ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

Cyberbullying Detection Using Machine Learning and Deep Learning Techniques: A Review

Sana Bano Khan[1], Dr. Mohammad Nasiruddin[2]

PG. Scholar, Professor
Department of Electronics and Telecommunications Engineering

Anjuman College of Engineering and Technology, Nagpur, Maharashtra, India

Abstract

The explosive growth of social media has made cyberbullying a widespread and serious concern, particularly for adolescents and young adults, with documented links to anxiety, depression, and, in extreme cases, self-harm. Because manual moderation cannot scale to the volume of content generated every day, automated detection using machine learning (ML) and deep learning (DL) has become an active research area. This paper reviews the evolution of cyberbullying detection research, from early lexicon- and rule-based methods through classical supervised classifiers such as Naïve Bayes, Support Vector Machines, and Random Forest, to modern deep architectures including Convolutional Neural Networks, Recurrent Networks, and transformer-based language models such as BERT. We summarise the typical detection pipeline, compare the strengths and weaknesses of the major algorithm families, catalogue commonly used public datasets, and discuss the open challenges that remain — including context and sarcasm understanding, multilingual and code-mixed text, multimodal (text-image-video) content, class imbalance, and real-time, cross-platform deployment. The review is intended to orient new researchers and practitioners toward promising directions for building more robust, fair, and scalable cyberbullying detection systems.

Keywords: *Cyberbullying detection, machine learning, deep learning, natural language processing, social media, text classification, hate speech, transformers.*

Introduction

Social media platforms have transformed the way people communicate, but the same anonymity and reach that make them powerful also make them a fertile ground for online harassment. Cyberbullying — repeated, intentional harm delivered through digital channels — is widely regarded as more damaging than traditional bullying because abusive content can be viewed by an unbounded audience, can persist indefinitely, and can reach a victim at any time or place. Global surveys consistently report that a substantial share of teenagers have experienced some form of online harassment, and the psychological consequences range from anxiety and depression to, in severe cases, self-harm and suicidal ideation.

Because the volume of user-generated content on platforms such as Twitter/X, Instagram, and Facebook makes manual moderation impractical, researchers have turned to automated approaches. Early systems

relied on keyword and lexicon matching, which is simple but brittle: it misses paraphrased abuse, sarcasm, and evolving slang, and it produces many false positives on benign text that merely contains a flagged word. This limitation motivated a shift toward statistical machine learning, where a classifier is trained on labelled examples to recognise the linguistic patterns associated with bullying, and more recently toward deep learning and transformer-based language models that capture contextual meaning rather than surface keywords.

This paper reviews the machine-learning and deep-learning literature on cyberbullying detection published over roughly the last decade. Section 2 summarises representative studies. Section 3 presents a generic block diagram of a detection system. Section 4 describes the typical methodology pipeline, including the flow of data collection through classification, and compares the major algorithm families. Section 5 lists public datasets used in this domain. Section 6 discusses evaluation metrics. Section 7 outlines open challenges, Section 8 suggests future directions, and Section 9 concludes the review.

Literature Survey

Table 1 summarises a representative cross-section of cyberbullying detection studies spanning classical machine learning to graph- and transformer-based approaches. The studies illustrate a clear trajectory: early work (2017–2019) relied almost exclusively on bag-of-words features with Naïve Bayes, SVM, or Decision Tree/Random Forest classifiers; work from 2018 onward increasingly incorporated deep architectures (CNNs, then RNN/LSTM) to capture local and sequential text patterns; and the most recent studies (2022–2024) exploit transformer-based contextual embeddings, code-mixed and multilingual text handling, and graph-based modelling of user interaction networks.

Author(s) /Ref.)	Year	Algorithm(s	Data Source	Key Finding
Shin & Berrios [1]	2021	SVM	Twitter	SVM performs well in high-dimensional feature spaces but needs careful kernel tuning.
Kumar et al. [2]	2021	Naïve Bayes, SVM	Social media comments	Combining NB with SVM reduced overfitting and improved robustness on decision-tree baselines.
Islam et al. [3]	2021	Decision Tree, NB, RF, SVM	Social network posts	SVM outperformed the other classical classifiers on the benchmark set used.
Balakrishnan et al. [4]	2020	Naïve Bayes, Random Forest	Twitter (psychological features)	Random Forest gave higher accuracy than Naïve Bayes when psychological/user features were added.
Dalvi et al. [5]	2020	Naïve Bayes, SVM	Twitter	SVM reached ~71.25% accuracy, outperforming the Naïve Bayes baseline.
Hani et al. [6]	2019	SVM, Neural Network	Social media text	A shallow neural network outperformed SVM on the same feature set.
Al-Ajlan & Ykhlef [7]	2018	CNN	Twitter	A 4-layer CNN with a dense classification layer improved detection over shallow ML models.
Al-Ajlan & Ykhlef [8]	2018	CNN (optimised)	Twitter	Hyper-parameter optimisation of the CNN reduced training cost but increased parameter count.
Foong & Oussalah [9]	2017	Rule-based + ML hybrid	Multiple platforms	Early hybrid system for cyberbullying analysis; limited to lexical cues.
Balakrisnan & Kaity [10]	2023	Systematic review (68 studies)	Multi-platform, multi-language	Random Forest and SVM were the most frequently reported best performers across 2011–2022 literature.

Elsafoury et al. [15]	2021	Survey (pipeline stages)	Multi-platform	Identified pre-processing and feature representation as the biggest sources of performance variance.
Maity et al. [14]	2022	Transformer (code-mixed)	Hinglish social media	Adding emoji/sentiment cues to transformer models improved detection in code-mixed text.
Wang et al. [17]	2020	Graph Convolutional Network (SOSNet)	Twitter	Graph-based modelling of user interactions improved fine-grained bullying-role detection.
Yi & Zubiaga [18]	2024	Transformer (debiasing)	Session-based social media	Proposed a data-independent debiasing method to reduce unfair transformer predictions.

Table 1: Summary of Representative Literature on Cyberbullying Detection

System Architecture

Despite differences in the specific algorithms used, most cyberbullying detection systems reported in the literature share a common high-level architecture, shown in Figure 1. Raw text (and, increasingly, image or video content) is collected from a social platform, cleaned and normalised, converted into a numerical feature representation, passed to a trained classifier, and the resulting bullying/non-bullying decision is used to flag, filter, or moderate the content. A feedback loop allows misclassifications identified during moderation to be fed back into the training set, supporting periodic retraining as language and platform behaviour evolve.

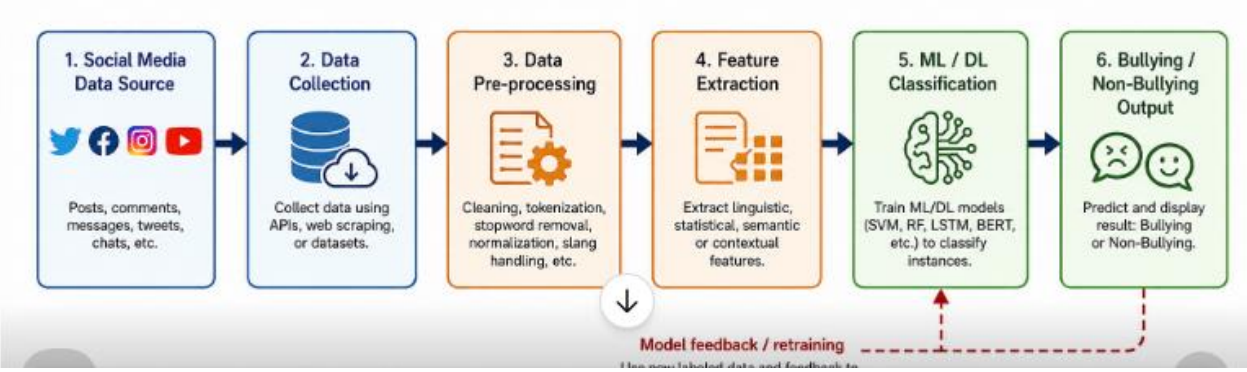

Figure 1: Block Diagram of a Generic Machine-Learning-Based Cyberbullying Detection System

Detection Methodology

Figure 2 expands the block diagram of Section 3 into a step-by-step flow diagram of the methodology commonly followed in the reviewed literature, using the pipeline adopted in this project as a concrete example (a Naïve Bayes / Logistic Regression classifier trained on a Kaggle-sourced harassment dataset).

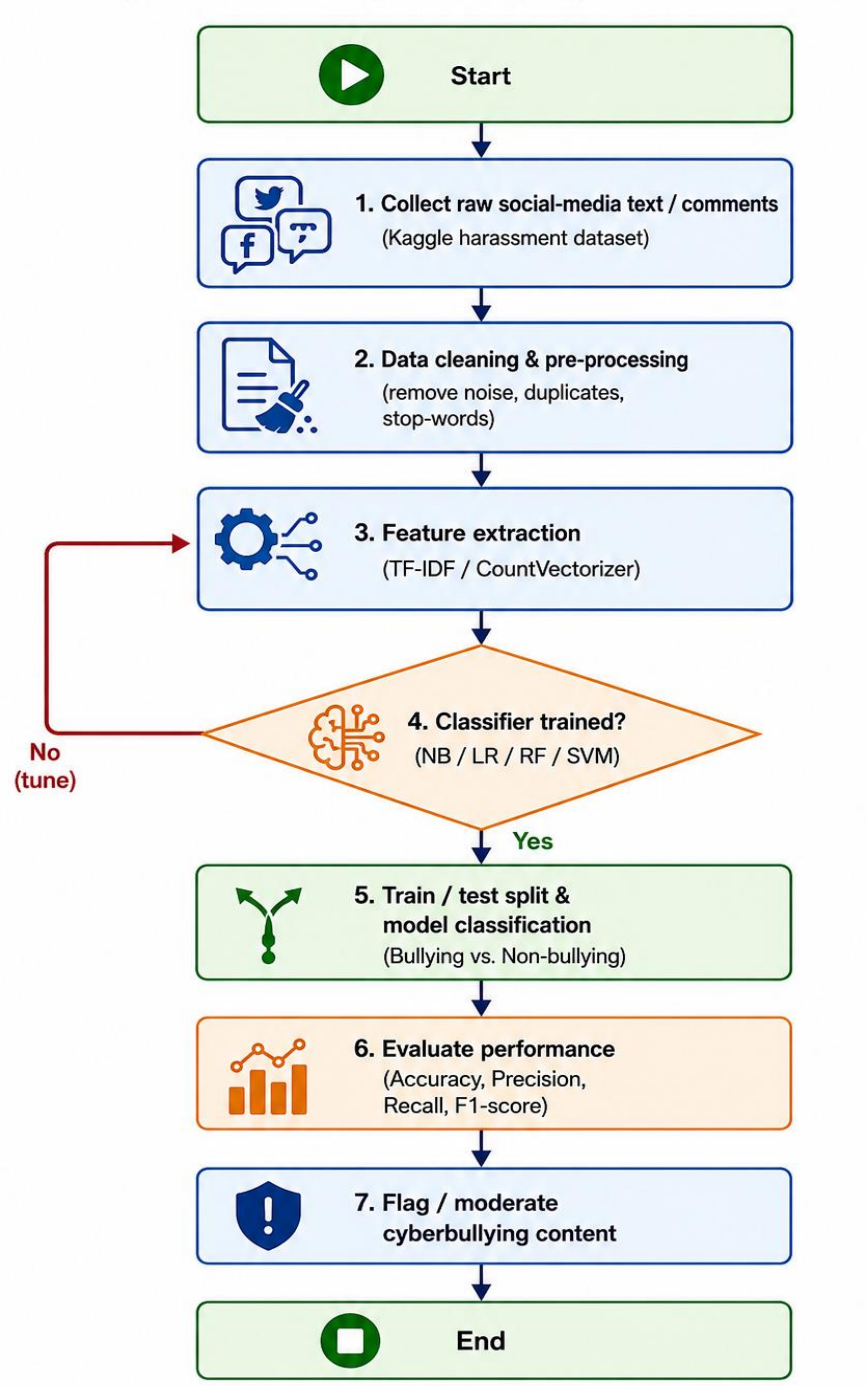

Figure 2: Flow Diagram of the Cyberbullying Detection Methodology

4.1 Data Collection

Data collection is the foundation of any ML-based system: the model can only be as good as the data it is trained on. Publicly available, pre-labelled corpora (Section 5) are commonly used so that results are comparable across studies. In the pipeline reviewed here, a Kaggle-hosted harassment dataset containing

several thousand social-media comments, each labelled as aggressive or non-aggressive, was used as the training source.

4.2 Data Pre-processing

Raw social-media text is noisy: it contains URLs, punctuation, emojis, misspellings, and duplicate entries. Pre-processing steps typically include lower-casing, removal of special characters and stop-words, de-duplication, and tokenisation. Libraries such as NumPy and pandas are commonly used for the underlying data handling in Python-based pipelines, with regular-expression-based cleaning routines used to strip unwanted symbols before the text is passed to feature extraction.

4.3 Feature Extraction

Since ML models operate on numerical inputs, extracted text must be vectorised. Classical pipelines commonly use Term Frequency–Inverse Document Frequency (TF-IDF) or CountVectorizer-style bag-of-words representations, which weight words by how distinctive they are to a document relative to the overall corpus. Deep-learning pipelines instead use dense word or sub-word embeddings (Word2Vec, GloVe, or contextual embeddings from transformer models), which better capture semantic similarity between words.

4.4 Classification Algorithms

A wide range of algorithms has been applied to cyberbullying detection, each with different trade-offs between accuracy, interpretability, and computational cost. Table 2 compares the algorithm families most frequently reported in the literature reviewed in Section 2.

Algorithm	Complexity	Advantage	Limitation	Typical Reported Accuracy
Naïve Bayes	Low	Fast, simple, effective on small text datasets	Assumes feature independence; weaker on nuanced/sarcastic text	~60–75%
Logistic Regression	Low	Interpretable, computationally cheap	Struggles with complex, non-linear language patterns	~75–85%
Support Vector Machine (SVM)	Medium	Strong in high-dimensional sparse feature spaces	Kernel/parameter tuning needed; slower on very large data	~71–90%
Random Forest	Medium	Reduces overfitting, handles mixed features well	Less interpretable; can be memory-intensive	~78–90%
CNN	High	Captures local text patterns (n-gram-like features)	Needs larger labelled datasets; more compute	~80–92%
RNN / LSTM	High	Captures sequential/contextual dependencies	Slower to train; can struggle with very long sequences	~80–91%
Transformer (BERT and variants)	Very High	State-of-the-art contextual language understanding, transfer learning	Computationally expensive; needs GPU resources and fine-tuning	~85–95%

Table 2: Comparative Analysis of Machine Learning and Deep Learning Algorithms for Cyberbullying Detection

In the specific pipeline referenced in this review, a Naïve Bayes classifier (GaussianNB, from scikit-learn's sklearn.naive_bayes module) was first applied and achieved close to 60% accuracy; Logistic Regression, evaluated subsequently alongside Random Forest, Decision Tree, and Gradient Boosting classifiers, achieved the best result at over 80%, consistent with the broader literature trend that ensemble and regression-based models tend to outperform a plain Naïve Bayes baseline on this task.

Datasets

Table 3 lists commonly referenced public datasets for cyberbullying and related abusive-language detection research. Dataset choice materially affects reported accuracy figures, since platforms, label definitions, class balance, and language coverage differ considerably between corpora — a key reason that accuracy numbers should not be compared directly across studies that use different datasets.

Dataset	Source / Platform	Approx. Size	Label Scheme
Kaggle Harassment (Parsed) Dataset	Twitter/social comments	~8,800	Binary (aggressive / not aggressive)
Formspring / Kaggle Bullying Corpus	Q&A social platform	~12,000	Binary (bullying / not bullying)
Wikipedia Talk Pages (Wulczyn et al.)	Wikipedia editor comments	~100,000+	Toxicity / personal attack labels
Hate Speech and Offensive Language (Davidson et al.)	Twitter	~25,000	Hate speech / offensive / neither
OLID / OffensEval	Twitter	~14,000	Offensive language, target, type

Table 3: Publicly Available Datasets Used in Cyberbullying Detection Research

. Performance Evaluation Metrics

Because cyberbullying datasets are typically imbalanced (non-bullying comments greatly outnumber bullying ones), accuracy alone can be a misleading metric — a classifier that always predicts "non-bullying" can still score a high accuracy while being practically useless. The literature therefore commonly reports a fuller set of metrics:

- Precision — the proportion of comments flagged as bullying that are genuinely bullying.
- Recall (Sensitivity) — the proportion of actual bullying comments that the system correctly identifies.
- F1-score — the harmonic mean of precision and recall, useful for imbalanced classes.
- Confusion matrix — a breakdown of true/false positives and negatives, used in this project's Naïve Bayes evaluation to understand what fraction of comments the model predicted correctly.
- ROC-AUC — the model's ability to discriminate between classes across decision thresholds.

Challenges and Open Research Issues

Despite steady progress, several challenges continue to limit real-world deployment of automated cyberbullying detection:

- Context and sarcasm: Bullying is frequently implicit, sarcastic, or dependent on conversational history, which purely text-level, single-message classifiers can miss.
- Evolving language: Slang, abbreviations, and coded terms change quickly, causing models trained on older data to degrade over time.
- Multilingual and code-mixed text: Most high-performing models and datasets are English-centric; regional languages and code-mixed text (e.g., Hinglish) remain comparatively under-served.
- Multimodal content: A growing share of harmful content is conveyed through images, memes, and video rather than text alone, requiring vision and OCR components in addition to NLP.
- Class imbalance and bias: Bullying examples are a minority class in most corpora, and models can inherit demographic or dialectal biases present in the training data.
- Real-time, cross-platform deployment: Production systems must classify content with low latency at scale, and must generalise across platforms with different content styles and moderation policies.

Future Research Directions

Based on the gaps identified above, promising directions for future work include: fine-tuning multilingual transformer models (e.g., mBERT, XLM-R, or region-specific models such as MuRIL) for low-resource languages and code-mixed text; combining text, image, and OCR-extracted text in unified multimodal architectures for meme- and image-based abuse; incorporating conversational/session-level context rather than classifying isolated messages; applying explainable-AI techniques so moderators can understand why content was flagged; and designing lightweight, distilled models suitable for real-time, on-device or edge deployment in messaging and gaming applications.

Conclusion

Automated cyberbullying detection has progressed substantially over the past decade, moving from simple keyword matching to classical machine learning and, most recently, to deep and transformer-based architectures capable of modelling context and semantics. This review has summarised the representative literature, presented a generic system architecture and detection methodology, compared the leading algorithm families, and catalogued public datasets used in the field. While classical classifiers such as Logistic Regression, SVM, and Random Forest remain competitive baselines — as reflected in the Logistic Regression pipeline reviewed here — deep and transformer-based approaches offer clear gains in contextual understanding, at the cost of greater computational demand. Closing the remaining gaps in multilingual coverage, multimodal content handling, and real-time deployment represents the most valuable direction for future research and for the practical goal of making online spaces safer.

References

[1] C. Shin and G. Berrios, "Improving cyberbullying detection using Twitter," 2021.

[2] R. Kumar, S. Parakh, and C. N. S. Vinoth Kumar, "Detection of cyberbullying using machine learning," 2021.

[3] M. M. Islam, M. A. Uddin, L. Islam, and A. Akter, "Cyberbullying detection on social networks using machine learning approaches," 2021.

[4] V. Balakrishnan, S. Khan, and H. Arabnia, "Improving cyberbullying detection using Twitter users' psychological features and machine learning," Computers & Security, 2020.

[5] R. R. Dalvi, S. B. Chavan, and A. Halbe, "Detecting a Twitter cyberbullying using machine learning," in Proc. 4th Int. Conf. Intelligent Computing and Control Systems (ICICCS), 2020, pp. 297–301.

[6] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 10, no. 5, 2019.

[7] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," Int. J. Adv. Comput. Sci. Appl. (IJACSA), 2018.

[8] M. A. Al-Ajlan and M. Ykhlef, "Optimized Twitter cyberbullying detection based on deep learning," 2018.

[9] Y. J. Foong and M. Oussalah, "Cyberbullying system detection and analysis," 2017.

[10] V. Balakrisnan and M. Kaity, "Cyberbullying detection and machine learning: a systematic literature review," Artificial Intelligence Review, vol. 56, pp. S1375–S1416, 2023.

[11] S. Suleiman, P. Taneja, and A. Nainwal, "Cyberbullying detection on Twitter using machine learning: a review," 2022.

[12] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop on Natural Language Processing for Social Media, 2017, pp. 1–10.

[13] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: a survey," arXiv:2207.10639, 2022.

[14] K. Maity, S. Saha, and P. Bhattacharyya, "Emoji, sentiment and emotion aided cyberbullying detection in Hinglish," IEEE Trans. Computational Social Systems, 2022.

[15] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, "When the timeline meets the pipeline: a survey on automated cyberbullying detection," IEEE Access, vol. 9, pp. 103541–103563, 2021.

[16] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," arXiv:1703.04009, 2017.

[17] J. Wang, K. Fu, and C.-T. Lu, "SOSNet: a graph convolutional network approach to fine-grained cyberbullying detection," in Proc. IEEE Int. Conf. Big Data, 2020, pp. 1699–1708.

[18] P. Yi and A. Zubiaga, "ID-XCB: data-independent debiasing for fair and accurate transformer-based cyberbullying detection," arXiv:2402.16458, 2024.

[19] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," Procedia Computer Science, vol. 181, 2021.

[20] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, "Automatic detection of cyberbullying and abusive language in Arabic content on social networks," 2021.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.

[22] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury, "A human-centered systematic literature review of cyberbullying detection algorithms," Proc. ACM Hum.-Comput. Interact., vol. 5, no. CSCW2, 2021.

[23] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on Web 2.0," in Proc. Content Analysis in the WEB 2.0, 2009, pp. 1–7.

[24] H. Allwaibed, M. Anbar, S. Manickam, and A. Bintang, "Cyberbullying detection approaches for Arabic texts: a systematic literature review," Frontiers in Artificial Intelligence, 2025.