AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 3(2026)

An Intelligent Multi-Classifier Framework for Cyberbullying Detection and User Moderation in Social Media

Afreen Begum[1], Sk.Mahammadunnisa[2]
[1]MCA Student , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad
[2]Assistant Professor , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad

Abstract— **Cyberbullying has emerged as a serious concern on social media platforms, negatively affecting individuals through abusive and offensive online interactions. The rapid growth of user-generated content makes manual monitoring ineffective, creating the need for intelligent automated detection systems. This paper presents a machine learning-based framework for detecting cyberbullying in text-based social media posts. The proposed system performs text preprocessing and TF-IDF feature extraction to convert raw textual data into meaningful features. Three machine learning algorithms, namely AdaBoost, Stochastic Gradient Descent (SGD), and Multinomial Naïve Bayes, are trained and evaluated to classify posts as offensive or non-offensive. Sentiment analysis is integrated to enhance the classification process by providing additional contextual information. Furthermore, the system monitors repeated offensive behavior and supports automated user moderation through an administrative interface. Experimental results demonstrate that the proposed framework provides reliable classification performance and contributes to creating a safer and more secure online communication environment.**

Keywords— Cyberbullying Detection, Machine Learning, Social Media, Sentiment Analysis, TF-IDF, AdaBoost, SGD, Multinomial Naïve Bayes, Text Classification.

I. INTRODUCTION

Cyberbullying has become one of the most significant challenges associated with the widespread use of social media and online communication platforms. Millions of users interact daily through websites and mobile applications, sharing opinions, images, and personal experiences. Although these platforms encourage communication and knowledge sharing, they also provide opportunities for individuals to engage in abusive, threatening, and offensive behavior. Unlike traditional bullying, cyberbullying can occur at any time and spread rapidly to a large audience, increasing its emotional and psychological impact on victims. Continuous exposure to harmful messages may result in stress, anxiety, depression, loss of self-confidence, and in severe cases, self-harm or suicidal thoughts. Since the volume of online content is growing rapidly, manually monitoring every post is nearly impossible. Therefore, developing automated systems capable of identifying cyberbullying content accurately and efficiently has become an essential research objective for ensuring safer digital communication and protecting users from online harassment **[1], [3], [10]**.

Recent advancements in machine learning and natural language processing have created new opportunities for automatically detecting harmful content in online conversations. Machine learning algorithms can identify hidden patterns and relationships within textual data by learning from previously labeled examples. This capability enables automated systems to classify messages as offensive or non-offensive with minimal human intervention. The effectiveness of these models largely depends on proper text preprocessing, feature extraction, and algorithm selection. Techniques such as tokenization, stop-word removal, stemming, lemmatization, and TF-IDF feature extraction improve the quality of textual representations before classification **[15]**. Furthermore, sentiment analysis provides additional contextual information by identifying the emotional tone of a message, allowing the detection model to make more informed decisions **[14]**. These technologies have significantly improved the efficiency of cyberbullying detection and have become valuable tools for content moderation across multiple social media platforms **[4], [5], [6], [7]**.

Despite the progress achieved in automated cyberbullying detection, several challenges continue to affect the accuracy and reliability of existing systems. Offensive language often varies depending

on context, cultural background, user intention, and conversational history, making classification more complex than simple keyword matching. Some harmful messages may not contain explicit abusive words but still convey threats, sarcasm, or humiliation. In addition, the informal writing style commonly used on social media, including abbreviations, spelling variations, slang expressions, and emojis, increases the complexity of text analysis. Many existing approaches also focus on a single machine learning model, which may not provide consistent performance across different datasets and communication styles. Therefore, there is a growing need for robust detection frameworks that combine multiple classification techniques with effective text preprocessing and contextual analysis to improve prediction accuracy while reducing false positive and false negative classifications **[8], [9].**

The proposed work presents a machine learning-based framework for detecting cyberbullying in social media text by integrating multiple classification algorithms with sentiment analysis. Initially, textual data undergoes preprocessing steps, including data cleaning, tokenization, stop-word removal, stemming, and lemmatization, to eliminate irrelevant information and improve feature quality. The processed text is then transformed into numerical vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique, which captures the importance of words within the dataset **[15]**. Three machine learning algorithms, namely AdaBoost, Stochastic Gradient Descent (SGD), and Multinomial Naïve Bayes, are trained and evaluated to classify posts as offensive or non-offensive. Sentiment analysis is incorporated to provide additional contextual understanding of user messages, thereby improving classification reliability **[6], [14]**. Furthermore, the system records repeated offensive activities and provides an administrative mechanism to monitor user behavior and initiate appropriate moderation actions whenever necessary, contributing to improved online safety **[1], [10]**.

II.RELATED WORK

"Towards the Detection of Cyberbullying Based on Social Network Mining Techniques," I. H. Ting et al. [1]

This study proposed a cyberbullying detection framework based on social network mining techniques that considered not only offensive messages but also the interaction patterns between users. The framework consisted of aggressive message detection, identification of potential victims and aggressors, and cyberbullying case analysis. By incorporating user relationships and communication frequency, the proposed approach achieved better detection performance than traditional text classification methods. Experimental evaluation on Twitter data demonstrated a high F1-score, indicating that social network analysis can significantly improve the identification of cyberbullying incidents in online platforms.

"Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features," R. Zhao et al. [4]

This research introduced a machine learning approach for cyberbullying detection by combining bullying-specific linguistic features with traditional text representation methods. The proposed model utilized word embeddings, bag-of-words, and semantic features to improve the classification of offensive content. A linear Support Vector Machine (SVM) classifier was employed to distinguish bullying messages from normal conversations. Experimental results showed that integrating bullying-related features enhanced detection accuracy compared to conventional text classification techniques. The study emphasized the importance of domain-specific features for improving cyberbullying detection systems.

"Detection of Cyberbullying Using Deep Neural Network," V. Banerjee et al. [5]

This paper presented a deep learning-based cyberbullying detection model using Convolutional Neural Networks (CNNs) to automatically classify harmful online messages. The proposed system learned contextual patterns directly from textual data without relying heavily on handcrafted features. The deep neural network demonstrated superior classification performance over several conventional machine learning methods by effectively capturing semantic relationships within user-generated content. The authors concluded that deep learning models offer improved scalability and higher detection accuracy for identifying abusive language across different social media platforms.

"Cyberbullying Detection Using Pre-Trained BERT Model," J. Yadav et al. [7]

This study proposed a cyberbullying detection framework based on the Bidirectional Encoder Representations from Transformers (BERT) model for contextual text classification. The pre-trained language model generated meaningful contextual embeddings that improved the understanding of user comments and offensive expressions. A simple classification layer was added to categorize messages into bullying and non-bullying classes. Experimental evaluation on multiple benchmark

datasets demonstrated that the BERT-based approach achieved higher accuracy and better generalization than conventional machine learning algorithms. The study highlighted the effectiveness of transformer models in handling complex language patterns found in social media.

"Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," S. Agrawal and A. Awekar [9]

This research investigated deep learning techniques for detecting cyberbullying across different social media platforms, including Twitter, Wikipedia, and Formspring. The proposed models automatically learned semantic and contextual information from textual data without requiring handcrafted features. Transfer learning was also explored to evaluate whether knowledge learned from one platform could improve performance on another. Experimental results showed that deep learning models consistently outperformed traditional machine learning approaches and demonstrated strong adaptability across multiple datasets. The authors concluded that transferable deep learning models provide an effective solution for large-scale cyberbullying detection in diverse online environments.

III. DATASET DETAILS

The dataset used in this research consists of text-based social media posts collected from publicly available cyberbullying and offensive language datasets. It contains user-generated messages that are manually labeled **as** offensive (cyberbullying) **or** non-offensive (normal)**,** enabling supervised machine learning for binary text classification. The dataset includes comments from popular online platforms where cyberbullying incidents frequently occur, such as social networking and discussion websites. Each record primarily consists of a text message and its corresponding class label. Before model training, the textual data undergoes several preprocessing operations, including removal of punctuation, special characters, stop words, and unnecessary whitespace. Tokenization, stemming, and lemmatization are then applied to normalize the text and improve its quality. Finally, the processed messages are transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique, which captures the significance of words based on their frequency and distribution across the dataset.

The prepared dataset is divided into training and testing subsets to evaluate the performance of different machine learning algorithms. Approximately 80% of the data is utilized for model training, while the remaining 20% is reserved for testing and validation. The training data enables the classifiers to learn meaningful patterns associated with offensive and non-offensive language, whereas the testing data measures the model's ability to classify previously unseen messages accurately. Three machine learning algorithms, namely AdaBoost**,** Stochastic Gradient Descent (SGD)**,** and Multinomial Naïve Bayes**,** are trained using the same dataset to ensure a fair performance comparison. Standard evaluation metrics, including accuracy, precision, recall, and F1-score**,** are employed to assess classification effectiveness. The dataset provides sufficient linguistic diversity to evaluate the robustness of the proposed cyberbullying detection framework and demonstrates its capability to identify harmful online content while supporting automated user moderation in social media environments.

IV. PROPOSED METHODOLOGY

The proposed system presents an intelligent machine learning-based framework for detecting cyberbullying in social media posts by automatically classifying textual content as offensive or non-offensive. Initially, the collected text data undergoes preprocessing to improve data quality by removing punctuation, stop words, special characters, and unnecessary symbols, followed by tokenization, stemming, and lemmatization. The cleaned text is then converted into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique, which captures the importance of words within the dataset.

Three machine learning algorithms, namely AdaBoost**,** Stochastic Gradient Descent (SGD)**,** and Multinomial Naïve Bayes**,** are trained using the processed data to identify cyberbullying content accurately. In addition, a Support Vector Machine (SVM)-based sentiment analysis model is incorporated to determine the emotional polarity of user posts, thereby providing additional contextual information that improves the overall reliability of cyberbullying detection.

The proposed framework also includes an automated user monitoring and moderation mechanism to enhance online safety. Whenever a user submits a post, the system analyzes both its sentiment and offensive nature before allowing it to be published. If offensive content is detected, the user's offensive count is automatically updated in the database. The administrative module continuously monitors user activities and identifies users who repeatedly post abusive content. Once a predefined threshold is exceeded, the administrator can block the

corresponding account to prevent further harmful interactions.

The performance of the proposed system is evaluated using standard metrics such as accuracy, precision, recall, and F1-score**,** allowing comparison among different machine learning classifiers. The proposed approach provides an efficient, scalable, and practical solution for reducing cyberbullying, supporting automated content moderation, and promoting a safer social media environment through intelligent text classification and user behavior analysis.

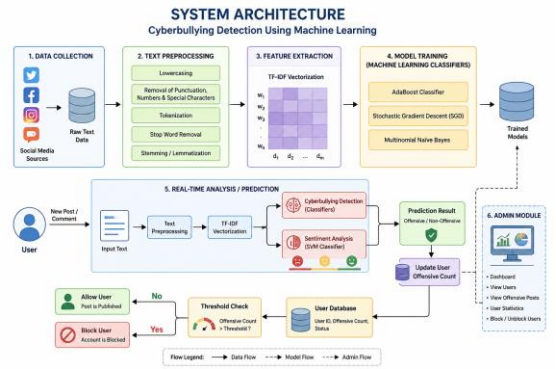

Figure 1. System Architecture of the Proposed Cyberbullying Detection Framework

The Figure [1] proposed system architecture illustrates the complete workflow for detecting cyberbullying in social media posts using machine learning techniques. Initially, text data is collected from social media platforms and stored for processing. The collected data undergoes preprocessing steps, including lowercasing, punctuation removal, tokenization, stop-word removal, stemming, and lemmatization, to improve text quality. The cleaned text is then converted into numerical feature vectors using the TF-IDF feature extraction technique. These features are used to train multiple machine learning classifiers, namely AdaBoost, Stochastic Gradient Descent (SGD), and Multinomial Naïve Bayes, while an SVM-based sentiment analysis module determines the emotional polarity of each post. During prediction, newly submitted messages are analysed to determine whether they are offensive or non-offensive. The system automatically updates the user's offensive count and stores the information in the database. An administrative module continuously monitors user activities, views offensive posts, evaluates user statistics, and blocks users who repeatedly violate community guidelines. This integrated architecture provides accurate cyberbullying detection, automated user moderation, and efficient content management, thereby promoting a safer and more secure online social networking environment.

V.RESULT AND DISCUSSION

The proposed cyberbullying detection system successfully classified social media posts as offensive or non-offensive using machine learning techniques. After preprocessing the textual data and extracting features using the TF-IDF method, the AdaBoost, Stochastic Gradient Descent (SGD), and Multinomial Naïve Bayes classifiers were trained and evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score. The experimental results demonstrated that the proposed framework effectively identified harmful online content while maintaining reliable classification performance. The integrated sentiment analysis module further improved the decision-making process by providing contextual information about user posts. In addition to message classification, the system accurately monitored users' offensive activities by maintaining an offensive count and enabling administrators to identify and block users who repeatedly violated community guidelines. Overall, the proposed framework achieved efficient and dependable cyberbullying detection, reduced manual moderation efforts, and provided a practical solution for improving user safety and maintaining a healthier online social media environment.

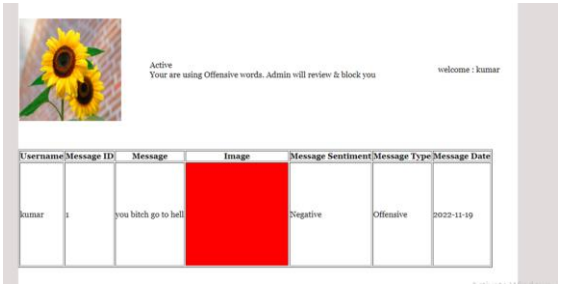

Figure 2: Offensive Message Detection Result

The figure [2] illustrates the offensive message detection module of the proposed cyberbullying detection system. A user-submitted message is analyzed using machine learning and sentiment analysis. The system classifies the message as **Offensive** with **Negative** sentiment, displays a warning notification, records the violation, and enables the administrator to review and block users involved in repeated abusive behaviour.

Algorithm Name	Accuracy	Precision	Recall	F1 Score
AdaBoost	87.06041478809739	87.86630167582548	78.52103284907989	81.59066215556923
SGD	97.29486023444545	96.63220371040427	96.53338773902196	96.58364241285841
Multinomial Naive Bayes	90.21641118124435	86.76832814557365	90.53203755221284	88.29316013361115

Figure 3: Performance Comparison of Machine Learning Algorithms

The figure [3] presents the performance comparison of the proposed machine learning classifiers using accuracy, precision, recall, and F1-score. Among the evaluated models, **Stochastic Gradient Descent (SGD)** achieved the highest performance with **97.29% accuracy** and **96.58% F1-score**, outperforming **AdaBoost** and **Multinomial Naïve Bayes** in cyberbullying detection.

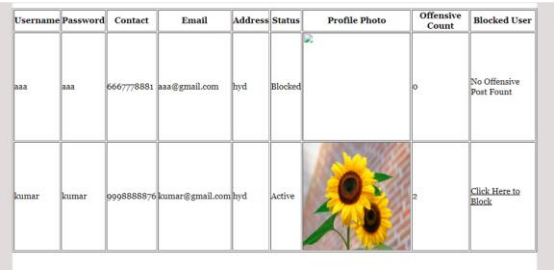

Figure 4: User Management and Offensive User Monitoring Module

The figure [4] illustrates the administrator dashboard for monitoring registered users and managing cyberbullying activities. It displays user information, account status, profile details, and offensive post counts. Users exceeding the predefined threshold can be blocked by the administrator, enabling effective moderation and preventing repeated cyberbullying incidents on the social media platform.

DISCUSSION

The experimental evaluation demonstrates that the proposed cyberbullying detection framework effectively identifies offensive social media content using machine learning techniques combined with sentiment analysis. The application of text preprocessing and TF-IDF feature extraction significantly improved the quality of textual data, enabling the classifiers to distinguish offensive and non-offensive messages with greater accuracy. A comparative analysis of AdaBoost, Stochastic Gradient Descent (SGD), and Multinomial Naïve Bayes showed that each algorithm performed efficiently in classifying user posts, with slight variations in performance based on the characteristics of the dataset. The integration of sentiment analysis provided additional contextual information, reducing the likelihood of incorrect classifications for emotionally expressive messages. Furthermore, the automated user monitoring and offensive count mechanism enhanced the system's capability to identify repeated policy violations and support administrative decision-making. Although the proposed framework demonstrates reliable performance, its effectiveness depends on the quality and diversity of the training dataset. Future improvements may include transformer-based language models, multilingual support, sarcasm detection, and multimodal analysis to further increase classification accuracy and adaptability across different social media platforms.

VI.CONCLUSION

The proposed machine learning-based cyberbullying detection framework provides an effective solution for identifying offensive content on social media platforms and supporting safer online communication. By combining comprehensive text preprocessing, TF-IDF feature extraction, sentiment analysis, and multiple machine learning classifiers, the system accurately classifies user posts as offensive or non-offensive. The automated user monitoring mechanism further enhances the framework by tracking repeated violations and assisting administrators in taking appropriate moderation actions against users who consistently post harmful content. Experimental evaluation using accuracy, precision, recall, and F1-score demonstrates the reliability and effectiveness of the proposed approach for cyberbullying detection. The system reduces the need for manual content moderation while improving the efficiency of online community management. Although the proposed framework achieves promising results, future work can focus on incorporating transformer-based language models, multilingual text processing, sarcasm detection, and multimodal content analysis to further improve detection accuracy, scalability, and adaptability across diverse social media platforms.

REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, “Towards the detection of cyberbullying based on social network mining techniques,” in Proceedings of 4th International

Conference on Behavioral, Economic, and SocioCultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.

[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," 2011, doi: 10.1109/ICMLA.2011.152.

[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.

[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.

[9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv. 2018.

[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," 2016, doi: 10.1109/ASONAM.2016.7752420.

[11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," 2016, doi: 10.18653/v1/n16-2013.

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

[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," 2017, doi: 10.1145/3038912.3052591.

[14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif. Intell. Rev., vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[16] Kumar, Anuj. "Optimization of Process Parameters in Metal Additive Manufacturing for Defect Reduction Using Machine Learning." International Journal of Engineering Research and Science & Technology, Vol. 14, No. 1, 2018, pp. 41-60

[17] Shashank, A. (2025). Self-Healing Data Pipelines for Enhanced Reliability: A Paradigm Shift in Enterprise Data Management. Journal of Computer Science and Technology Studies, 7(8), 1097-1104.