gaps persist. Improved standardization of evaluation metrics and preprocessing methods, coupled with clearer reporting of methodologies, would greatly enhance the robustness and comparability of findings in future research.

7.3 | RQ3: Based on the Reviewed Literature, What Emerging Trends and Future Directions Are Shaping the Development of Cyber Abuse Detection?

One key insight derived from the survey is that recent advancements in cyber abuse detection are focused more on applying standardized English-language models to new or underrepresented languages, rather than developing entirely new architectures. The predominant approach in these studies involves transfer learning and fine-tuning existing models, which allows for the adaptation of well-established models to different linguistic contexts. This trend reflects a shift toward leveraging the strengths of existing architectures, optimizing them for diverse languages rather than creating novel frameworks from scratch.

In cyber abuse detection, several evolving trends are reshaping the field. Traditionally, this domain has been treated primarily as a binary classification problem—identifying abusive or nonabusive content. However, more recent studies are shifting toward a more nuanced understanding, recognizing that cyber abuse involves multiple dimensions. These include factors such as the nature of the target, the directness of hate, the type of hostility, and whether the abuse is community-driven. This shift has led to the development of multi-task algorithms, capable of performing tasks such as sentiment analysis, community detection, and toxicity prediction simultaneously.

A growing trend is the incorporation of social network analysis and traditional data mining techniques such as sentiment analysis, methods that have been demonstrated valid in similar domains as crime detection (Oatley and Ewart 2011). These methods are increasingly acknowledged as powerful tools for analyzing the complex, relational aspects of cyber abuse, such as how it spreads within communities or networks and the influence of key actors in propagating hate speech. We identified four insightful papers that highlight the value of unsupervised techniques in this area. These works underscore the necessity of analyzing cyber abuse from multiple perspectives, beyond simple classification, to capture its broader dynamics and impact.

If we stick only to the classification domain of the problem, the majority of efforts are focused on text preprocessing and the creation of embeddings to capture the nuances of language. Among various models, the combination of BERT and LSTMs has emerged as one of the most powerful tools for this purpose.

Another significant observation is that the majority of current datasets are heavily based on Twitter, while other social platforms, such as Reddit, remain underexplored. This reliance on a single platform limits the generalizability and versatility of existing models. Expanding datasets to include diverse sources of social media data is crucial for enhancing the robustness and adaptability of cyber abuse detection systems.

Another important insight derived from our research is related to XAI techniques. We have observed an increasing trend in the use of such methods, as highlighted by Saleh et al. (2023), Garg et al. (2024), and Yadav and Singh (2024). These studies demonstrate a shift in focus—not only on achieving high accuracy in hate speech detection but also on ensuring interpretability. This shift is crucial, as it enables moderators and end-users to better understand and trust the decisions made by AI-driven tools.

Finally, there is a growing call for the standardization of evaluation processes across studies. Many papers claim to outperform baseline models, but these claims often rely on custom datasets or datasets limited to a single platform, making meaningful comparisons difficult. A standardized approach to reporting preprocessing methods, evaluation metrics, and dataset usage is essential. This would improve the comparability of findings across different studies and support the development of best practices in the field, ensuring more reliable and generalizable models for detecting cyber abuse.

7.4 | RQ4: Which Types of Cyber Abuse Detection Are Most Significantly Influenced by LLMs?

As demonstrated in Table 6, BERT-based techniques continue to dominate across various types of cyber abuse detection. However, when focusing solely on LLMs, certain areas—such as shaming, doxing, emotional abuse, trolling, and impersonation—remain largely unexplored. In the domain of cyberbullying, only two papers (Ferrer et al. 2024; Ahmed et al. 2022) have utilized LLMs, while in hate speech detection, six papers (Wullach et al. 2022; Pan et al. 2024; Ramos et al. 2024; Chhikara et al. 2024; García-Díaz, Pan, et al. 2023; Ahmed et al. 2022) are based on LLMs.

These findings lead us to two conclusions. First, hate speech is the form of cyber abuse most influenced by LLMs. This may, however, be partially attributed to the huge focus on hate speech within the research community, making it the most extensively studied area in cyber abuse detection. Second, the application of LLMs in cyber abuse detection is still in its infancy and warrants further exploration, especially given the potential of LLMs demonstrated in other fields.

7.5 | RQ5: Are There Specific Forms of Cyber Abuse That Remain Underexplored and Lack Adequate Detection Solutions?

To conclude this section, it is crucial to recognize the significant gaps in research concerning other kinds of cyber abuse as emotional abuse and doxing within the broader landscape of textual

cyber abuse detection and analysis. Despite the serious harm that these forms of cyber abuse can inflict, they remain underexplored and are often overshadowed by a predominant focus on hate speech. While comprehensive studies on hate speech may touch upon related forms of abuse, there is a pressing need for more specialized and nuanced approaches tailored to address these specific issues.

The situation is further exacerbated in other forms of cyberbullying, such as dogpiling—an online harassment tactic where groups of individuals collectively target the same victim. This type of cyberbullying could benefit significantly from social network analysis and data mining techniques, yet it remains largely unexamined in the literature.

Similarly, in the domains of doxing and emotional or psychological abuse, the specific targets and contexts have received insufficient attention. Most existing research relies heavily on classification techniques; however, traditional unsupervised methods could be very valuable for analyzing psychological (Rosenbusch et al. 2021) and emotional abuse. Techniques such as similarity detection and correlation analysis could also enhance our understanding of doxing and its implications (Alonso-Fernandez et al. 2021).

These conclusions are further supported by studies with qualitative results and surveys in the fields of psychology and mental health. There is substantial research addressing issues such as emotional abuse (de la Villa Moral and Prieto 2022; Adebayo et al. 2021; Stephenson et al. 2018) and doxxing (Anderson and Wood 2021; Eckert and Metzger-Riftkin 2020), providing intriguing insights into how individuals are emotionally abused through social networks. Relevant insights are derived from (de la Villa Moral and Prieto 2022), where the authors analyzed 627 different cases, uncovering statistically significant and positive relationships between emotional dependence, cyber control, and online aggression victimization. Based on the findings from these studies, which confirm that emotional abuse and doxxing are prevalent in our daily interactions, we can conclude that the lack of artificial intelligence systems capable of addressing these nuanced forms of cyber abuse needs to be further explored in future research.

7.6 | Future Directions

To conclude our discussion, we find it essential to outline several future directions identified in our review. These aim to open new research avenues or further promote promising existing ones.

First, we found that the vast majority of research focuses predominantly on cyberbullying and hate speech. However, based on objective studies from fields such as psychology and mental health, which highlight the diverse forms of cyber abuse that individuals experience, we identified significant opportunities for improvement and exploration in this area. For instance, emotional cyber abuse, where the perpetrator is closely connected to the victim (e.g., partners or close acquaintances), requires unique approaches. In such cases, techniques can be applied within the victim's immediate network. These strategies differ substantially from those used to analyze hate speech, where, in many instances, the perpetrator is a stranger to the victim. This distinction highlights the need for tailored methods to address the specific dynamics of different types of cyber abuse.

Another promising future direction lies in addressing the challenge of real-time detection. Many of the studies we reviewed rely on static datasets, which, while yielding strong results during training, often face difficulties when applied to real-world scenarios. In environments such as high schools or social networks, there is a critical need for real-time systems capable of identifying and addressing issues as they arise, enabling interventions at the moment the problem begins to unfold (Kumar and Sachdeva 2022).

Regarding the intrinsic features of LMs and LLMs, there remains considerable room for research. One promising avenue is exploring LLMs that incorporate context beyond text alone, such as tone, platform-specific dynamics, and user behavior. Additionally, addressing and mitigating biases in LLMs is crucial to prevent misidentification or unequal treatment of abusive content based on demographics, written language, or regional variations. This challenge ties closely to the ethical implications of deploying LLMs in moderation systems—balancing user freedom with safety while ensuring transparency in AI-driven interventions. Furthermore, improving accuracy in detection remains a vital goal. This can be achieved by leveraging new models, combining multiple approaches, and continuously updating existing ones to better tackle the evolving landscape of cyber abuse.

Finally, although it falls outside the scope of our study, it is evident that much of today's cyber abuse takes the form of multimodal abuse. This includes voice messages, images, and AI-generated content, among others. Consequently, there is significant potential for research in the area of multimodal hate speech detection, which will require substantial advancements in the coming years to effectively address these challenges. This research area has been gaining significant attention in recent years (Arya et al. 2024; Thapa et al. 2024; Wang and Markov 2024).

8 | Conclusion

In this paper, we have reviewed the latest research on textual cyber abuse detection, with a particular focus on studies that leverage LMs and LLMs. This survey is the first to extend its scope beyond commonly studied areas such as hate speech and cyberbullying, encompassing a broader range of cyber abuse typologies.

We analyzed 87 papers, rigorously categorizing them according to different forms of cyber abuse and their harmful objectives. From this body of work, we identified critical insights into novel methodologies and the increasing role of language models, which have become essential in contemporary cyber abuse detection.

Our findings emphasize the widespread dominance of BERT-based approaches, particularly in the detection of hate speech, where these models consistently achieve state-of-the-art results.

BERT has emerged as the most extensively used model for text extraction and classification in the field.

Additionally, emerging LLMs like GPT, Gemini, and FLAN-T5 show significant promise, especially in low-resource settings and few-shot learning scenarios, where they excel at capturing unseen linguistic nuances. Generative models like GPT also hold potential for data augmentation, offering a dual benefit: the creation of synthetic datasets for training and mitigating issues of class imbalance.

Author Contributions

J. Angel Diaz-Garcia: conceptualization (equal), data curation (equal), formal analysis (equal), investigation (equal), methodology (equal), resources (equal), software (equal), validation (equal), visualization (equal), writing – original draft (equal). **Joao Paulo Carvalho:** conceptualization (equal), data curation (equal), formal analysis (equal), funding acquisition (lead), investigation (equal), methodology (equal), project administration (lead), resources (equal), software (equal), supervision (lead), validation (equal), visualization (equal), writing – review and editing (lead).

Acknowledgments

The research reported in this paper was supported by the DesinfoScan project: Grant TED2021-129402B-C21 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and FederaMed project: Grant PID2021-123960OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. Finally, the research reported in this paper is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309). This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020), and by the Recovery and Resilience Plan (RRP) and Next Generation EU European Funds through project C644865762-00000008 Accelerat.AI.

Consent

All authors have reviewed and approved the manuscript and consent to its publication.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Related WIREs Articles

Social network analysis: An overview

Role of sentiment analysis in social media security and analytics

References

Aarthi, B., and B. J. Chelliah. 2022. "Deep Recurrent Neural Network-Based Aquila Optimization-Based Online Shaming Emotion Analysis." *Concurrency and Computation: Practice and Experience* 34: e6882.

Aarthi, B., and B. J. Chelliah. 2023. "Hatdo: Hybrid Archimedes Tasmanian Devil Optimization CNN for Classifying Offensive Comments and Non-Offensive Comments." *Neural Computing and Applications* 35: 18395–18415.

Abarna, S., J. Sheeba, and S. Pradeep Devaneyan. 2023. "A Novel Ensemble Model for Identification and Classification of Cyber Harassment on Social Media Platform." *Journal of Intelligent & Fuzzy Systems* 45: 13–36.

Abualigah, L., D. Yousri, M. Abd Elaziz, A. A. Ewees, M. A. Al-Qaness, and A. H. Gandomi. 2021. "Aquila Optimizer: A Novel Meta-Heuristic Optimization Algorithm." *Computers & Industrial Engineering* 157: 107250.

Adebayo, O., O. Omojola, and K. Evbuoma. 2021. "Social Media and Emotional Abuse in Marriage: Evidence From Lagos State, Nigeria." *African Journal of Reproductive Health* 25: 68–79.

Ahmed, T., S. Ivan, M. Kabir, H. Mahmud, and K. Hasan. 2022. "Performance Analysis of Transformer-Based Architectures and Their Ensembles to Detect Trait-Based Cyberbullying." *Social Network Analysis and Mining* 12: 99.

Ahn, H., Y. Kim, J. Kim, and Y.-S. Han. 2024. "SharedCon: Implicit Hate Speech Detection Using Shared Semantics." In *Findings of the Association for Computational Linguistics: ACL 2024*, edited by L.-W. Ku, A. Martins, and V. Srikumar, 10444–10455. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.622.

Aitken, M., F. Perquier, B. Park, et al. 2025. "Trajectories of Parent Criticism Across Treatment for Youth Self-Harm." *Journal of Child Psychology and Psychiatry*.

Akar, E. 2025. "Exploring the Impact of Social Network Structures on Toxicity in Online Mental Health Communities." *Computers in Human Behavior* 165: 108542.

Albadi, N., M. Kurdi, and S. Mishra. 2018. "Are They Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere." In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 69–76. IEEE.

Albladi, A., M. Islam, A. Das, et al. 2025. "Hate Speech Detection Using Large Language Models: A Comprehensive Review." *IEEE Access* 13: 20871–20892.

Alemu, Y., H. Chen, C. Duan, D. Caulley, R. Arriaga, and E. Sezgin. 2023. "Detecting Clinically Relevant Emotional Distress and Functional Impairment in Children and Adolescents: Protocol for an Automated Speech Analysis Algorithm Development Study." *JMIR Research Protocols* 12: 46970. https://doi.org/10.2196/46970.

Alfurayj, H. S., S. L. Lutfi, and R. Perumal. 2024. "A Chained Deep Learning Model for Fine-Grained Cyberbullying Detection With Bystander Dynamics." *IEEE Access* 12: 105588–105604.

Al-Harigy, L. M., H. A. Al-Nuaim, N. Moradpoor, and Z. Tan. 2024. "Towards a Cyberbullying Detection Approach: Fine-Tuned Contrastive Self-Supervised Learning for Data Augmentation." *International Journal of Data Science and Analytics* 19: 22.

Ali, R., U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg. 2022. "Hate Speech Detection on Twitter Using Transfer Learning." *Computer Speech & Language* 74: 101365.

Alkomah, F., S. Salati, and X. Ma. 2022. "A New Hate Speech Detection System Based on Textual and Psychological Features." *International Journal of Advanced Computer Science and Applications* 13: 860–869.

Almaliki, M., A. M. Almars, I. Gad, and E.-S. Atlam. 2023. "Abmm: Arabic Bert-Mini Model for Hate-Speech Detection on Social Media." *Electronics* 12: 1048.

Alonso-Fernandez, F., N. M. S. Belvisi, K. Hernandez-Diaz, N. Muhammad, and J. Bigun. 2021. "Writer Identification Using Microblogging Texts for Social Media Forensics." *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3: 405–426.

Alrowais, F., A. A. Jamjoom, H. Karamti, et al. 2024. "Robertanet: Enhanced Roberta Transformer Based Model for Cyberbullying Detection With Glove Features." *IEEE Access* 12: 58950–58959.

Althobaiti, M. J. 2022. "Bert-Based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis." *International Journal of Advanced Computer Science and Applications* 13: 5.

Altinel, A. B., S. Sahin, M. Z. Gurbuz, and G. K. Baydogmus. 2024. "So-Hatred: A Novel Hybrid System for Turkish Hate Speech Detection in Social Media With Ensemble Deep Learning Improved by Bert and Clustered-Graph Networks." *IEEE Access* 12: 86252–86270.

Altobelli, T. 2010. "Cyber-Abuse—A New Worldwide Threat to Children's Rights." *Family Court Review* 48: 459–481.

Anderson, B., and M. A. Wood. 2021. *Doxxing: A Scoping Review and Typology*, 205–226. Emerald international handbook of technology-facilitated violence and abuse.

Anjum, A., and R. Katarya. 2024. "Hatedetector: Multilingual Technique for the Analysis and Detection of Online Hate Speech in Social Networks." *Multimedia Tools and Applications* 83: 48021–48048.

Ansari, G., P. Kaur, and C. Saxena. 2024. "Data Augmentation for Improving Explainability of Hate Speech Detection." *Arabian Journal for Science and Engineering* 49: 3609–3621.

Appel, M., and F. Prietzel. 2022. "The Detection of Political Deepfakes." *Journal of Computer-Mediated Communication* 27: zmac008.

Arcila-Calderón, C., J. J. Amores, P. Sánchez-Holgado, L. Vrysis, N. Vryzas, and M. Oller Alonso. 2022. "How to Detect Online Hate Towards Migrants and Refugees? Developing and Evaluating a Classifier of Racist and Xenophobic Hate Speech Using Shallow and Deep Learning." *Sustainability* 14: 13094.

Arshad, M. U., R. Ali, M. O. Beg, and W. Shahzad. 2023. "Uhated: Hate Speech Detection in Urdu Language Using Transfer Learning." *Language Resources and Evaluation* 57: 713–732.

Arteaga, C. B. 2024. "Deepfakes sexuales: impacto, prevención y perspectivas de g'enero en el entorno digital." *Miguel Hernández Communication Journal* 15: 229–244.

Arya, G., M. K. Hasan, A. Bagwari, et al. 2024. "Multimodal Hate Speech Detection in Memes Using Contrastive Language-Image Pre-Training." *IEEE Access* 12: 22359–22375.

Asif, M., M. Al-Razgan, Y. A. Ali, and L. Yunrong. 2024. "Graph Convolution Networks for Social Media Trolls Detection Use Deep Feature Extraction." *Journal of Cloud Computing* 13: 33.

Badri, N., F. Kboubi, and A. Habacha Chaibi. 2024. "Abusive and Hate Speech Classification in Arabic Text Using Pre-Trained Language Models and Data Augmentation." *ACM Transactions on Asian and Low-Resource Language Information Processing* 23: 1–28.

Bakar, J. A., N. Yusoff, N. H. Harun, M. M. Nadzir, and S. Omar. 2023. "Text Simplification Using Hybrid Semantic Compression and Support Vector Machine for Troll Threat Sentences." *International Journal of Advanced Computer Science and Applications* 14: 10.

Barari, S., C. Lucas, K. Munger, et al. 2021. Political Deepfake Videos Misinform the Public, but no More Than Other Fake Media. OSF Preprints, 13.

Basak, R., S. Sural, N. Ganguly, and S. K. Ghosh. 2019. "Online Public Shaming on Twitter: Detection, Analysis, and Mitigation." *IEEE Transactions on Computational Social Systems* 6: 208–220.

Bathaei, A., and D. Štreimikienė. 2023. "A Systematic Review of Agricultural Sustainability Indicators." *Agriculture* 13: 241.

Bauman, S., R. B. Toomey, and J. L. Walker. 2013. "Associations Among Bullying, Cyberbullying, and Suicide in High School Students." *Journal of Adolescence* 36: 341–350.

Bedi, P., and C. Sharma. 2016. "Community Detection in Social Networks." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 6: 115–135.

Been Choi, E., J. Kim, D. Jeong, E. Park, and A. P. del Pobil. 2024. "Detecting Agro: Korean Trolling and Clickbaiting Behaviour in Online Environments." *Journal of Information Science* 50: 3–16.

Beres, N. A., J. Frommel, E. Reid, R. L. Mandryk, and M. Klarkowski. 2021. "Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. Association for Computing Machinery.

Bilal, M., A. Khan, S. Jan, S. Musa, and S. Ali. 2023. "Roman Urdu Hate Speech Detection Using Transformer-Based Model for Cyber Security Applications." *Sensors* 23: 3909.

Bishop, J. 2013. "The Art of Trolling Law Enforcement: A Review and Model for Implementing 'Flame Trolling' Legislation Enacted in Great Britain (1981–2012)." *International Review of Law, Computers & Technology* 27: 301–318.

Blancaflor, E., J. I. Garcia, F. D. Magno, and M. J. Vilar. 2024. "Deepfake Blackmailing on the Rise: The Burgeoning Posterity of Revenge Pornography in the Philippines." In *Proceedings of the 2024 9th International Conference on Intelligent Information Technology*, 295–301. Association for Computing Machinery.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Bottlapally, T., and L. Zhou. 2025. "Beyond Privacy: Understanding and Mitigating Doxing in the Digital Environment." In *Proceedings of the Hawaii International Conference on System Sciences*. HICSS Publishing.

Bullock, J., and M. Luengo-Oroz. 2019. Automated Speech Generation From un General Assembly Statements: Mapping Risks in AI Generated Texts arXiv preprint arXiv:1906.01946.

Calabrese, A., L. Neves, N. Shah, et al. 2024. "Explainability and Hate Speech: Structured Explanations Make Social Media Moderators Faster." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, edited by L.-W. Ku, A. Martins, and V. Srikumar, 398–408. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-short.38.

Caselli, T., V. Basile, J. Mitrović, and M. Granitzer. 2021. "HateBERT: Retraining BERT for Abusive Language Detection in English." In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, edited by A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, 17–25. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.woah-1.3.

Cassiani-Miranda, C. A., A. Campo-Arias, and C. C. Caballero-Domínguez. 2022. "Factors Associated With Cyberbullying Victimisation Among Colombian High-School Adolescents." *Journal of Child & Adolescent Trauma* 15: 27–36.

Centers for Disease Control and Prevention. 2009. Technology and Youth: Protecting Your Child From Electronic Aggression. Injury Prevention & Control: Violence Prevention.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "Smote: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–357.

Chen, S., J. Wang, and K. He. 2024. "Chinese Cyberbullying Detection Using Xlnet and Deep Bi-LSTM Hybrid Model." *Information* 15: 93.

Chen, Y., X. Zhang, L. Lu, et al. 2022. "Machine Learning Methods to Identify Predictors of Psychological Distress." *Processes* 10: 1030. https://doi.org/10.3390/pr10051030.

Chhikara, M., S. K. Malik, and V. Jain. 2024. "Identification of Social Network Automated Hate Speech Using Gltr With Bert and Gpt-2: A Novel Approach." *Journal of Information & Optimization Sciences* 45: 315–335.

Chung, H. W., L. Hou, S. Longpre, et al. 2024. "Scaling Instruction-Finetuned Language Models." *Journal of Machine Learning Research* 25: 1–53.

Cranmer, G. A., D. Linvill, H. Smith, et al. 2024. "Social Media Trolls as Faux Third-Party Agents of Image Repair: China's Disinformation Campaign and Statecraft in the Daryl Morey Affair." *Journal of Applied Communication Research* 52: 5–26.

de la Vega, L. G. M., and V. Ng. 2018. "Modeling Trolling in Social Media Conversations." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association.

de la Villa Moral, M., and P. Prieto. 2022. "Emotional Dependence and Partner Cyber Abuse Through Social Networks in Spanish University Students." *Revista Iberoamericana de Psicolog'ıa y Salud* 32: 15–27.

del Valle-Cano, G., L. Quijano-Sánchez, F. Liberatore, and J. Gómez. 2023. "Socialhaterbert: A Dichotomous Approach for Automatically Detecting Hate Speech on Twitter Through Textual Analysis and User Profiles." *Expert Systems With Applications* 216: 119446.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 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, Volume 1 (Long and Short Papers)*, edited by J. Burstein, C. Doran, and T. Solorio, 4171–4186. Min-nesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Divon, T., C. Are, and P. Briggs. 2025. "Platform Gaslighting: A User-Centric Insight Into Social Media Corporate Communications of Content Moderation." *Platforms & Society* 2: 29768624241303109.

Divya, V., and A. U. Mirza. 2024. "Transforming Content Creation: The Influence of Generative AI on a New Frontier." In *Exploring the Frontiers of Artificial Intelligence and Machine Learning Technologies*, 143. San International Scientific Publication.

Dobrobaba, M. 2022. "Deepfakes as a Threat to Human Rights." *Lex Russica* 75: 112–119.

Dobrojevic, M., L. Jovanovic, L. Babic, et al. 2024. "Cyberbullying Sexism Harassment Identification by Metaheurustics-Tuned Extreme Gradient Boosting." *Computers, Materials & Continua* 80: 4997–5027.

Dunn, S. 2024. "Legal Definitions of Intimate Images in the Age of Sexual Deepfakes and Generative AI." *McGill Law Journal* 69.

Eckert, S., and J. Metzger-Riftkin. 2020. "Doxxing." In *The International Encyclopedia of Gender, Media, and Communication*, 1–5. John Wiley & Sons, Inc.

Ezzeddine, F., O. Ayoub, S. Giordano, et al. 2023. "Exposing Influence Campaigns in the Age of LLMS: A Behavioral-Based Ai Approach to Detecting State-Sponsored Trolls." *EPJ Data Science* 12: 46.

Faria, F. T. J., L. H. Baniata, and S. Kang. 2024. "Investigating the Predominance of Large Language Mod-Els in Low-Resource Bangla Language Over Transformer Models for Hate Speech Detection: A Comparative Analysis." *Mathematics* 12: 3687.

Ferrer, R., K. Ali, and C. Hughes. 2024. "Using Ai-Based Virtual Companions to Assist Adolescents With Autism in Recognizing and Addressing Cyberbullying." *Sensors* 24: 3875.

Firmino, A. A., C. de Souza Baptista, and A. C. de Paiva. 2024. "Improving Hate Speech Detection Using Cross-Lingual Learning." *Expert Systems With Applications* 235: 121115.

Fisher, B. W., J. H. Gardella, and A. R. Teurbe-Tolon. 2016. "Peer Cybervictimization Among Adolescents and the Associated Internalizing and Externalizing Problems: A Meta-Analysis." *Journal of Youth and Adolescence* 45: 1727–1743.

Florido-Benítez, L. 2025. "The Role of Cybersecurity as a Preventive Measure in Digital Tourism and Travel: A Systematic Literature Review." *Discover Computing* 28: 28.

Fortuna, P., and S. Nunes. 2018. "A Survey on Automatic Detection of Hate Speech in Text." *ACM Computing Surveys* 51: 1–30.

Franks, M. A., and A. E. Waldman. 2018. "Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions." *Maryland Law Review* 78: 892.

García-Díaz, J. A., S. M. Jiménez-Zafra, M. A. García-Cumbreras, and R. Valencia-García. 2023. "Evaluating Feature Combination Strategies for Hate-Speech Detection in Spanish Using Linguistic Features and Transformers." *Complex & Intelligent Systems* 9: 2893–2914.

García-Díaz, J. A., R. Pan, and R. Valencia-García. 2023. "Leveraging Zero and Few-Shot Learning for Enhanced Model Generality in Hate Speech Detection in Spanish and English." *Mathematics* 11: 5004.

Garg, P., M. Sharma, and P. Kumar. 2024. "Improving Hate Speech Classification Through Ensemble Learning and Explainable Ai Techniques." *Arabian Journal for Science and Engineering*: 1–14.

Gharawi, M. A., A. Badawy, D. E. Ramadan, and S. Elsayed. 2021. "Social Media Impersonation in the Virtual World." *Al-Hikmah International Journal of Islamic Studies and Human Sciences* 4: 57–65.

Ghosh, K., and A. Senapati. 2024. "Hate Speech Detection in Low-Resourced Indian Languages: An Analysis of Transformer-Based Monolingual and Multilingual Models With Cross-Lingual Experiments." *Natural Language Processing* 31, no. 2: 1.

Gupta, S., M. Kaur, and S. Lakra. 2021. "Bert-bu12 Hate Speech Detection Using Bidirectional Encoder-Decoder." *International Journal of System Dynamics Applications* 11: 1–16.

Hamm, M. P., A. S. Newton, A. Chisholm, et al. 2015. "Prevalence and Effect of Cyberbullying on Children and Young People: A Scoping Review of Social Media Studies." *JAMA Pediatrics* 169: 770–777.

He, B., C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar. 2021. "Racism Is a Virus: Anti-Asian Hate and Counterspeech in Social Media During the Covid-19 Crisis." In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 90–94. IEEE.

He, P., X. Liu, J. Gao, and W. Chen. 2020. Deberta: Decoding-Enhanced Bert With Disentangled Attention arXiv preprint arXiv:2006.03654.

Herrera Montano, I., J. J. García Aranda, J. Ramos Diaz, S. Molina Cardín, I. De la Torre Díez, and J. J. Rodrigues. 2022. "Survey of Techniques on Data Leakage Protection and Methods to Address the Insider Threat." *Cluster Computing* 25: 4289–4302.

Hossain, E., O. Sharif, M. M. Hoque, M. A. A. Dewan, N. Siddique, and M. A. Hossain. 2022. "Identification of Multilingual Offense and Troll From Social Media Memes Using Weighted Ensemble of Multimodal Features." *Journal of King Saud University, Computer and Information Sciences* 34: 6605–6623.

Huang, W., W. Tang, H. Jiang, J. Luo, and Y. Zhang. 2021. "Stop Deceiving! An Effective Defense Scheme Against Voice Impersonation Attacks on Smart Devices." *IEEE Internet of Things Journal* 9: 5304–5314.

Huang, Z. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging arXiv preprint arXiv:1508.01991.

Husain, F., and O. Uzuner. 2022. "Investigating the Effect of Preprocessing Arabic Text on Offensive Language and Hate Speech Detection." *ACM Transactions on Asian and Low-Resource Language Information Processing* 21: 1–20.

Iyer, R., M. Nedeljkovic, and D. Meyer. 2022. "Using Vocal Characteristics to Classify Psychological Distress in Adult Helpline Callers: Retrospective Observational Study." *JMIR Formative Research* 6: e42249. https://doi.org/10.2196/42249.

Jang, J., S. Yoon, G. Son, M. Kang, J. Y. Choeh, and K.-H. Choi. 2022. "Predicting Personality and Psychological Distress Using Natural Language Processing: A Study Protocol." *Frontiers in Psychology* 13: 865541. https://doi.org/10.3389/fpsyg.2022.865541.

Jiang, A. Q., A. Sablayrolles, A. Mensch, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Karayi˘git, H., A. Akdagli, and C. İ. Aci. 2022. "Homophobic and Hate Speech Detection Using Multilingual-Bert Model on Turkish Social Media." *Information Technology and Control* 51: 356–375.

Karimi, Y., A. Squicciarini, and S. Wilson. 2022. "Automated Detection of Doxing on Twitter." *Proceedings of the ACM on Human-Computer Interaction* 6: 1–24.

Keya, A. J., M. M. Kabir, N. J. Shammey, M. F. Mridha, M. R. Islam, and Y. Watanobe. 2023. "G-Bert: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media." *IEEE Access* 11: 79697–79709.

Khafajeh, H. 2024. "Cyberbullying Detection in Social Networks Using Deep Learning." *International Arab Journal of Information Technology* 21, no. 6: 1054–1063.

Khan, S., M. Fazil, V. K. Sejwal, et al. 2022. "Bichat: Bilstm With Deep CNN and Hierarchical Attention for Hate Speech Detection." *Journal of King Saud University, Computer and Information Sciences* 34: 4335–4344.

Kim, J., S. Jin, S. Park, S. Park, and K. Han. 2024. "Label-Aware Hard Negative Sampling Strategies With Momentum Contrastive Learning for Implicit Hate Speech Detection." In *Findings of the Association for Computational Linguistics: ACL 2024*, edited by L.-W. Ku, A. Martins, and V. Srikumar, 16177–16188. Association for Computational Linguistics. URL. https://doi.org/10.18653/v1/2024.findings-acl.957.

Kim, J., B. Lee, and K. A. Sohn. 2022. "Why Is It Hate Speech? Masked Rationale Prediction for Explainable Hate Speech Detection." In *Proceedings of the 29th International Conference on Computational Linguistics*, edited by N. Calzolari, C. R. Huang, and H. Kim, 6644–6655. Republic of Korea: International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.577.

Kim, S., M. Kimber, M. H. Boyle, and K. Georgiades. 2019. "Sex Differences in the Association Between Cyberbullying Victimization and Mental Health, Substance Use, and Suicidal Ideation in Adolescents." *Canadian Journal of Psychiatry* 64: 126–135.

Koivukari, K., and P. Korpisaari. 2021. "Online Shaming-A New Challenge for Criminal Justice." In *Perspectives on Platform Regulation*, 473–488. Nomos Verlagsgesellschaft mbH & Co. KG.

Kowalski, R. M., G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner. 2014. "Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research Among Youth." *Psychological Bulletin* 140: 1073–1137.

Kumar, A., and N. Sachdeva. 2022. "Multi-Input Integrative Learning Using Deep Neural Networks and Transfer Learning for Cyberbullying Detection in Real-Time Code-Mix Data." *Multimedia Systems* 28: 2027–2041.

Kumar, P. 2024. "Large Language Models (Llms): Survey, Technical Frameworks, and Future Challenges." *Artificial Intelligence Review* 57: 260.

Kumar, Y., K. Huang, A. Perez, et al. 2024. "Bias and Cyberbullying Detection and Data Generation Using Transformer Artificial Intelligence Models and Top Large Language Models." *Electronics* 13: 3431.

Kwan, I., K. Dickson, M. Richardson, et al. 2020. "Cyberbullying and Children and Young People's Mental Health: A Systematic Map of Systematic Reviews." *Cyberpsychology, Behavior, and Social Networking* 23: 72–82.

Li, L., W. Zhang, and M. Barni. 2022. "Covert Task Embedding: Turning a Dnn Into an Insider Agent Leaking out Private Information." *IEEE Transactions on Neural Networks and Learning Systems* 35, no. 7: 10159–10166.

Li, T., Z. Zeng, Q. Li, and S. Sun. 2024. "Integrating Gin-Based Multimodal Feature Transformation and Multi-Feature Combination Voting for Irony-Aware Cyberbullying Detection." *Information Processing & Management* 61: 103651.

Li, Y., Y. Huang, M. E. Ildiz, A. S. Rawat, and S. Oymak. 2024. "Mechanics of Next Token Prediction With Self-Attention." In *International Conference on Artificial Intelligence and Statistics*, 685–693. PMLR.

Liu, C., Z. Liu, and G. Yuan. 2020. "The Longitudinal Influence of Cyberbullying Victimization on Depression and Posttraumatic Stress Symptoms: The Mediation Role of Rumination." *Archives of Psychiatric Nursing* 34: 206–210.

Liu, X., F. Shen, J. Zhao, and C. Nie. 2024. "Eap: An Effective Black-Box Impersonation Adversarial Patch Attack Method on Face Recognition in the Physical World." *Neurocomputing* 580: 127517.

Liu, Y., M. Ott, N. Goyal, et al. 2019. Roberta: A Robustly Optimized Bert Pretraining Approach. arXiv preprint arXiv:1907.11692.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS 2017*, 4768–4777. Curran Associates Inc.

Luo, L., T. Yang, X. Zheng, et al. 2022. "Altered Centromedial Amygdala Functional Connectivity in Adults Is Associated With Childhood Emotional Abuse and Predicts Levels of Depression and Anxiety." *Journal of Affective Disorders* 303: 148–154. https://doi.org/10.1016/j.jad.2022.02.023.

MacDermott, Á., M. Motylinski, F. Iqbal, K. Stamp, M. Hussain, and A. Marrington. 2022. "Using Deep Learning to Detect Social Media 'Trolls'." *Forensic Science International: Digital Investigation* 43: 301446.

Madhavi, M., S. Agal, N. D. Odedra, et al. 2024. "Elevating Offensive Language Detection: Cnn-Gru and Bert for Enhanced Hate Speech Identification." *International Journal of Advanced Computer Science and Applications* 15, no. 5: 1164–1172.

Mahdi, M. A., S. M. Fati, M. A. G. Hazber, S. Ahamad, and S. A. Saad. 2024. "Enhancing Ara-Bic Cyberbullying Detection With End-To-End Transformer Model." *Computer Modeling in Engineering and Sciences* 141, no. 2: 1651–1671. https://doi.org/10.32604/cmes.2024.052291.

Mahmud, T., M. Ptaszynski, and F. Masui. 2024. "Exhaustive Study Into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts." *Electronics* 13: 1677.

Maity, K., A. Kumar, and S. Saha. 2022. "A Multitask Multimodal Framework for Sentiment and Emotion-Aided Cyberbullying Detection." *IEEE Internet Computing* 26: 68–78.

Maity, K., A. Poornash, S. Bhattacharya, et al. 2024. "Hatethaisent: Sentiment-Aided Hate Speech Detection in Thai Language." *IEEE Transactions on Computational Social Systems* 11, no. 5: 5714–5727.

Malik, M. S. I., A. Nawaz, and M. M. Jamjoom. 2024. "Hate Speech and Target Community Detection in Nastaliq Urdu Using Transfer Learning Techniques." *IEEE Access* 12: 116875–116890.

Marcondes, F. S., J. J. Almeida, and P. Novais. 2024. "An Exploratory Design Science Research on Troll Factories." *Integrated Computer-Aided Engineering* 31, no. 1: 95–115.

Mathew, B., P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. 2021. "Hatexplain: A Benchmark Dataset for Explainable Hate Speech Detection." *Proceedings of the AAAI Conference on Artificial Intelligence* 35: 14867–14875.

Mazari, A. C., N. Boudoukhani, and A. Djeffal. 2024. "Bert-Based Ensemble Learning for Multi-Aspect Hate Speech Detection." *Cluster Computing* 27: 325–339.

McLoughlin, C., and J. Burgess. 2009. "Texting, Sexting and Social Networking Among Australian Youth and the Need for Cyber Safety Education." In *Proceedings of AARE International Education Research Conference*. Australian Association for Research in Education.

Mehta, H., and K. Passi. 2022. "Social Media Hate Speech Detection Using Explainable Artificial Intelligence (Xai)." *Algorithms* 15: 291.

Mewada, A., and R. K. Dewang. 2024. "Cipf: Identifying Fake Profiles on Social Media Using a Cnn-Based Communal Influence Propagation Framework." *Multimedia Tools and Applications* 83: 29419–29454.

Miao, W., X. Zhao, Y. Zhang, S. Chen, X. Li, and Q. Li. 2024. "A Deep Learning-Based Method for Preventing Data Leakage in Electric Power Industrial Internet of Things Business Data Interactions." *Sensors* 24: 4069.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, and PRISMA Group. 2010. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The Prisma Statement." *International Journal of Surgery* 8: 336–341.

Mollas, I., Z. Chrysopoulou, S. Karlos, and G. Tsoumakas. 2022. "Ethos: A Multi-Label Hate Speech Detection Dataset." *Complex & Intelligent Systems* 8: 4663–4678.

Montesinos-Cánovas, E., F. García-Sánchez, J. A. García-Díaz, G. Alcaraz-Mármol, and R. Valencia-García. 2023. "Spanish Hate-Speech Detection in Football." *Procesamiento Del Lenguaje Natural* 71: 15–27.

Moreno-Sandoval, L. G., A. Pomares-Quimbaya, S. A. Barbosa-Sierra, and L. M. Pantoja-Rojas. 2024. "Detection of Hate Speech, Racism and Misogyny in Digital Social Networks: Colombian Case Study." *Big Data and Cognitive Computing* 8: 113.

Muir, S. R., L. D. Roberts, L. Sheridan, and A. R. Coleman. 2023. "Examining the Role of Moral, Emotional, Behavioural, and Personality Factors in Predicting Online Shaming." *PLoS One* 18: e0279750.

Muneer, A., A. Alwadain, M. G. Ragab, and A. Alqushaibi. 2023. "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced Bert." *Information* 14: 467.

Nalawade, M. P. D., and D. Kulkarni. 2021. "A Survey on Identification of Online Public Shaming Using Machine Learning Framework." *International Journal of Advance Scientific Research and Engineering Trends* 6, no. 4: 2456–2774.

Nandi, A., K. Sarkar, A. Mallick, and A. De. 2024. "Combining Multiple Pre-Trained Models for Hate Speech Detection in Bengali, Marathi, and Hindi." *Multimedia Tools and Applications* 83: 1–25.

Nepomuceno, M. V., H. Rahemi, T. Cenesizoglu, and L. Charlin. 2023. "Should We Feed the Trolls? Using Marketer-Generated Content to Explain Average Toxicity and Product Usage." *Journal of Interactive Marketing* 58: 440–462.

Nguyen, D. Q., T. Vu, and A. T. Nguyen. 2020. "Bertweet: A Pre-Trained Language Model for English Tweets." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14. Association for Computational Linguistics.

Oatley, G., and B. Ewart. 2011. "Data Mining and Crime Analysis." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1: 147–153.

Ollagnier, A., E. Cabrio, and S. Villata. 2023. "Unsupervised Fine-Grained Hate Speech Target Community Detection and Characterisation on Social Media." *Social Network Analysis and Mining* 13: 58.

Pan, R., J. A. García-Díaz, and R. Valencia-García. 2024. "Comparing Fine-Tuning, Zero and Few-Shot Strategies With Large Language Models in Hate Speech Detection in English." *CMES Computer Modeling in Engineering and Sciences* 140, no. 3: 2849–2868.

Paul, S., S. Saha, and J. P. Singh. 2023. "Covid-19 and Cyberbullying: Deep Ensemble Model to Identify Cyberbullying From Code-Switched Languages During the Pandemic." *Multimedia Tools and Applications* 82: 8773–8789.

Paz, M. A., J. Montero-Díaz, and A. Moreno-Delgado. 2020. "Hate Speech: A Systematized Review." *SAGE Open* 10: 2158244020973022.

Penedo, G., Q. Malartic, D. Hesslow, et al. 2023. The Refinedweb Dataset for Falcon Llm: Outperforming Curated Corpora With Web Data, and Web Data Only arXiv preprint arXiv:2306.01116.

Pikuliak, M., M. Šimko, and M. Bieliková. 2021. "Cross-Lingual Learning for Text Processing: A Survey." *Expert Systems With Applications* 165: 113765.

Pookpanich, P., and T. Siriborvornratanakul. 2024. "Offensive Language and Hate Speech Detection Using Deep Learning in Football News Live Streaming Chat on Youtube in Thailand." *Social Network Analysis and Mining* 14: 18.

Pranesh, R. R., A. Shekhar, and A. Kumar. 2021. Towards Automatic Online Hate Speech Intervention Generation Using Pretrained Language Model.

Putra, C. D., and H.-C. Wang. 2024. "Semi-Meta-Supervised Hate Speech Detection." *Knowledge-Based Systems* 287: 111386.

Putri, S. D., M. P. D. Saragi, and M. P. Dinata Saragi. 2025. "The Impact of Parental Verbal Abuse on Adolescent Decision-Making in North Binjai, Indonesia." *Ascarya: Journal of Islamic Science, Culture, and Social Studies* 5: 13–26.

Radford, A., K. Narasimhan, T. Salimans, et al. 2018. Improving Language Understanding by Generative Pre-Training.

Raffel, C., N. Shazeer, A. Roberts, et al. 2020. "Exploring the Limits of Transfer Learning With a Unified Text-To-Text Transformer." *Journal of Machine Learning Research* 21: 5485–5551.

Rajeshwari, B. S., and I. Divya. 2025. "Combating Cyberbullying With Generative Ai: Large Language Models (Llms) in Cyberbullying Prevention." In *Combating Cyberbullying With Generative AI*, 329–364. IGI Global Scientific Publishing.

Rajput, G., N. S. Punn, S. K. Sonbhadra, and S. Agarwal. 2021. "Hate Speech Detection Using Static Bert Embeddings." In *Big Data Analytics: 9th International Conference, BDA 2021, Virtual Event, December 15–18, 2021, Proceedings 9*, 67–77. Springer.

Ramos, G., F. Batista, R. Ribeiro, et al. 2024. "Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts." *IEEE Access* 12: 101374–101389.

Razi, F., and N. Ejaz. 2024. "Multilingual Detection of Cyberbullying in Mixed Urdu, Roman Urdu, and English Social Media Conversations." *IEEE Access* 12: 105201–105210.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. *"Why Should i Trust You?" Explaining the Predictions of any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. Association for Computing Machinery.

Rini, R., and L. Cohen. 2022. "Deepfakes, Deep Harms." *Journal of Ethics and Social Philosophy* 22: 143.

Romero Moreno, F. 2024. "Generative Ai and Deepfakes: A Human Rights Approach to Tackling Harmful Content." *International Review of Law, Computers & Technology* 38, no. 3: 1.

Ronneberg, C. R., N. Lv, O. A. Ajilore, et al. 2024. "Study of a Pst-Trained Voice-Enabled Artificial Intelligence Counselor for Adults With Emotional Distress (Speac-2): Design and Methods." *Contemporary Clinical Trials* 142: 107574. https://doi.org/10.1016/j.cct.2024.107574.

Rosenbusch, H., F. Soldner, A. M. Evans, and M. Zeelenberg. 2021. "Supervised Machine Learning Meth-Ods in Psychology: A Practical Introduction With Annotated r Code." *Social and Personality Psychology Compass* 15: e12579.

Rousay, V. 2023. "Sexual Deepfakes and Image-Based Sexual Abuse: Victim-Survivor Experiences and Embodied Harms." Master's thesis, Harvard University.

Saini, H., H. Mehra, R. Rani, G. Jaiswal, A. Sharma, and A. Dev. 2023. "Enhancing Cyberbullying Detection: A Comparative Study of

Ensemble Cnn–Svm and Bert Models." *Social Network Analysis and Mining* 14, no. 1.

Salawu, S., Y. He, and J. Lumsden. 2017. "Approaches to Automated Detection of Cyberbullying: A Survey." *IEEE Transactions on Affective Computing* 11: 3–24.

Saleh, H., A. Alhothali, and K. Moria. 2023. "Detection of Hate Speech Using Bert and Hate Speech Word Embedding With Deep Model." *Applied Artificial Intelligence* 37: 2166719.

Samee, N. A., U. Khan, S. Khan, M. M. Jamjoom, M. Sharif, and D. H. Kim. 2023. "Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and Emotional Features for Cyberbullying Detection." *IEEE Access* 11: 124524–124554.

Sanh, V., A. Webson, C. Raffel, et al. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv preprint arXiv:2110.08207*.

Schillaci, Z. 2024. "Llm Adoption Trends and Associated Risks." In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, edited by Z. Schillaci, 121–128. Springer Nature Switzerland.

Schmidt, A., and M. Wiegand. 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*, 1–10. Association for Computational Linguistics.

Shan, G., W. Pu, J. B. Thatcher, and P. Roth. 2024. "How Doxing on Social Media Leads to Social Stigma and Perceived Dignity." In *Proceedings of the Hawaii International Conference on System Sciences (HICSS-57) 2*. University of Colorado Boulder. https://aisel.aisnet.org/hicss-57/cl/social_media/2.

Shukla, A. K. 2023. "Detecting Impersonation Episode Using Teaching Learning-Based Optimization and Support Vector Machine Techniques." *Expert Systems* 40: e13290.

Siddiqui, J. A., S. S. Yuhaniz, G. Mujtaba, S. A. Soomro, and Z. A. Mahar. 2024. "Fine-Grained Multilingual Hate Speech Detection Using Explainable Ai and Transformers." *IEEE Access* 12: 143177–143192.

Singh, P., N. K. Singh, S. Chand, et al. 2023. "Mbert-Gru Multilingual Deep Learning Framework for Hate Speech Detection in Social Media." *Journal of Intelligent & Fuzzy Systems* 44: 8177–8192.

Solo, A. M. 2025. "Preventing Deepfakes From Being Used for Impersonation and Defamation." In *Deepfakes and Their Impact on Business*, 267–284. IGI Global Scientific Publishing.

Sreelakshmi, K., B. Premjith, B. R. Chakravarthi, and K. Soman. 2024. "Detection of Hate Speech and Offensive Language Codemix Text in Dravidian Languages Using Cost-Sensitive Learning Approach." *IEEE Access* 12: 20064–20090.

Ștefăniță, O., and D.-M. Buf. 2021. "Hate Speech in Social Media and Its Effects on the Lgbt Community: A Review of the Current Research." *Romanian Journal of Communication and Public Relations* 23: 47–55.

Stephenson, V. L., B. M. Wickham, and N. M. Capezza. 2018. "Psychological Abuse in the Context of Social Media." *Violence and Gender* 5: 129–134.

Su, X., Y. Li, P. Branco, and D. Inkpen. 2023. "Ssl-Gan-Roberta: A Robust Semi-Supervised Model for Detecting Anti-Asian Covid-19 Hate Speech on Social Media." *Natural Language Engineering* 30: 1–20.

Sugarman, D. B., M. Nation, N. P. Yuan, G. P. Kuperminc, L. Hassoun Ayoub, and S. Hamby. 2018. "Hate and Violence: Addressing Discrimination Based on Race, Ethnicity, Religion, Sexual Orientation, and Gender Identity." *Psychology of Violence* 8: 649–656.

Suhas Bharadwaj, R., S. Kuzhalvaimozhi, and N. Vedavathi. 2022. "A Novel Multimodal Hybrid Classifier Based Cyberbullying Detection for Social Media Platform." In *Proceedings of the Computational Methods in Systems and Software*, 689–699. Springer.

Sultan, T., M. A. T. Rony, M. S. Islam, S. Aldosary, and W. El-Shafai. 2024. "Memesvita: A Novel Multimodal Fusion Technique for Troll Memes Identification." *IEEE Access* 12: 177811–177828.

Surani, M., and R. Mangrulkar. 2021. "Comparative Analysis of Deep Learning Techniques to Detect Online Public Shaming." In *ITM Web of Conferences*, vol. 40, 03030. EDP Sciences.

Suryawanshi, S., B. R. Chakravarthi, M. Arcan, and P. Buitelaar. 2023. "Trollswithopinion: A Taxonomy and Dataset for Predicting Domain-Specific Opinion Manipulation in Troll Memes." *Multimedia Tools and Applications* 82: 9137–9171.

Swed, O., S. Dassanayaka, and D. Volchenkov. 2024. "Keeping It Authentic: The Social Footprint of the Trolls' Network." *Social Network Analysis and Mining* 14: 38.

Tabassum, I., and V. Nunavath. 2024. "A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data." *Applied Sciences* 14: 12007.

Team, M. N. 2023. Introducing Mpt-7b: A New Standard for Open-Source, Commercially Usable LLMS.

Thanh Nguyen, L. 2024. "ViHateT5: Enhancing Hate Speech Detection in Vietnamese With a Unified Text-To-Text Transformer Model." In *Findings of the Association for Computational Linguistics: ACL 2024*, edited by L.-W. Ku, A. Martins, and V. Srikumar, 5948–5961. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.355.

Thapa, S., K. Rauniyar, F. A. Jafri, et al. 2024. "Extended Multimodal Hate Speech Event Detection During Russia-Ukraine Crisis-Shared Task at Case 2024." In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events From Text, CASE 2024*, 221–228. Association for Computational Linguistics.

Thulin, E. J., and J. E. Heinze. 2025. "Gaslighting in Teen Dating Violence: Links to Anxiety and Depression." *Journal of Interpersonal Violence*: 1–20.

Toliyat, A., S. I. Levitan, Z. Peng, and R. Etemadpour. 2022. "Asian Hate Speech Detection on Twitter During Covid-19." *Frontiers in Artificial Intelligence* 5: 932381.

Tonneau, M., P. Quinta De Castro, K. Lasri, et al. 2024. "NaijaHate: Evaluating Hate Speech Detection on Nigerian Twitter Using Representative Data." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by L.-W. Ku, A. Martins, and V. Srikumar, 9020–9040. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.488.

Touvron, H., L. Martin, K. Stone, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models arXiv preprint arXiv:2307.09288.

Umer, M., E. A. Alabdulqader, A. A. Alarfaj, L. Cascone, and M. Nappi. 2024. "Cyberbullying Detection Using Pca Extracted Glove Features and Robertanet Transformer Learning Model." *IEEE Transactions on Computational Social Systems*.

Uyheng, J., J. Moffitt, and K. M. Carley. 2022. "The Language and Targets of Online Trolling: A Psy-Cholinguistic Approach for Social Cybersecurity." *Information Processing & Management* 59: 103012.

van Huijstee, M., W. Nieuwenhuizen, M. Sanders, E. Masson, and P. van Boheemen. 2022. Harmful Behaviour Online: An Investigation of Harmful and Immoral Behaviour Online in The Netherlands.

Vandebosch, H., and K. Van Cleemput. 2009. "Cyberbullying Among Youngsters: Profiles of Bullies and Victims." *New Media & Society* 11: 1349–1371.

Wach, K., C. D. Duong, J. Ejdys, et al. 2023. "The Dark Side of Generative Artificial Intelligence: A Critical Analysis of Controversies and Risks of Chatgpt." *Entrepreneurial Business and Economics Review* 11: 7–30.

Wachs, S., M. Gámez-Guadix, and M. F. Wright. 2022. "Online Hate Speech Victimization and Depressive Symptoms Among Adolescents: The Protective Role of Resilience." *Cyberpsychology, Behavior, and Social Networking* 25: 416–423.

Wang, Y., and I. Markov. 2024. "Cltl@ Multimodal Hate Speech Event Detection 2024: The Winning Approach to Detecting Multimodal Hate Speech and Its Targets." In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events From Text (CASE 2024)*, 73–78. Association for Computational Linguistics.

Weber, M., D. Fu, Q. Anthony, et al. 2024. Redpajama: An Open Dataset for Training Large Language Models arXiv preprint arXiv:2411.12372.

Wullach, T., A. Adler, and E. Minkov. 2022. "Character-Level Hypernetworks for Hate Speech Detection." *Expert Systems With Applications* 205: 117571.

Wypych, M., and M. Bilewicz. 2022. "Psychological Toll of Hate Speech: The Role of Acculturation Stress in the Effects of Exposure to Ethnic Slurs on Mental Health Among Ukrainian Immigrants in Poland." *Cultural Diversity and Ethnic Minority Psychology* 30: 0522.

Xu, L., Y. tong, Q. Dong, et al. 2020. Cluener2020: Fine-Grained Named Entity Recognition Dataset and Benchmark for Chinese. arXiv:2001.04351. https://arxiv.org/abs/2001.04351.

Yadav, A., and V. Singh. 2024. "Hatefusion: Harnessing Attention-Based Techniques for Enhanced Filtering and Detection of Implicit Hate Speech." *IEEE Transactions on Computational Social Systems*.

Yadav, N., S. Masud, V. Goyal, M. S. Akhtar, and T. Chakraborty. 2024. "Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech." In *Findings of the Association for Computational Linguistics: ACL 2024*, edited by L.-W. Ku, A. Mar-Tins, and V. Srikumar, 13967–13983. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.831.

Yu, L., G. Mottola, C. N. Kieffer, et al. 2023. "Vulnerability of Older Adults to Government Impersonation Scams." *JAMA Network Open* 6: e2335319.

Yun, S., S. Kang, and H. Kim. 2023. "Bert-Based Logits Ensemble Model for Gender Bias and Hate Speech Detection." *Journal of Information Processing Systems* 19, no. 5: 641–651.

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)*. Association for Computational Linguistics.

Zampieri, M., S. Rosenthal, P. Nakov, A. Dmonte, and T. Ranasinghe. 2023. "Offenseval 2023: Offensive Language Identification in the Age of Large Language Models." *Natural Language Engineering* 29: 1416–1435. https://doi.org/10.1017/S1351324923000517.

Zarei, K., R. Farahbakhsh, N. Crespi, and G. Tyson. 2020. "Impersonation on Social Media: A Deep Neural Approach to Identify Ingenuine Content." In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 11–15. IEEE.

Zhang, M., J. He, T. Ji, and C.-T. Lu. 2024. "Don't Go to Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by L.-W. Ku, A. Martins, and V. Srikumar, 12073–12086. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.652.

Zhou, L., A. Caines, I. Pete, and A. Hutchings. 2023. "Automated Hate Speech Detection and Span Extraction in Underground Hacking and Extremist Forums." *Natural Language Engineering* 29: 1247–1274.