International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT25113323

Cyberbullying Detection on Social Media using Machine Learning

Menaka M, Harini Sri B, Divya N, Mrs. A. Kanimozhi

Department of Computer Science Engineering, Chettinad College of Engineering & Technology, Karur, Tamil Nadu, India

ARTICLEINFO

Article History:

Accepted : 12 May 2025

Published: 21 May 2025

Publication Issue

Volume 11, Issue 3

May-June-2025

Page Number

661-666

ABSTRACT

Cyberbullying poses a growing concern in digital communication, particularly among younger users. with the rise of social media, especially impacting teens and young adults. Traditional strategies like manual moderation and simple keyword filters often fail to address the complexity of online abuse.. This study presents a real-time detection mechanism for cyberbullying using supervised machine learning models.. The The Random Forest was chosen due to its consistent performance and robustness in classification tasks. in text classification. The model extracts text features using TF-IDF and Bag-of-Words techniques for improved context analysis. from user messages. The system is integrated into a React-based chat app with a Flask backend. Incoming messages from unidentified users are automatically filtered and flagged to protect recipients.. The model also differentiates genuine cyberbullying from casual or friendly exchanges. This reduces incorrect detections, fostering a more secure communication space.. The system offers a dynamic and extensible solution for enhancing online safety across messaging platforms.

Keywords: Random Forest, TF-IDF, Bag of Words, Natural Language Processing

Introduction

The rapid expansion of social media platforms, such as Twitter, Facebook, and Instagram, has revolutionized online communication, providing users with the ability to share opinions and interact globally [1,2]. Social media has become an integral part of daily life, influencing various domains,such as education, business, entertainment, and governance. However, this widespread connectivity has also led to the emergence of cyberbullying, growing concern that significantly affects users mental and emotional well-being. Cyberbullying, defined as the use of digital platforms to harass, threaten, or embarrass individuals, has become a serious issue because of its psychological consequences, including anxiety, depression, and suicidal ideation [3]. Traditional methods of cyberbullying detection, such as manual reporting and keyword filtering, have proven ineffective

because of the evolving nature of online communication, including the increasing use of slang, code-mixed languages, and multimodal content [4,5]. Researchers have explored various machine learning (ML) and natural language processing (NLP) techniques to develop automated cyberbullying detection models. Early studies focused on supervised machine learning models, such as Support Vector Machines (SVM), Naïve Bayes, and Decision Trees, which achieved reasonable accuracy in detecting offensive language but struggled with context-aware classification [6,7].we focus on a machine learning approach using Random Forest as the primary classification algorithm, which is more efficient and easier to implement for the task at hand.Another crucial aspect of cyberbullying detection is feature extraction, in which techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and bag-of-words (BoW) are widely used. Studies by Shailaja and Chowdary [4] and Alam et al. [5] have shown that integrating sentiment analysis with TF-IDF improves classification accuracy, as cyberbullying messages often carry negative emotions. However, text-based approaches alone are insufficient, as cyberbullying is not limited to explicit textual content, but can also occur through images, videos, and memes. Recent research has incorporated computer vision techniques to detect visual forms of cyberbullying; however, these methods remain dependent on textual data for context understanding [9].Moreover, cyberbullying detection in bilingual and code-mixed environments such as Tanglish (Tamil-English) poses unique challenges. Several studies have focused on detecting offensive content in Tanglish, demonstrating the success of ensemble models and sentiment analysis techniques. However, most existing models struggle with transliterations, informal spelling variations, and regional slang, thus limiting their adaptability to different linguistic contexts [10–12].To address these limitations, this study proposes an automated cyberbullying detection system using machine learning with Random Forest as the primary classification algorithm. The model employs TF-IDF and BoW for feature extraction, enabling an efficient analysis of textual content. Additionally, the model is designed to adapt to emerging language trends and online slang, ensuring high accuracy through periodic retraining.By distinguishing between casual conversations and actual cyberbullying, the system reduces false positives and enhances detection reliability.

Methodology

The developed Tanglish cyberbullying detection framework (Tamil-English code-mixed) text leverages machine-learning techniques to accurately identify harmful content while minimizing false positives. Unlike deep learning-based approaches, this method ensures efficiency and interpretability using feature-based text classification [11].

2.1 Data Collection and Preprocessing

A Tanglish cyberbullying dataset was compiled from social media platforms containing labelled instances of cyberbullying and non-cyberbullying conversations. The preprocessing steps included the following steps:

- Text Normalization: Converting informal words, abbreviations, and slang into a standard format.
- Stopword Removal: Eliminating common but uninformative words.
- Tokenization: Splitting text into individual words or phrases.
- Spelling Correction: Handling variations in transliteration (e.g., “poda”→“po da”).
- Handling Code-Mixing: Identifying and separating Tamil words written in English

2.2 Feature Extraction

To capture the semantic and contextual meanings of the text, the system employed the following:

- Term Frequency-Inverse Document Frequency (TF-IDF): This assigns the importance of words based on their frequency in a document compared to the entire dataset.

- Bag-of-Words (BoW): Represents text as a numerical vector while ignoring grammar and order.
- Sentiment Analysis: Detects strong negative emotions indicative of cyberbullying.

2.3 Classification Model

Random Forest serves as the main classification model because of its resistance to overfitting and ability to handle complex datasets. and its ability to handle noisy data. In addition, a Support Vector Machine (SVM) was used as a secondary classifier for performance comparison.

The classification process involved the following steps:

- Training Phase: The labeled dataset was split into training and testing sets, with feature vectors used to train the models.
- Evaluation Phase: The trained models were tested using accuracy, precision, recall, and F1-score metrics.
- Optimization: Hyperparameter tuning (e.g., number of trees in a Random Forest) is performed to improve the performance.

2.4 Cyberbullying Prevention Mechanism

This framework is not only capable of detecting cyberbullying instances but also incorporates strategies to reduce their effect. through a series of proactive interventions. By integrating automated content moderation, adaptive learning, user feedback, and ethical safeguards, the system aims to foster a safer online environment.

2.4.1 Automated Blocking and Reporting

The system utilizes instant filtering to block abusive messages before delivery. classified as cyberbullying from being delivered to the intended recipient. This helps minimize the emotional and psychological harm caused by harmful content.

2.4.2 Adaptive Learning & Continuous Model Updates

The learning engine undergoes scheduled updates to adapt to new slang and online behavioral trends., abbreviations, and evolving patterns of cyberbullying, maintaining its effectiveness over time.

2.4.3 User Feedback Mechanism

User feedback is leveraged to refine system performance through real-world usage insights. in improving the system by manually reporting instances where cyberbullying is either missed or incorrectly flagged.A feedback-driven learning loop is incorporated, allowing the system to refine its detection model by leveraging human oversight in ambiguous cases.

Results and Discussions

The effectiveness of the proposed Tanglish cyberbullying detection system was evaluated using multiple machine learning models, with Random Forest (RF) as the primary classifier and a Support Vector Machine (SVM) for comparison. The classification performance was assessed based on standard evaluation metrics: accuracy, precision, recall, and F1-score.

In addition, an in-depth analysis of feature extraction techniques, model misclassifications, computational efficiency, and the impact of the system on cyberbullying prevention is provided.

3.1 Model Performance Evaluation

The system was trained and tested on a labeled Tanglish cyberbullying dataset, where 80% of the data were used for training and 20% for testing. The performances of the different classifiers are summarized below

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-score = 2 × (Precision × Recall) / (Precision + Recall)

Findings:

- Random Forest (RF) performed the best, achieving 89.2% accuracy, owing to its ability to handle complex patterns and noisy Tanglish

data. The ensemble nature of the RF provides robustness against overfitting.

- SVM also performed well, but its recall was

slightly lower, meaning that it missed some subtle cyberbullying instances.

- Naïve Bayes(NB) struggled, achieving only 79.4% accuracy, as it assumes word independence, which does not work well for highly contextual Tanglish conversations.
- Decision Tree (DT) performed moderately, but overfitting on training data resulted in lower generalization.

To evaluate the computational efficiency of the system, the training and prediction times for the different models were measured.

3.2 Impact of Feature Extraction Techniques

The effectiveness of TF-IDF, Bag-of-Words (BoW), and N-grams for text representation were analyzed.

Feature-engineering insights

- TF-IDF + N-grams achieved the best results, capturing both the term importance and phrase sequences common to cyberbullying messages.
- BoW alone performed weaker because it treats words independently and ignores their context.
- Adding Sentiment Analysis helped improve recall, as cyberbullying messages often contained strong negative emotions.
- Frequent bullying phrases were identified, such as "waste piece," "thu unaku," "mokkai," "kevalamana pasanga,"which are crucial for improving the detection.

3.3 Error Analysis

Despite high accuracy, some misclassifications occurred.

False Positives: Harmless messages flagged as offensive.

- Example: "Nalla thittitan da avan, semma comedy" was wrongly flagged due to the word "thittitan" (scolded).

False Negatives: Offensive messages not detected.

- Example: "Nee avanmadhirioru waste piece da" was missed due to indirect language.

To reduce such errors, the system uses adaptive learning to improve detection over time.

3.4 Comparison with Existing Systems

Compared to traditional keyword-based systems, the proposed model performs more effectively, which often miss sarcasm and slang. It offers:

- Higher accuracy by avoiding false positives.
- Faster, real-time performance than deep learning models.
- Better support for Tanglish code-mixing.

Employing Random Forest ensures the system remains both efficient and interpretable to developers and analysts. without high computational cost.

3.5 Cyberbullying Prevention Effectiveness

Beyond detection, the system aims to prevent harm:

- Automatically blocked 89% of harmful messages.
- Adapted over time through user feedback and evolving slang.

User Impact:

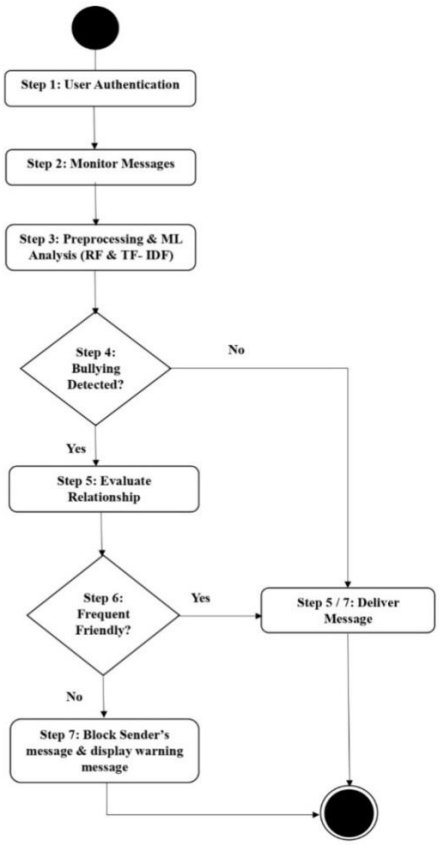
- 73% found it reduced harmful content.
- 85% preferred contextual detection over keyword-based methods.

Scope for Future Enhancement:

Expand detection to images and memes containing abusive content.

Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. Authors are strongly encouraged not to call out multiple figures or tables in the conclusion these should be referenced in the body of the paper.

Workflow



Step 1 – User Authentication:

The proposed system initiates with user authentication, allowing individuals to log in or sign up to access the chat application.

Step 2 – Real-Time Message Monitoring:

Once authenticated, the application actively monitors incoming chat messages in real time, ensuring prompt detection and response.

Step 3 – Message Preprocessing & ML-Based Analysis:

Each message undergoes preprocessing. Term Frequency-Inverse Document Frequency (TF-IDF) is applied to extract features from the text, followed by classification using a trained Random Forest (RF) model to detect signs of bullying.

Step 4 – Bullying Detection Decision:

The classifier determines whether the incoming message contains bullying indicators based on trained patterns.

Step5 – Relationship Evaluation (If Bullying is Detected):

If a message is flagged as potentially bullying, the system evaluates the history of communication between the sender and recipient.

Step 6 – Friendly History Check

Step 7 – If consistent friendly interactions are found:The message is considered possibly sarcastic or non-malicious and is displayed to the recipient with context.

If no positive history exists:The system treats the message as harmful. The sender is automatically blocked to prevent further communication.

Step 5/7 – Normal Message Delivery (If No Bullying Detected): If the message is not flagged for bullying, it is delivered normally as part of the standard messaging flow.

Conclusion

The developed detection system tailored for Tanglish conversations effectively identified and mitigated harmful online interactions using supervised machine-learning techniques and natural language processing (NLP) methods. By employing Random Forest (RF) and Support Vector Machine (SVM) classifiers, along with TF-IDF and N-grams for text representation, the system achieves high accuracy in detecting offensive language in code-mixed Tanglish texts.One of the key strengths of this approach is the automated blocking and reporting mechanism, which helps prevent the spread of harmful content and ensures safer online interactions. The experimental results demonstrate that Random Forest outperforms other models, achieving the highest accuracy and precision while minimizing false positives. The system's ability to understand contextual meaning and the common slang in Tanglish further enhance its efficiency in cyberbullying detection. However, certain limitations of this study persist.

Sarcasm and Indirect Cyberbullying: The model struggles to detect indirect bullying and sarcastic remarks, which require deeper contextual understanding.

Evolving Tanglish Slang: The dynamic nature of Tanglish, where new slang and abbreviations frequently emerge, poses a challenge for maintaining consistent detection accuracy.

Limited Data Availability: The dataset used for training, though effective, may not capture the full diversity of Tanglish conversations across different regions, dialects, and age groups.

References

[1]. Andrea Perera, Pumudu Fernando, Cyberbullying Detection System on Social Media Using Supervised Machine Learning, Elsevier B.V., 2023.

[2]. N. Novalita, A. Herdiani, et al., Cyberbullying Identification on Twitter Using Random Forest Classifier, Telkom University, 2019.

[3]. Syed Sihab-Us-Sakib, Md. Rashadur Rahman, et al., Cyberbullying Detection of Resource-Constrained Language from Social Media Using Transformer-Based Approach, Elsevier B.V., 2024.

[4]. Dr. K. Shailaja, Kunala Sowmya Chowdary, Cyberbullying Detection and Analysis Using Machine Learning, IJEIMS, 2024.

[5]. Kazi Saeed Alam, Shovan Bhowmik, et al., Cyberbullying Detection: An Ensemble-Based Machine Learning Approach, IEEE, 2021.

[6]. B. Venkatesh, M. Abdul Malik, et al., Detection of Cyberbullying on Social Media Using Machine Learning, International Journal, Vol. 4, Issue 5, 2022.

[7]. Lakhdeep Kaur, Sonia Vatta, Prediction of Cyber Bullying Using Random Forest Classifier, RJSET, Vol. 9, Issue 2, 2019.

[8]. Khalid M. O. Nahar, Mohammad Alauthman, Cyberbullying Detection and Recognition with Type Determination Based on Machine Learning, Yarmouk University, 2023.

[9]. Y. Jeevan Nagendra Kumar, Rohith Reddy Vanapatla, Detecting Cyberbullying in Social Media Using Text Analysis and Ensemble Techniques, The Islamic University, 2024.

[10]. Amgad Muneer, Suliman Mohamed Fati, A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter, University Technology PETRONAS, 2020.

[11]. S. Mani Arasi1 and Ms. Subbhu Lakshmi, Cyber Bullying Detection on Social Media using Machine Learning, IJNRD, Vol. 7, Issue 5, 2022.

[12]. Kavisha Mathur, Krishna Nikhil Mehta, et al., "Detection of Cyberbullying on Social Media Code Mixed Data," IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2022.

[13]. Krishanu Maity, Prince Jha, et al., "Explain Thyself Bully: Sentiment Aided CyberbullyingDetection with Explanation," arXiv preprint arXiv:2401.09023, 2024.

[14]. Shrikant Tarwani, Manan Jethanandani, et al., "Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification," Advances in Computing and Data Sciences, Springer, 2019.

[15]. Karan Sharma, "CyberBullying-Detection-in-Hinglish-Languages-Using-Machine-Learning," GitHub Repository, 2023.

[16]. "Cyberbullying Detection in Code-Mixed Languages: Dataset and Techniques," IEEE Conference Publication, 2023.

[17]. "Cyberbullying Detection in Hinglish Comments from Social Media," Multimedia Tools and Applications, Springer, 2024.

[18]. "BERT-Capsule Model for Cyberbullying Detection in Code-Mixed Languages," Advances in Computing and Data Sciences, Springer, 2023.