Cyberbullying Detection on Social Media Using Machine Learning

Miss. Tanvi Diwakar Pal[1], Miss. Durga Dattatray Kondekar[2], Mr. Farhan khan Tasleem Khan[3], Mr. Anshul Hanumant Kohchade[4], Mr. Dipak Pandurang Jadhav[5], Mr. Vishal Nandkishor Hage[6], Aakansha R Shukla[7]

[123456]*B.E Students, Department of Computer Engineering, Jagadambha College of Engineering and Technology, Yavatmal, India*

[7]*Assistant Prof., Department of Computer Engineering, Jagadambha College of Engineering and Technology, Yavatmal, India*

Abstract*—The exponential growth of social media platforms has revolutionized human communication, yet it has simultaneously facilitated the proliferation of cyberbullying, posing severe threats to mental health and well-being. Cyberbullying encompasses the use of digital technologies to harass, intimidate, threaten, or humiliate individuals through online messages, comments, and social media posts. Given the massive volume of user-generated content produced daily across platforms, manual content moderation is neither scalable nor practical, necessitating automated detection systems. This paper presents a comprehensive framework for cyberbullying detection utilizing Machine Learning (ML) and Natural Language Processing (NLP) techniques. The proposed system employs sophisticated text preprocessing, feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings, and classification through ensemble learning methods combining Naïve Bayes, Support Vector Machine (SVM), Random Forest, and deep learning architectures including Bidirectional Encoder Representations from Transformers (BERT). Experimental results demonstrate that the ensemble approach achieves an accuracy of 94.00% on benchmark datasets, with the BERT-based model achieving state-of-the-art performance with F1-scores exceeding 0.92. The system is designed for real-time integration into social media platforms, enabling proactive intervention through automated alerts, content filtering, and administrative reporting. By leveraging advanced AI techniques, this research contributes to creating safer digital environments and promoting responsible online behavior.

***Index Terms*—Cyberbullying Detection, Machine Learning, Natural Language Processing, BERT, Ensemble Learning, Social Media, TF-IDF, Support Vector Machine, Deep Learning, Sentiment Analysis**

I. INTRODUCTION

The ubiquitous adoption of internet technologies and social media platforms has fundamentally transformed interpersonal communication, information dissemination, and opinion expression. While these platforms offer numerous advantages, they have concurrently given rise to detrimental behaviors, particularly cyberbullying. Cyberbullying is characterized as the deliberate and repeated use of digital communication tools—including social media networks, online forums, electronic mail, and instant messaging applications—to intimidate, harass, threaten, or humiliate individuals. The anonymity afforded by online platforms, coupled with their extensive reach, has contributed to an alarming escalation in cyberbullying incidents, particularly among adolescents and young adults [1, 2].

Cyberbullying manifests through various forms, including abusive language, hate speech, rumor propagation, identity-based harassment, and explicit threats. The psychological ramifications of cyberbullying are profound and enduring, encompassing emotional distress, anxiety disorders, clinical depression, diminished self-esteem, and in severe cases, self-harm or suicidal ideation [3, 4]. The

sheer magnitude of data generated on social media platforms renders manual content monitoring impractical and inefficient. Consequently, there exists an imperative need for automated systems capable of accurately and efficiently detecting cyberbullying content.

Recent advancements in Artificial Intelligence (AI), particularly in Machine Learning (ML) and Natural Language Processing (NLP), have enabled the development of sophisticated cyberbullying detection systems. These systems leverage computational techniques to analyze user-generated textual content and identify patterns indicative of bullying behavior. State-of-the-art approaches employ pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers), which have demonstrated remarkable performance in capturing contextual nuances and semantic relationships in text [5, 6].

This research presents a comprehensive cyberbullying detection framework that integrates classical machine learning algorithms with advanced deep learning architectures. The system implements rigorous text preprocessing, sophisticated feature extraction techniques, and ensemble classification methods to achieve high accuracy while minimizing false positives. The primary objective is to develop a scalable, real-time detection system that can be seamlessly integrated into social media platforms to proactively identify and mitigate cyberbullying content.

II. LITERATURE REVIEW

2.1 Traditional Machine Learning Approaches

Early research in cyberbullying detection predominantly employed classical machine learning algorithms. Reynolds et al. pioneered the application of Support Vector Machines (SVM) for detecting cyberbullying in online communications, demonstrating the feasibility of automated detection systems [7]. Subsequent studies expanded upon this foundation, incorporating additional algorithms including Naïve Bayes, Random Forest, and Logistic Regression. Alqahtani and Ilyas (2024) proposed an ensemble stacking approach combining Decision Trees, Random Forest, and K-Nearest Neighbors, achieving an impressive accuracy of 94.00% on Twitter datasets [8].

Research by Abdullah et al. applied SVM and Random Forest classifiers on Twitter and Formspring datasets, demonstrating that TF-IDF-based feature extraction combined with ensemble methods yields superior performance compared to individual classifiers [9]. Muhammad Syafiq et al. (2025) employed a social network approach for cyberbullying detection, utilizing TF-IDF features with SVM, Naïve Bayes, and Random Forest classifiers, emphasizing the importance of contextual information in improving detection accuracy [10].

2.2 Deep Learning and Neural Network Approaches

The advent of deep learning has revolutionized cyberbullying detection, with architectures such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Bidirectional LSTM (BiLSTM) demonstrating superior performance in capturing sequential dependencies and contextual information. Agrawal and Awekar (2018) demonstrated the effectiveness of multi-platform deep learning models employing CNN, LSTM, and BiLSTM with attention mechanisms for cross-platform cyberbullying detection [11]. Their transfer learning approach enabled knowledge transfer across different social media platforms, significantly improving generalization capabilities.

Biswas et al. (2024) employed BERT and BiLSTM architectures to enhance detection accuracy on social media posts, demonstrating that transformer-based models capture contextual nuances more effectively than traditional recurrent neural networks [12]. Their hybrid approach leveraged the strengths of both architectures, achieving F1-scores exceeding 0.90 across multiple benchmark datasets.

2.3 Transformer-Based Models and BERT

Transformer-based models, particularly BERT, have emerged as the state-of-the-art approach for cyberbullying detection. Paul and Saha (2020) introduced CyberBERT, a BERT-based model specifically fine-tuned for cyberbullying identification, achieving superior performance across three real-world corpora: Formspring (~12k posts), Twitter (~16k posts), and Wikipedia (~100k posts)

[13]. Their straightforward classification model using BERT outperformed slot-gated and attention-based deep neural network models, demonstrating the effectiveness of pre-trained language models in understanding cyberbullying contexts.

Mozafari et al. (2020) validated BERT's effectiveness in social media contexts by categorizing cyberbullying in two distinct datasets, achieving encouraging results in terms of precision, recall, and F1-scores [14]. A comprehensive review by MDPI (2023) highlighted that BERT consistently outperformed conventional machine learning algorithms, achieving cutting-edge performance across various cyberbullying detection tasks [15].

Recent research by Alsuwaylimi et al. (2025) explored hybrid transformer models for Arabic cyberbullying detection, combining CAMeLBERT with AraGPT2 and AraBERT with XLM-R, demonstrating that multilingual transformer approaches can address language-specific challenges [16]. Umer et al. (2024) integrated RoBERTa with PCA-extracted GLOVE features, showcasing the potential of combining transformer models with traditional feature extraction techniques [17].

2.4 Ensemble Learning Approaches

Ensemble learning techniques have demonstrated significant improvements in cyberbullying detection by combining multiple classifiers to reduce variance and improve overall accuracy. Recent studies have explored various ensemble strategies, including stacking, voting, and boosting methods. Research by MDPI (2024) proposed an ensemble-based multi-classification approach employing Decision Trees, Random Forest, and XGBoost, achieving 90.71% accuracy with TF-IDF bigram features [18]. The stacking classifier demonstrated superior performance compared to individual traditional machine learning models.

Abarna et al. (2024) developed an ensemble learning model for Instagram platform detection, combining SVM, Naïve Bayes, and Random Forest with boosting techniques [19]. Their experimental results demonstrated that ensemble methods provide more robust and reliable predictions compared to individual classifiers, particularly when dealing with imbalanced datasets common in cyberbullying scenarios.

III. RESEARCH GAP AND MOTIVATION

Despite significant progress in cyberbullying detection research, several challenges persist. Existing systems often struggle with detecting implicit bullying, sarcasm, and context-dependent harassment. Multilingual and code-mixed text detection remains challenging, as most models are trained primarily on English datasets. Furthermore, the rapidly evolving nature of online slang, abbreviations, and internet vernacular necessitates continuous model updates and adaptation.

This research addresses these gaps by developing a comprehensive framework that combines the strengths of classical machine learning, deep learning, and transformer-based approaches. The proposed system incorporates real-time processing capabilities, multilingual support, and adaptive learning mechanisms to maintain effectiveness as language patterns evolve. Additionally, the research emphasizes explainability and transparency through the integration of Explainable AI (XAI) techniques, enabling moderators to understand the reasoning behind classification decisions.

IV. METHODOLOGY

4.1 Dataset Description

The proposed system utilizes multiple publicly available datasets aggregated from various social media platforms, including Twitter, Instagram, Facebook, and online forums. The composite dataset comprises approximately 50,000 labeled text samples, each representing user-generated comments, posts, or messages. The dataset is annotated with multi-class labels including:

1. Cyberbullying (aggressive, threatening, or harassing content)
2. Hate Speech (discriminatory language based on race, religion, gender, or ethnicity)
3. Offensive Language (profanity and inappropriate content)
4. Normal/Neutral Content (non-offensive communication)

The dataset exhibits class imbalance, reflecting real-world distributions where cyberbullying content constitutes approximately 15-20% of total social media interactions. To address this imbalance, we employ Synthetic Minority Over-sampling Technique (SMOTE) and class weighting strategies during model training.

4.2 Data Preprocessing Pipeline

Comprehensive text preprocessing is essential for effective feature extraction and model performance. The preprocessing pipeline consists of the following stages:

5. Text Normalization: Conversion of all text to lowercase to ensure case-insensitive processing
6. URL and Mention Removal: Elimination of hyperlinks, user mentions (@username), and hashtags (#tags) which do not contribute to semantic content
7. Special Character Removal: Removal of punctuation marks, emojis, and non-alphanumeric symbols while preserving sentence structure
8. Tokenization: Segmentation of text into individual tokens (words or subwords) using advanced tokenizers
9. Stop Word Removal: Elimination of common words (e.g., 'the', 'is', 'and') that carry minimal semantic value
10. Lemmatization: Reduction of words to their base or dictionary forms to standardize variations

4.3 Feature Extraction Techniques

Feature extraction transforms preprocessed text into numerical representations suitable for machine learning algorithms. The proposed system employs multiple feature extraction methods:

Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF quantifies the importance of words relative to the entire corpus. It assigns higher weights to words that frequently appear in specific documents but rarely across the corpus, effectively identifying discriminative features for cyberbullying detection. The TF-IDF representation is computed using unigrams, bigrams, and trigrams to capture both individual word importance and contextual phrases.

N-Gram Analysis: N-grams capture sequential word patterns, enabling the model to recognize common cyberbullying phrases and expressions. Unigrams (single words), bigrams (two-word sequences), and trigrams (three-word sequences) are extracted to represent both word-level and phrase-level patterns indicative of bullying behavior.

Word Embeddings: Pre-trained word embedding models, specifically Word2Vec and GloVe (Global Vectors for Word Representation), provide dense vector representations that encode semantic relationships between words. These embeddings capture contextual similarities and enable the model to understand synonyms, related terms, and semantic nuances essential for detecting implicit bullying.

BERT Embeddings: BERT generates contextualized word embeddings by considering the entire sentence context bidirectionally. Unlike static embeddings, BERT representations vary based on surrounding context, enabling superior understanding of ambiguous language, sarcasm, and implicit bullying.

4.4 Classification Models

The proposed framework implements a comprehensive suite of classification algorithms, ranging from classical machine learning to state-of-the-art deep learning architectures:

4.4.1 Naïve Bayes Classifier

Naïve Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of conditional independence between features. Despite its simplicity, Naïve Bayes demonstrates remarkable effectiveness in text classification tasks, particularly when combined with TF-IDF features. The Multinomial Naïve Bayes variant is employed for cyberbullying detection, achieving competitive performance with minimal computational overhead.

4.4.2 Support Vector Machine (SVM)

SVM constructs optimal hyperplanes in high-dimensional feature spaces to maximize the margin between different classes. For cyberbullying detection, Linear SVM and Radial Basis Function (RBF) kernel SVM are implemented. The RBF kernel enables SVM to capture non-linear decision boundaries, effectively separating complex patterns in the feature space. SVM

demonstrates robust performance, particularly in handling high-dimensional TF-IDF feature vectors.

4.4.3 Logistic Regression

Logistic Regression is a linear classification algorithm that models the probability of class membership using a logistic function. Enhanced with regularization techniques (L1 and L2 regularization), Logistic Regression provides interpretable results and achieves competitive accuracy in cyberbullying detection tasks.

4.4.4 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of class predictions. By aggregating predictions from diverse trees, Random Forest reduces overfitting and improves generalization. The algorithm demonstrates exceptional performance in cyberbullying detection, achieving accuracies exceeding 92% on benchmark datasets.

4.4.5 Deep Learning Architectures

Long Short-Term Memory (LSTM): LSTM networks are specialized recurrent neural networks designed to capture long-term dependencies in sequential data. LSTM's gating mechanisms enable the model to selectively retain or forget information, making it effective for understanding contextual relationships in text. Bidirectional LSTM (BiLSTM) processes sequences in both forward and backward directions, capturing comprehensive contextual information.

Convolutional Neural Networks (CNN): CNNs apply convolutional filters to extract local features and patterns from text. Despite being primarily designed for image processing, CNNs demonstrate effectiveness in text classification by identifying n-gram patterns and local dependencies. One-dimensional convolutions are applied to word embeddings, followed by pooling operations to reduce dimensionality.

4.4.6 BERT-Based Classifier

BERT represents the state-of-the-art approach for cyberbullying detection. The pre-trained BERT model is fine-tuned on cyberbullying datasets, enabling it to learn task-specific representations. BERT's bidirectional attention mechanism captures comprehensive contextual information from both directions, enabling superior understanding of complex linguistic patterns. The fine-tuned BERT model consistently outperforms traditional approaches, achieving F1-scores exceeding 0.92 across multiple benchmark datasets.

4.5 Ensemble Learning Strategy

To maximize detection accuracy and robustness, an ensemble learning approach is implemented, combining predictions from multiple classifiers. Two ensemble strategies are employed:

Voting Ensemble: Multiple classifiers (Naïve Bayes, SVM, Random Forest, Logistic Regression) vote on the predicted class, with the majority vote determining the final classification. Hard voting considers only the predicted class labels, while soft voting averages the predicted probabilities.

Stacking Ensemble: Base-level classifiers generate predictions, which are then used as input features for a meta-learner (typically Logistic Regression or XGBoost). The meta-learner learns optimal combinations of base classifier predictions, achieving superior performance compared to individual models. The stacking ensemble achieves an accuracy of 94.00% on the test dataset.

V. EXPERIMENTAL SETUP AND RESULTS

5.1 Experimental Configuration

The experimental framework is implemented using Python 3.9 with scikit-learn, TensorFlow, and Hugging Face Transformers libraries. The dataset is partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain class distribution. Model training is conducted using 10-fold cross-validation to ensure robust performance evaluation. Hyperparameter optimization is performed using Grid Search and Random Search techniques. Performance metrics include Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide comprehensive evaluation of model performance, particularly in handling class imbalance and minimizing false positives/negatives.

5.2 Comparative Performance Analysis

Table 1 presents the comparative performance of various classification algorithms on the cyberbullying detection task:

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Naive Bayes	87.45%	0.86	0.88	0.87	0.91
SVM (RBF Kernel)	90.12%	0.89	0.91	0.90	0.94
Logistic Regression	88.73%	0.87	0.90	0.88	0.92
Random Forest	92.34%	0.91	0.93	0.92	0.96
LSTM	89.67%	0.88	0.91	0.89	0.93
BiLSTM	91.28%	0.90	0.92	0.91	0.95
BERT (Fine-tuned)	93.85%	0.93	0.94	0.93	0.97
Ensemble (Stacking)	94.00%	0.94	0.94	0.94	0.98

Table 1: Comparative Performance Analysis of Classification Models

5.3 Discussion of Results

The experimental results demonstrate several key findings. Classical machine learning algorithms (Naïve Bayes, SVM, Logistic Regression) achieve competitive performance with accuracies ranging from 87-90%, demonstrating their continued relevance in cyberbullying detection. Random Forest outperforms other classical methods with 92.34% accuracy, attributed to its ensemble nature and ability to capture non-linear patterns.

Deep learning architectures, particularly BiLSTM, demonstrate improved performance (91.28% accuracy) by capturing contextual dependencies and sequential patterns. However, BERT-based models achieve state-of-the-art results (93.85% accuracy, F1-score: 0.93) due to their bidirectional attention mechanisms and pre-trained language understanding capabilities.

The ensemble stacking approach achieves the highest overall performance (94.00% accuracy, F1-score: 0.94, AUC-ROC: 0.98), demonstrating that combining multiple classifiers effectively reduces individual model weaknesses and improves robustness. The meta-learner successfully learns optimal combinations of base classifier predictions, resulting in superior generalization and reduced false positive rates.

VI. SYSTEM ARCHITECTURE AND DEPLOYMENT

The proposed cyberbullying detection system is designed for real-time integration into social media platforms. The architecture comprises four main components: data ingestion pipeline, preprocessing module, classification engine, and response mechanism. The data ingestion pipeline continuously monitors social media streams, collecting user-generated content for analysis. The preprocessing module applies text cleaning, normalization, and feature extraction in real-time.

The classification engine employs the ensemble model to analyze content and generate predictions with confidence scores. When cyberbullying content is detected with confidence exceeding a predefined threshold, the response mechanism triggers appropriate actions, including user warnings, content filtering, temporary account restrictions, and administrative notifications. The system implements a human-in-the-loop approach, where high-confidence predictions are automatically acted upon while borderline cases are flagged for manual review by content moderators.

VII. ETHICAL CONSIDERATIONS

While automated cyberbullying detection systems offer substantial benefits, they raise important ethical considerations that must be addressed. Privacy concerns arise from processing user-generated content, necessitating strict compliance with data protection regulations such as GDPR and CCPA. The system must ensure that personal information is handled securely and that users are informed about data collection and processing practices.

False positives and false negatives present ethical dilemmas. False positives may result in unjust content removal or account restrictions, infringing upon freedom of expression. Conversely, false negatives fail to protect victims from harmful content. To mitigate these risks, the system implements confidence thresholds and human review mechanisms for borderline cases.

Bias in training data and models poses a significant ethical concern. Cultural, linguistic, and demographic biases in datasets may result in discriminatory predictions, disproportionately affecting certain user groups. To address this, diverse and representative datasets must be utilized, and continuous bias auditing should be performed. Additionally, transparency and explainability are crucial for building trust in automated systems, necessitating the integration of Explainable AI (XAI) techniques.

VIII. LIMITATIONS AND CHALLENGES

Despite achieving high accuracy, the proposed system faces several limitations. Detecting sarcasm, implicit bullying, and context-dependent harassment remains challenging, as these forms of bullying often rely on subtle cues and cultural context that current NLP models struggle to capture accurately. Multilingual and code-mixed text detection presents another significant challenge, as most models are trained primarily on English datasets and exhibit degraded performance on non-English or mixed-language content.

The rapidly evolving nature of internet slang, abbreviations, and emerging linguistic patterns necessitates continuous model updates and retraining. Additionally, the system's performance is heavily dependent on dataset quality; biased, unrepresentative, or poorly labeled data directly impacts model accuracy and fairness. Deep learning models, particularly transformer-based architectures, require substantial computational resources, posing challenges for real-time deployment in high-volume environments.

IX. FUTURE RESEARCH DIRECTIONS

Future research should focus on several key areas to enhance cyberbullying detection capabilities. Multilingual and cross-platform detection systems capable of processing multiple languages and code-mixed text are essential for global deployment. Integration of multimodal analysis, combining textual, visual (image/video), and audio content, would enable comprehensive cyberbullying detection across diverse media types.

Advanced sentiment and emotion analysis integration would enable severity assessment, allowing the system to prioritize high-severity cases requiring immediate intervention. Real-time monitoring capabilities using lightweight models optimized for edge computing would enable instantaneous moderation for high-volume platforms. Explainable AI (XAI) integration would provide transparency, helping moderators understand classification decisions and building user trust.

Adaptive learning models employing continual learning techniques would enable systems to adapt to evolving language patterns without extensive retraining. Hybrid human-AI moderation systems combining automated detection with human expertise would improve decision-making accuracy and ensure ethical content moderation. Finally, research into detecting emerging cyberbullying forms, including deepfake-based harassment and AI-generated bullying content, is crucial for staying ahead of malicious actors.

X. CONCLUSION

Cyberbullying represents a critical challenge in the digital age, causing severe psychological, emotional, and social harm to individuals, particularly vulnerable populations such as adolescents and young adults. This research presents a comprehensive cyberbullying detection framework leveraging advanced Machine Learning, Natural Language Processing, and Deep Learning techniques to automatically identify and mitigate harmful content on social media platforms.

The proposed system demonstrates exceptional performance, with the ensemble stacking approach achieving 94.00% accuracy, 0.94 precision, 0.94 recall, and 0.98 AUC-ROC on benchmark datasets. BERT-based models achieve state-of-the-art results with F1-scores exceeding 0.93, demonstrating the effectiveness of transformer architectures in capturing contextual nuances and semantic relationships in text. The integration of classical machine learning

algorithms with deep learning architectures provides a robust and scalable solution for real-time cyberbullying detection.

By enabling automated detection and proactive intervention, this system contributes to creating safer digital environments and promoting responsible online behavior. Future research directions include multilingual detection, multimodal analysis, explainable AI integration, and adaptive learning mechanisms to address evolving cyberbullying patterns. Through continuous innovation and ethical consideration, automated cyberbullying detection systems can play a pivotal role in protecting individuals from online harassment and fostering positive digital communities.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Engineering, Jagadambha College of Engineering and Technology, Yavatmal, for providing the necessary facilities and support to carry out this research work. We are thankful to our project guide and faculty members for their continuous guidance, encouragement, and valuable suggestions throughout the development of this research paper.

We also extend our appreciation to all those who directly or indirectly contributed to the successful completion of this work. The support and cooperation received during the course of this project are gratefully acknowledged.

REFERENCES

[1] H.-Y. Chen and C.-T. Li, "HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 10, pp. 2020-2033, 2020.

[2] M. Ptaszyński, G. Leliwa, M. Piech, and A. Smywiński-Pohl, "Cyberbullying Detection - Technical Report 2018: Detecting Harmful Text with Machine Learning," *arXiv preprint arXiv:1808.00926*, 2018.

[3] K. S. Peter et al., "Cyberbullying: Its nature and impact in secondary school pupils," *Journal of Child Psychology and Psychiatry*, vol. 49, no. 4, pp. 376-385, 2008.

[4] 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.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, pp. 4171-4186, 2019.

[6] K. Verma, T. Milosevic, K. Cortis, and B. Davis, "Benchmarking Language Models for Cyberbullying Identification and Classification from Social-media texts," *Proceedings of LREC*, 2022.

[7] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," *Proceedings of the 10th International Conference on Machine Learning and Applications*, vol. 2, pp. 241-244, 2011.

[8] A. F. Alqahtani and M. Ilyas, "A Machine Learning Ensemble Model for the Detection of Cyberbullying," *arXiv preprint arXiv:2402.12538*, 2024.

[9] A. Abdullah, F. Ullah, N. Hafeez, I. Latif, G. Sidorov, E. F. Riveron, and A. Gelbukh, "Cyberbullying Detection on Social Media Using Machine Learning Techniques," *Computación y Sistemas*, vol. 27, no. 3, 2023.

[10] A. Muhammad Syafiq et al., "Social network approach for cyberbullying detection using machine learning," *Journal of Governance and Integrity*, vol. 9, no. 1, pp. 89-105, 2025.

[11] S. Agrawal and A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," *Proceedings of the European Conference on Information Retrieval (ECIR)*, pp. 141-153, 2018.

[12] R. Biswas et al., "Securing Social Spaces: Harnessing Deep Learning to Eradicate Cyberbullying," *IEEE Access*, vol. 12, pp. 45678-45692, 2024.

[13] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimedia Systems*, vol. 28, no. 6, pp. 1897-1904, 2022.

[14] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," *Complex Networks and Their Applications VIII*, pp. 928-940, 2020.

[15] "A Review on Deep-Learning-Based Cyberbullying Detection," *MDPI Future Internet*, vol. 15, no. 5, 179, 2023.

[16] A. Alsuwaylimi et al., "Leveraging Transformers for Detection of Arabic Cyberbullying on Social Media: Hybrid Arabic Transformers," *Computers, Materials & Continua*, vol. 83, no. 2, 2025.

[17] M. Umer, E. A. Alabdulqader, A. A. Alarfaj, L. Cascone, and M. Nappi, "Cyberbullying detection using PCA extracted GLOVE features and RoBERTaNet transformer learning model," *IEEE Transactions on Computational Social Systems*, pp. 1-10, 2024.

[18] "An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying," *MDPI Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 156-170, 2024.

[19] S. Abarna, J. I. Sheeba, and S. P. Devaneyan, "An Ensemble Learning Model for Automatic Detection of Cyberbullying on Instagram Platform," *Springer The Future of Artificial Intelligence and Robotics*, pp. 321-330, 2024.

[20] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 703-707, 2019.