IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Detection of Cyberbullying and Hate Speech on Social Media Platforms using Machine Learning A Review of Techniques, Challenges, and Future Directions

Dr. Santosh Barku Pawar[1], Prof. S. R. Phapale[2], Miss. Puja Goraksh Nalage[3]
Department of Master of Computer Applications (MCA)
Vidya Niketan Institute of Management and Computer Applications,
Bota, Sangamner, Maharashtra, India

Abstract : *The rapid growth of social media platforms has transformed the way people create, access, and share information. However, the same platforms have facilitated the widespread dissemination of cyberbullying, hate speech, offensive language, and online harassment. Hostile online content can inflict severe psychological trauma, divide communities, damage personal reputations, and undermine trust in digital public spaces. Traditional manual content moderation methods are often time-consuming, emotionally draining, and impossible to scale to the enormous volume of daily user-generated content. Machine Learning (ML) and Artificial Intelligence (AI) offer promising approaches to identifying toxic online interactions through automated content analysis, natural language processing, pattern recognition, and social network analysis.*

Contextual language representations and the integration of textual, visual, and social network information can improve detection capabilities, but important challenges remain, including limited multilingual resources, code-mixed regional languages, model explainability, dataset bias, adversarial manipulation, and reliable real-time detection. The review concludes that future detection systems should combine AI-based screening with contextual verification, transparent explanations, human oversight, and privacy-preserving practices.

Key Words: Cyberbullying Detection, Hate Speech, Social Media, Machine Learning, Deep Learning, Natural Language Processing, Transformers, Multimodal Learning, Toxicity Moderation.

1. INTRODUCTION

Social media platforms have become important sources of news, public discussion, and real-time personal interaction. Users can publish and share text, images, and videos rapidly, often without the editorial review associated with traditional media channels. Although this openness provides substantial social benefits, it also allows abusive, discriminatory, or hostile behaviors to spread widely.

Cyberbullying generally refers to intentional, repeated harm inflicted on individuals through digital communication. Related concepts include hate speech, which involves offensive language targeted at protected groups based on race, ethnicity, religion, or gender, and general toxicity, which covers profanity and personal insults. These distinctions matter because the intent of a post, its severity, and its potential real-world harm require tailored moderation approaches.

Artificial Intelligence provides techniques for analyzing large amounts of online text and media. Machine learning models identify statistical patterns in textual and behavioral features, while deep learning models learn complex contextual representations from sequence data. Natural language processing supports semantic analysis, and graph-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

based approaches examine relationships among users, comments, and propagation channels. Multimodal systems combine text with images, video, and social context.

1.1 Objectives of the Review

- To examine the role of AI in detecting cyberbullying and hate speech on social media.
- To review traditional machine learning, deep learning, transformer, and graph-based methods. To investigate multimodal detection and evidence-aware verification techniques.
- To summarize important benchmark datasets and evaluation metrics.
- To identify the limitations and operational challenges of existing approaches.
- To propose future research directions for reliable and explainable detection systems.

II. REVIEW METHODOLOGY

This paper adopts a narrative literature-review approach. It synthesizes established research on cyberbullying and hate speech detection and incorporates survey papers addressing deep learning, multimodal systems, and large language models. Relevant studies were identified through scholarly search resources associated with IEEE Xplore, ACM Digital Library, Springer, and related academic publication sources. Search concepts included cyberbullying detection, AI-based hate speech classification, deep learning for toxic text, transformer-based moderation, multimodal abuse detection, and social network analysis.

Background and Conceptual Framework

2.1 Cyberbullying and Related Concepts

Cyberbullying and hate speech detection are commonly treated as classification or verification problems, but several related tasks should be distinguished:

Cyberbullying classification: Predicts whether a social media post constitutes personal harassment or targeted abuse.
Hate speech detection: Identifies discriminatory or incitement language directed at specific demographic groups.
Profanity vs. toxicity: Distinguishes casual colloquial swearing from harmful targeted abuse.
Target identification: Determines whether the toxic behavior targets an individual, an organization, or a protected class.
Multimodal abuse detection: Identifies altered, synthetic, or misleading text embedded in visual meme content.
Stance and sentiment analysis: Evaluates whether responses in a comment thread support, oppose, or escalate toxic behavior.

2.2 Why Social Media Makes Detection Difficult

Social media creates a dynamic information environment in which content is copied, edited, translated, and shared rapidly. High engagement does not establish acceptability, and emotional post virality may reflect outrage rather than benign communication. Additional complications include automated bot accounts, coordinated harassment campaigns, sarcasm, impersonation, manipulated screenshots, and AI-generated toxic text or deepfake media.

III. ARTIFICIAL INTELLIGENCE TECHNIQUES FOR TOXICITY DETECTION

3.1 Traditional Machine Learning

Early automated systems commonly used manually designed textual and statistical features, including word frequency, TF-IDF, sentiment scores, punctuation density, length characteristics, writing style, and user-related metadata. Common algorithms include Logistic Regression, Naive Bayes, Support Vector Machines, Decision Trees, Random Forests, and gradient-boosting methods.

The general workflow consists of collecting labeled examples, cleaning and preprocessing text, extracting features, training a classifier, and evaluating predictions on held-out test data. Traditional machine learning offers relatively low

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

computational cost and useful baselines, but struggles with context, irony, complex slang, and claims that are abusive despite appearing linguistically neutral.

3.2 Deep Learning

Deep learning models automatically learn representations from large datasets and capture complex linguistic patterns without extensive feature engineering. Convolutional Neural Networks (CNNs) learn local phrase patterns in text and visual features in images. Recurrent Neural Networks (RNNs) process sequences, while Long Short-Term Memory (LSTM) networks retain relevant contextual information over longer comment threads. Attention mechanisms help models emphasize relevant abusive words within a post.

Deep learning reduces dependence on manually selected features, but effectiveness depends on training-data quality, model design, and evaluation methodology. Models can learn accidental correlations, such as demographic keywords, instead of genuine indicators of harassment.

3.3 Transformer-Based Models

Transformer architectures have substantially influenced NLP by enabling contextual representations of words and long-range relationships within text. Models such as BERT, RoBERTa, and HateBERT can be fine-tuned on labeled toxic language corpora. Unlike simple word-frequency methods, contextual models represent a word differently depending on its surrounding sentence structure

3.4 Graph Neural Networks and Social Context

Abusive content spreads through networks of users, posts, replies, retweets, and community groups. Graph-based models represent these relationships as nodes and edges. A graph may contain users and interactions, posts and replies, sharing relationships expressing agreement or outrage, and connections between users and historical violation records.

Graph Neural Networks (GNNs) aggregate information from connected nodes to learn representations of user behavioral structures and propagation networks. These methods detect coordinated harassment campaigns that text-only models miss. However, legitimate viral discussions can spread rapidly, and coordinated false flags may resemble ordinary debate. Social context should be treated as supporting evidence, not definitive proof.

3.5 Multimodal Cyberbullying and Hate Speech Detection

Online harassment often combines written claims with photographs, screenshots, video clips, and internet memes. A text-only classifier may miss an offensive contradiction between an innocent caption and an accompanying image. Multimodal systems integrate text, images, video, audio, social context, and knowledge bases.

Feature fusion can occur at the input level, representation level, or decision level. For example, a system may encode a caption with a language model, encode a meme image with a vision transformer, and combine both representations to evaluate toxicity .

3.6 Large Language Models and Retrieval-Augmented Verification

Large Language Models (LLMs) can assist with interpreting conversational context, extracting target entities, identifying implicit insults, and summarizing thread histories. Retrieval-augmented systems search platform community guidelines, toxic phrase bases, and past moderation precedents, then compare retrieved rules with a post:

Extract the principal target and assertion from a social media comment. Identify mentioned groups, individuals, and slurs.

Retrieve relevant moderation policies and contextual precedents.

Assess whether content violates safety policies or constitutes targeted abuse. Generate an explanation with references to community standards.

Refer uncertain or high-impact moderation cases to human reviewers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

IV. LITERATURE REVIEW

4.1 Deep Learning for Toxicity Detection

Hu et al. reviewed deep learning-based text classification, including supervised, weakly supervised, and unsupervised frameworks . Their review examines how text content, social context, and external knowledge bases contribute to detection accuracy.

4.2 Multimodal Abuse Detection

Comito et al. examined multimodal toxicity detection and highlighted the importance of combining information from different media types . Hangloo and Arora reviewed methods, datasets, and future perspectives for combating multimodal online hate

4.3 Social Context and Dataset Development

Shu et al. introduced data repositories designed to support research using text content, social context, and spatiotemporal information . Such datasets help researchers investigate how user interaction patterns relate to harassment propagation.

4.4 Systematic Reviews and Emerging Models

Mostafa et al. reviewed unimodal and multimodal deep learning methods, feature extraction, evaluation measures, and challenges posed by large language models . Elgammal and Alhajj examined deep learning for textual toxicity detection and emerging LLM moderation applications . The literature indicates growing interest in contextual language models, domain transfer learning, multimodal meme analysis, and policy-grounded methods.

4.5 Summary of the Literature

From handcrafted features to learned representations: Deep learning and transformers provide more flexible representations of informal social media text.

From isolated content to contextual analysis: User interaction networks, historical behavior, and propagation patterns provide vital auxiliary signals.

From binary classification to explainable verification: Retrieval, multimodal reasoning, and human oversight help distinguish benign profanity from substantiated harassment.

These developments are complementary. The appropriate approach depends on the moderation task, available evidence, computational budget, and the consequences of a false positive prediction.

V. COMPARISON OF AI TECHNIQUES

Technique	Main Input	Advantages	Limitations
Traditional ML	Text features & metadata	Efficient; useful baseline model.	Limited contextual & slang understanding.
CNN	Text or images	Learns local phrase & visual patterns.	May miss broader conversational context.
LSTM / RNN	Text sequences	Models sequential word dependencies.	Training latency & long-range memory limits.
Transformers	Contextual text	Strong semantic representations.	Computational cost; risk of learning dataset bias.
GNNs	User & interaction networks	Models harassment propagation networks.	Sensitive to network quality & platform changes.
Multimodal AI	Text, images, video media	Combines complementary evidence.	Data alignment & feature fusion complexity.
LLMs with retrieval	Posts & policy rules	Can explain decisions & reference policies.	Hallucinations, retrieval errors, & latency.

No single method is universally best. Results depend on the dataset, task definition, class distribution, temporal setting, and evaluation protocol. Accuracyfigures should not be compared directlyunless experimental conditions are identical.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

Datasets Used in Cyberbullying and Hate Speech Detection

Dataset	Main Characteristics	Research Application
Davidson Hate Speech	Twitter posts annotated as hate, offensive, or neutral.	Separating explicit hate from general profanity.
Kaggle Cyberbullying	Multi-class social media posts covering identity types.	Categorizing age, ethnicity, gender & religion abuse.
Impermium Dataset	Online comments labeled for personal insults.	Harassment and personal insult detection.
Fakeddit / Hateful Memes	Multimodal social media posts and image memes.	Joint visual-textual abuse analysis.
ETHOS Dataset	Multi-label hate speech dataset from social media.	Fine-grained demographic target classification.

The datasets differ in collection methods, label definitions, time periods, and content formats. Researchers should examine licensing, availability, platform restrictions, and label provenance. They should check for duplicate posts, class imbalance, and temporal leakage. Cross-domain evaluations provide a realistic assessment of model generalization.

VI. EVALUATION METRICS

6.1 Accuracy

Accuracy measures the proportion of correctly classified examples:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

6.2 Precision

Precision measures the proportion of predicted toxic cases that are actually true toxic content:

Precision = TP / (TP + FP). High precision reduces false accusations against benign user posts.

6.3 Recall

Recall measures the proportion of actual toxic cases correctly identified by the model:

Recall = TP / (TP + FN). High recall reduces the number of abusive posts that escape detection filters.

6.4 F1 -Score

The F1 -score is the harmonic mean of precision and recall:

F1 = 2 × (Precision × Recall) / (Precision + Recall)

6.5 Additional Evaluation Measures

ROC-AUC: Ranking performance across various classification thresholds. Precision-Recall AUC: Useful when abusive positive examples are relatively rare. Calibration: Whether predicted confidence corresponds to observed correctness. Detection Latency: Real-time processing speed required to flag toxic posts before propagation.

Major Challenges

6.6 Data Quality and Label Reliability

Toxicity datasets contain labels assigned by human annotators, platform moderators, or crowd workers. Subjective standards differ, and context can be interpreted differently across cultures. A possible direction is to record annotator demographics, preserve uncertainty categories, and periodically re-evaluate outdated labels.

6.7 Multilingual and Low-Resource Languages

Many pretrained language models have strong coverage for high-resource languages like English. Regional languages, code-switching (e.g., Hinglish or Marathi-English), transliteration, and local slang create major challenges. Future work must develop regional-language datasets and evaluation protocols.

6.8 Changing Harassment Patterns

Abusive slurs, writing styles, memes, and evasion tactics evolve continuously. Models trained on historical data may fail when user behaviors shift. Time-based evaluation, continuous monitoring, and carefully reviewed retraining address this issue.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

6.9 Model Explainability

A classification label without an explanation provides limited value for appeal processes. Systems should link decisions to specific abusive spans and community guidelines, distinguish targeted harassment from benign profanity, and communicate confidence levels.

6.10 Adversarial Attacks and AI-Generated Abuse

Offenders deliberately change spelling, substitute lookalike symbols, or use AI tools to generate subtle toxic text to evade filters. Combining text analysis with source verification, media forensics, and robust adversarial training improves resilience.

VII. PROPOSED FRAMEWORK FOR AI-BASED MODERATION

This paper proposes a conceptual framework combining machine learning, multimodal analysis, policy knowledge bases, and human oversight:

Stage 1: Data Collection. Ingest public social media posts, comments, media attachments, and user interaction metadata in accordance with privacy regulations and platform terms.

Stage 2: Preprocessing. Remove noise, normalize text, identify languages, extract claims/slurs, and process meme images.

Stage 3: Feature Extraction. Use a fine-tuned transformer for textual representations, a vision model for images, and graph features for propagation patterns.

Stage 4: Initial Classification. Use a lightweight classifier to estimate whether content requires deep analysis; treat this as an initial triage signal.

7.1 Conceptual Workflow:

Raw social media stream → preprocessing & normalization → lightweight AI screening → multimodal & social context analysis → policy guideline retrieval → confidence assessment → human review when required → transparent decision.

7.2 Expected Benefits

Improves early identification of toxic harassment and hate speech. Combines textual and visual signals with relevant conversational context.

Reduces false positive bans by grounding assessments in clear platform policies. Communicates uncertainty rather than forcing every post into a rigid binary label.

VIII. FUTURE RESEARCH DIRECTIONS

8.1 Explainable and policy-grounded AI: Identify specific abusive spans under assessment, cite platform policies, and distinguish verified violations from uncertain predictions.

8.2 Multilingual toxicity detection: Prioritize underrepresented regional languages, code-switching, and transliterated text with culturally appropriate native-speaker annotations.

8.3 Multimodal cross-platform learning: Jointly analyze text, images, audio, and video while accounting for platform-specific comment structures.

8.4 Detection of AI-generated toxic content: Study how synthetic text and deepfakes affect online moderation, integrating forensic verification tools.

8.5 Early proactive detection: Intercept targeted harassment campaigns before widespread viral dissemination while accounting for limited early signals.

IX. DISCUSSION

The reviewed research shows that AI offers complementary ways to address online harassment and hate speech. Traditional machine learning provides efficient baselines, deep learning learns complex sequence representations, transformers support contextual language understanding, and graph models capture user network relationships.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Multimodal systems examine inconsistencies between text and meme graphics, while retrieval-augmented LLMs assist with policy-grounded moderation.

A crucial distinction exists between detecting suspicious language and determining targeted malicious intent. A model may learn correlations associated with previously labeled toxic posts without evaluating context. High post engagement or slang usage does not inherently establish abusive intent.

X. CONCLUSION

Artificial Intelligence has become an essential research domain for detecting cyberbullying and hate speech on social media. Machine learning and deep learning analyze large volumes of content, transformer architectures improve contextual representation, graph methods examine propagation patterns, and multimodal models combine text with visual memes. Retrieval-augmented language models offer further opportunities to support explainable moderation grounded in community standards.

Despite these advances, reliable detection remains challenging because online toxicity evolves, datasets contain human biases, and models can produce incorrect predictions. Multilingual coverage, explainability, user privacy, real-time latency, and resistance to adversarial evasion remain vital areas for future research.

REFERENCES

1. [1] L. Hu et al. (2022)
- Title: Deep learning for toxic text detection: A comprehensive survey
- Journal: AI Open, Vol. 3, pp. 133–155.
- DOI / Link: https://doi.org/10.1016/j.aiopen.2022.09.001

2. [2] M. Mostafa et al. (2024)
- Title: Modality Deep-learning Frameworks for Cyberbullying Detection on Social Networks: A Systematic LITERATURE REVIEW
- Journal: ACM Computing Surveys, Vol. 57, No. 3, Article 77.
- DOI / Link: https://doi.org/10.1145/3700748

3. [3] C. Comito, L. Caroprese, & E. Zumpano (2023)
- Title: Multimodal hate speech detection on social media: a survey of deep learning techniques
- Journal: Social Network Analysis and Mining, Vol. 13, Article 101.
- DOI / Link: https://doi.org/10.1007/s13278-023-01104-w

4.[4] S. Hangloo & B. Arora (2022)
- Title: Combating multimodal online toxicity on social media: methods, datasets, and future perspective
- Journal: Multimedia Systems, Vol. 28, pp. 2321–2351.
- DOI / Link: https://doi.org/10.1007/s00530-022-00966-y

5. [5] Z. Elgammal & R. Alhajj (2026)
- Title: Evolution of Deep Learning Models for Misinformation and Toxicity Detection in Social Media Textual Data: Background, Architectures, Datasets, and Emerging LLM Applications
- Journal: Social Science Computer Review.
- DOI / Link: https://doi.org/10.1177/08944393261432625

6. [6] K. Shu et al. (2020)
- Title: CyberBullyNet: A Data Repository with Text Content, Social Context, and Spatiotemporal Information for Studying Online Harassment on Social Media
- Journal: Big Data, Vol. 8, No. 3, pp. 171–188.
- DOI / Link: https://doi.org/10.1089/big.2020.0062