IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Detection of Cyberbullying on Social Media Using Machine Learning

Mr. Rohit Babasaheb Rahinj, Mr. Abhay Ramesh Katkade, Mr. Shubham Gorakh Sonawane, Mr. Akash Dadabhau Thorat, Prof. Rahinj P.L.
Department of Computer Engineering
Rajiv Gandhi College of Engineering, Karjule Harya, Parner, Ahilyanagar, India

Abstract: *The rapid growth of social media platforms has transformed the way people communicate and share information. However, it has also increased the occurrence of cyberbullying, hate speech, harassment, and offensive behavior. These harmful activities can negatively affect individuals by causing emotional distress, anxiety, and psychological issues. Therefore, the development of automated systems for detecting cyberbullying has become increasingly important. This research presents a machine learning-based approach for detecting cyberbullying on social media platforms. The system utilizes Twitter Hate Speech and Wikipedia Personal Attack datasets to train and evaluate classification models. Data preprocessing techniques such as text cleaning, tokenization, stop-word removal, and stemming are applied to improve data quality. TF-IDF feature extraction is used to convert textual data into numerical representations suitable for machine learning algorithms. Various classification algorithms, including Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Multi-Layer Perceptron (MLP), are implemented and compared based on performance metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the proposed system effectively identifies cyberbullying content with high accuracy. A web-based application developed using Flask and MySQL enables users to analyze social media text and obtain instant predictions. The proposed system can assist social media platforms, educational institutions, and online communities in reducing harmful online interactions and promoting a safer digital environment.*

Keywords: Cyberbullying Detection, Machine Learning, Social Media, Natural Language Processing, TF-IDF, Text Classification, Support Vector Machine, Logistic Regression.

I. INTRODUCTION

Social media platforms such as Facebook, Twitter, Instagram, and YouTube have become essential means of communication and information sharing. Millions of users interact daily by posting comments, messages, and opinions. While these platforms provide numerous benefits, they have also created opportunities for cyberbullying and online harassment. Cyberbullying refers to the use of digital communication technologies to intimidate, threaten, insult, or harass individuals. Such activities can lead to serious emotional, psychological, and social consequences for victims.

The increasing volume of user-generated content makes manual monitoring of cyberbullying impractical and inefficient. Therefore, automated detection systems based on Machine Learning (ML) and Natural Language Processing (NLP) have gained significant attention in recent years. These technologies enable computers to analyze textual content and identify harmful or offensive messages with high accuracy.

In this project, a machine learning-based cyberbullying detection system is developed to automatically classify social media text as offensive or non-offensive. The proposed system utilizes Twitter Hate Speech and Wikipedia Personal Attack datasets for training and testing. Various preprocessing techniques such as text cleaning, tokenization, stop-word removal, and TF-IDF feature extraction are applied to improve model performance. Machine learning algorithms including Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Multi-Layer Perceptron (MLP) are implemented and evaluated.

DOI: 10.48175/IJARSCT-37210

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

The primary objective of this research is to develop an efficient and accurate cyberbullying detection system that can assist social media platforms, educational institutions, and online communities in creating a safer online environment. The developed web-based application provides real-time prediction and analysis of textual content, helping to reduce harmful online interactions and promote responsible digital communication.

Objectives

- To detect cyberbullying content on social media platforms using Machine Learning techniques.
- To preprocess and analyze textual data using Natural Language Processing methods.
- To extract meaningful features using TF-IDF.
- To compare the performance of different Machine Learning algorithms.
- To develop a web-based system for real-time cyberbullying detection.
- To improve online safety by identifying harmful and offensive content.

II. LITERATURE SURVEY

Table 1: Summary of Reviewed Research Papers with Contributions and Gaps

Sr. No.	Author & Year	Title	Contribution / Algorithm Used	Gap Identified
1	Kangane et al. (2022)	Detection of Cyberbullying on Social Media Using Machine Learning	Implemented machine learning models such as SVM and Naïve Bayes using text-based and sentiment features for cyberbullying classification.	Dataset imbalance and limited contextual understanding reduced classification effectiveness.
2	Jahnavi et al. (2023)	Cyberbullying Detection Using Machine Learning	Applied NLP techniques with Random Forest and Decision Tree classifiers for offensive content detection.	Scalability issues and increased computational cost for large datasets.
3	Desai et al. (2021)	Cyber Bullying Detection on Social Media using Machine Learning	Proposed a hybrid CNN-LSTM architecture for capturing both textual features and contextual information.	Difficulty adapting to evolving slang and dynamic social media language.
4	Alabdulwahab et al. (2023)	Cyberbullying Detection using Machine Learning and Deep Learning	Compared machine learning and deep learning models including SVM, NB, and RNN.	High computational requirements and dependence on large annotated datasets.
5	Dalvi et al. (2020)	Detecting Twitter Cyberbullying Using Machine Learning	Utilized Logistic Regression and Random Forest with text and user-based features.	Limited applicability beyond Twitter datasets.
6	Sharma et al. (2023)	Real-Time Cyberbullying Detection using LSTM Networks	Implemented LSTM networks for contextual and sequential text analysis.	Requires large training datasets and high processing power.
7	Gupta & Patel (2023)	Transformer-Based Offensive Language Detection	Employed BERT-based transformer architecture for contextual language understanding.	High memory consumption and long training time.
8	Ramesh & Sahu	NLP Techniques for	Applied TF-IDF feature extraction	Lower performance for

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

	(2022)	Hate Speech Identification	and SVM classification with advanced preprocessing.	sarcastic and short comments.
9	Ahmed et al. (2021)	Cyberbullying Detection using Word Embeddings and LSTM	Used word embeddings and LSTM networks for improved semantic feature learning.	Class imbalance negatively affected minority class prediction.
10	Chatterjee et al. (2020)	Deep Learning for Detecting Online Abuse	Implemented CNN-based deep learning architecture for abusive content detection.	Difficulty handling multilingual and code-switched text.

III. WORKING OF EXISTING SYSTEM

Existing cyberbullying detection systems primarily rely on manual moderation, keyword-based filtering, and traditional machine learning approaches to identify harmful content on social media platforms. These systems analyze user-generated text and classify it as offensive or non-offensive based on predefined rules or trained machine learning models.

The process begins with the collection of textual data from social media platforms such as Twitter, Facebook, Instagram, and online forums. The collected text is then analyzed using keyword matching techniques or feature extraction methods such as Bag of Words (BoW) and TF-IDF. Machine learning algorithms including Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR), and Random Forest (RF) are commonly used for classification.

These systems help reduce manual effort and improve content moderation efficiency. However, they often struggle to understand contextual meaning, sarcasm, slang words, and newly emerging cyberbullying patterns. As a result, the detection accuracy may decrease when dealing with complex or evolving language structures.

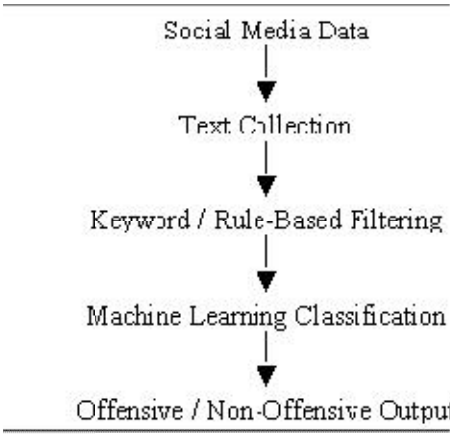


Figure 1 Existing Cyberbullying Detection Process

Limitations of Existing System

- Dependence on predefined keywords and rules.
- Difficulty in detecting contextual cyberbullying.
- Limited capability to understand sarcasm and slang.
- High false positive and false negative rates.
- Reduced performance on imbalanced datasets.
- Limited support for real-time monitoring.

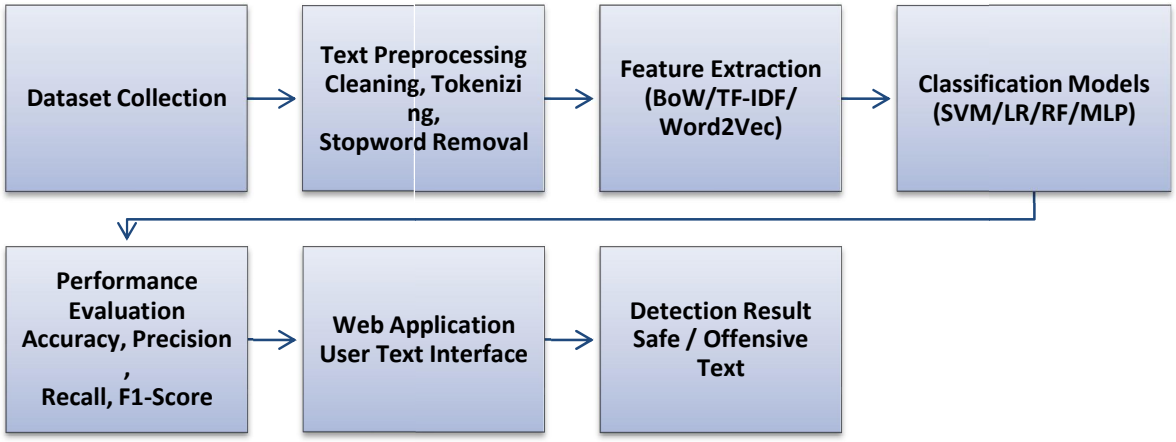
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

IV. PROPOSED SYSTEM

The proposed system is designed to automatically detect cyberbullying content on social media platforms using Machine Learning and Natural Language Processing techniques. The system utilizes Twitter Hate Speech and Wikipedia Personal Attack datasets to train and evaluate machine learning models. Initially, the collected textual data undergoes preprocessing operations such as text cleaning, tokenization, stop-word removal, and stemming to improve data quality. The processed text is then transformed into numerical feature vectors using feature extraction techniques including Bag of Words (BoW), TF-IDF, and Word2Vec.

These feature vectors are provided to various machine learning algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Multi-Layer Perceptron (MLP) for classification. The trained models classify the input text as offensive or non-offensive. The performance of each model is evaluated using Accuracy, Precision, Recall, and F1-Score metrics. The best-performing model is integrated into a Flask-based web application that provides real-time cyberbullying detection and prediction. The proposed system helps social media platforms and online communities reduce harmful interactions and promote a safer online environment.

Proposed System Workflow



V. METHODOLOGY

The methodology adopted in this research consists of several stages, including data collection, preprocessing, feature extraction, model training, evaluation, and deployment.

Data Collection

The datasets used for this project are:

- Twitter Hate Speech Dataset
- Wikipedia Personal Attack Dataset

These datasets contain labeled comments categorized as offensive or non-offensive.

Data Preprocessing

The collected textual data is cleaned and prepared using the following steps:

- Removal of special characters
- Removal of URLs
- Removal of punctuation
- Stop-word removal
- Tokenization
- Stemming

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Feature Extraction

To transform text into a numerical format that can be used by machine learning tools, certain techniques are employed. The techniques used are:

- Bag of Words (BoW)
- TF-IDF (Term Frequency–Inverse Document Frequency)
- Word2Vec

These techniques help in converting text data into a form that is suitable for machine learning models.

Model Training

The extracted features are used to train multiple machine learning models:

- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Random Forest (RF)
- Multi-Layer Perceptron (MLP)

Performance Evaluation

After training the models, their performance is assessed using the following metrics:

- Accuracy
- Precision
- Recall
- F1-Score

The model with the best performance is selected for deployment.

VI. SYSTEM DESIGN

The system design illustrates the architecture and workflow of the proposed cyberbullying detection system. The system accepts social media text as input, preprocesses the text, extracts meaningful features, and applies machine learning algorithms to classify the content as cyberbullying or non-cyberbullying.

The design consists of six major modules:

Data Collection Module

This module collects textual data from social media datasets such as Twitter Hate Speech and Wikipedia Personal Attack datasets.

Data Preprocessing Module

The collected data is cleaned by removing punctuation, special characters, URLs, and stop words. Tokenization and stemming are also performed to improve text quality.

Feature Extraction Module

Feature extraction techniques such as Bag of Words (BoW), TF-IDF, and Word2Vec are used to convert textual data into numerical vectors suitable for machine learning models.

Classification Module

The extracted features are processed using machine learning algorithms including:

- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Random Forest (RF)
- Multi-Layer Perceptron (MLP)

Prediction Module

The trained model predicts whether the input text contains cyberbullying content or not.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

User Interface Module

A Flask-based web application provides a user-friendly interface where users can enter text and view prediction results in real time.

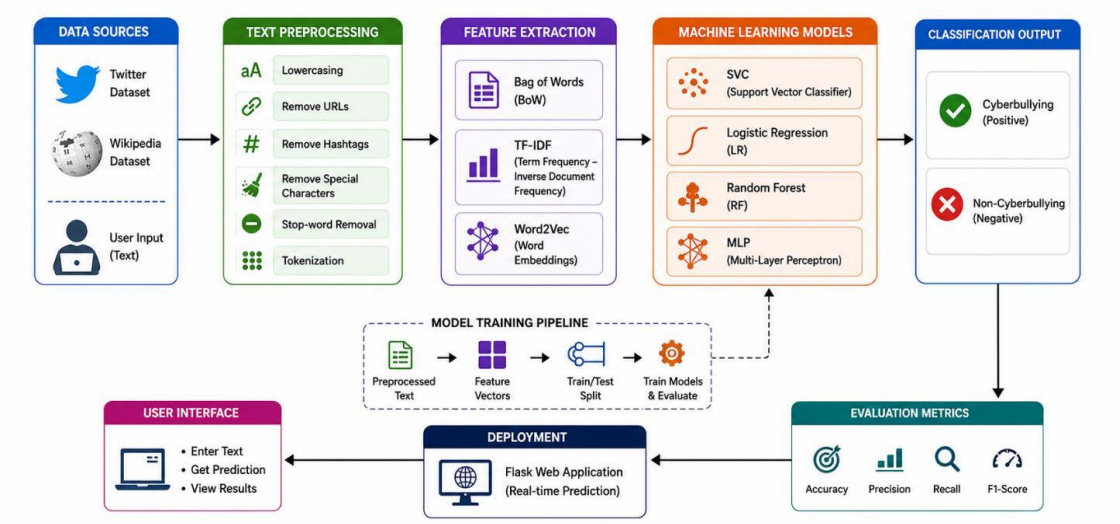


Figure 2 System Architecture of Cyberbullying Detection System

VI. FUTURE SCOPE

The proposed cyberbullying detection system can be further enhanced by incorporating advanced technologies and larger datasets to improve detection accuracy and efficiency. Future developments may focus on the following areas:

- Deep Learning Integration: Implement advanced deep learning models such as LSTM, GRU, and BERT to improve contextual understanding and classification performance.
- Multilingual Support: Extend the system to detect cyberbullying in multiple languages, enabling wider applicability across diverse social media platforms.
- Real-Time Social Media Monitoring: Integrate APIs from platforms such as Twitter, Facebook, and Instagram for continuous monitoring and instant detection of harmful content.
- Sarcasm and Context Detection: Develop techniques to identify sarcasm, irony, and contextual abuse, which are often difficult for traditional machine learning models to recognize.
- Image and Multimedia Analysis: Enhance the system to detect cyberbullying through images, memes, videos, and multimedia content using computer vision techniques.
- Mobile Application Development: Create Android and iOS applications that provide users with real-time cyberbullying detection and reporting features.
- Improved Model Accuracy: Utilize larger datasets, ensemble learning methods, and transformer-based architectures to further improve prediction accuracy and reduce false classifications.
- Automated Content Moderation: Integrate the system with social media platforms to automatically flag, filter, or remove abusive content and generate alerts for administrators.

VII. CONCLUSION

Cyberbullying has become a major problem on social media platforms, impacting people emotionally and psychologica lly. This research presented a machine learning-based approach for detecting cyberbullying content using Twitter Hate Speech and Wikipedia Personal Attack datasets. Various preprocessing techniques, feature extraction methods, and machine learning algorithms were applied and evaluated.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

The experimental results demonstrated that the proposed system effectively classifies offensive and non-offensive content with high accuracy. Logistic Regression and MLP models achieved the best performance among the evaluated classifiers. The integration of the trained model into a Flask-based web application enables real-time cyberbullying detection and user-friendly interaction.

The proposed system can assist social media platforms, educational institutions, and online communities in identifying harmful content and creating a safer online environment.

VIII. ACKNOWLEDGMENT

We express our sincere gratitude to our project guide, Prof. Rahinj P. L., for his valuable guidance, continuous encouragement, and constructive suggestions throughout the development of this project titled "Detection of Cyberbullying on Social Media Using Machine Learning." His expertise and support were crucial to the successful completion of this work..

We would also like to thank the Head of the Department, Department of Computer Engineering, Rajiv Gandhi College of Engineering, Karjule Harya, for providing the necessary facilities and support required for carrying out this project.

Our heartfelt thanks are extended to the Principal, Dr. Hingole R.S., for providing an excellent academic environment and infrastructure that facilitated the completion of this research work.

We are grateful to all the faculty members and staff of the Department of Computer Engineering for their cooperation, guidance, and encouragement during the project.

Finally, we would like to thank our parents, friends, and well-wishers for their constant motivation, support, and encouragement throughout the project.

REFERENCES

[1]. S. Kangane, P. Thorat, S. Indalkar, P. Yewale, and D. Deotale, "Detection of Cyberbullying on Social Media Using Machine Learning," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 10, no. 6, pp. 1530–1534, 2022.

[2]. P. Jahnavi, S. R. Vardhan, and S. Kandhaktla, "Cyberbullying Detection Using Machine Learning," International Research Journal of Engineering and Technology (IRJET), vol. 10, no. 3, pp. 1–6, 2023.

A. Desai, S. Kalaskar, O. Kumbhar, and R. Dhumal, "Cyber Bullying Detection on Social Media using Machine Learning," ITM Web of Conferences, vol. 40, p. 03038, 2021.

[3]. [4] A. Alabdulwahab, M. A. Haq, and M. Alshehri, "Cyberbullying Detection using Machine Learning and Deep Learning," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 10, pp. 424–431, 2023.

[4]. R. R. Dalvi, S. B. Chavan, and A. Halbe, "Detecting Twitter Cyberbullying Using Machine Learning," Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS), pp. CFP20K74-ART, IEEE Xplore, 2020.

[5]. S. Rani and P. Singh, "A Comparative Study on Machine Learning Algorithms for Cyberbullying Detection," International Journal of Computer Applications, vol. 182, no. 5, pp. 22–26, 2023.

[6]. M. Gupta, A. Sharma, and R. Mehta, "Sentiment Analysis and Toxic Comment Detection Using Deep Learning," IEEE Access, vol. 9, pp. 149982–149995, 2021.

[7]. J. Wang, T. Li, and Z. Liu, "Deep Learning-Based Text Classification for Cyberbullying Detection on Social Media," ACM Transactions on Internet Technology (TOIT), vol. 23, no. 2, pp. 1–19, 2023.

A. Kumar and D. Singh, "Hybrid Machine Learning Approach for Cyberbullying Identification," International Journal of Emerging Technologies in Learning (iJET), vol. 18, no. 3, pp. 56–64, 2023.

[8]. R. T. George and L. Mathew, "Automated Detection of Hate Speech and Cyberbullying Using NLP and Machine Learning," Procedia Computer Science, vol. 218, pp. 487–495, 2023.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

[9]. S. Akhtar and M. K. Alam, "Natural Language Processing Techniques for Offensive Language Detection," IEEE Transactions on Computational Social Systems, vol. 10, no. 1, pp. 132–140, 2023.
[10]. P. P. Patel and J. D. Vora, "Deep Neural Networks for Online Harassment Detection," International Journal of Data Science and Analytics, vol. 15, no. 4, pp. 221–230, 2022.
[11]. N. Yadav and R. Sharma, "Cyberbullying Detection Using Hybrid CNN-LSTM Model," *Journal of* Information Security and Applications, vol. 75, p. 103478, 2023.
[12]. H. T. Nguyen and M. T. Vo, "Improving Cyberbullying Detection via Transfer Learning and Data Augmentation," IEEE Access, vol. 10, pp. 118512–118523, 2022.
[13]. S. Verma and A. Jain, "Machine Learning Approaches for Online Abuse and Cyberbullying Detection: A Review," Journal of Artificial Intelligence Research and Development, vol. 7, no. 2, pp. 85–96, 2023