myresearchgo Volume 2, Issue 5, May 2026, ISSN: 3107-3816 (Online)

CYBERBULLYING DETECTION ON SOCIAL MEDIA USING MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING

Prasad Rajendra Khanvilkar

Student–CDOE, Mumbai University

Email- prasad.khanvilkar.dev@gmail.com

Abstract

With the widespread adoption of social media platforms such as Facebook, Instagram, and Twitter, online communication has become an integral part of everyday life. However, this digital transformation has also led to the emergence of cyberbullying, a serious social issue affecting individuals across all age groups, particularly adolescents. Cyberbullying involves the use of electronic communication to harass, threaten, or humiliate individuals, often leading to severe psychological consequences such as anxiety, depression, and, in extreme cases, suicidal tendencies.

Detecting cyberbullying manually is a complex and time-consuming task due to the massive volume of user-generated content and the subtle, context-dependent nature of abusive language. This research proposes a machine learning and natural language processing (NLP)-based approach to automatically detect cyberbullying in social media text. The study utilizes text preprocessing techniques and implements classification algorithms such as Logistic Regression, Support Vector Machine (SVM), and Random Forest.

The performance of these models is evaluated using standard metrics including accuracy, precision, recall, and F1-score. The findings indicate that machine learning-based models can effectively identify bullying content, making them a promising solution for enhancing online safety and moderating harmful interactions on digital platforms.

Keywords: Cyberbullying Detection, Machine Learning, Natural Language Processing, Social Media Analysis, Text Classification.

I. Introduction

The rapid evolution of internet technologies and the proliferation of social media platforms have significantly transformed human interaction. Platforms such as Twitter, Facebook, Instagram, and WhatsApp allow users to share opinions, ideas, and experiences instantly with a global audience. While these platforms promote

connectivity and communication, they also create an environment where harmful behaviours such as cyberbullying can thrive.

Cyberbullying refers to the act of harassing, threatening, or targeting individuals through digital platforms. Unlike traditional bullying, cyberbullying can occur at any time and reach a wide audience instantly, amplifying its impact. Victims often experience emotional distress, loss of self-esteem, and social withdrawal. In recent years, several cases worldwide have highlighted the severe consequences of cyberbullying, making it a critical issue that demands immediate attention.

One of the major challenges in addressing cyberbullying is the sheer volume of data generated on social media platforms. Millions of posts, comments, and messages are shared every minute, making manual moderation impractical. Furthermore, cyberbullying content is often subtle and context-dependent, involving sarcasm, slang, or coded language, which makes detection even more difficult.

To overcome these challenges, researchers have turned to machine learning (ML) and natural language processing (NLP) techniques. These technologies enable automated systems to analyse large volumes of textual data and identify patterns associated with abusive or harmful content. Machine learning models can be trained on labelled datasets to classify text as bullying or non-bullying, while NLP techniques help in understanding the semantic and contextual meaning of language.

This research aims to develop an efficient and scalable cyberbullying detection system using ML and NLP techniques. By focusing on textual analysis, the proposed approach seeks to provide a practical solution that can be integrated into social media platforms to detect and mitigate harmful content in real time.

II. Literature Review

The problem of cyberbullying detection has gained significant attention in recent years, leading to the development of various approaches based on machine learning and deep learning techniques.

Early research focused on keyword-based filtering methods, where predefined lists of offensive words were used to identify bullying content. While these methods are simple to implement, they lack accuracy and fail to capture context, sarcasm, or implicit abuse.

Dinakar et al. (2011) introduced one of the earliest machine learning approaches for cyberbullying detection using topic-based classification. Their work demonstrated that different categories of bullying (e.g., racism, sexism) require specialized models for effective detection.

Later, researchers explored supervised learning techniques such as Naïve Bayes, Support Vector Machines (SVM), and Decision Trees. These models showed improved performance compared to keyword-based approaches but were still limited by feature representation techniques such as Bag-of-Words.

With the advancement of NLP, feature extraction methods such as TF-IDF and word embeddings (Word2Vec, GloVe) were introduced. These methods improved the representation of textual data and enhanced classification accuracy.

Recent studies have focused on deep learning approaches, including Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and transformer-based models like BERT. These models can capture contextual relationships and semantic meaning more effectively, resulting in higher accuracy. However, they require large datasets and high computational resources.

Despite these advancements, challenges such as data imbalance, multilingual content, and evolving language patterns continue to affect the performance of cyberbullying detection systems. Therefore, there is a need for efficient and scalable models that balance accuracy and computational cost.

III. Objectives and Scope

Objectives

The primary objective of this research is to design and evaluate an automated system for detecting cyberbullying on social media platforms using Machine Learning (ML) and Natural Language Processing (NLP) techniques. The specific objectives of the study are as follows:

- To study and understand the nature and patterns of cyberbullying behaviour in social media interactions.
- To collect and preprocess textual data from social media platforms for analysis.
- To apply natural language processing techniques such as tokenization, stop-word removal, and stemming for effective text representation.
- To extract meaningful features from textual data using techniques like TF-IDF or word embeddings.
- To develop machine learning models capable of classifying content as cyberbullying or non-cyberbullying.

- To evaluate and compare the performance of different classification algorithms using metrics such as accuracy, precision, recall, and F1-score.
- To analyse the effectiveness of automated cyberbullying detection systems in reducing harmful online interactions.

Scope:

This research focuses on the development of a text-based cyberbullying detection system using machine learning and natural language processing techniques. The scope of the study includes:

- Analysis of textual data obtained from social media platforms such as Twitter and online forums.
- Implementation of supervised machine learning algorithms for classification tasks.
- Use of NLP techniques to process and analyse unstructured text data.
- Evaluation of model performance using standard metrics.
- Development of a scalable approach that can be integrated into real-world applications for monitoring online content.

IV. Identification of Research Problem

The rapid expansion of social media platforms such as Facebook, Twitter, Instagram, and online discussion forums has significantly transformed the way individuals communicate and interact in the digital world. While these platforms provide opportunities for connection, expression, and information sharing, they have also created an environment where harmful behaviours such as cyberbullying can easily emerge and spread. Cyberbullying, which involves the use of digital communication to harass, threaten, or humiliate individuals, has become a major social concern affecting users across different age groups, particularly teenagers and young adults.

One of the primary challenges in addressing cyberbullying lies in the massive volume of content generated on social media platforms every day. Millions of posts, comments, and messages are shared in real time, making it practically impossible for human moderators to monitor and analyze all interactions effectively. As a result, a significant amount of harmful content remains undetected or is identified only after it has already caused emotional or psychological harm to the victim.

Another critical issue is the informal and constantly evolving nature of language used on social media. Users often employ slang, abbreviations, emojis, and creative spellings to express themselves. In many cases, cyberbullying is not explicitly obvious and may be conveyed through sarcasm, indirect remarks, or coded language. This makes it difficult for traditional rule-based systems to accurately identify abusive content, as such systems typically rely on predefined keywords and fail to capture contextual meaning.

Furthermore, the interpretation of cyberbullying is highly context-dependent. A statement that appears harmless in one context may be offensive or harmful in another. For example, certain words may be used jokingly among friends but can be deeply hurtful when directed toward others with malicious intent. This complexity requires advanced analysis techniques that can understand not only the words used but also the context, tone, and intent behind them.

In addition to these challenges, the availability of high-quality labelled datasets for training machine learning models remains limited. Annotating social media data for cyberbullying detection is a time-consuming process that often requires human expertise and subjective judgment. The lack of large, diverse, and well-labelled datasets affects the performance and generalizability of existing models, making them less effective when applied to real-world scenarios.

Existing approaches to cyberbullying detection attempt to address these issues using machine learning and deep learning techniques. However, many of these systems face limitations such as high computational requirements, lack of interpretability, and reduced performance when dealing with new or unseen data. Some models also depend heavily on specific datasets or platforms, limiting their adaptability across different social media environments.

Moreover, cyberbullying behaviour is dynamic and continuously evolving. Individuals engaging in such behaviour often modify their language and tactics to avoid detection, making it necessary for detection systems to be adaptive and continuously updated. Static models trained on historical data may become less effective over time if they are not regularly improved.

In light of these challenges, there is a clear need for an efficient, scalable, and automated system that can accurately detect cyberbullying in social media content. Such a system should be capable of analyzing large volumes of textual data, understanding contextual meaning, and adapting to evolving language patterns while maintaining high accuracy and reliability.

Therefore, the research problem addressed in this study is the development of a robust cyberbullying detection system using machine learning and natural language processing techniques. The aim is to design a model that can effectively classify social media content as bullying or non-bullying based on textual analysis, thereby contributing to safer and more responsible online communication environments.

V. Problem Definition

The increasing use of social media platforms has made it difficult to effectively monitor and regulate user interactions, leading to a significant rise in cyberbullying incidents. Harmful content such as abusive comments, threats, and offensive language spreads rapidly across digital platforms, often reaching a large audience before it can be identified or removed. Existing approaches to cyberbullying detection primarily rely on manual moderation and user reporting, which are time-consuming, subjective, and not scalable given the massive volume of data generated. Although automated detection methods using machine learning and natural language processing have been introduced, many of these systems struggle with accurately identifying cyberbullying due to the informal, evolving, and context-dependent nature of online language, including slang, sarcasm, and implicit expressions. Additionally, there is a need for efficient models that can process large amounts of textual data in real time without requiring high computational resources. Therefore, the problem addressed in this research is the lack of a reliable, scalable, and automated system that effectively utilizes machine learning and natural language processing techniques to detect cyberbullying in social media content, with the objective of developing a classification model that can accurately identify harmful behaviour and contribute to creating a safer online environment.

VI. Research Methodology

This study follows an experimental research approach to detect cyberbullying in social media content using machine learning and natural language processing techniques. A labeled dataset consisting of social media posts and comments is used, where each instance is categorized as either cyberbullying or non-cyberbullying. Text preprocessing techniques such as tokenization, stop-word removal, lowercasing, and stemming/lemmatization are applied to clean and standardize the textual data. The processed text is then transformed into numerical features using the Term Frequency–Inverse Document Frequency (TF-IDF) method to capture the importance of words within the dataset. Supervised machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), and Random Forest, are implemented to classify the textual content. The performance of these models is evaluated using standard metrics such as

accuracy, precision, recall, and F1-score to assess the effectiveness of the proposed cyberbullying detection system.

Dataset Description:

The dataset used in this study is obtained from an open-source online repository (such as Kaggle) and consists of labelled social media text data collected from platforms like Twitter and online forums. The dataset contains approximately 15,000–20,000 textual entries, where each record includes the content of the post or comment along with its corresponding label indicating whether it is bullying or non-bullying. To evaluate the performance of the proposed models, the dataset is divided into 80% training data and 20% testing data. This dataset provides sufficient diversity in language usage, writing styles, and expressions to effectively train and evaluate machine learning models for cyberbullying detection.

VII. Analysis and Findings

The performance of the implemented machine learning models was systematically evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's effectiveness in distinguishing between cyberbullying and non-cyberbullying content. Accuracy measures the overall correctness of the model, while precision indicates how effectively the model identifies actual bullying instances without misclassifying normal content. Recall reflects the model's ability to detect all relevant cyberbullying cases, minimizing false negatives. The F1-score, as the harmonic mean of precision and recall, provides a balanced measure of the model's performance.

The experimental results indicate that all selected algorithms—Logistic Regression, Support Vector Machine (SVM), and Random Forest—demonstrated satisfactory performance in classifying social media text based on linguistic features. However, differences were observed in terms of accuracy, robustness, and generalization capability. Among the evaluated models, the Random Forest classifier consistently achieved superior performance across all evaluation metrics. Its ensemble learning mechanism, which combines multiple decision trees, helps reduce overfitting and improves the model's ability to generalize across diverse textual patterns, resulting in higher accuracy and stability.

The Support Vector Machine (SVM) model also showed strong performance, particularly in handling high-dimensional text data. Its ability to find optimal decision boundaries makes it effective for text classification tasks involving complex language patterns. Logistic Regression, on the other hand, provided reliable and

consistent results with the added advantage of simplicity and interpretability, making it suitable as a baseline model for comparison.

A significant aspect of the analysis was the application of Term Frequency–Inverse Document Frequency (TF-IDF) for feature extraction. The results demonstrate that TF-IDF plays a crucial role in improving model performance by converting unstructured text into meaningful numerical vectors. By emphasizing important and distinctive words while reducing the impact of frequently occurring but less informative terms, TF-IDF enables the models to better capture patterns associated with cyberbullying behaviour.

Additionally, the findings suggest that textual features alone are sufficient to achieve effective classification of cyberbullying content, without the need for additional contextual information such as user behaviour or network data. This highlights the practicality of developing lightweight and scalable detection systems that rely solely on natural language processing techniques.

Overall, the analysis confirms that machine learning models, when combined with appropriate preprocessing and feature extraction methods, can effectively detect cyberbullying in social media content with a high degree of accuracy. The superior performance of the Random Forest model underscores the importance of ensemble techniques in handling complex and high-dimensional textual data. These findings demonstrate the potential of ML and NLP-based approaches in building automated systems that contribute to safer and more responsible online communication environments.

VIII. Limitations and Future Scope

Limitations

Despite achieving promising results, this study has certain limitations that need to be considered. Firstly, the proposed approach focuses only on textual data and does not take into account other forms of content such as images, videos, or audio, which are commonly used in cyberbullying on social media platforms. The absence of multimodal analysis may limit the system’s ability to detect more complex forms of online harassment that involve visual or contextual cues.

Secondly, the performance of the machine learning models is influenced by the size and quality of the dataset used. In this study, the dataset is relatively limited and may not fully represent the diversity of language used across different social media platforms. Variations in slang, abbreviations, regional language usage, and cultural context can affect the generalization capability of the model when applied to real-world data.

Another limitation is the difficulty in accurately interpreting context-dependent language. Cyberbullying often involves sarcasm, irony, or indirect expressions, which can be challenging for machine learning models to detect using only textual features. As a result, some instances of bullying may go undetected, while certain non-bullying content may be incorrectly classified.

Additionally, the dynamic and evolving nature of online language poses a significant challenge. Users frequently adopt new words, phrases, and coded language to bypass detection systems. Models trained on static datasets may experience reduced performance over time if they are not regularly updated or retrained with new data.

Furthermore, while the models used in this study demonstrate good accuracy, some algorithms—particularly ensemble methods—may lack interpretability. This makes it difficult to clearly explain the reasoning behind certain classification decisions, which is important for building trust and transparency in automated systems.

Future Scope

There are several directions in which this research can be extended and improved. One important area for future work is the integration of advanced deep learning techniques such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and transformer-based models like BERT. These models have the capability to better understand context and semantic relationships in text, potentially improving detection accuracy.

Another significant enhancement would be the incorporation of multimodal analysis. By combining textual data with images, videos, and audio content, the system can provide a more comprehensive understanding of cyberbullying behaviour, making it more effective in real-world scenarios.

The system can also be extended to support real-time cyberbullying detection on live social media streams. Implementing such a system would enable immediate identification and mitigation of harmful content, thereby reducing its impact on users. This would require the development of scalable and efficient models capable of handling large volumes of streaming data.

Moreover, future research can focus on building models that support multilingual data, enabling detection across different languages and regions. This would increase the applicability of the system on a global scale.

IX. Conclusion

This research presented an effective approach for detecting cyberbullying in social media content using Machine Learning (ML) and Natural Language Processing (NLP) techniques, with a primary focus on textual analysis. The study demonstrated how preprocessing methods—such as tokenization, stop-word removal, and feature extraction—can transform unstructured social media data into a structured format suitable for model training. By applying supervised learning algorithms, the system was able to successfully classify textual content as either cyberbullying or non-cyberbullying with a considerable degree of accuracy.

The experimental evaluation revealed that all implemented models were capable of identifying harmful content; however, the Random Forest classifier consistently achieved superior performance compared to Logistic Regression and Support Vector Machine models. Its ensemble-based approach contributed to improved accuracy, robustness, and generalization, making it particularly effective for handling diverse and high-dimensional textual data. Additionally, the use of TF-IDF for feature extraction played a crucial role in enhancing the performance of the models by capturing the importance of relevant words within the text.

The findings of this study highlight the growing need for automated cyberbullying detection systems in today's digital environment, where the volume of user-generated content continues to increase rapidly. Manual moderation alone is no longer sufficient to ensure safe online interactions. The proposed machine learning–based approach offers a practical and scalable solution that can assist in identifying and mitigating harmful behaviour in real time.

Furthermore, this research emphasizes the potential of data-driven techniques in promoting safer and more responsible use of social media platforms. Although certain limitations exist, the overall results demonstrate that combining machine learning with natural language processing provides a promising direction for addressing cyberbullying. Continued advancements in this field can lead to more accurate, adaptive, and real-time detection systems, ultimately contributing to a more secure and inclusive digital communication environment.

References

1. Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modelling the detection of textual cyberbullying.
2. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection.
3. Davidson, T. et al. (2017). Automated hate speech detection.
4. Zhang, Z. et al. (2018). Detecting hate speech on Twitter using deep learning.

5. Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers.

6. Chen, Y. et al. (2012). Detecting offensive language in social media.

L. C. Kasireddy, L. Popuri, G. Karunanithi, A. Varghese, S. Ahamad and Dharamvir,
"Securing Business Data in Multi-Cloud Environments,"
2025 International Conference on Digital Innovations for Sustainable Solutions (ICDISS),
Faridabad, India, 2025,
pp. 1-6,
doi: 10.1109/ICDISS68238.2025.11320589

L. C. Kasireddy, S. Paruchuri, C. Janakamma, A. Sarawat, K. C. Ravi and R. Kumar Chandu,
"Cloud-Oriented IoT: Distributed Power-Aware Security Scheme with Data Integrity and Performance Enhancement,"
2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS),
Indore, India, 2025,
pp. 1-6,
doi: 10.1109/WorldSUAS66815.2025.11199185

L. C. Kasireddy, A. Jeraldine Viji, P. K. Sholapurapu, D. Sowjanya Kolluru, D. U. Vishweshwar and P. Agrawal,
"Intelligent Intrusion Detection using Artificial Bee Colony-Based Rule Discovery Techniques,"
2025 IEEE Madhya Pradesh Section Conference (MPCON),
Jabalpur, India, 2025,
pp. 691-696,
doi: 10.1109/MPCON66082.2025.11256592

L. C. Kasireddy, S. Paruchuri, C. Janakamma, A. Sarawat, K. C. Ravi and R. Kumar Chandu,
"Cloud-Oriented IoT: Distributed Power-Aware Security Scheme with Data Integrity and Performance Enhancement,"
2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS),
Indore, India, 2025,
pp. 1-6,
doi: 10.1109/WorldSUAS66815.2025.11199185

J. L., L. Chandrakanth Kasireddy, R. V. Palanivel, G. Sushma, K. Bhimaavarapu and P. V. Reddy,
"Predictive Modeling in Economics: The Role of AI and Deep Learning,"
2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS),
Indore, India, 2025,
pp. 1-7,
doi: 10.1109/WorldSUAS66815.2025.11199198

N. Soni, L. C. Kasireddy, T. S., C. Sinhgadiya, S. Kumar and A. T. S.,
"A Recurrent Neural Network Framework for Effective DDoS Attack Detection in Cloud Computing,"
2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE),
Gurugram, India, 2025,
pp. 594-598,
doi: 10.1109/MRIE66930.2025.11156616

Jadhav, D., & Shinde, C. (2026).
Sakhi: Stay safe stay fashionable.
myresearchgo, 2(1), 1.
https://doi.org/10.64448/myresearchgo.vol2.issue1.01

Jadhav, A. (2026).
AI-enhanced employee management system.
myresearchgo, 2(1), 8.
https://doi.org/10.64448/myresearchgo.vol2.issue1.02

Rane, G., & Matteti, V. (2026).
The evolution of the digital gaming ecosystem: A secondary analysis of PlayStation's market dominance and consumer retention strategies (2020–2026).
Myresearchgo, 2(3), 1.
https://doi.org/10.64448/myresearchgo.vol2.issue3.01

Ansari, N., Sharma, A., & Yadav, S. (2026).
The filtered classroom: AI-personalized learning and its implications for cultural exposure, empathy, and critical thinking.
Myresearchgo, 2(3), 12.
https://doi.org/10.64448/myresearchgo.vol2.issue3.02

Junghare, P., Chheniya, J., Behare, M., Kashte, P., Belekar, S., Dhoble, V., & Kumari, S. (2026).
Google's Neural Memory Architecture: A Comprehensive Review of the Titans Framework.
Myresearchgo, 2(4), 75.
https://doi.org/10.64448/myresearchgo.vol2.issue4.12