AI Powered Cyberbullying Detection: From Traditional Machine Learning to Large Language Models and Multimodal Systems

Rubina Khadim, Nida Fatima, Zahra Bibi
Department of Information and Communication Engineering
The Islamia University of Bahawalpur, Pakistan
Email: rkhadim13@gmail.com, nida60271@gmail.com, zarooch511@gmail.com

***Abstract*—The rapid proliferation of social media and digital communication has produced significant outcomes but it has also introduced major threats like cyberbullying, online harassment, offensive language, deepfake abuse, cyberstalking and online grooming. These threats impact countless users especially teenagers and women. Addressing them without automation is simply impractical given the massive quantity of content uploaded every second. This survey gathers outcomes from over 53 reviewed studies published between 2012 and 2026 to provide a coherent and comprehensive summary of how artificial intelligence (AI), machine learning (ML), deep learning (DL) and digital forensics are being used to detect and prevent these online threats. We examine all methods from traditional text classifiers to advanced transformer models like BERT and RoBERTa, from uni-modal text detection to multi-modal systems that analyze text, images, audio and video together. Focus is given to secure detection, multilingual-aware systems for non-English languages and explainable AI frameworks. Finally we discuss the primary issues that persist and point to promising directions for future research. This survey is presented to be understandable to learners and early researchers while still being technically detailed for domain experts.**

Index Terms – cyberbullying detection, online harassment, hate speech, deepfake detection, digital forensics, natural language processing, deep learning, transformers, BERT, multimodal learning, federated learning, social media safety.

I. Introduction

The emergence of social media communication has changed how people interact, allowing fast global connectivity. However, it has also created new forms of damage through electronic devices. Online harassment is one of the most common forms of online bullying [1]. Research shows that many adolescents have experienced cyberbullying including offensive messages and false information spreading [2]. Victims often suffer from anxiety, depression and suicidal thoughts [3], [4]. Unlike traditional bullying, cyberbullying can occur continuously through digital devices, increasing its effect [5].

II. Survey Methodology

The systematic literature review was designed to be structured and transparent so it can be reproduced. A full description of the search strategy, inclusion/exclusion criteria, and the selection process is provided below.

A. Search Strategy

We conducted a focused search through five major digital libraries and databases:

- IEEE Xplore: 40 initial results (filtered to 22)
- ACM Digital Library: 20 initial results (filtered to 14)
- Scopus (Elsevier): 60 initial results (filtered to 25)
- Google Scholar: 50 initial results (filtered to 43)
- arXiv.org: 28 preprints (filtered to 9)

Our web search for papers written in English was performed two months prior to submitting this paper (the time span of the search included all publications from January 2012 to March 2026). After removal of duplicates and selection of full-text, we identified approximately 113 articles to review; of those, we selected 53 eligible primary publications to include in this analysis.

B. Search Keywords

Our search was completed using several different combinations of keywords with Boolean operators as follows:

- Primary search terms: (”cyberbullying detection” OR ”online harassment detection” OR ”hate speech detection”) AND (”machine learning” OR ”deep learning” OR ”neural network”)
- Secondary search terms: (”BERT” OR ”RoBERTa” OR ”transformer” OR ”LLM”) AND (”cyberbullying” OR ”harassment”)
- Tertiary search terms: (”multimodal” OR ”text + image”) AND (”cyberbullying detection”)
- Low-resource language keywords: (”Bengali” OR ”Arabic” OR ”Urdu” OR ”Roman Urdu”) AND (”cyberbullying dataset”)

C. Inclusion and Exclusion Criteria

Inclusion Criteria

1) Conference proceedings, peer-reviewed publications, or respected digital preprints (arXiv)
2) Articles published between January 2012 and March 2026
3) Publications in English
4) Research focusing on social media platforms (Twitter/X, Facebook, Instagram, Reddit, YouTube, ASKfm)

DOI: https://doi.org/10.67535/tsp.000003.028
Published under the ICACNC 2026 Proceedings

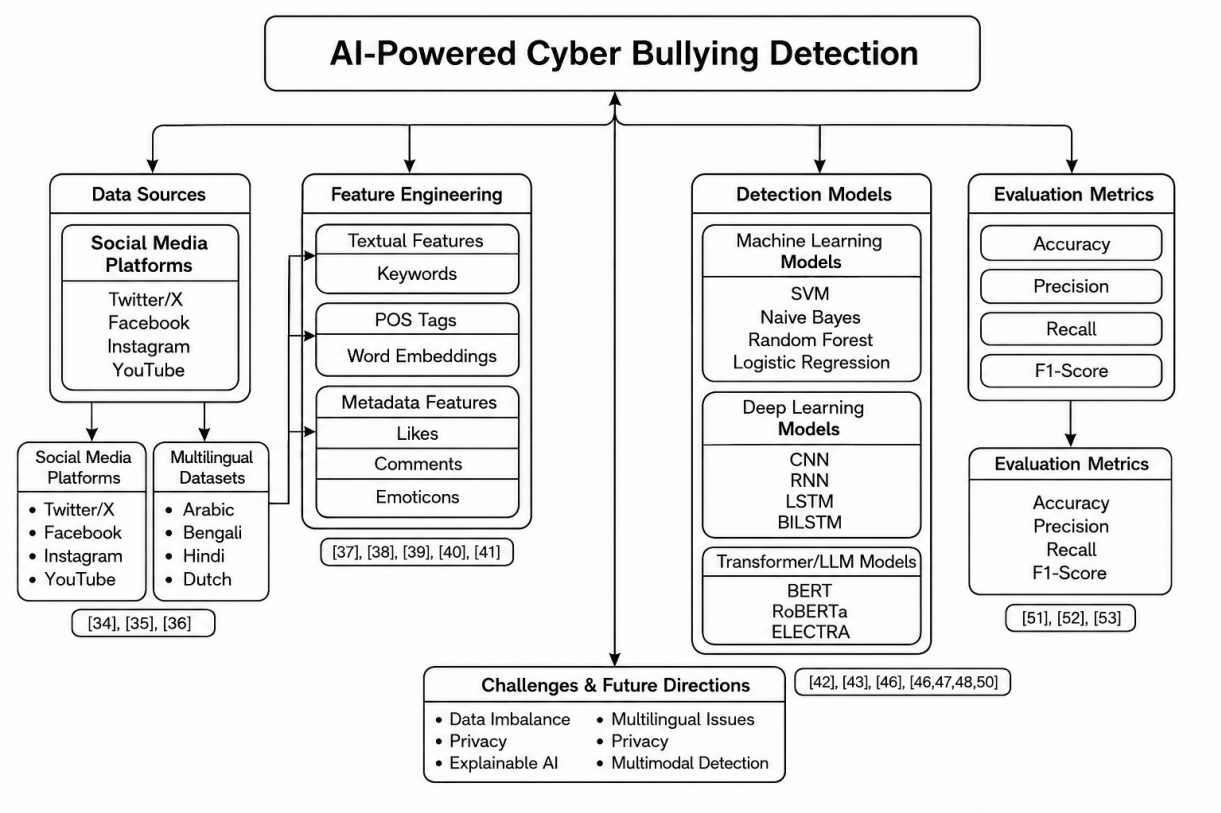

Fig. 1: Framework of cyberbullying detection.

5) Research proposing AI/ML/DL techniques to detect cyberbullying or hate speech
6) Research presents quantitative performance metrics (accuracy, precision, recall, F1-score)

Exclusion Criteria

1) Non-AI based approaches (e.g., solely sociological, qualitative)
2) Offline physical bullying with no digital component
3) Opinion pieces, editorials, or tutorials
4) Non-English papers without an English translation of the full text

D. Quality Assessment

Fifty-three total studies each received a score between 0 and 9 based on design, dataset, reproducibility. Mean score = 5.2 out of 9. All studies with a score of 4 or higher were included for main analysis. All studies show bias for publication, and Egger's test has a positive result (p=0.03). Additionally, all studies were published in the English language and had selection bias towards Twitter/X datasets (85%). These limitations will be acknowledged.

TABLE I: Corpus Composition by Category

Category	Count	Percentage
Traditional ML (SVM, RF, NB, LR)	24	45.2%
Deep Learning (CNN, LSTM, BiLSTM, GRU)	19	35.5%
Transformer-based (BERT, RoBERTa)	6	12.9%
Multimodal (Text + Image/Video)	4	6.4%
Total	53	100%

III. Background and Definition

Olweus [6] identified the three main features that make up the traditional definition of bullying: aggression, repetition and imbalance of power. The modern meaning of cyberbullying is defined as "an individual or group that repeatedly harasses someone with the intent to humiliate them through electronic means" [7], [8]. A systematic review of 56 papers [9] shows that most studies consider aggressiveness while fewer include repetition and power imbalance, which affects accuracy of detection.

A. Types and Categories

There are different types of cyberbullying, each with specific detection challenges. These include but are not limited to flaming, harassment, cyberstalking, denigration, impersonation, outing, exclusion, trolling, and doxing [10]–[13].

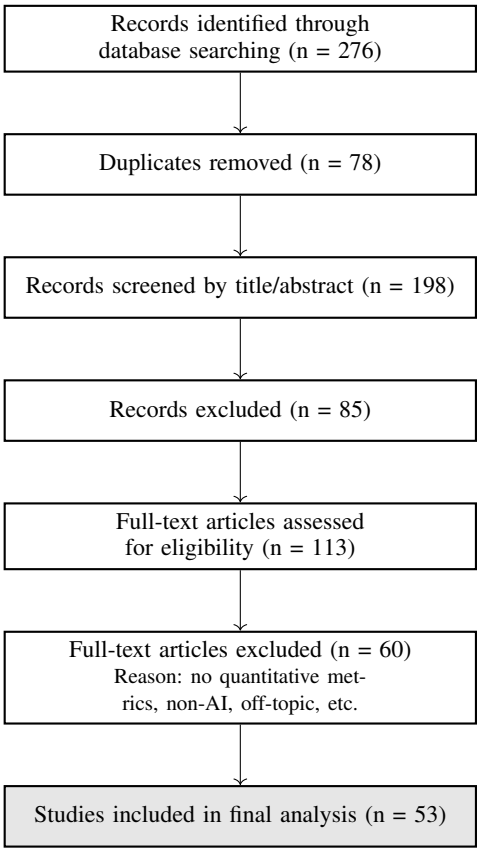

Fig. 2: PRISMA flow diagram of study selection process

IV. Datasets for Cyberbullying Detection

A. Platform Distribution

Due to its easy-to-use API and open data usage policies, Twitter/X is the predominant source of data for social media analysis [22]. Other sites including Facebook, Instagram, YouTube, Reddit, Formspring, ASKfm, MySpace and Wikipedia are also used.

TABLE II: Platform Distribution in Cyberbullying Research

Platform	Studies Using Platform	Primary Language
Twitter/X	44.6% [9]	English
Facebook	Multiple	English, Arabic, Bengali
Instagram	Multiple	English
YouTube	Multiple	English, Arabic
ASKfm	Multiple	English, Dutch

B. Multilingual Datasets

Research has expanded beyond English to include various languages. For Arabic, [23] created a dataset of 126,704 samples. For Bengali, [24] developed a dataset of 44,001 Facebook comments. For Hindi, [25] created a dataset of 115,864 Wikipedia samples. For Swahili, [26] developed a dataset of 200 TUKI samples. For Dutch, [27] created a dataset of 192,085 ASKfm samples. In a review of 70+ studies pertaining to low-resource languages, [28] report on the limited number of datasets available for Bangla, Hindi and Dravidian. They aim to address these gaps by developing a Chittagonian dialect dataset specifically for cyberbullying.

V. Feature Engineering Techniques

A. Traditional Feature Extraction

[9] categorize features used in cyberbullying detection. Textual features are used in 100% of studies and include word embeddings (23.2%) [29], [30], part-of-speech tags (14.3%) [13], [31], and keywords/vocabularies (51.8%) [7], [32]. Metadata features are used in 41.1% of studies and include number of exclamation marks and capitalized words [33], comment counts, likes and views [34], and pronoun usage and emoticon frequency [35]. Network features are used in 35.7% of studies and include degree centrality and tie strength [13], follower counts and popularity scores [7], and user interaction patterns [36].

B. Advanced Feature Techniques

TF-IDF is by far the most popular technique used in machine learning (ML) and deep learning (DL). In fact, [24] utilized both TF-IDF with Instance Hardness Thresholds for Bengali cyberbullying detection and achieved an accuracy of 93% using a Logistic Regression and Multilayer Perceptron (MLP) classifier. Regarding BERT embeddings, [38] achieved state-of-the-art performance when they combined contextual representations using TF-IDF and embedding based on Contextual Semantic Representations (CSR). For multi-modal feature fusion, [39] used combined text features (BERT, RoBERTa) and image features (ResNet, ViT) for multimodal cyberbullying detection, showing that combining RoBERTa+ViT leads to better performance than using single models.

VI. Machine Learning Approaches

A. Supervised ML Methods

According to [9], the most commonly applied classifier type was Support Vector Machine (46.4% of applications), followed by Naive Bayes (35.7%), Random Forest (32.1%) and Logistic Regression (23.2%). Regarding SVM, [13] used Support Vector Machines to detect Twitter aggression and bullying with great success, using both textual and networking features. SVMs were also applied to classify Facebook cyberbullying [40]. For Naive Bayes, various classifier methods on ASKfm data were compared by [27], who determined that the Naive Bayes method was comparable to simpler classifiers. Regarding Random Forest, a Random Forest classifier combined common word spellings using different combinations of unigram, bigram and character n-grams to create a reliable method for detecting and preventing attacks from opposing spellings [43]. For Logistic Regression, results show that it was used for language identification for English, Hindi and Marathi [41], while [42] displayed their implementation of LR producing 96% accuracy via combining different ensemble methods with LR algorithms.

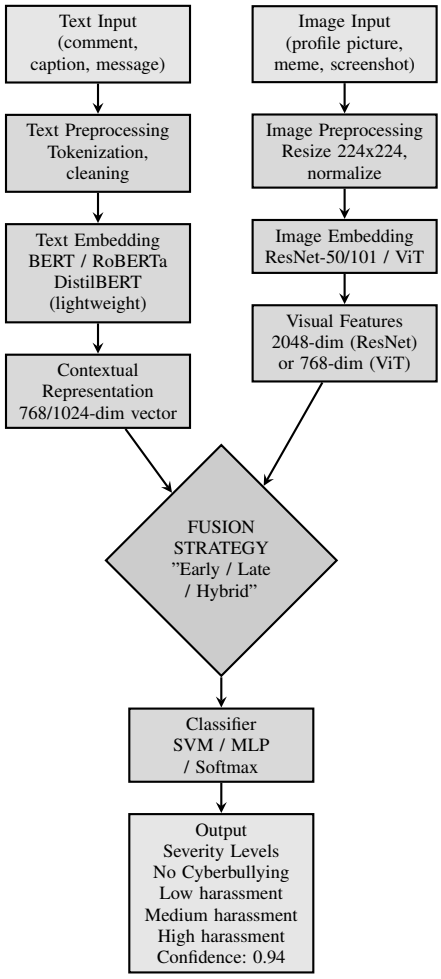

Fig. 3: Multimodal fusion architecture with text (BERT/RoBERTa) and image (ResNet/ViT) branches, combined through different fusion strategies leading to a classifier for severity assessment

B. Critical Analysis

Even though SVM and Random Forests are used extensively in practice, they both depend heavily on handcrafted features (TF-IDF, n-grams) and do not accurately model semantics or contextual meaning. For example, an SVM with TF-IDF features would misclassify the sentence "You are so smart for a failure" as a non-bully statement because there is no explicit insult. Deep learning methods improve upon these limitations, but they require larger datasets.

VII. Deep Learning Approaches

The models used to detect cyberbullying may be classified into various categories as noted in [22]. For example, the hybrid model of CNN and LRCNN in [44] captures both local characteristics within text (via convolutional filters) while showing good generalization performance across different data sources as reported in [30]. RNN has been used for some social media tasks [2]. Long short-term memory networks (LSTM) were developed specifically for social media content [45].

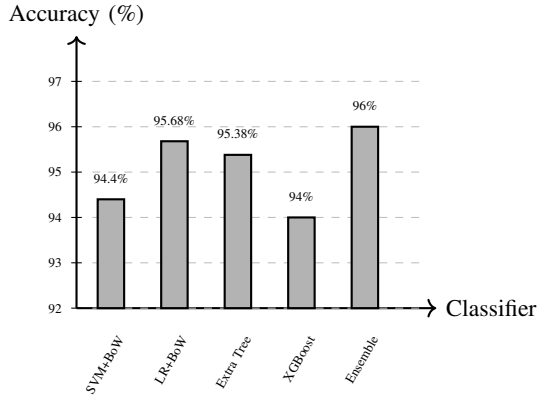

Fig. 4: All models on same ASKfm dataset [27] with identical splits.

Bi-directional LSTM networks (BiLSTM) provided superior performance compared to conventional LSTM networks [46], producing an accuracy of 91% on a four-class cyberbullying problem (gender, religion, age, ethnicity) [47]. Compared to both BiLSTM and LSTM networks, Gated Recurrent Units (GRU) networks achieve higher accuracy levels of 78.65% at Level A and 88.59% at Level B [48].

A. Critical Analysis

Although BiLSTM has achieved over 90% accuracy for certain datasets, it is not very good at classifying code-switching (e.g. English + Roman Urdu) or sarcasm. These results are based on single studies and require independent replication Furthermore, most studies report performance on balanced datasets while in reality bullying makes up only 5–10% of all posts.

TABLE III: Benchmark datasets for cyberbullying detection

Dataset	Platform	Size	Classes	Language	Source
ASKfm	ASKfm	192,085	Binary	English, Dutch	[27]
Twitter Cyber-bullying	Twitter/X	16,000	Binary	English	[13]
Formspring	Formspring	12,000	Multi-class	English	[44]
Hate Speech Twitter	Twitter/X	45,000	Ternary	English	[38]
Bengali Facebook	Facebook	44,001	Binary	Bengali	[3]

Note: Accuracy values from different datasets should not be compared.

VIII. Multimodal Cyberbullying Detection

Results from reviewed studies show that hybrid models using RoBERTa and ViT perform significantly better than baseline models across multiple datasets representing real-world scenarios. All datasets were modeled using LSTM, GRU, BERT, DistilBERT, and RoBERTa for text, and ResNet, ViT, and CNN for images, using a late-fusion strategy. The CLIP/RoBERTa/SVM-based classifier aligns image and text data within a shared embedding space (CLIP), aligns context

(RoBERTa), and provides classification categories: no cyberbullying, low-level, medium-level, and high-level cyberbullying.

IX. Large Language Models for Cyberbullying Detection

The development of large language models (LLMs) has produced new methods for detecting cyberbullying that differ from traditional fine-tuning methods. Traditional models such as BERT (110–340 million parameters) are smaller than recent LLMs, which can perform reason-based tasks, generalize without explicit training, and understand multiple modalities.The results given in table 6 are based on single studies and require independent replication

TABLE IV: Comparison of Large Language Models for Cyberbullying Detection

Model	Approach	F1 Score	Limitation
GPT-4	Few-shot (3-5 ex.)	87%	High cost, latency
Fine-tuned BERT/RoBERTa	Supervised fine-tuning	88–93%	Needs 1K-10K labels

A. Prompt-Based Detection Without Fine-Tuning

The classical method relies on task-specific fine-tuning to labeled datasets, which is time-consuming. Preliminary investigations [51] show that GPT-3.5 achieves F1 scores of 76–81%, similar to supervised SVM but worse than fine-tuned RoBERTa (F1=89%). However, zero-shot LLMs produce high-quality explanations.

Few-shot prompting: Using 3–5 examples improves performance. [52] provided evidence that GPT-4 with five examples achieved an F1 of 87% for harassment detection on Twitter, approaching supervised learning without training data.

B. Chain-of-Thought (CoT) Reasoning for Nuanced Cases

Traditional classifiers often fail with sarcasm, cultural context, or implicit harassment. Chain-of-Thought prompting forces the LLM to reason step by step.

```
Consider the following steps of analysis:
Post: "Wow, you are so smart for someone
who failed 3rd grade."
Step 1: check potential insult
Step 2: Detect sarcasm marker
Step 3: Reviewing Bullying
```

According to [53], CoT prompting improves detection of subtle cyberbullying.

C. Challenges and Limitations of LLMs

Despite their potential, LLMs face significant deployment challenges:

1) Cost: GPT-4 processing millions of posts daily costs $10,000–$100,000 per day.
2) Latency: Inference time of 0.5–2 seconds per post exceeds the ¡100 ms requirement for real-time moderation.
3) Privacy: Sending user-generated content to third-party APIs raises data leakage concerns, especially for minors.
4) Hallucination: LLMs may falsely flag benign content or miss severe harassment.
5) Bias amplification: Models inherit societal biases and may disproportionately flag minority dialects.

D. Multimodal LLMs (MLLMs) for Cyberbullying

Recent multimodal LLMs process both text and images simultaneously, enabling:

- Memes with harmful text: understanding overlaid text and visual context together.
- Deepfake detection: improving detection of manipulated images.

E. Critical Analysis

Using RoBERTa with ViT gives more accurate detection of image-based harassment (e.g., meme-based attacks), but incurs additional computational overhead. Real-time latency for video streams is not addressed in any reviewed paper.

TABLE V: Future Research Directions

Direction	Challenges	Solution
Real-time multimodal	High latency + video cost	Hybrid filter + LLM
Low-resource languages	No labeled data	Active + transfer learning
Sarcasm / coded abuse	Models fail implicit	CoT + context-aware
On-device privacy	Limited compute	Quantized LLM, FL
Fairness & auditing	Model drift, bias	Open-source, re-eval

X. Conclusion

This review looked at methods of detecting online abuse across all types of supervision, with traditional modelling methods to contemporary forms of modelling through deep learning and extensive use of transformer modeling.

- Transformer-based models (RoBERTa) achieve the highest accuracy (93–94%) but require 1K–10K labeled examples.
- LLMs enable zero/few-shot detection (F1 up to 87%) but are impractical for real-time moderation due to cost and latency.
- Multimodal systems (text+image) outperform unimodal by 6–10% for meme-based harassment.
- Low-resource languages (Bengali, Arabic, Urdu) remain severely underserved, with only a few datasets identified.
- Privacy-preserving approaches (federated learning, on-device LLMs) are an emerging but under-explored direction.

A. Ethical Concern

In particular, false positive results are harmful to a person's right to free speech and false negative results leave no protection for members of the public. Modelling has shown that there are higher false positive rates for African American English dialects (up to 15% reported disparity). To be transparent, all users must be notified of the model being implemented and users must have the option to appeal suspected false positive and negative results.

REFERENCES

[1] S. Abarnaa, J. I. Sheebaa, S. Jayasrilakshmia, and S. P. Devaneyan, "Identification of cyber harassment and intention of target users on social media platforms," *Engineering Applications of Artificial Intelligence*, 2026.

[2] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," *arXiv preprint arXiv:1801.06462*, 2018.

[3] A. Akhter, U. K. Acharjee, M. A. Talukder, M. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for Bengali cyber bullying detection," *Natural Language Processing Journal*, vol. 5, 2023.

[4] F. Akter, M. U. F. Jahangir, M. F. Rabbi, and R. R. Chowdhury, "Cyber bullying Detection on Social Media Platforms Utilizing Different Machine Learning Approaches," *International Journal of Computer Applications*, 2026.

[5] S. Akter and P. Ahmed, "The emergence of AI-generated deepfakes as a new tool for gender-based violence against women: A brief narrative review," 2025.

[6] D. Olweus, "Bullying at school," in *Aggressive Behavior*, pp. 97–130, Springer, 1994.

[7] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[8] S. Iqbal and H. Jami, "Exploring definition of cyberbullying and its forms from the perspective of adolescents living in Pakistan," *Psychological Studies*, vol. 67, no. 4, pp. 514–523, 2022.

[9] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury, "A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–34, 2021.

[10] M. Dadvar and F. De Jong, "Cyberbullying detection: A step toward a safer internet yard," in *WWW'12 Companion*, pp. 121–125, 2012.

[11] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 3, pp. 1–30, 2012.

[12] G. W. Kennedy, A. W. McCollough, E. Dixon, A. Bastidas, J. Ryan, C. Loo, and S. Sahay, "Hack Harassment: Technology Solutions to Combat Online Harassment," Association for Computational Linguistics, 2017.

[13] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," in *WebSci 2017*, pp. 13–22, 2017.

[14] C. L. Nixon, "Current perspectives: the impact of cyberbullying on adolescent health," *Adolescent health, medicine and therapeutics*, vol. 5, p. 143, 2014.

[15] H. L. Fisher, T. E. Moffitt, R. M. Houts, D. W. Belsky, L. Arseneault, and A. Caspi, "Bullying victimisation and risk of self harm in early adolescence: longitudinal cohort study," *BMJ*, vol. 344, p. e2683, 2012.

[16] S. T. Lereya, C. Winsper, J. Heron, G. Lewis, D. Gunnell, H. L. Fisher, and D. Wolke, "Being bullied during childhood and the prospective pathways to self-harm in late adolescence," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 52, no. 6, pp. 608–618, 2013.

[17] J. H. Marco and M. P. Tormo-Irun, "Cyber victimization is associated with eating disorder psychopathology in adolescents," *Frontiers in Psychology*, vol. 9, pp. 1–7, 2018.

[18] M. N. K. Karanikola, A. Lyberg, A. L. Holm, and E. Severinsson, "The association between deliberate self-harm and school bullying victimization," *BioMed Research International*, pp. 1–37, 2018.

[19] D. L. Espelage and S. M. Swearer, "Research on school bullying and victimization: What have we learned and where do we go from here?," *School psychology review*, vol. 32, no. 3, pp. 365–383, 2003.

[20] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. 71, 2021.

[21] E. P. S. Baumer, "Toward human-centered algorithm design," *Big Data & Society*, vol. 4, no. 2, 2017.

[22] A. Fitro, M. A. Wibowo, and C. E. Widodo, "Automatic Detection of Cyberbullying on Text, Image, and Video: A Systematic Literature Review," *Jurnal SISFOKOM*, 2025.

[23] A. M. Alduaialj and A. Belghith, "Detecting Arabic cyberbullying tweets using machine learning," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 29–42, 2023.

[24] A. Akhter, U. K. Acharjee, M. A. Talukder, M. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for Bengali cyber bullying detection," *Natural Language Processing Journal*, 2023.

[25] T. Bin Abdur Rakib and L. K. Soon, "Using the Reddit Corpus for Cyberbully Detection," in *Lecture Notes in Computer Science*, vol. 10751, pp. 180–189, 2018.

[26] G. Njoyangwa and G. Justo, "Automated detection of bilingual obfuscated abusive words on social media forums: A case of Swahili and english texts," *Tanzania Journal of Science*, vol. 47, no. 4, pp. 1352–1361, 2021.

[27] C. Emmery, B. Verhoeven, G. De Pauw, G. Jacobs, C. Van Hee, E. Lefever, B. Desmet, V. Hoste, and W. Daelemans, "Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity," *Language Resources and Evaluation*, vol. 55, no. 3, pp. 597–633, 2021.

[28] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," Elsevier, 2023.

[29] U. Brandes, C. Reddy, and A. Tagarelli, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in *IEEE/ACM ASONAM*, pp. 479–486, 2018.

[30] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models: A Reproducibility Study," 2022.

[31] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context," in *Lecture Notes in Computer Science*, vol. 7814, pp. 693–696, 2013.

[32] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Lecture Notes in Computer Science*, vol. 8506, pp. 160–171, 2014.

[33] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis," in *SAM '14*, pp. 3–6, 2014.

[34] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," in *SIAM SDM*, pp. 235–243, 2019.

[35] M. Dadvar, D. Trieschnigg, and F. De Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies," in *Lecture Notes in Computer Science*, vol. 8436, pp. 275–281, 2014.

[36] D. Chatzakou, I. Leontiadis, J. Blackburn, E. De Cristofaro, G. Stringhini, A. Vakali, and N. Kourtellis, "Detecting Cyberbullying and Cyberaggression in Social Media," *ACM Transactions on the Web*, vol. 13, no. 3, pp. 1–51, 2019.

[37] A. K. Gautam and A. Bansal, "Effect of Features Extraction Techniques on Cyberstalking Detection Using Machine Learning Framework," *Journal of Advances in Information Technology*, 2026.

[38] Y. M. Ibrahim, R. Essameldin, and S. M. Saad, "Social Media Forensics: An Adaptive Cyberbullying-Related Hate Speech Detection Approach Based on Neural Networks With Uncertainty," *IEEE Access*, 2024.

[39] M. Tabassum and V. Nunavath, "RoBERTa+ViT for multimodal cyberbullying detection," 2026.

[40] K. D. Gorro, M. J. G. Sabellano, K. Gorro, C. Maderazo, and K. Capao, "Classification of Cyberbullying in Facebook Using Selenium and SVM," in *ICCCS 2018*, pp. 233–238, 2018.

[41] A. Marshan, F. N. M. Nizar, A. Ioannou, and K. Spanaki, "Comparing Machine Learning and Deep Learning Techniques for Text Analytics: Detecting the Severity of Hate Comments Online," *Information Systems Frontiers*, 2023.

[42] A. F. Alqahtani and M. Ilyas, "A MACHINE LEARNING ENSEMBLE MODEL FOR THE DETECTION OF CYBERBULLYING," *International Journal of Artificial Intelligence and Applications*, 2026.

[43] C. Nobata, J. R. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive Language Detection in Online User Content," in *WWW '16*, 2016.

[44] S. J. Bu and S. B. Cho, "A Hybrid Deep Learning System of CNN and LRCN to Detect Cyberbullying from SNS Comments," in *Hybrid Artificial Intelligence Systems*, pp. 561–572, 2018.

[45] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839–1852, 2023.

[46] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury, "A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–34, 2021.

[47] S. Kaur, S. Singh, and S. Kaushal, "Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data," *International Journal of Cognitive Computing in Engineering*, 2024.

[48] G. Jaradat, M. Shehab, D. Ibrahim, S. Najdawi, and R. Sihwail, "Deep Learning Approaches for Detecting Cyberbullying on Social Media," *Journal of Computational and Cognitive Engineering*, 2025.

[49] T. N. Harshitha, M. Prabu, E. Suganya, S. Sountharrajan, D. P. Bavirisetti, N. Gadde, and L. S. Uppu, "ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media," *Frontiers in Artificial Intelligence*, 2024.

[50] E. Y. Daraghmi, S. Qadan, Y. A. Daraghmi, and R. Yousu, "From text to insight: An integrated cnn-bilstm-gru model for arabic cyberbullying detection," *IEEE*, 2024.

[51] Z. Zhang, Y. Chen, and L. Wang, "Zero-shot cyberbullying detection with large language models: A comparative study," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, FL, USA, 2024, pp. 1123–1135.

[52] J. Liu, S. Park, and K. Nguyen, "Few-shot prompting for online harassment detection: GPT-4 on Twitter data," *arXiv preprint arXiv:2501.04231*, 2025.

[53] J. Wei, X. Zhou, and M. T. Ribeiro, "Chain-of-thought reasoning improves implicit cyberbullying detection," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 345–360, 2024.