International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

Toxicity and Offensive Word Detection

Shivam , Shresth Sharma , Shrey Swami , Shubham Yadav

Ankush Gupta

***Abstract*—The exponential growth of online communication platforms has led to a significant rise in toxic and offensive language, posing serious challenges for content moderation and user safety. Detecting such harmful text requires intelligent models capable of understanding linguistic subtleties, cultural context, and user intent. This paper presents a comprehensive review of research developments in toxic and offensive language detection, spanning traditional machine learning approaches, deep learning architectures, transformer models, and recent advancements in large language models (LLMs). The survey emphasizes the evolution of methods from lexical feature engineering to contextual representation learning and multilingual modeling. Key issues such as dataset imbalance, bias mitigation, interpretability, and efficiency are critically analyzed. Furthermore, the review highlights the importance of fairnessaware learning, explainable AI, and parameter-efficient finetuning for scalable real-world applications. The findings underline the ongoing transition toward hybrid, interpretable, and ethically aligned systems that integrate automation with human oversight to ensure safe and responsible online communication.**

I. INTRODUCTION

The rapid expansion of social media platforms, online forums, and digital communication tools has revolutionized global connectivity but has also amplified the prevalence of hate speech, cyberbullying, and offensive language. The widespread dissemination of such harmful content not only disrupts healthy communication but also contributes to psychological distress, misinformation, and social polarization. Detecting and moderating toxic language has therefore become a pressing necessity for maintaining safe digital environments and fostering responsible user interaction.

Traditional approaches to offensive language detection primarily relied on rule-based systems and manually curated lexicons that flagged predefined offensive keywords. While these systems offered simplicity and interpretability, they failed to capture contextual nuances such as sarcasm, implicit hate, or culturally dependent expressions. Machine learning (ML) algorithms, including Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM), marked the first step toward automation by learning statistical patterns from annotated datasets. However, their dependence on handcrafted features limited their ability to generalize across diverse linguistic styles and social domains.

With the advent of deep learning (DL), researchers began to leverage neural architectures capable of learning hierarchical text representations. Models such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) achieved substantial improvements by capturing

syntactic and semantic dependencies. The introduction of contextual word embeddings like Word2Vec, fastText, and GloVe further enhanced model understanding of relationships

between words. Subsequent developments in transformerbased architectures, particularly BERT, RoBERTa, and ALBERT,

revolutionized toxicity detection through selfattention mechanisms, enabling models to understand context bidirectionally and handle complex linguistic structures.

Recent advancements in large language models (LLMs) such as GPT, LLaMA, and Mistral have expanded the horizon of toxic language detection by integrating multilingual, crossdomain, and explainability capabilities. These models demonstrate strong generalization and can be fine-tuned using parameter efficient methods like LoRA and QLoRA, which reduce computational requirements while maintaining high accuracy. Moreover, modern research emphasizes interpretability, fairness, and ethical AI, ensuring that detection systems not only perform effectively but also avoid reinforcing societal bias or discrimination.

Despite considerable progress, several challenges persist. Dataset imbalance, subjectivity in annotation, and bias across languages and demographics continue to affect model reliability. Furthermore, adversarial attacks and code-mixed text introduce additional complexity. Addressing these issues requires the integration of explainable AI, fairness-driven evaluation, and continual learning strategies. The goal of contemporary research is to design intelligent systems that can detect, explain, and mitigate offensive content while preserving freedom of expression understanding.

II. LITERATURE SURVEY

The detection of toxic and offensive language has become a central focus in Natural Language Processing (NLP), especially with the surge in social media usage and the increasing incidents of online harassment and hate speech. Researchers have progressively evolved from statistical machine learning (ML) methods to deep learning (DL), transformer-based models, and most recently, large language models (LLMs). Each stage has

improved contextual understanding, multilingual coverage, and explainability in text moderation systems.

Early studies primarily relied on handcrafted linguistic features and traditional classifiers. Gaydhani et al. [3] utilized n-gram and TF-IDF features with Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machines (SVM) for detecting offensive and hate speech on Twitter, achieving an accuracy of 95.6%. Similarly, Pradhan et al. [7] implemented lexical dictionaries and bag-ofwords models for identifying abusive language, which performed well on explicit text but failed to capture hidden toxicity or sarcasm. Bonetti et al. [1] compared these classical approaches with emerging neural models and found that while SVM and LR were efficient for short-text classification, their dependency on feature engineering limited adaptability to dynamic social media language.

With the rise of deep learning, contextual models began to dominate toxicity detection research. D'Sa et al. [5] developed CNN and BiLSTM architectures combined with BERT embeddings, achieving an F1-score of 97% on hate speech datasets. Their findings demonstrated that fine-tuned BERT models outperform statistical classifiers by learning deep semantic associations. Jose et al. [9] also explored transformer based architectures such as RoBERTa and BERTweet, reporting significant accuracy improvements due to attention mechanisms that capture implicit hate and contextual aggression. Hussain et al. [12] proposed a hybrid model named ORUD-Detect, integrating Bi-LSTM with fastText and Word2Vec embeddings for Roman Urdu text. This approach reached an F1-score of 98%, confirming the robustness of hybrid embedding-based systems in handling code-mixed and morphologically rich languages.

As the field expanded, researchers addressed challenges in multilingual and domainspecific toxicity detection. Alansari and Luqman [8] designed a Multi-Task Learning (MTL) framework enhanced by Active Learning for Arabic offensive speech classification, achieving a macro-F1 of 85.4% on the OSACT dataset. Their system integrated auxiliary tasks such as detecting vulgar or violent expressions to improve contextual coverage. Meanwhile, Bhat et al. [2] introduced ToxiScope, a dataset tailored for workplace communication, including corporate emails and professional chat data. They demonstrated that models trained on social media data perform poorly on formal communication, underscoring the importance of domain adaptation. Garg et al. [6] examined bias in toxic-speech datasets, identifying lexical and sampling biases as primary causes of misclassification against minority dialects. Their research emphasized fairness-driven evaluation and data balancing for ethical deployment.

Interpretability has also become a major research direction in toxicity detection. Risch et al. [10] introduced explainability metrics such as the Explanatory Power Index (EPI) and compared LIME, Layerwise Relevance Propagation (LRP), and attention-based explanation methods for LSTM networks. Their results confirmed that interpretable models can maintain high performance while providing human-understandable justifications. Sharma et al. [4] and Banerjee et al. [15] expanded this work, analyzing explainable transformers and adversarially robust learning. They concluded that explainability and fairness must coexist with accuracy to ensure ethical AI moderation practices.

Recent developments in large language models (LLMs) have significantly reshaped offensive content detection. Hussain et al. [11] fine-tuned Meta-LLaMA-3-8B using QLoRA (Quantized Low-Rank Adaptation) on Roman Urdu–English code-mixed data, achieving an F1-score of 91.45%. This efficient adaptation method allowed high performance under limited computational resources. Jahan et al. [13] and Achintalwar et al. [14] extended this research by testing models such as GPT3.5, Mistral, and Falcon for multilingual offensive speech and counter-speech generation. Their findings revealed that LLMs surpass previous transformerbased architectures in semantic understanding and cultural adaptability but require fine-tuning and bias mitigation for safe real-world application.

The continuous evolution of toxicity detection methods is illustrated in Figure 1, which visualizes the transition from rule based and statistical ML approaches to deep learning, transformer architectures, and fine-tuned LLMs. Early models relied heavily on surface features and lacked contextual reasoning, while modern LLMs demonstrate an advanced understanding of language semantics, enabling more reliable moderation of toxic content across domains and languages.

Ref	Authors/Year	Model/Techniques	Dataset / Language	Accuracy / F1	Key Findings
[3]	Gaydhani et al., 2018	SVM, LR (TF-IDF)	Twitter	95.6%	Lexical baselines effective for explicit hate
[5]	D'Sa et al., 2018	CNN, Bi-LSTM, BERT	Hate Speech Dataset	97% F1	BERT contextual embeddings outperform ML
[8]	Alansari & Luqman, 2025	MTL + Active Learning	Arabic (OSACT)	85.4% F1	Auxiliary tasks enhance multilingual performance
[2]	Bhat et al., 2021	BERT (ToxiScope)	Workplace Emails	79% F1	Domain adaptation improves contextual accuracy
[6]	Garg et al., 2022	Bias Survey	Multiples	—	Framework for fairness-aware detection
[10]	Risch et al., 2020	LSTM + LIME/LRP	TRAC	—	Introduced Explanatory Power Index
[11]	Hussain et al., 2025	LLaMA-3-8B + QLoRA	Roman Urdu–Eng	91.45% F1	Efficient LL M fine-tuning
[13]	Jahan et al., 2025	GPT-3.5, Mistral	Multilingual	89-90% F1	LLM models improve semantic understanding

Table 1: Comparative Summary of Offensive Language Detection Studies.
Detailing models, datasets, and outcomes across fifteen studies. It illustrates the clear progression from traditional ML approaches to advanced transformer- and LLM-based systems, reflecting continuous improvements in accuracy and contextual awareness.

Model Type	Representative Works	Core Techniques	Strengths	Limitations
Traditional ML	[1], [2], [3]	TF-IDF, SVM, Naïve Bayes	Fast, interpretable, requires small data	Fails on implicit or contextual toxicity
Deep Learning	[4], [5], [6]	CNN, Bi-LSTM, fastText	Learns contextual features, high recall	Requires large datasets
Transformerbased	[7], [8], [9]	BERT, RoBERTa, ALBERT	Handles semantics, robust cross-domain	Computationally expensive
LLM-based	[13], [14], [15]	GPT, LLaMA, Mistral	Multilingual, explainable, scalable	Sensitive to bias and resource heavy

Table 2: Comparison of Model Architectures Used in Toxic Language Detection.

Machine learning, deep learning, transformer, and LLM frameworks—used in offensive language detection. The table highlights how complexity and contextual capability increase with each generation while also noting the trade-offs in computational cost.

Dataset	Language(S)	Source Platform	No. Of Samples	Category	Reference
TRAC 2020	English, Hindi, Bengali	Social Media	40K	Aggression, Toxicity	[5]
OSACT	Arabic	Twitter	20K	Offensive, Hate	[7]
ToxiScope	English	Workplace Emails, Chats	10K	Professional Toxicity	[8]
HateEval	English, Spainish	Twitter	18K	Hate, Offensive	[3]
OLID	English	Twitter	14K	Offensive, Targeted/Untargeted	[1]
Kaggle Toxic Comments	Multilingual	Wikipedia, Reddit	160K	Toxic, Obscene, Threat	[9]

Table 3: Key Datasets for Offensive and Toxic Language Detection.

Lists benchmark datasets frequently employed for offensive language identification across multiple languages and domains. It emphasizes the dominance of English and social-media-based corpora and underlines the need for culturally diverse, balanced, and multilingual datasets for fair evaluation.

Metric	Purpose	Interpretation	Common Range	Used In
Accuracy	Overall performance	Percentage of correct classifications	70-95%	[1], [4], [13]
Precision	Measures false positives	High = Few nontoxic labeled as toxic	60-95%	[3], [7], [9]
Recall	Measures false negatives	High = fewer missed toxic samples	65-90%	[4], [8], [12]
F1 Score	Balance of precision & recall	Harmonic mean of precision and recall	70-95%	[5], [13], [14]
AUC-ROC	Binary discrimination ability	1.0 = perfect classifier	0.6-0.95	[6], [9], [15]

Table 4. Evaluation Metrics Commonly Used in Toxic Language Detection.

Presents key evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC used to assess model performance. It shows that researchers increasingly prefer F1-Score and AUC-ROC for handling class imbalance and comparing models on fairness and robustness.

REFERENCES

[1] S. Gaydhani et al., "Detecting Hate Speech and Offensive Language on Twitter Using Machine Learning," IEEE SIIE, 2020.

[2] S. Sadiq et al., "Automated Classification of Harmful Online Content Using Machine Learning," Proceedings of ICDIS, 2019.

[3] D. Bonetti et al., "Rule-Based and Statistical Models for Hate Speech Detection," Information Journal, 2023.

[4] D. D'Sa et al., "Deep Learning Architectures for Hate and Offensive Speech Classification," EMNLP Findings, 2021.

[5] J. Risch et al., "Offensive Language Identification Using Deep Neural Networks," ACL Workshop on TRAC, 2020.

[6] R. Sharma et al., "Multilingual Transfer Learning for Offensive Language Detection," Applied Sciences, 2023.

[7] A. Alansari and M. Luqman, "Active Learning for Arabic Offensive Speech Classification," Applied Sciences, 2025.

[8] A. Bhat et al., "ToxiScope: Workplace Toxicity Dataset and BERT-Based Classification," Information, 2024.

[9] N. Garg et al., "Bias and Fairness in Hate Speech Detection," Mathematics, 2023.

[10] J. Risch et al., "Explainability in Toxic Language Detection," Findings of EMNLP, 2021.

[11] A. Banerjee et al., "Adversarial Robustness and Interpretability in Neural Models," IEEE Access, 2022.

[12] R. Sharma et al., "Attention Visualization for Explainable Transformers," Applied Sciences, 2024.

[13] T. Hussain et al., "LLaMA-3-8B with QLoRA for Roman Urdu Toxicity Detection," arXiv preprint, 2025.

[14] M. Jahan et al., "Advancements in Offensive Language Detection Using GPT and Mistral," ACL Long Papers, 2025.

[15] M. Luqman et al., "Comprehensive Review and Experimental Analysis of Offensive Language Detection," Applied Sciences, 2025