



International Journal of Advanced Research in Education and Technology (IJARETY)

Volume 13, Issue 3, May - June 2026

Impact Factor: 9.971



Cyber-Bullying Detection in Hinglish Languages using Machine Learning

Aashna Mainoddin Mulla

Dept. of Master of Computer Applications, CMR Institute of Technology, Bengaluru, India

ABSTRACT: Cyberbullying is one of the most pressing challenges arising from the widespread use of social media. Statistics indicate that approximately 36.5% of people believe they have experienced cyberbullying—more than double the figures recorded in 2007, with a further increase observed between 2018 and 2019. While existing solutions partially address this issue, they are limited in scope or employ insufficiently efficient algorithms. This paper investigates fundamentally new approaches to automatically detect cyberbullying incidents across tweets, comments, and messages on social media networks. A real-time dataset consisting of Twitter data, WhatsApp chats, and YouTube comments in English and Hinglish language was collected and labelled for cyberbullying. Using Machine Learning (ML) combined with Natural Language Processing (NLP), the proposed system identifies bullying content at its origin to achieve superior detection performance.

KEYWORDS: Cyberbullying, Machine Learning, Natural Language Processing, Hinglish Languages, Social Media.

I. INTRODUCTION

The internet is the world's largest platform for communicating, connecting, and sharing ideas, content, photos, videos, views, and daily updates globally. Social media platforms such as Twitter, YouTube, Instagram, LinkedIn, Facebook, and WhatsApp are among the most extensively used. With almost every person engaging on social media today, the freedom of expression is increasingly being exploited, giving rise to cyberbullying as a major societal concern.

Jaradat et al. [1] proposed a deep learning framework using BiLSTM, CNN-BiLSTM, BiLSTM-GRU, and ANN models to detect multiple categories of cyberbullying from a dataset of 47,692 tweets, with the BiLSTM model achieving approximately 91% accuracy. Perera and Fernando [2] developed a supervised machine learning approach using SVM, Logistic Regression, and Naïve Bayes, achieving over 90% accuracy with TF-IDF features and sentiment analysis. Jha et al. [3] introduced MemeGuard, a framework combining LLMs and vision-language models to detect toxic memes. Menaka et al. [4] addressed code-mixed Tanglish language, with Random Forest achieving 89.2% accuracy.

Large-scale behavioral analysis studies have also been conducted to understand cyberbullying dynamics.[5] For instance, research using 10 Weibo datasets and five public datasets containing over 1.47 million posts employed deep learning models including CNN, LSTM, and transformer-based architectures to identify different cyberbullying roles and behaviors across social networks. Hybrid machine learning and deep learning models have also been explored to improve robustness and fairness in detection systems. [6] proposed a hybrid framework combining CNN, LSTM, ensemble learning, and feature engineering techniques. Their approach achieved classification accuracy above 90% and demonstrated that synthetic data generated using large language models can enhance fairness and model performance. In addition to technical detection models, several studies have focused on the psychological and behavioral aspects of cyberbullying

This paper contributes by developing an efficacious technique integrating Machine Learning and NLP to detect offensive or hateful words in both English and Hinglish languages. The proposed system builds a chat application that predicts whether text entered in group chats is bullying or non-bullying in real time, raising awareness and preventing the devastating after-effects of cyberbullying.

II. RELATED WORK

Apart from algorithmic approaches, several studies have investigated the psychological and behavioral aspects of cyberbullying. El Mezayen et al. [8] conducted a cross-sectional study involving 400 nursing students and identified a significant statistical relationship between cyber-aggression and demographic characteristics ($p < 0.05$). Kathuria et al. [9] investigated cyberbullying detection using TF-IDF features combined with machine learning models such as Naïve Bayes, SVM, and ensemble methods on the Daturks tweet dataset. The study showed that ensemble models

outperformed individual classifiers in detecting abusive content. In addition to technical detection systems, some studies have examined the psychological effects of cyberbullying. Riska et al. [10] analyzed the relationship between emotion regulation ability and cyberbullying behavior among vocational school students and found a significant negative correlation ($r = -0.483$) between emotional regulation and cyberbullying tendencies. Jesseline et al. [11] explored the psychological impact of cyberbullying among young adult victims using a qualitative phenomenological approach. Their study reported that victims often experience stress, anxiety, and reduced self-esteem as a result of online harassment. Recent studies have also focused on improving detection accuracy using advanced deep learning techniques. Jesseline et al. [12] proposed a contrastive self-supervised learning framework combined with BERT and Bi-LSTM models, achieving a macro F1-score of 0.9231, which significantly outperformed baseline deep learning models.

Big data analytics has also been applied to cyberbullying detection. A study published in IJARETY [13] utilized NLP, sentiment analysis, CNN, RNN, and graph-based machine learning techniques to analyze large-scale social media datasets and improve cyberbullying prediction across platforms. Comparative studies have also evaluated the effectiveness of machine learning and deep learning models. A review published in ANJS [14] compared algorithms such as SVM, Random Forest, Naïve Bayes, LSTM, BiLSTM, and GRU, concluding that deep learning models generally outperform traditional machine learning algorithms in cyberbullying detection tasks. Researchers have also proposed transformer-based models for cyberbullying detection. A study published in the Information journal [15] combined XLNet with Deep Bi-LSTM, achieving improved performance compared to traditional methods such as BERT and SVM for Chinese cyberbullying datasets. Another recent study [16] proposed a Modified DeBERT a model combined with a Gated Broad Learning System, achieving 95.41% accuracy on the SOSNet dataset and over 91% accuracy on other benchmark datasets, demonstrating the effectiveness of transformer-based models.

Explainable AI techniques have also been applied for multilingual cyberbullying detection. A study using MURIL with Explainable AI techniques achieved high accuracy across multiple Hindi-English datasets, including 94.63% accuracy on the Mendeley Indo-Hate Speech dataset [17]. Optimization techniques have also been explored to improve model performance. A study published in IEEE Access [18] proposed an Elman Recurrent Neural Network optimized using the Dolphin Echolocation Algorithm, achieving 90.45% accuracy with improved precision and recall. Topic modeling and deep learning combinations have also shown promising results. A study published in Multimedia Tools and Applications [19] combined FAEO topic modelling with Extended CNN, achieving 92.91% accuracy on the CB-MNDLY dataset and 91.89% accuracy on RW-CB-Twitter dataset.

Explainable AI techniques have also been applied for multilingual cyberbullying detection. A study using MURIL with Explainable AI techniques achieved high accuracy across multiple Hindi-English datasets, including 94.63% accuracy on the Mendeley Indo-Hate Speech dataset [17]. Optimization techniques have also been explored to improve model performance. A study published in IEEE Access [18] proposed an Elman Recurrent Neural Network optimized using the Dolphin Echolocation Algorithm, achieving 90.45% accuracy with improved precision and recall. Topic modeling and deep learning combinations have also shown promising results. A study published in Multimedia Tools and Applications [19] combined FAEO topic modelling with Extended CNN, achieving 92.91% accuracy on the CB-MNDLY dataset and 91.89% accuracy on RW-CB-Twitter dataset.

III. RESEARCH METHODOLOGY

The proposed system consists of two major phases: (1) Natural Language Processing and (2) Machine Learning. Real-time data is collected from Twitter, WhatsApp chats, and YouTube comments in English and Hinglish. The pipeline covers data extraction, preprocessing, feature selection, and classification using multiple ML algorithms.

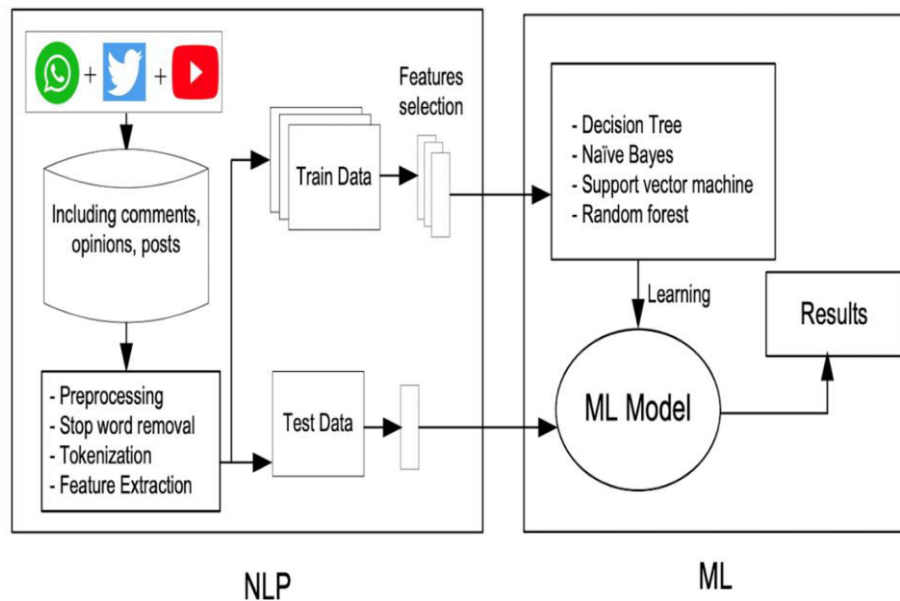


Fig.1. Functional Block Diagram

3.1 Data Extraction

The English dataset was scraped from Twitter and supplemented with Kaggle data, consisting of approximately 15,307 rows. The Hinglish dataset was extracted from Twitter, WhatsApp chats, and YouTube comments, containing approximately 3,000 rows, then merged into a larger corpus. The dataset contains two columns: 'Tweets' and 'Label', where -1 indicates offensive (toxic) and 0 indicates non-offensive (non-toxic) content.

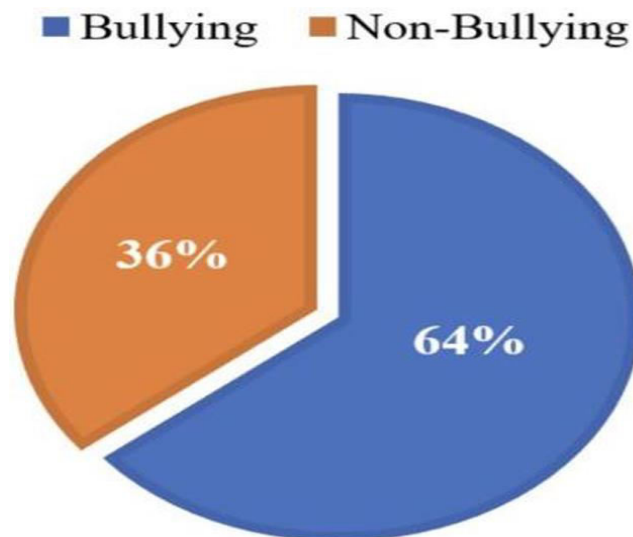


Fig.2. Data Classification (Bullying: 64%, Non-Bullying: 36%)

3.2 Data Cleaning

The cleaning phase removes punctuation marks, special characters, retweet symbols, hashtags, numeric values, hexadecimal values, and URLs, as they do not influence the meaning of the sentence. Words shorter than three letters are eliminated. All sentences are converted to lowercase to avoid duplication. Manually curated stop-word lists for English and Hinglish are applied.

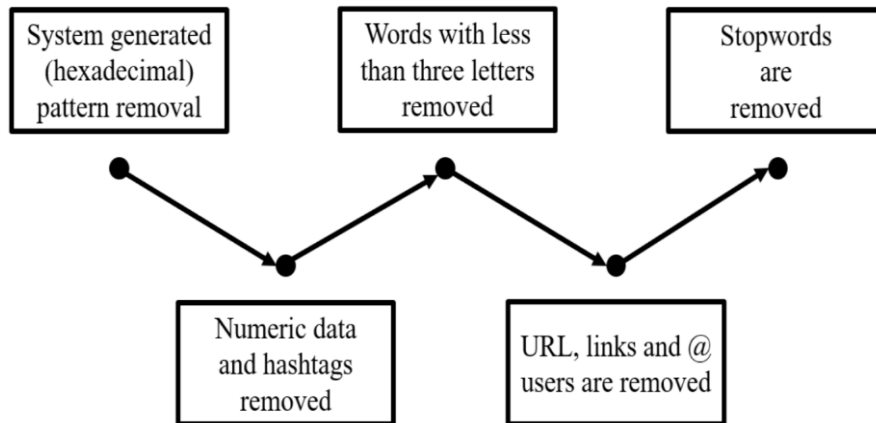


Fig.3. Process of Data Cleaning

3.3 Preprocessing Techniques

After cleaning, NLP techniques are applied to transform text into a machine-understandable format:

- **Tokenization:** The process of splitting a text sequence into smaller chunks—sentences, words, terms, or symbols—called tokens.
- **Lemmatization:** Applied after tokenization to reduce the inflectional forms of each word to its root word.
- **Vectorization:** Assigns weight (probability) to each word in a document, enabling word predictions and semantic analysis.

3.4 Feature Selection

Two feature extraction techniques are employed and compared:

- **Count Vectorization (CV):** Converts a collection of words into a vector of term counts, producing a word matrix in which each cell shows the frequency of that word in a particular document.
- **TF-IDF:** A scoring measure that evaluates how relevant a word is within a document by multiplying term frequency and inverse document frequency. Values closer to 0 indicate more common words; higher values indicate rarer, more significant terms.

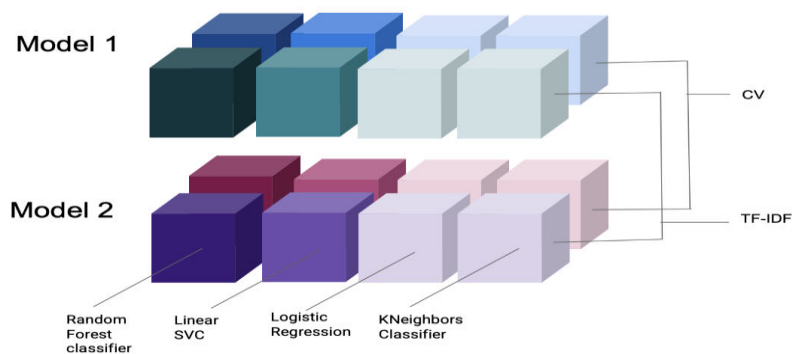


Fig.5. Comparison between CV and TF-IDF Feature Models

3.5 Machine Learning Algorithms

Eight machine learning classifiers were studied and compared based on popularity, ease of use, training time, and prediction time:

- **Support Vector Machine (SVM) :-**
A classification algorithm that finds the best-fitting hyperplane to categorize data into classes. SVM performs well with high-dimensional data and is less prone to overfitting.
- **K-Nearest Neighbors (KNN) :-**
Classifies new data by comparing it with all available data using a similarity measure, placing each object into the class most frequent among its K nearest neighbors.
- **Logistic Regression :-**

A supervised algorithm used to predict categorical dependent variables, outputting a probability that a given input belongs to a particular class.

- **Random Forest Classifier :-**

Consists of a large number of decision trees; the class with the most votes becomes the model's prediction. Fast, scalable, and robust to noise.

- **Bagging Classifier :-**

An ensemble method where multiple models are trained individually and combined via an averaging process to reduce overfitting.

- **Stochastic Gradient Descent (SGD) Classifier :-**

An optimization algorithm for discriminative learning of linear classifiers under convex loss functions such as SVM and Logistic Regression.

- **AdaBoost Classifier :-**

An ensemble boosting classifier that iteratively focuses on correcting the errors of weaker models using one-level decision trees as base learners.

- **Multinomial NB Classifier :-**

A probabilistic algorithm based on Bayes' theorem, commonly used for NLP. Calculates the probability that a given sample belongs to a class, assuming feature independence.

3.6 Evaluation Metrics

Model performance is evaluated using:

- **Precision:** $TP / (TP + FP)$ — the proportion of predicted positives that are actually positive.
- **Recall:** $TP / (TP + FN)$ — the proportion of actual positives that are predicted positive.
- **F-Measure:** $2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$ — the harmonic mean.
- **Accuracy:** $(TP + TN) / (TP + FP + TN + FN)$ — the proportion of correctly classified instances.

IV. RESULTS AND DISCUSSION

A comparative analysis was conducted between Count Vectorizer (CV) and TF-IDF as feature extraction models, using four algorithms on the English dataset.

Table 1: Comparative Accuracy – English Dataset (CV vs. TF-IDF)

Algorithm	Logistic Regression	Linear SVC	Random Forest
CV	0.936	0.958	0.965
TF-IDF	0.940	0.967	0.971

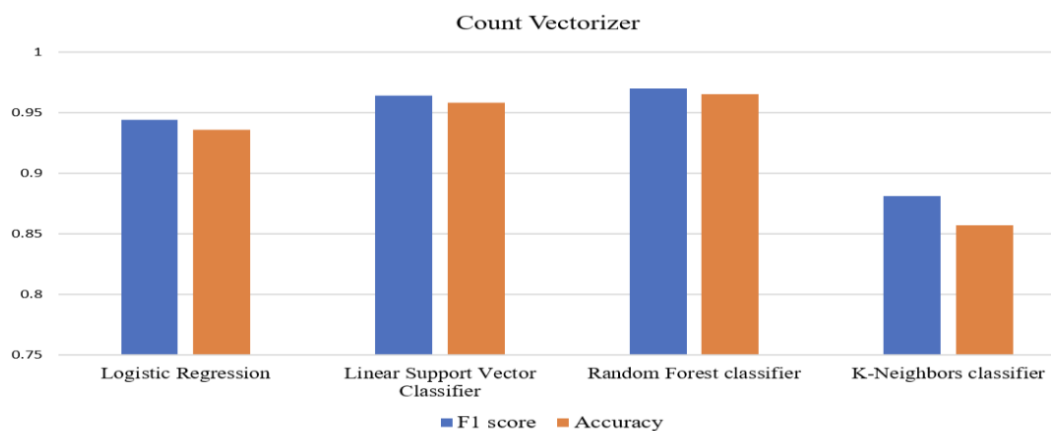


Fig.6. Comparison of Algorithms with Count Vectorizer

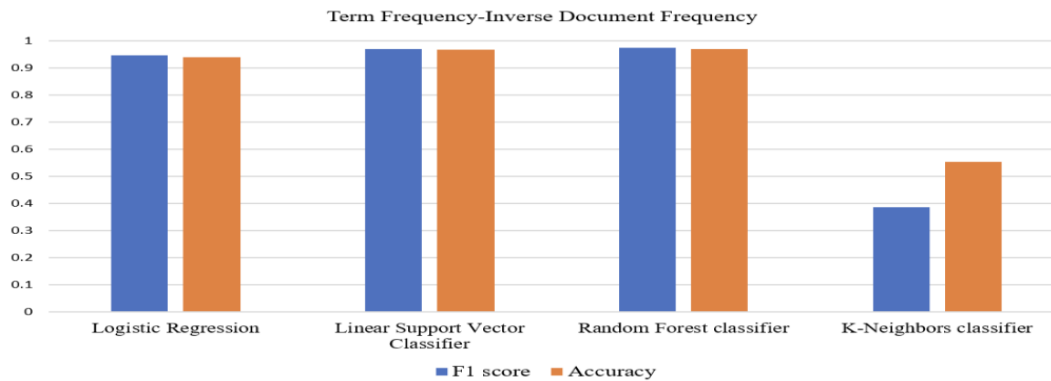


Fig.7. Comparison of Algorithms with Term Frequency-Inverse Document Frequency

Random Forest achieves the best accuracy of 96.5% (CV) and 97.1% (TF-IDF). Linear SVC records the best recall score and fastest training and prediction time. TF-IDF yields slightly better accuracy than CV by prioritizing token importance rather than mere frequency.

For the Hinglish dataset, additional algorithms—Multinomial NB, Decision Tree, AdaBoost, and Bagging—were applied using TF-IDF as the feature extraction model:

Table 2: Algorithm Performance on Hinglish Dataset (CV vs. TF-IDF)

Algorithm	CV Accuracy	TF-IDF Accuracy	CV F1	TF-IDF F1
Decision Tree	0.955	0.962	0.965	0.968
Linear SVC	0.940	0.958	0.954	0.966
Bagging	0.955	0.956	0.961	0.965
Logistic Reg.	0.935	0.944	0.949	0.949
SGD	0.933	0.943	0.942	0.947
Multinomial NB	0.890	0.907	0.903	0.918
AdaBoost	0.827	0.830	0.832	0.851

Decision Tree Classifier provides the highest accuracy (TF-IDF: 0.962) but requires the most time. Linear SVC and Logistic Regression offer a strong balance between accuracy and speed. Linear SVC and SGD Classifier emerge as the best overall choices—competitive accuracy with the least training and prediction time.

VI. CONCLUSION

This paper successfully demonstrated a complete pipeline for cyberbullying detection in English and Hinglish languages. The system performs data extraction, cleaning, and visualization using Python libraries, and applies NLP techniques including tokenization, lemmatization, and vectorization. A comparative analysis established TF-IDF as the superior feature extraction model, particularly for Hinglish data.

Among the eight ML algorithms evaluated, Linear SVC and SGD Classifier provide the best combination of accuracy and speed for classifying and predicting bullying messages in Hinglish languages. The system is deployed as an interactive chat application capable of detecting and blocking bullying content in real time.

Future work will focus on improving model accuracy using deep learning techniques, extending language support to Gujarati, Marathi, Tamil, Telugu, and Kannada, and developing a multimodal model capable of detecting cyberbullying in images and videos.

VIII. FUTURE SCOPE

This research work can be improved in the future by doing the following work. Firstly, the accuracy of the models can further be increased to get better results by using deep learning. Next, detection can be done in more languages such as Gujarati, Marathi, Tamil, Telugu, Kannada etc. In short, the project can be made more diverse to make it applicable for multiple applications. For this diverse datasets for the required languages are required and a list of stopwords is required. Then a similar procedure can be implemented. We have developed a model which classifies cyberbullying through text only. In the future, we can develop a model which can classify cyberbullying through images and videos.

REFERENCES

- [1] G. Jaradat et al., "Deep Learning-Based Multiclass Cyberbullying Detection," Journal of Computational and Cognitive Engineering, 2025.
- [2] A. Perera and P. Fernando, "Cyberbullying Detection System on Social Media Using Supervised Machine Learning," Procedia Computer Science, vol. 239, pp. 506–516, 2024.
- [3] P. Jha et al., "MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention," arXiv:2406.05344, 2024.
- [4] M. Menaka et al., "Cyberbullying Detection on Social Media using Machine Learning," IJSRCSEIT, vol. 11, no. 3, pp. 661–666, 2025.
- [5] Almufareh et al., IEEE Access, 2025. DOI: 10.1109/ACCESS.2025.3558843.
- [6] Wang et al., "Cyberbullying Role Identification using Deep Learning and Clustering," arXiv:2408.03502, 2024.
- [7] Kumar et al., "Hybrid Deep Learning Model for Cyberbullying Detection," Electronics, vol. 13, 2024.
- [8] El Mezayen et al., BMC Nursing, 2025. DOI: 10.1186/s12912-025-03855-7.
- [9] Kathuria et al., "Cyberbullying Detection using Machine Learning," 2023.
- [10] Riska et al., "Emotion Regulation and Cyberbullying Behavior," 2025.
- [11] Jesseline et al., "Psychological Impact of Cyberbullying on Young Adults," 2024.
- [12] Jesseline et al., "Contrastive Self-Supervised Learning for Cyberbullying Detection," 2024.
- [13] IJARETY, "Big Data Analytics for Cyberbullying Detection," 2025.
- [14] ANJS, "Comparative Analysis of ML and DL Models for Cyberbullying Detection," 2023.
- [15] "Cyberbullying Detection in Chinese Text using XLNet and Deep Bi-LSTM," Information, 2024.
- [16] "Modified DeBERTa for Cyberbullying Detection," arXiv, 2025.
- [17] "Explainable AI for Hindi-English Cyberbullying Detection using MURIL," arXiv, 2025.
- [18] "DEA-RNN Optimized Cyberbullying Detection," IEEE Access, 2022.
- [19] "Topic Modeling and Extended CNN for Cyberbullying Detection," Multimedia Tools and Applications, 2023.
- [20] "Real-Time Cyberbullying Detection using Naïve Bayes and SVM," IJACSA, 2022.



International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 9.971