International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

Detecting Hate Speech and Cyberbullying on Social Media: A Comparative Analysis of Machine Learning, Deep Learning, Transformer and Multimodal Approaches

Himanshu Singh[1], Mani Pareek[2]

[12]*Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur, India*

Abstract - *Hate speech and cyberbullying have become a growing issue with social media. As human moderators are limited in what they can read, many studies have emerged to address this issue using different methods. Twenty studies from 2024 to 2026 were compared to show how classical ML, DL, transformers, multimodal and LLM based agents can help. Data was available for English, Indonesian, Thai, Tamil, Urdu and Roman Urdu languages, as well as a 13 language corpus. All twenty studies were able to be grouped together into a six-stage pipeline. The results, shown in table 1, ranged from 86% to 97.5%. Many different types of methods were used which meant that the datasets and evaluation methods were more important than the model type. Multimodal approaches helped in video and meme tasks, gaining an additional 1.5% relative F1 over the best single modality model in Thai memes. Adding bystander roles helped improve the F1 score by 25.35% in one cyberbullying study. This isn't always the case when a stricter evaluation approach is taken. The precision score dropped to 1.9% when applied to a live bluesky stream. A graph hybrid had a large drop in score (94.55% to approximately 86% macro-F1) when the test nodes were hidden from the graph. Finally, possible issues and future directions are discussed.*

Key Words: Hate Speech Detection, Cyberbullying, Transformers, Multimodal, Low Resource Languages, Content Moderation.

1. INTRODUCTION

With social media being the main place for public discussion, there is an increase in hate speech and cyberbullying due to the open and anonymous nature of these platforms [10][15]. Hate speech is aggressive language targeted toward individuals or groups based on factors such as race, religion, gender or sexual orientation [18]. Whilst this differs slightly, cyberbullying is defined as repeated aggressive behaviours directed toward an individual who is unable to easily defend themselves using electronic means [20]. This can cause victims to suffer from stress, anxiety and depression, with some cases leading to self harm. On average 41% of adults in the U.S have been harassed online [20]. A report found that cyberbullying was the top threat faced by children in Thailand online (49%) [6].

Due to the high volumes of content, automatically detecting hate speech and cyberbullying is an important task within the field of Natural Language Processing (NLP) and Multimedia Analytics. Early systems used keyword lists and traditional machine learning models such as Naïve Bayes, SVM and Logistic Regression [7][11]. More recent works have used Recurrent and Convolution Neural Networks [9][14][15], Pre-Trained Transformers [3][18][19], Multi-modal models [1][6] and LLM based Agentic Systems [16]. Work has also been done on non-English languages such as Tamil, Urdu, Roman Urdu, Thai and Indonesian (Low Resource and Code-Mixed Languages) [4][16]-[18].

There are plenty of papers claiming a high accuracy (>90%), however there are many differences between the data sets, labels, languages and evaluation methods. This paper provides three key aspects. Firstly, twenty of the most recent studies are combined to form one detection pipeline (i). Secondly, the data sets, models and results are compared (ii). Finally, the issues with deploying systems to moderate are discussed (iii).

The main contributions of this paper can be summarised as follows:

- A single six-stage detection pipeline, built from twenty recent studies, that covers data collection, pre-processing, feature representation, model learning, classification and explanation (Section 2).
- A side-by-side comparison of classical ML, deep learning, transformer or LLM-based and multimodal approaches, based on the scores reported by the original authors (Sections 3 and 4).
- Discussion of the hidden factors behind the high level results (e.g. poor evaluation protocols, noisy labelling, class imbalance) and challenges and future research directions (see Sections 5 and 6).

The remainder of this paper is structured as follows. Section 2 outlines how the studies were selected and a unified pipeline was created. Section 3 provides a synthesis of the literature. Section 4 discusses the findings. Challenges and limitations are provided in Section 5. Future research directions are outlined in Section 6. Finally, Section 7 provides a conclusion.

2. METHODOLOGY

2.1 Selection of studies and data extraction

I reviewed 20 peer-reviewed papers published between 2024 and 2026. These papers were all published in IEEE conferences, IEEE Access and IEEE Intelligent Systems. Of these, 9 studies discussed hate speech [1]-[5], [16]-[19], 9 discussed cyberbullying [6]-[9], [11], [13]-[15], [20] and 2 studies were review papers [10], [12]. For each study, information was recorded regarding: the task, language, modality, amount of data, pre-processing steps, model type, evaluation metric, highest performing result and any limitations mentioned. No models were re-trained. All the numbers used within this paper are those which were stated by the original authors.

2.2 Unified detection pipeline

Nearly all systems followed the same 6 steps shown in Fig. 1. Data was first collected from a variety of sources including Twitter/X, Facebook, Instagram, Reddit and YouTube. Second, pre-processing was undertaken (cleaning, tokenisation, removing stop words, lemmatising text, normalising slang and emojis, transliteration, audio and image processing using ASR or OCR). Next, feature extraction was performed (TF-IDF, n-gram, GloVe/FastText, BERT contextual embeddings, audio features using MFCC and visual features using CLIP or ViT). After this stage, a model was learnt. Classification was performed (binary, multi-class or severity). Finally, explainability/counter measures were applied (LIME or SHAP, warnings, blocking, legal advice or counter speech).

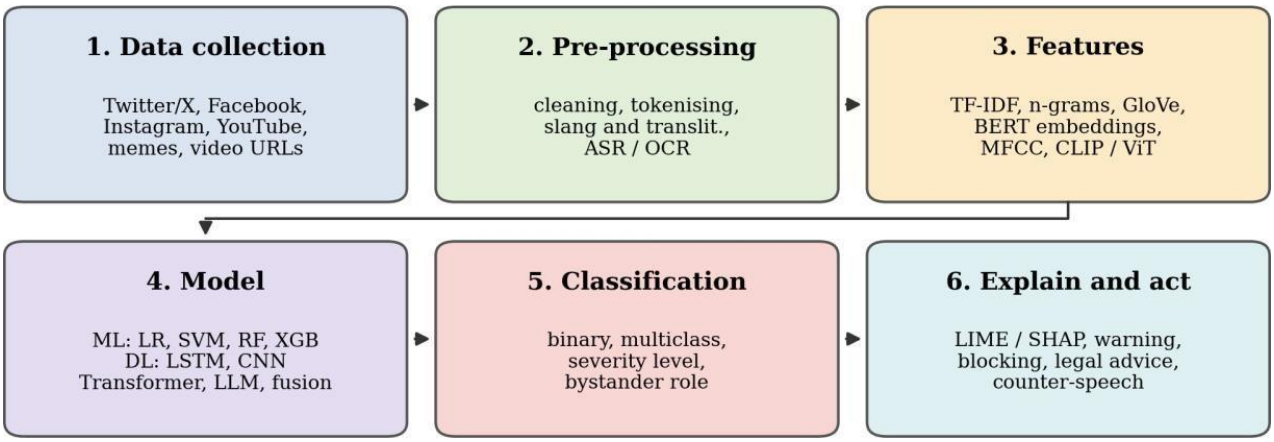

Fig. 1. Unified six-stage hate speech and cyberbullying detection pipeline built from the twenty reviewed studies.

2.3 Metrics

Almost every paper gives information on accuracy, precision, recall and F1 scores:

$$\text{Accuracy} = TP + TN \,/\, TP + TN + FP + FN$$

$$\text{Precision} = P = TP \,/\, TP + FP \quad \text{Recall} = R = TP \,/\, TP + FN$$

$$\text{F1 score} = 2PR \,/\, P + R \qquad (1)$$

2.4 Evaluation

The papers were classified into four categories: classical ML, deep learning (LSTM/CNN), transformer or LLM based and multimodal fusion. From every paper, the highest result in terms of accuracy or F1 scores were noted (Fig. 2). As all papers use different datasets and splits, this is just indicative of how well each model performed and not a true leaderboard.

3. LITERATURE SYNTHESIS

3.1 Classical Machine Learning

Despite the advances in deep learnings, logistic regression, support vector machines, naïve bayes and random forests are still good baseline models to use. These models were complemented with TF-IDF and feature selection. LR achieved 94% accuracy on 115,863 rows [7], 91% accuracy on a 6-class dataset of tweets [11] and RF achieved 94% accuracy after oversampling with SMOTE [14]. Classical models can still be optimised, such as XGBoost with adaptive particle swarm optimisation to achieve 97.52% accuracy and 97.62% F1 [8]. Bag of words models struggle to detect subtle abuse. For instance, RF achieved 78% accuracy but only 15% recall for non-bullying tweets [11].

3.2 Deep Learning

An LSTM model using an embedding layer from Keras achieved 96.64% accuracy on Twitter data, 94.49% accuracy on Instagram data and 91.26% accuracy on

Facebook data [15]. CyberGuard had a BiLSTM with 87.77% accuracy, the system also had additional layers to determine severity and include a legal advice chatbot [9]. An CNN-BiLSTM hybrid with GloVe failed to outperform an oversampled RF with 94% accuracy and only achieved 83% accuracy [14]. It is important to note that deeper models do not always outperform simpler models with small amounts of data.

3.3 Transformers and LLMs

The fine-tuned transformer models produced the most consistent results. RoBERTa achieved 92.8% accuracy [3], XLM-RoBERTa and mBERT both achieved 95.79% accuracy on Urdu, Roman Urdu and English [18]. A BERT-LSTM classifier chain achieved 89% F1 on fine-grained cyberbullying [20]. This study also found that including bystander roles (defender, instigator, neutral) helped improve F1 by 25.35%. There is now beginning to be some work looking at LLM agents. IRAIVI is a detection system that identifies targets, includes RAG for legal advice and generates counter speech in Tamil. It used Tree-of-Thought prompting and achieved an F1 score of 0.91 with GPT-3.5 and 0.85 with Qwen-4B [16]. However, a review in [10] argued that BERT based models (98-99%) outperformed ML/DL models (94-96%). However, these were different datasets.

3.4 Multimodal

Hate and bullying often comes through the tone, speech and images in the text. AGMF uses text (Whisper/Sentence-BERT), audio (MFCC) and image (CLIP) information and combines them using a learnable sigmoid gate. This achieved 91.33% accuracy on the HateMM dataset compared to 71-86% for classical classifiers [1]. Text only models outperformed image only models on Thai memes. WangchanBERTa combined with ViT-B/32 using late fusion achieved the highest bully F1 score of 0.861. Text only models achieved a F1 score of 0.845 compared to image only models at 0.726 [6]. Systems have been developed to detect hate in memes, images, videos and audio. An image model achieved an F1 score of 84.21% [13].

3.5 Multilingual and Low Resource

Transliteration and code mixing was required to handle the data in Tamil [16], Urdu and Roman Urdu [17][18], Thai [6] and Indonesian [4]. The corpus of 11,245 video URLs from YouTube [17] achieved an accuracy of 0.90 with classical models on Urdu and 0.96 with DL on Roman Urdu. A study on 13 languages achieved an F1 score of 0.95 on Roman Urdu, 0.90 on GermEval-2018 and 0.93 on GermEval-2021. LIME is used to explain the models [19]. An IndoBERT-GCN hybrid achieved a macro-F1 score of 94.55% using a transductive approach [4]. However, this decreased to around 85-87% using an inductive approach. This was similar to IndoBERT on its own (86.18%).

3.6 Explainable, Ethical and Intervention

Some systems include explanations for why the text was flagged. MUST and the 13 language study both include explainable AI (XAI) to help moderators understand why a text was flagged [18][19]. SHAP was used to show that their models rely heavily on explicit identity words and fail to detect negated hate [2]. Some systems provide interventions after detection. Pre posting warnings, severity and temporary bans [9][13]. Others include legal advice and counter speech, with unclear cases being sent to human reviewers [16]. This reflects the ethical concern of losing genuine speech due to false positives.

TABLE I. SUMMARY OF THE TWENTY REVIEWED STUDIES (SCORES IN %, AS REPORTED BY THE AUTHORS)

Ref.	Task / language / data	Method	Best reported result
[1]	Video hate speech; HateMM (750 videos)	Whisper+SBERT, MFCC, CLIP with gated fusion	Acc 91.33; F1 91.39
[2]	Sports-domain hate; MetaHate, EURO 2024 Bluesky/YouTube	HateBERT, DistilBERT, DeBERTa; confident learning, SHAP	Acc 90.5 (MetaHate); Bluesky P 1.9
[3]	English social media hate	BERT, RoBERTa, DistilBERT, XLNet, T5, ELECTRA; LR	RoBERTa Acc 92.8; tuned 94.89
[4]	Indonesian hate; 14,306 posts	IndoBERT + GCN hybrid	Macro-F1 86.18 (ind.); 94.55 (trans.)
[5]	Text, image and meme hate	CNN + BERT with OCR, fusion	F1 89 (text); 86 (image/meme)
[6]	Thai meme cyberbullying; 5,974 images	WangchanBERTa + ViT-B/32 late fusion	Bully F1 86.1
[7]	Cyberbullying text; 115,863 rows	Naïve Bayes, LR, SVM	LR Acc 94
[8]	Cyberbullying among students	XGBoost + adaptive PSO; IG, Chi-square	Acc 97.52; F1 97.62
[9]	Cyberbullying + severity + legal chatbot	BiLSTM	Acc 87.77; F1 87.81

Ref.	Task / language / data	Method	Best reported result
[10]	Review: generative AI for cyberbullying	GAN, VAE, diffusion, transformers / LLMs	Review (BERT 98–99 vs ML/DL 94–96)
[11]	Multiclass cyberbullying; 41,983 tweets	LR, Naïve Bayes, RF	LR test Acc 91
[12]	Review: ML for cyberbullying	SVM, NB, DT, RF; feature engineering	Review (no experiment)
[13]	Text, image, video, audio cyberbullying + prevention	RF+TF-IDF, VGG16, CNN-GRU, BERT severity	Text Acc 91.97; image F1 84.21
[14]	Multiclass cyberbullying; 47,692 tweets	Six ML models, CNN-BiLSTM, GloVe, SMOTE	RF Acc 94; CNN-BiLSTM 83
[15]	Cyberbullying on Twitter, Instagram, Facebook	LSTM with embedding layer	Acc 96.64 / 94.49 / 91.26
[16]	Tamil gender-based hate + counter-speech	Agentic LLMs (ToT, RAG, GA prompts)	F1 0.91 (GPT-3.5); CS quality 0.86
[17]	Urdu / Roman Urdu misogyny; 11,245 video URLs	Classical ML, DL, BERT	Acc 0.90 (Urdu); 0.96 (Roman)
[18]	Urdu, Roman Urdu, English hate (MURE)	XLM-RoBERTa, mBERT, XAI	Acc/F1 95.79
[19]	Hate speech in 13 languages	ML, DL, transformers, FastText+SVM, LIME	F1 0.95 (Roman Urdu); 0.90–0.93 (German)
[20]	Fine-grained cyberbullying + bystander roles	BERT + LSTM classifier chain	F1 89 (+25.35% with bystanders)

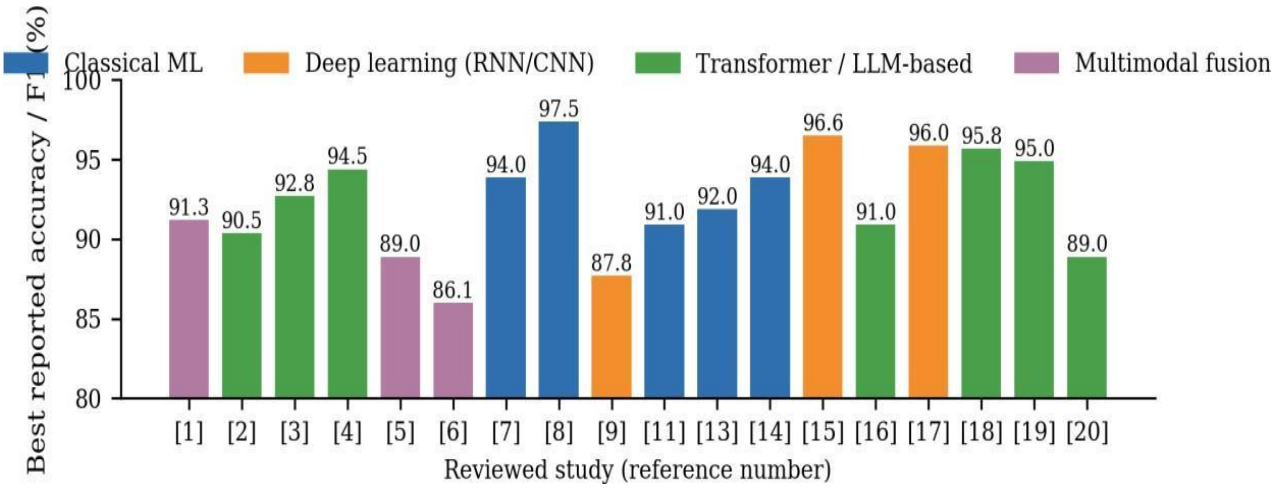

Fig. 2: Best accuracy achieved/f1 achieved for the experimental studies, grouped by method family. The y-axis is not zeroed but begins at 80%. Note that [10] and [12] have been excluded. [4] is the score on the transductive data (around 86% on the inductive).

4. RESULTS AND DISCUSSION

In table II, the headline results have been grouped based upon the technique family used. All four families achieve similar scores between ~86% and 97.5%. However, it can be seen that classical machine learning techniques outperform the others (91%-97.5%). This is largely due to most of the results being achieved using single relatively clean benchmarks with often extensive oversampling and fine-tuning. The multimodal techniques receive lower results due to them being applied to more difficult video and meme data.

TABLE II. HEADLINE SCORES BY METHOD FAMILY

Family	Studies	Range (%)	Observation
Classical ML	[7],[8],[11],[13],[14]	91.0–97.5	Strong on clean benchmarks; weak on implicit abuse and minority classes
Deep learning	[9],[15],[17]	87.8–96.6	Good with embeddings; no gain over tuned RF on small data [14]
Transformer / LLM	[2]–[4],[16],[18]–[20]	89.0–95.8	Most consistent across languages; supports context and explanation
Multimodal fusion	[1],[5],[6]	86.1–91.3	Beats unimodal baselines on video/memes; needs richer data

4.1 Evaluation protocol matters more than architecture.

While HateBERT achieves a high accuracy score of 90.5% on the MetaHate dataset, it performs poorly on real-world data such as Bluesky (1.9% precision, only five posts contain clear hate). It also performed poorly on YouTube (macro-F1 = 0.53) [2]. The graph hybrid approach in [4] also suffers from poor performance when the test nodes are removed from the graph (~8-9 point decrease in performance). In both cases the headline score came from a favourable test setting, and the weaker result appeared only when the evaluation was made stricter. Scores from different studies should therefore be read together with how the test data was collected and split.

4.2 Label quality limits every model.

The confident learning stage identified 75,483 MetaHate labels to be potentially incorrect.

These were manually checked on a sample of 600 cases. It was found that 12.7% of the labels were incorrect. If the highest 25% of the cases identified were removed from consideration, 92.1% accuracy could be achieved compared to the original 89.7% [2]. This shows that part of the reported error comes from the data rather than from the model. Removing the top 25% of label issues raised accuracy from 89.7% to 92.1% (Table III), so cleaning the labels helped without changing the model.

4.3 Context and modality add real value.

Various multi-modal approaches such as bystander roles (F1+25.35%) [20], late fusion (Text-Image fusion) (relative F1+1.5%) [6] and learned gates [1] outperformed the individual modalities (Table III). However, in the case of Thai meme harm, the text-based model is slightly better (0.955 vs. 0.948) [6]. The size of the gain depends on the task: it was large for bystander roles (F1+25.35%), small for late fusion on Thai memes (relative F1+1.5%), and the text-only model was slightly ahead for Thai meme harm.

TABLE III. REPORTED GAINS FROM CONTEXT, MODALITY OR DATA QUALITY

Study	Intervention	Reported effect
[1]	Gated tri-modal fusion	Acc 91.33 vs 85.71 for best classical baseline
[6]	Text + image late fusion	Bully F1 0.861 vs 0.848 (best unimodal)
[20]	Bystander-role features	+25.35% F1; final F1 89
[2]	Remove top 25% label issues	Acc 89.7 to 92.1
[4]	Global graph (transductive)	Macro-F1 94.55 vs about 86 inductive

4.4 Accuracy can be misleading

The data is imbalanced – In the case of Thai memes, 92.8% of the data is non-harmful [6]. In [11] they achieved an accuracy of 78%, but achieved a recall of 15% on one of the classes. Allowing results to be seen in terms of precision and recall for each class and macro-F1. For this reason, accuracy alone should not be used to compare systems on imbalanced data, and per-class precision, recall and macro-F1 give a more honest picture.

4.5 Results for low resource languages

Thanks to multilingual transformers, improved transliteration handling and new corpora results have improved to between 0.90-0.96 on Urdu and Roman Urdu [17]-[19]. Tamil is also starting to be covered using LLM agents [16]. These results are encouraging, but they were reported on different datasets and scoring methods, so they show progress for these languages rather than a direct ranking of the models.

4.6 Deployment considerations

Overall it suggests that if you are developing or assessing a moderation system then you should ensure that: your results show precision, recall and macro-F1 for each class. You should test on data from a different source or time than your training data. If you are using a graph to test then ensure the test nodes/posts are hidden.

Ensure labelling is audited to understand where the model may be failing. Human-in-the-loop labelling will be required for ambiguous examples. However multimodal models provide a good investment in video and meme modalities. A well tuned classical baseline provides a fair comparison on clean text. Reporting the split, the test source and the metrics for every class would also make future results easier to reproduce and compare, which links back to the reproducibility problem discussed in Section 3.

5. CHALLENGES AND LIMITATIONS

There are several challenges with this task. (i) Labelling can be subjective and ambiguous whether a statement is considered hate speech or just offensive. (ii) Implicit hate speech, sarcasm, negation, cryptic slurs and code mixing are challenging to detect [2][16]. (iii) Generalisation across different platforms and over time is difficult. Real data has very little hate speech [2]. (iv) Multimodal data is limited. Only 750 videos in HateMM [1]. (v) Poor reproducibility. Many papers do not publish their splits or code. Some have inconsistencies, for example the accuracy on Instagram in [15] is reported as 94.49% and 95.49%. (vi) Ethical concerns. False positives may stifle real speech. Automatic moderation should be balanced with human-in-the-loop [5][13].

6. FUTURE RESEARCH

Considering the issues discussed, the most promising future research directions are: (1) Common inductive benchmarks with fixed splits and class specific metrics across platforms. (2) Larger multimodal and video corpora. Transliteration aware multi-lingual pretraining. (3) Conversation level models using bystander and thread context [20]. (4) Explainable and calibrated models with human-in-the-loop escalation for ambiguous examples [16][18][19]. (5) Safe large language models or agents as moderators with retrieval, severity scoring and counter speech [9][16]. (6) Regular auditing of labelling quality [2]. Testing for fairness across communities.

7. CONCLUSION

This paper investigated twenty recent studies of hate speech and cyberbullying detection using a common six stage pipeline. Transformers provided the most consistent results. Multimodal fusion helps with video and meme modalities. Classical ML models also provide a competitive baseline. The high headline accuracies (86%-97.5%) are not comparable and may overstate the performance due to the significant drop in performance when evaluated more strictly. More accurate labelling, realistic benchmarking across platforms, more context and multimodal signals, and open human oversight.

REFERENCES

1) M. S. Lekshmi, S. Malathy, S. S. Suni, A. P. Mathew, S. Babu, and A. Muhammed, “Adaptive gated multimodal fusion network for video-based hate speech detection,” in Proc. ICIET, 2026.

2) B. Keehnen, S. K. Metin, Ç. Kentmen-Çin, H. Aka, and A. A. Salah, “Hate speech detection on social media: A case study on the sports domain,” in Proc. SIU, 2026.

3) R. K. Ayyasamy et al., “A comparative study of fine-tuned transformer models for social media hate speech detection,” in Proc. IEEE ICIH, 2025.

4) S. P. Harlin and Fitriyani, “Hate speech detection on social media using a hybrid IndoBERT and graph convolutional network model,” in Proc. ICCRAIDS, 2026.

5) K. Pushpa Rani et al., “Hate speech detection based on deep learning by fusion approach,” in Proc. ASSIC, 2025.

6) S. Phosit, S. Kongsamlit, and K. Pasupa, “Detecting cyberbullying in Thai memes: A multimodal approach using deep learning,” in Proc. IEEE CINLP-SoMe, 2025.

7) A. R. Tripathi et al., “A data-driven retrospective analysis of cyberbullying using machine learning algorithms,” in Proc. AUTOCOM, 2025.

8) P. J. Vijayakumar et al., “Detecting and preventing cyberbullying among higher education students on social media with XGBoost and particle swarm optimization,” in Proc. GINOTECH, 2025.

9) N. Benny et al., “CyberGuard: Cyberbullying detection system using BiLSTM for enhancing digital safety and justice in online communities,” in Proc. ACCESS, 2025.

10) A. Khan et al., “Generative artificial intelligence (GAI) models for cyberbullying detection and prevention,” in Proc. SATC, 2025.

11) S. Gupta et al., "Cyber bullying detection and classification using machine learning algorithms," in Proc. CYBERCOM, 2024.

12) A. Toktarova, D. Sultan, and Z. Azhibekova, "Review of machine learning models in cyberbullying detection problem," in Proc. IEEE SIST, 2024.

13) V. J. Sharmila et al., "Cyberbullying detection and prevention in social media using machine learning and deep learning techniques," in Proc. DABCon, 2024.

14) P. G. Balaji, P. P. Katariya, S. S, and M. Venugopalan, "Cyberbullying detection on multiclass data using machine learning and a hybrid CNN-BiLSTM architecture," in Proc. ICKECS, 2024.

15) M. H. Obaida, S. M. Elkaffas, and S. K. Guirguis, "Deep learning algorithms for cyber-bulling detection in social media platforms," IEEE Access, vol. 12, pp. 76901–76908, 2024.

16) K. Pannerselvam and S. Rajiakodi, "IRAIVI: Agentic framework for Tamil hate speech detection and counter-speech generation toward responsible AI intervention," IEEE Access, vol. 14, pp. 132668–, 2026.

17) Z. A. Khan, Y. Xia, F. Khaliq, W. Jiang, and M. S. Anwar, "AI-based hate speech detection system using video URLs for effective content moderation," IEEE Intell. Syst., Nov./Dec. 2025.

18) I. Hussain, M. R. Rizvi, Z. Abbas, A. N. Cheema, and I. M. Almanjahie, "MUST: An explainable AI-based framework for multilingual hate speech detection," IEEE Access, vol. 13, pp. 202758–, 2025.

19) E. Hashmi et al., "Enhancing multilingual hate speech detection: From language-specific insights to cross-linguistic integration," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3452987.

20) H. S. Alfurayj, S. L. Lutfi, and R. Perumal, "A chained deep learning model for fine-grained cyberbullying detection with bystander dynamics," IEEE Access, vol. 12, pp. 105588–105604, 2024.