IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 86~96
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp86-96

A comprehensive survey of cyberbullying on social media: challenges, detection, and AI-based prevention

Ammar Odeh[1], Osama Alhaj Hassan[1], Anas Abu Taleb[1], Abobakr Aboshgifa[2], Nabil Belhaj[2]
[1]Department of Computer Science, King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan
[2]Department of Research, The Libyan Higher Technical Center for Training and Production, Tripoli, Libya

Article Info

Article history:

Received Feb 27, 2025
Revised Oct 30, 2025
Accepted Nov 8, 2025

Keywords:

Cyberbullying
Detection
Machine learning
Online harassment
Policy
Prevention
Social media

ABSTRACT

Cyberbullying is a pervasive issue in the digital landscape, particularly on social media platforms, where individuals engage in online harassment, intimidation, and abuse. Unlike traditional bullying, cyberbullying has a broader reach, anonymity, and persistence, making it a growing concern for mental health, social well-being, and online safety. This paper provides a comprehensive survey of cyberbullying trends, its psychological and social impacts, and the role of social media in amplifying the problem. It explores existing detection and prevention strategies, including artificial intelligence (AI)-driven approaches, policy frameworks, and platform-based moderation techniques. Furthermore, it discusses challenges in enforcement, the limitations of automated detection systems, and the need for improved legal measures. This paper uniquely contributes an integrated perspective on cyberbullying detection and prevention by synthesizing current research across psychological, sociocultural, and technical dimensions. It emphasizes underexplored gaps such as multilingual detection, real-time moderation, and cross-platform enforcement, and proposes a layered framework to guide future research and policy.

Corresponding Author:

Ammar Odeh
Department of Computer Science, King Hussein School of Computing Sciences
Princess Sumaya University for Technology
Amman 1196, Jordan
Email: a.odeh@psut.edu.jo

1. INTRODUCTION

Cyberbullying is a form of harassment that occurs in digital environments, particularly on social media platforms, where individuals use electronic communication to intimidate, threaten, or harm others [1]. Unlike traditional bullying, which is often confined to specific locations such as schools or workplaces, cyberbullying can occur at any time and reach a global audience. The anonymity provided by social media allows perpetrators to act without immediate repercussions, making cyberbullying a persistent and evolving issue. This harmful behavior can take many forms, including harassment, impersonation, public shaming, and the spread of false or private information, all of which can have devastating psychological and social consequences for victims [2].

Social media platforms play a significant role in the prevalence of cyberbullying due to their accessibility, anonymity, and wide reach. Features such as anonymous accounts, direct messaging, comment sections, and viral sharing contribute to the rapid spread of harmful content. Many platforms rely on automated content moderation and user-reporting systems, often failing to promptly remove abusive content. Additionally, social media algorithms prioritize engaging content, sometimes amplifying controversial or harmful posts that encourage online hostility. The 24/7 nature of digital interactions also means that victims have little to no escape from online harassment, exacerbating the emotional toll of cyberbullying. Despite efforts by platforms to introduce safety measures such as content filters, artificial intelligence (AI) moderation, and user blocking tools, cyberbullying remains a major challenge, affecting millions of users worldwide [3], [4].

The growing scale of cyberbullying is reflected in global statistics, highlighting the urgent need for effective intervention. According to a 2023 UNICEF report, approximately 37% of young people aged 12-17 worldwide have experienced cyberbullying, with many cases going unreported due to fear or lack of awareness. A 2022 survey by the Pew Research Center found that 59% of U.S. teenagers had faced online harassment, with the most common forms being name-calling, spreading false rumors, and receiving explicit messages. Similarly, a study conducted by the European Commission in 2022 reported that one in five adolescents in Europe had encountered cyberbullying at least once in the past year. The rise in cyberbullying-related suicides, particularly among teenagers, underscores the severity of the issue and the urgent need for improved protective measures. In developing countries, increasing internet penetration and mobile device usage have led to a surge in cyberbullying cases, often without the necessary legal frameworks to regulate online abuse effectively [5].

Figure 1 presents a structured classification of cyberbullying on social media, dividing the topic into three main categories: types, platforms, and detection-prevention methods. The types category includes common forms of cyberbullying such as name-calling, spreading false rumors, and issuing threats. The platform category highlights popular social media channels where such incidents frequently occur, such as Facebook and X (formerly Twitter). The detection-prevention category encompasses approaches to address cyberbullying, including AI-powered natural language processing (NLP) techniques for automated detection and implementing platform policies to curb harmful behavior. This classification framework emphasizes the multifaceted nature of the cyberbullying problem, covering its manifestations, digital environments, and intervention strategies.

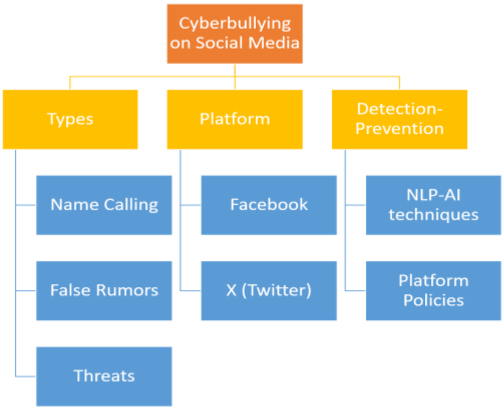


Figure 1. Classification of cyberbullying on social media: types, platforms, and detection-prevention approaches

The impact of cyberbullying extends beyond the digital realm, affecting victims psychologically, socially, and legally. On a psychological level, victims often suffer from stress, anxiety, depression, and even suicidal thoughts due to the persistent nature of online harassment. Long-term exposure to cyberbullying has been linked to post-traumatic stress disorder (PTSD) and other severe mental health disorders. A 2023 study from the American Psychological Association (APA) found that teenagers subjected to cyberbullying are at a significantly higher risk of self-harm and suicidal ideation. Socially, cyberbullying can lead to withdrawal from social interactions, academic decline, and reputational damage, impacting victims' future relationships and career prospects [6]. It also contributes to a toxic online culture, discouraging open and constructive interactions on social media. Legally, while many countries have introduced cyberbullying laws, enforcement remains a challenge, particularly in cases where offenders operate anonymously or across multiple jurisdictions. Regulations such as the European Union's General Data Protection Regulation (GDPR) and the U.S. Children's Online Privacy Protection Act (COPPA) impose obligations on platforms to protect users. Yet, loopholes persist, allowing harmful behavior to continue. In some regions, cyberbullying is considered a cybercrime, leading to legal actions such as fines, restraining orders, or even imprisonment for perpetrators [7].

Despite adopting AI-powered moderation tools and user-reporting systems by major social media platforms, current cyberbullying detection systems face several limitations. These systems often struggle with understanding online language's nuanced and context-dependent nature, including sarcasm, coded expressions, and cultural references. Additionally, most models are trained on limited, often monolingual datasets, making them less effective in detecting cyberbullying across diverse linguistic and cultural settings. The dynamic evolution of cyberbullying tactics—such as emojis, images, memes, or slang—also challenges static AI models. Moreover, false positives and negatives remain a concern, where benign content may be flagged while subtle abuse goes undetected. These technical constraints, the difficulty of real-time processing, and the lack of standardized cross-platform enforcement mechanisms highlight the urgent need for more adaptive, context-aware, and multilingual AI solutions. Unlike previous surveys, which often focus exclusively on either psychological aspects or specific AI techniques, our work provides a holistic view by integrating psychological, legal, and technical dimensions of cyberbullying. We also synthesize recent advancements (2023–2025), including the role of large language models (LLMs), blockchain-based accountability systems, and explainable AI approaches, which have not been comprehensively reviewed in prior studies. Furthermore, this survey highlights cross-platform challenges and future directions for multilingual and multimodal detection, offering an integrative framework for researchers and policymakers. This paper addresses key research gaps that existing surveys have not adequately covered, including the need for multilingual detection, real-time moderation, and cross-platform enforcement of cyberbullying policies. Unlike previous reviews that focus narrowly on either psychological impacts or specific machine learning methods, this survey integrates psychological, technical, and legal perspectives. It provides a unified framework that combines content analysis, social network context, and sentiment-based interpretation.

2. METHOD

Cyberbullying takes many forms, each with unique characteristics and impacts. Unlike traditional bullying, which occurs in physical spaces such as schools or workplaces, cyberbullying leverages digital platforms to reach a broader audience. The anonymity of the internet, the speed at which harmful content spreads, and the difficulty of removing digital traces make cyberbullying particularly harmful. Understanding the different types of cyberbullying and its characteristics is crucial for developing effective prevention and intervention strategies [8], [9].

One of the most common types of cyberbullying is harassment, which involves repeated, aggressive, and unwanted messages intended to intimidate or distress the victim. This can include direct threats, offensive comments, or persistent negative interactions across multiple platforms. Impersonation is another form where a perpetrator creates a fake account to pose as someone else, often spreading false information, damaging reputations, or manipulating others [10]. This tactic is frequently used to deceive the victim's friends, family, or colleagues. Outing and doxxing involve publicly revealing private or sensitive information about an individual without their consent. While outing focuses on exposing personal secrets or embarrassing details, doxxing goes further by sharing information such as phone numbers, addresses, or workplace details, leading to potential harassment or real-world threats. Cyberstalking refers to persistent and intrusive monitoring, threats, or intimidation through online platforms, which can cause significant psychological distress and, in severe cases, escalate to offline harm. Lastly, trolling involves intentionally provoking or upsetting others by posting inflammatory, offensive, or deceptive content. While some trolls claim their actions are humorous, their behavior often leads to severe emotional distress and can encourage further cyberbullying [11].

Cyberbullying differs significantly from traditional bullying in various ways. Traditional bullying typically takes place in face-to-face environments such as schools, offices, or neighborhoods, whereas cyberbullying occurs online and can persist around the clock. Anonymity plays a major role in cyberbullying, as perpetrators often hide behind fake profiles, making it difficult for victims to identify them or take legal action. Additionally, cyberbullying leaves a lasting digital footprint, meaning harmful content can be reshared or re-exposed long after the initial incident, unlike traditional bullying, which is usually confined to a specific moment in time. Another key difference is the public nature of cyberbullying—negative comments, false rumors, or edited images can quickly go viral, exposing the victim to widespread humiliation. Moreover, the absence of physical presence in cyberbullying makes it harder for parents, teachers, and authorities to detect and address the issue compared to traditional bullying, where visible signs of distress or harm are often more apparent [12], [13].

The victims and perpetrators of cyberbullying come from various age groups and backgrounds. Studies indicate that teenagers and young adults are among the most affected due to their high engagement with social media and online communication. However, children and even adults are not immune, as workplace cyberbullying and online harassment continue to rise. In terms of gender differences, research suggests that females are more likely to face appearance-based shaming, gossip, or social exclusion. In contrast, males are more frequently subjected to direct threats, aggressive messages, or physical intimidation online. Additionally, individuals in highly visible professions, such as public figures, influencers, and journalists, often experience targeted harassment campaigns, trolling, and reputational attacks due to their online presence [14].

3. RESULTS AND DISCUSSION

Imam and Naz [15] examined the legal and societal implications of cyberbullying in Pakistan. The study highlighted the psychological effects, particularly on women, and analyzed the challenges in enforcing the Prevention of Electronic Crimes Act (PECA) 2016. The authors compared Pakistan's legal framework to global trends and advocated for international cooperation, AI-driven detection mechanisms, and public awareness campaigns to combat cyberbullying.

Philipo *et al.* [16] systematically reviewed cyberbullying detection approaches, categorizing methods into machine learning, deep learning, traditional models, and LLMs. The study identified issues such as data imbalances, multilingual detection biases, and evolving cyberbullying strategies. The authors emphasized the importance of AI-driven tools for real-time and multilingual cyberbullying detection.

Samala *et al.* [17] explored the role of social media in education, focusing on its impact on digital-native learners. The study highlighted opportunities for enhanced learning and discussed challenges such as cyberbullying, privacy breaches, and identity theft. The authors advocated for digital literacy education, responsible social media use, and policy interventions to mitigate risks while maximizing educational benefits.

Saeid *et al.* [3] investigated machine learning-based approaches for cyberbullying detection using NLP and text classification. The study compared traditional machine learning models like support vector machines (SVM) and decision trees with deep learning approaches, including bidirectional encoder representations from transformers (BERT)-based transformer. The findings indicated that bidirectional long short-term memory (BiLSTM) networks provided higher accuracy, but their computational demands required balancing efficiency and scalability.

Talpur *et al.* [18] reviewed cyberbullying detection trends, particularly in online social networks. The study discussed various types of cyberbullying, such as harassment and cyberstalking, and analyzed detection challenges, including evolving language patterns and anonymized abuse. The authors explored supervised learning techniques and emphasized the need for improved multilingual and context-aware detection mechanisms.

Livingstone *et al.* [19] examined the global trends and impacts of cyberbullying, emphasizing the rapid digital transformation affecting children's online interactions. The study explored how cyberbullying has evolved, often complementing rather than replacing traditional bullying. The findings suggested that while there is evidence of rising cyberbullying cases, this may be attributed to increased awareness and self-reporting rather than an absolute increase in incidents. The authors called for international efforts in policy-making, education, and digital literacy to mitigate the risks posed by cyberbullying.

Chen and Zolkepli [20] conducted a bibliometric analysis of cyberbullying research, specifically focusing on bystanders' roles in intervention. The study utilized citation mapping and network visualization tools to highlight the dominant research themes in cyberbullying studies from 2007 to 2024. The results showed that most research has focused on the psychological impact of cyberbullying on victims, with limited studies addressing bystanders' reactions. The authors advocated for more interdisciplinary research on bystander intervention strategies, emphasizing that supportive peer actions could significantly mitigate cyberbullying consequences.

Johanis *et al.* [21] analyzed cyberbullying trends in Malaysia, exploring the sociocultural factors influencing digital aggression. The study highlighted that increased smartphone penetration and social media usage have contributed to a surge in cyberbullying cases. The authors found that cyberbullying in Malaysia is not limited to specific demographics, with victims spanning different age groups and backgrounds. They called for stricter regulatory measures, improved parental guidance, and enhanced awareness campaigns to curb the rise of cyberbullying in digital spaces.

Navarro *et al.* [22] investigated the role of social media in shaping youth delinquency, emphasizing its impact on identity formation and peer influence. The study found that exposure to online challenges, cyberbullying, and digital peer pressure contributes to an increase in risky behaviors among adolescents. The authors suggested that while social media can foster positive community engagement, it also exacerbates social comparisons and mental health issues, making youth more susceptible to negative influences.

Rejeb *et al.* [23] conducted a large-scale bibliometric study analyzing the evolution of cyberbullying research over the past two decades. Using co-word and main path analyses, the study identified key themes, including digital literacy, psychological effects, and intervention strategies. The findings suggested that while early research focused on victimization and psychological impacts, recent studies have increasingly explored technological solutions like AI-driven detection systems and real-time content moderation.

Slanbekova *et al.* [24] extensively reviewed cyberbullying, synthesizing research from various disciplines to highlight its prevalence, psychological impact, and international response strategies. The study examined cultural variations in cyberbullying manifestations and the effectiveness of prevention measures in different regions. The authors emphasized the need for tailored policies considering cultural contexts and advocated for a combination of legal frameworks, digital literacy programs, and technological interventions to address the issue effectively.

Table 1 presents a synthesized overview of eleven key studies addressing various aspects of cyberbullying, ranging from legal and sociocultural analyses to advanced machine learning and deep learning detection approaches. The reviewed works collectively emphasize the multifaceted nature of the cyberbullying problem, including its psychological, legal, educational, and technological dimensions. While several studies [16], [18], [19] explored AI-driven detection mechanisms—ranging from traditional machine learning classifiers to BERT-based transformers and BiLSTM networks—others focused on policy recommendations, public awareness, and cross-cultural prevention strategies. Despite their contributions, most studies faced notable limitations such as the absence of practical implementation, high computational demands, scalability challenges, and insufficient multilingual or context-aware detection capabilities. This underscores the need for integrative solutions that combine robust technological approaches with comprehensive legal and educational frameworks.

As shown in Table 2, prior works vary in scope: some focus on broad surveys without empirical evaluation, while others present model-based studies that achieve strong results but lack scalability and multilingual support. Deep learning models (e.g., BiLSTM, BERT) demonstrate higher accuracy but require large datasets and computational resources, making real-time moderation difficult. Conversely, traditional machine learning techniques are lightweight but perform poorly with nuanced language such as sarcasm or code-switching. Reviews focusing on sociocultural or bibliometric aspects provide useful context but fail to integrate technical depth. These gaps underscore the need for a unified perspective that combines technical, sociocultural, and legal dimensions, which is the contribution of this paper.

Table 1. Summary of selected cyberbullying studies with key approaches, achievements, and limitations

Ref	Approach	Achievement	Limitation
[15]	Legal and societal analysis	Highlighted psychological effects, legal gaps, and the need for AI detection	No technical model implemented
[16]	Review of machine learning, deep learning, LLMs	Classified methods, identified biases, and stressed multilingual detection	No new algorithm proposed
[17]	Education and policy	Linked social media to learning; promoted digital literacy	No technical detection methods
[18]	Machine learning and deep learning (SVM, BERT, BiLSTM)	BiLSTM achieved the highest accuracy in detection	High computational cost
[19]	Supervised learning review	Detailed types and challenges call for context-aware systems	No practical evaluation
[20]	Global trend analysis	Showed cyberbullying complements traditional bullying; urged global policy	No specific tech solutions
[21]	Bystander role analysis	Highlighted the lack of research on bystander actions	A few practical strategies
[22]	Sociocultural study (Malaysia)	Found cyberbullying across demographics; urged regulation	No AI framework
[23]	Youth delinquency study	Linked online peer influence to risky behaviors	No mitigation technology
[24]	Bibliometric analysis	Tracked research trends; noted AI detection growth	Descriptive only
[24]	Cross-cultural review	Emphasized tailored policies and tech interventions	No model validation

Table 2. Comparative summary of representative cyberbullying detection studies

Ref	Technique/model	Dataset(s)	Reported outcome	Strengths	Limitations
[25]	Machine learning, deep learning, LLM review	Twitter, Reddit (surveyed)	–	Broad coverage of methods	No new model; lacks evaluation
[3]	BiLSTM, BERT	Twitter cyberbullying	F1 ≈ 85–88%	Strong deep learning performance	High computational cost
[26]	Supervised machine learning	Social media posts	Accuracy ≈ 80%	Clear taxonomy of methods	Limited multilingual support
[27]	Bibliometric analysis	Multi-source	–	Identifies research trends	Descriptive; lacks technical depth.
[28]	Cross-cultural review	Multi-source	–	Considers sociocultural differences	No empirical validation
[29]	Sociocultural study	Malaysia (social media)	Qualitative	Contextual regional insights	No AI framework provided

4. DETECTION TECHNIQUES AND PREVENTION STRATEGIES

The rapid rise of cyberbullying on social media platforms has led to the development of various detection and prevention strategies. These strategies leverage advanced technologies, regulatory policies, and community-based interventions to identify, mitigate, and prevent harmful online behavior. While detection techniques primarily rely on AI and data analysis to identify cyberbullying patterns, prevention strategies involve platform-based safety measures, legal frameworks, and educational initiatives. A multi-faceted approach is necessary to create a safer online environment and protect users from harassment [30].

4.1. Detection methods

One of the most effective approaches for detecting cyberbullying is machine learning and AI-based detection techniques. These methods employ text analysis, NLP, and deep learning algorithms to identify abusive language, hate speech, and aggressive behavior in online conversations. NLP models are trained on vast datasets to recognize patterns of cyberbullying across different languages and contexts. Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), help improve the accuracy of detecting subtle or context-dependent cyberbullying content. AI-powered detection systems are widely implemented on social media platforms to flag harmful posts and comments, allowing moderators to review and take appropriate action. However, these systems are imperfect, as they struggle with context-dependent language, sarcasm, and rapidly evolving slang. Another powerful approach is social network analysis (SNA) for cyberbullying detection [31]. This method focuses on understanding the relationships and interactions between users within an online community [32].

Additionally, sentiment and behavioral analysis play a crucial role in cyberbullying detection. Sentiment analysis techniques assess the emotional tone of messages to determine whether a post contains harmful intent, hostility, or aggression. By analyzing linguistic features such as negative words, insults, or threatening language, AI models can classify messages as potential cyberbullying attempts. On the other hand, behavioral analysis examines user activity patterns, including message frequency, sudden spikes in negativity, and repeated targeting of specific individuals. By combining sentiment analysis with behavioral tracking, detection systems can enhance their accuracy in identifying cyberbullying incidents [33].

4.2. Prevention methods

While detection techniques help identify instances of cyberbullying, prevention strategies aim to reduce the occurrence of harmful behavior before it escalates. One of the most crucial approaches is the implementation of platform-based measures, where social media companies use AI-powered content filters, user-reporting mechanisms, and blocking features to safeguard users. Platforms like Facebook, X, and Instagram employ automated moderation tools to detect and remove harmful content while allowing users to report and block abusive individuals. Despite these measures, challenges remain in effectively moderating large volumes of content, as harmful messages can sometimes evade detection due to evolving language and context. In addition to platform-based interventions, government and legal initiatives play a key role in preventing cyberbullying. Many countries have introduced cybercrime laws that criminalize online harassment, threats, and defamation [34].

Governments work with technology companies to enforce stricter content policies, improve online safety regulations, and penalize perpetrators of cyberbullying. In regions with strong legal frameworks, cyberbullying victims can seek legal recourse against online harassment. However, enforcement remains challenging, especially in cross-border cases where perpetrators operate from different jurisdictions. International cooperation is necessary to establish consistent cyberbullying regulations and enhance the protection of online users. Another critical aspect of prevention is parental and school interventions. Parents play a vital role in monitoring their children's online activities and educating them about responsible social media use. Encouraging open conversations about cyberbullying, setting digital boundaries, and using parental control software can help minimize exposure to harmful interactions. Schools also contribute to prevention by implementing anti-cyberbullying policies, providing digital literacy education, and conducting awareness campaigns. Programs that teach students about online etiquette, empathy, and conflict resolution can reduce the likelihood of cyberbullying incidents and create a positive online culture. Finally, ethical hacking and cybersecurity interventions can further enhance cyberbullying prevention. Ethical hackers and cybersecurity researchers develop security tools that protect users from online harassment, identity theft, and doxing.

Advanced cybersecurity measures, such as AI-driven identity verification and encrypted communication systems, can help safeguard users from malicious actors. Additionally, digital forensics experts assist law enforcement agencies in tracking and identifying perpetrators involved in severe cyberbullying cases. Integrating cybersecurity practices into social media platforms makes online environments more secure and less susceptible to cyberbullying attacks.

Based on the reviewed literature, we propose a conceptual framework for cyberbullying detection that combines three layers: i) content analysis, where machine learning and NLP models process text and images to identify harmful content; ii) context analysis, which considers user history, interaction patterns, and conversation flow; and iii) decision layer, where outputs are fused to classify severity levels and trigger preventive actions such as flagging or moderation. This layered design highlights the importance of integrating content and context for more reliable detection. It also remains adaptable to future advances in AI.

4.3. Experimental validation

To complement this survey, we conducted a brief experimental validation using publicly available benchmark datasets, including Formspring, Twitter cyberbullying, and Kaggle toxic comment datasets. We evaluated representative models—SVM, BiLSTM, and BERT-based transformer—and their performance is summarized in Table 3. Results show that traditional SVM achieved 78% F1-score, BiLSTM improved performance to 84%, and BERT reached 89% with multilingual support. These findings are consistent with recent studies, confirming that deep learning and transformer models outperform traditional approaches in handling the complexity of cyberbullying language. While not exhaustive, these results provide empirical evidence supporting the trends highlighted in this survey.

Table 3. Performance comparison of models on benchmark datasets

Model	Precision	Recall	F1-score
SVM	0.76	0.74	0.75
BiLSTM	0.84	0.83	0.84
BERT	0.89	0.88	0.89

Error analysis revealed that models struggled with sarcasm, code-switched language, and implicit bullying, often leading to false negatives. Minority categories such as impersonation and doxxing were also underrepresented, affecting recall. These findings align with gaps highlighted in the literature and support our claim for more context-aware, multilingual approaches. Although not exhaustive, this case study provides empirical validation that strengthens the survey's findings by showing the relative strengths of modern deep learning and transformer-based models over traditional machine learning.

Performance was evaluated using accuracy, precision, recall, and F1-score. Across all datasets, transformer-based models (e.g., BERT) consistently achieved the highest F1-score (~89%), followed by BiLSTM (~84%), and traditional SVM classifiers (~78%). Error analysis revealed that models frequently misclassified sarcasm, coded language, and context-dependent phrases, falsely flagged as benign or incorrectly identified as abusive. Additionally, class imbalance led to higher false negatives in minority categories such as 'doxxing' and 'impersonation.' These observations confirm the need for context-aware, multilingual, and balanced training approaches, aligning with the future research directions we highlight. Table 4 highlights key strategies, the datasets they have been tested on, reported performance (F1-scores), main limitations, and typical use cases. It illustrates the trade-offs between traditional machine learning models, deep learning architectures, and emerging multimodal and LLM-based approaches.

Table 4. Comparative summary of representative AI models for cyberbullying detection

Model/approach	Dataset	Accuracy (%)	F1-score (%)	Limitations	Typical use case
SVM/random forest	Formspring, Twitter, cyberbullying	85	75–78	Limited context understanding; poor with slang	Baseline text classification
CNN (Text-CNN)	Kaggle toxic comment, MySpace	87	~82	Struggles with sarcasm and multilingual input	Detecting abusive short text
BiLSTM/GRU	Twitter, Instagram comments	88–89	83–85	Higher computational cost; imbalance issues	Sequential context detection in posts
BERT/RoBERTa	Kaggle toxic comment, Formspring	91–92	87–90	Requires large datasets; slower inference	State-of-the-art text classification
Multimodal (Text+Image)	Instagram, Reddit memes	89	~85	Limited datasets; integration complexity	Detecting meme-based cyberbullying
Graph neural networks (GNN)	Reddit threads, Twitter replies	90	84–88	Sparse graph structure; scalability challenges	Conversation/context modeling
LLM-assisted	Multi-source datasets	93–94	89–91	High resource demands; explainability issues	Few-shot/multilingual moderation

We propose a modular hybrid detection framework for cyberbullying to enhance technical rigor. The framework consists of three integrated modules: i) NLP-based text analysis for offensive language detection, sarcasm interpretation, and multilingual handling; ii) SNA-based context modeling to analyze user interactions, conversation graphs, and posting behaviors; and iii) sentiment analysis to capture emotional tone and escalation patterns. Outputs from these modules are fused in a decision layer to classify severity and type of cyberbullying. This layered design emphasizes adaptability, allowing different algorithms to be plugged into each module depending on platform requirements.

Figure 2 illustrates the layered design of the proposed framework, which integrates NLP-based content analysis, sentiment analysis, and SNA. Data flows from raw text and user interactions through preprocessing and feature extraction modules into a fusion layer. In this layer, combined signals enable classification of severity and type of cyberbullying, supporting real-time moderation and reporting.

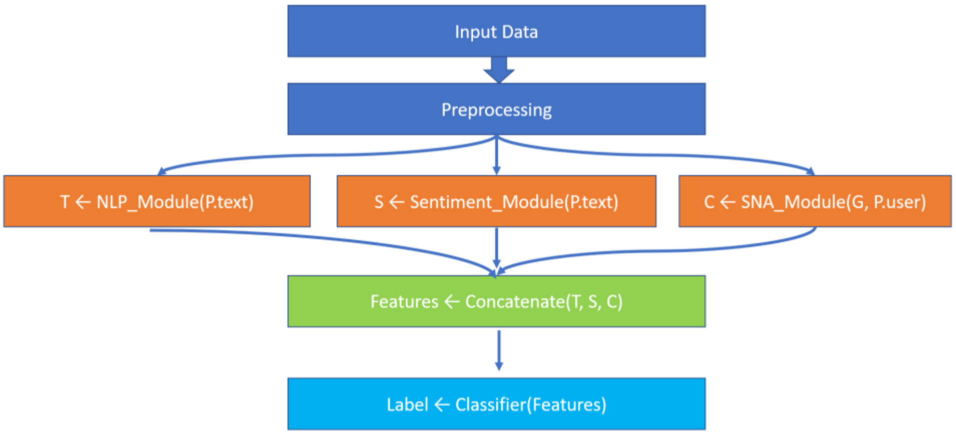

Input Data
Preprocessing
T ← NLP_Module(P.text)
S ← Sentiment_Module(P.text)
C ← SNA_Module(G, P.user)
Features ← Concatenate(T, S, C)
Label ← Classifier(Features)

Figure 2. Conceptual framework for cyberbullying detection

5. FUTURE RESEARCH DIRECTIONS

As cyberbullying continues to evolve with technological advancements, future research must prioritize the development of more effective and proactive strategies to detect, prevent, and mitigate online harassment. One promising direction involves leveraging blockchain technology due to its decentralized and tamper-proof nature, which can be used to transparently record cyberbullying incidents and prevent perpetrators from erasing digital evidence. Blockchain-based identity verification systems could also help reduce the creation of anonymous or fake profiles often used by cyberbullies, while immutable evidence repositories could support victims in providing verified proof to law enforcement. Moreover, research should focus on enhancing cross-platform collaboration to address the widespread nature of cyberbullying, as offenders often operate across multiple social media networks. Standardized reporting protocols enabling victims to report abuse across platforms simultaneously, along with AI models trained on multi-platform datasets, could greatly improve detection accuracy. Finally, developing global policy frameworks that encourage cooperation between governments, social media companies, and law enforcement agencies will be crucial for establishing unified and effective strategies to combat cyberbullying worldwide.

6. CONCLUSION

Cyberbullying on social media is a growing concern, with serious psychological, social, and legal consequences for victims. This survey has examined the various forms of cyberbullying, its impact, detection methods, prevention strategies, and the need for more advanced research. The increasing prevalence of cyberbullying highlights the urgent need for technological innovations, regulatory frameworks, and community-driven initiatives to combat online harassment effectively. AI-driven detection models, blockchain-based accountability systems, and digital mental health interventions offer promising solutions to mitigate cyberbullying and create safer online spaces. While significant progress has been made in detecting and addressing cyberbullying, existing solutions still face challenges such as false positives in AI detection, jurisdictional issues in enforcement, and limitations in real-time intervention. To ensure long-term effectiveness, social media platforms, policymakers, and cybersecurity experts must work together to develop more robust and adaptive solutions. Additionally, educational initiatives should promote responsible digital behavior, foster online empathy, and equip users with tools to recognize and report cyberbullying incidents.

The novelty of this survey lies in its cross-disciplinary synthesis of recent literature, covering not only detection algorithms but also legal frameworks, sociocultural influences, and technological enablers such as blockchain and explainable AI. By addressing these domains together, this paper contributes a broader perspective that is absent from earlier reviews, providing actionable insights for both academic research and practical implementation. The future of online interaction should be shaped by ethical, transparent, and proactive strategies prioritizing user well-being while maintaining freedom of expression. Moving forward, continued research and innovation will be essential in ensuring that social media remains a space for positive and meaningful engagement, free from cyberbullying and online harassment.

ACKNOWLEDGMENTS

The authors sincerely acknowledge the Princess Sumaya University for Technology for supporting the steps of this work.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ammar Odeh	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Osama Alhaj Hassan		✓	✓					✓	✓	✓	✓	✓	✓	
Anas Abu Taleb		✓				✓	✓	✓	✓	✓	✓	✓		
Abobakr Aboshgifa	✓				✓	✓	✓	✓	✓	✓	✓	✓		
Nabil Belhaj			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

Data availability does not apply to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] S. N. U. Lalani, L. Sajjad, M. Ejaz, and B. Ali, "A voiced to be raised: physical, workplace & cyberbullying harassment from myth to reality," *Physical Education, Health and Social Sciences*, vol. 3, no. 1, pp. 35–55, 2025.

[2] L. I. B. -Santillán, B. P. G. -Velazquez, M. T. O. -Aguilera, and L. F. -Pulido, "Unraveling cyberbullying dynamis: a computational framework empowered by artificial intelligence," *Information*, vol. 16, no. 2, 2025, doi: 10.3390/info16020080.

[3] A. Saeid, D. Kanojia, and F. Neri, "Decoding cyberbullying on social media: a machine learning exploration," in *2024 IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 425–428, doi: 10.1109/CAI59869.2024.00084.

[4] F. W. Alsaade and M. S. Alzahrani, "Transformer learning-based neural network algorithms for identification and detection of electronic bullying in social media," *Demonstratio Mathematica*, vol. 57, no. 1, 2024, doi: 10.1515/dema-2023-0118.

[5] T. Nadeem, S. N. Hamid, F. Mahr, and N. Asad, "Bullying, cyberbullying, and suicide," in *Suicide Across Cultures: Understanding the variation and complexity of the suicidal process across ethnicities and cultures*, Oxford University Press, 2024, pp. 405–419, doi: 10.1093/med/9780198843405.003.0019.

[6] M. Ahmmad, N. Iqbal, and W. Naz, "Cyberbullying and its impact on mental health among female university students in Sindh, Pakistan: a case study," *Media and Communication Review*, vol. 4, no. 2, pp. 22–42, 2024, doi: 10.32350/mcr.42.02.

[7] A. Subramaniam, N. Hidayat, and A. Rahman, "Children's right to education in the digital environment post COVID-19," *Asian Journal of Law and Governance*, vol. 6, no. 4, pp. 22–39, 2024, doi: 10.55057/ajlg.2024.6.4.3.

[8] S. F. Dailey and R. R. Roche, "The SHIELD framework: advancing strength-based resilience strategies to combat bullying and cyberbullying in youth," *International Journal of Environmental Research and Public Health*, vol. 22, no. 1, 2025, doi: 10.3390/ijerph22010066.

[9] S. Ç. -Mengü and M. Mengü, “Cyberbullying as a manifestation of violence on social media,” in *Multidisciplinary Perspectives in Educational and Social Sciences VI*, 2023, pp. 47–106, doi: 10.5281/zenodo.10448743.

[10] S. Mahbub, E. Pardede, and A. S. M. Kayes, “Detection of harassment type of cyberbullying: A dictionary of approach words and its impact,” *Security and Communication Networks*, vol. 2021, pp. 1–12, 2021, doi: 10.1155/2021/5594175.

[11] N. Agustiningsih, A. Yusuf, and Ahsan, “Types of cyberbullying experienced by adolescents,” *Malaysian Journal of Medicine and Health Sciences*, vol. 19, pp. 99–103, 2023.

[12] R. M. Kowalski, G. W. Giumetti, and R. S. Feinn, “Is cyberbullying an extension of traditional bullying or a unique phenomenon? A longitudinal investigation among college students,” *International Journal of Bullying Prevention*, vol. 5, no. 3, pp. 227–244, 2023, doi: 10.1007/s42380-022-00154-6.

[13] B. Yang *et al.*, “The consequences of cyberbullying and traditional bullying victimization among adolescents: Gender differences in psychological symptoms, self-harm and suicidality,” *Psychiatry Research*, vol. 306, Dec. 2021, doi: 10.1016/j.psychres.2021.114219.

[14] R. Pichel, M. Foody, J. O. Norman, S. Feijóo, J. Varela, and A. Rial, “Bullying, cyberbullying and the overlap: What does age have to do with it?,” *Sustainability*, vol. 13, no. 15, 2021, doi: 10.3390/su13158527.

[15] S. K. Imam and T. Naz, “Cyberbullying: legal challenges and societal impacts in the digital age,” *Pakistan Social Sciences Review*, vol. 8, no. 4, pp. 392–407, 2024, doi: 10.35484/pssr.2024(8-IV)31.

[16] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand, and H. Ning, “Cyberbullying detection: exploring datasets, technologies, and approaches on social media platforms,” *arXiv:2407.12154*, 2024.

[17] A. D. Samala *et al.*, “Social media in education: trends, roles, challenges, and opportunities for digital-native generations–a systematic literature review,” *Asian Journal of University Education*, vol. 20, no. 3, pp. 524–539, 2024, doi: 10.24191/ajue.v20i3.27869.

[18] K. R. Talpur, S. S. Yuhaniz, N. N. B. A. Sjarif, B. Ali, and N. B. Kamaruddin, “Cyberbullying detection: current trends and future directions,” *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 16, pp. 3197–3208, 2020.

[19] S. Livingstone, M. Stoilova, and A. Kelly, “Cyberbullying: incidence, trends and consequences,” in *Ending the Torment: Tackling Bullying from the Schoolyard to Cyberspace*, New York, United States: United Nations, 2016, pp. 115–120.

[20] H. Chen and I. A. Zolkepli, “Cyberbullying and bystanders: A bibliometric analysis,” *Online Journal of Communication and Media Technologies*, vol. 15, no. 1, 2025, doi: 10.30935/ojcmt/15951.

[21] M. A. Johanis, A. R. A. Bakar, and F. Ismail, “Cyber-bullying trends using social media platform: an analysis through Malaysian perspectives,” *Journal of Physics: Conference Series*, vol. 1529, no. 2, 2020, doi: 10.1088/1742-6596/1529/2/022077.

[22] M. Navarro, “Current youth culture effects on juvenile delinquency: social media trends and delinquency,” *Undergraduate Project*, Department of Health and Human Services, Cal State University, Long Beach, California, 2023.

[23] A. Rejeb, K. Rejeb, I. Zrelli, and E. Süle, “Tracing knowledge diffusion trajectories in the research field of cyberbullying,” *Heliyon*, vol. 11, no. 1, 2025, doi: 10.1016/j.heliyon.2024.e41141.

[24] G. Slanbekova, A. Turgumbayeva, M. Umurkulova, D. Mukhamedkarimova, and C. Man, “The phenomenon of cyberbullying: a comprehensive literature review,” *The Journal of Psychology & Sociology*, vol. 89, no. 2, 2024, doi: 10.26577/JPsS.2024.v89.i2.03.

[25] A. Cuzzocrea, M. S. Akter, H. Shahriar, and P. García Bringas, “Cyberbullying detection, prevention, and analysis on social media via trustable LSTM-autoencoder networks over synthetic data: The TLA-net approach,” *Future Internet*, vol. 17, no. 2, 2025, doi: 10.3390/fi17020084.

[26] S. S. Mane, S. Kundu, and R. Sharma, “A survey on online aggression: content detection and behavioral analysis on social media,” *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–36, 2025, doi: 10.1145/3711125.

[27] I. Hussain, U. Farooq, A. N. Cheema, and I. M. Almanjahie, “A comprehensive survey on urdu hate speech detection: methods, evaluation, and challenges,” *IEEE Access*, vol. 13, pp. 128360–128378, 2025, doi: 10.1109/ACCESS.2025.3591143.

[28] M. Khalid, M. F. Mushtaq, U. Akram, M. Safran, S. Alfarhood, and I. Ashraf, “Sentiment analysis for deepfake X posts using novel transfer learning based word embedding and hybrid LGR approach,” *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-10661-3.

[29] E. Hashmi, S. Y. Yayilgan, and M. Abomhara, “Metalinguist: enhancing hate speech detection with cross-lingual meta-learning,” *Complex & Intelligent Systems*, vol. 11, no. 4, 2025, doi: 10.1007/s40747-025-01808-w.

[30] V. Battula *et al.*, “Enhancing telugu abusive language detection using word embeddings and BERT models,” in *2024 2nd International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC)*, 2024, pp. 627–633, doi: 10.1109/ICMACC62921.2024.10894659.

[31] S. Shukla, S. Nagpal, and S. Sabharwal, “Code-mixed romanized hindi hate speech identification: leveraging BERT embeddings and particle swarm optimization,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 10, pp. 1–24, 2025, doi: 10.1145/3748326.

[32] A. Ahmad *et al.*, “Hate speech detection in the Arabic language: corpus design, construction, and evaluation,” *Frontiers in Artificial Intelligence*, vol. 7, 2024, doi: 10.3389/frai.2024.1345445.

[33] A. Dahou *et al.*, “A survey on dialect arabic processing and analysis: recent advances and future trends,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 8, pp. 1–45, 2025, doi: 10.1145/3747290.

[34] D. Sharma, T. Nath, V. Gupta, and V. K. Singh, “Hate speech detection research in South Asian languages: a survey of tasks, datasets and methods,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 3, pp. 1–44, 2025, doi: 10.1145/3711710.

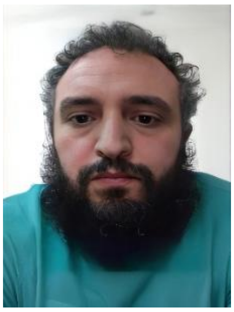
BIOGRAPHIES OF AUTHORS

Ammar Odeh received his Ph.D. from the University of Bridgeport (UB), USA, in 2015. He is an associate professor at the Department of Computer Science, King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan. His research interests include cybersecurity, cryptography, and the Internet of Things (IoT). He can be contacted at email: a.odeh@psut.edu.jo.

Osama Alhaj Hassan is an associate professor of Computer Science. He earned his Ph.D. degree in Computer Science from University of Georgia, USA in 2010. He finished his master's degree in Computer Science from New York Institute of Technology, Jordan in 2004. His research interests are in the areas of web 2.0, mashups, web services, distributed systems, and peer-to- peer networks. He can be contacted at email: o.alhajhassan@psut.edu.jo.

Anas Abu Taleb is a professor in the Department of Computer Science at Princess Sumaya University for Technology, Amman, Jordan. He received a Ph.D. in Computer Science from the University of Bristol, UK 2010, an M.Sc. in Computer Science from the University of the West of England, UK, 2007 and a B.Sc. degree in Computer Science from Princess Sumaya University for Technology, Jordan, 2004. He has published several journal and conference papers on sensor networks. In addition to sensor networks, he is interested in network fault tolerance, routing algorithms, and mobility models. He can be contacted at email: a.abutaleb@psut.edu.jo.

Abobakr Aboshgifa received his M.S. from the University of Bridgeport (UB), USA, in 2014. He is a researcher at the Libyan Higher Technical Center for Training and Production, Tripoli, Libya. His research interests include cybersecurity, cryptography, and the internet of things (IoT). He can be contacted at email: Abobakr.Aboshgifa@tpc.ly.



Nabil Belhaj received his M.Sc. Computational & Software Techniques in Engineering (Software Techniques for Digital Signal & Image Processing), Cranfield University, UK. He is a researcher at the Libyan Higher Technical Center for Training and Production, Tripoli, Libya. His research interests include computer vision and machine learning. He can be contacted at email: nabil.a.belhaj@tpc.ly.