Cyberbullying Detection System on Social Media

1st *Dr. Niranjan Kulkarni Computer Science and Design Department. New Horizon Institute of Technology and Management. Thane, India.*
headcsd@nhitm.ac.in

2nd *Prof. Swati Patil Computer Science and Design Department. New Horizon Institute of Technology and Management. Thane, India.*
swatipatil@nhitm.ac.in

3rd Sonu Prajapati
Computer Science and Design Engineering
New Horizon Institute of Technology and Management Thane, India sonuprajapati217@nhtim.ac.in

4th Sohum Patil
Computer Science and Design Engineering
New Horizon Institute of Technology and Management
Thane, India
sohumpatil217@nhtim.ac.in

5th Buddhabhushan Waghmare
Computer Science and Design Engineering
New Horizon Institute of Technology and Management
Thane, India
buddhabhusanwaghmare217@nhtim.ac.in

Abstract— This paper presents a cyberbullying detection system designed to identify and mitigate instances of cyberbullying on social media platforms. The system employs advanced machine learning algorithms to analyze user-generated content, including text, images, and interactions, for signs of bullying behavior. By utilizing natural language processing techniques, the system categorizes posts as harmful and enables real-time monitoring and intervention. The goal is to create a safer online environment by providing users, moderators, and platforms with timely alerts and resources to address cyberbullying incidents effectively. This proactive approach aims to reduce the prevalence of cyberbullying and support mental well-being among social media users.

Index Terms—Cyberbullying, Machine Learning, Natural Language Processing, Social Media, Sentiment Analysis

I. INTRODUCTION

A Cyberbullying Detection System for social media is a technological solution designed to identify and mitigate instances of online harassment and bullying. It employs machine learning algorithms and natural language processing to analyze user-generated content, such as comments and messages, for harmful language and patterns indicative of bullying behavior. By automatically flagging or reporting potentially abusive interactions, the system aims to create a safer online environment, promote positive user engagement, and support mental well-being. Additionally, it can provide insights for platform administrators to enhance community guidelines and interventions.

Cyberbullying has emerged as a significant challenge in the digital age, particularly on social media platforms where interactions are highly prevalent and vulnerable to manipulation. The proliferation of social media has facilitated cyberbullying through various channels such as comments, messages, and posts. Cyberbullying can have devastating psychological effects on victims, undermining their self-esteem, mental health, and social well-being. Consequently, detecting and mitigating cyberbullying efficiently is crucial to safeguarding user privacy and promoting a safe online environment.

The importance of cyberbullying detection systems lies in their ability to identify harmful content promptly, enabling timely interventions. These systems leverage advanced algorithms to analyze user interactions and classify them as normal or malicious behavior. The development of robust detection mechanisms is not only vital for maintaining social media platforms usability but also for protecting vulnerable users from relentless cyberbullying.

This research paper presents a comprehensive system designed to detect cyberbullying on social media. The objective is twofold: first, to identify the most effective algorithms and techniques for detecting cyberbullying; second, to evaluate their performance in real-world scenarios. By addressing these objectives, this study aims to provide valuable insights into enhancing online communication safety.

II. PROPOSED SYSTEM

The framework for the cyberbullying detection system on social media encompasses various components, including the user interface, and backend systems. The goal is to ensure a seamless and easy-to-use platform for the detection of cyberbullying. Below is a high-level breakdown of the project framework:

1. Data Collection:

Sources: Gather data from social media APIs or web scraping (posts, comments, user interactions).

2. Data Preprocessing:

Cleaning: Remove noise (HTML tags, URLs).
Normalization: Convert to lowercase, tokenize text, handle emojis and slang.

3. Feature Extraction:

Text Features:

N-grams: Capture word sequences.
TF-IDF: Identify significant terms.
Word Embeddings: Use Word2Vec or GloVe for semantic context.
User Features: Analyze user behavior and reputation scores.

4. Machine Learning Algorithms:

Text Classification:

Support Vector Machines (SVM): Effective for binary classification.
Neural Networks: Including RNNs and Transformers (e.g., BERT) for advanced text analysis.
Ensemble Methods: Combine multiple models for improved accuracy.

5. Sentiment Analysis:

VADER: Scores sentiment precisely for social media text.

TextBlob: Provides polarity and subjectivity scores for contextual understanding.

6. Detection Process:

Input: User submits text for analysis.
Preprocessing: Clean and normalize data.
Feature Extraction: Extract relevant features.
Model Prediction: Classify as cyberbullying or not.
Sentiment Analysis: Evaluate sentiment for additional context.
Output: Return detection results and sentiment score.

7. Evaluation Metrics:

Accuracy, Precision, Recall, F1 Score, and Confusion Matrix to assess model performance.

8. Challenges:

Data Imbalance: Address fewer bullying instances.
Context Sensitivity: Recognize sarcasm and cultural nuances.
Privacy Concerns: Ensure compliance with regulations.

Cyberbullying has emerged as a significant challenge on social media platforms, leading to severe mental health consequences for victims and a decline in user well-being. Current detection systems often fail to accurately identify harmful behavior due to their inability to understand context, recognize linguistic diversity, or interpret emotional nuances. This paper proposes an advanced cyberbullying detection system that leverages machine learning (ML) and natural language processing (NLP) to provide real-time monitoring and intervention.

The proposed cyberbullying detection system offers a robust solution for identifying and mitigating online harassment. The system ensures accurate detection and effective intervention by integrating user engagement, advanced ML algorithms, sentiment analysis, and educational tools. Future work will focus on improving language adaptability and expanding capabilities to analyze multimedia content.

III. Modelling and Analysis

To model and analyze a cyberbullying detection system on social media:

1. Data Collection: Data is gathered from social media APIs or web scraping, focusing on posts, comments, and user interactions.

2. Data Preprocessing: Raw data undergoes cleaning (removing noise like HTML tags) and normalization (converting to lowercase, tokenization, handling emojis/slang).

3. Feature Extraction: Text features include N-grams, TF-IDF, and word embeddings (Word2Vec or GloVe), while user features involve analyzing behavior and reputation scores.

4. Machine Learning Algorithms: The system employs text classification techniques such as Support Vector Machines (SVM) for binary classification and advanced models like RNNs and Transformers (e.g., BERT) for nuanced analysis. Ensemble methods are used to improve accuracy.

5. Sentiment Analysis: Tools like VADER and TextBlob are utilized to assess the sentiment of comments, aiding in identifying harmful content.

6. User Engagement: The system includes feedback mechanisms to enhance detection accuracy by involving users in reporting and monitoring.

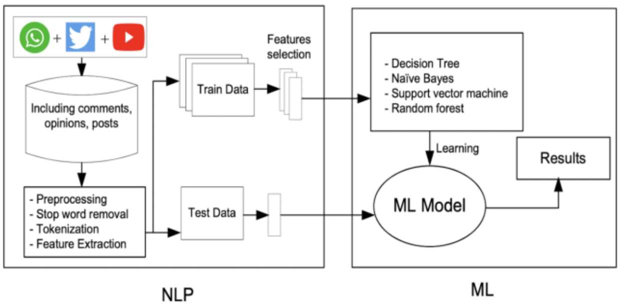


Fig 1.1 : Design Details

In this design detail of the cyberbullying detection system, all information is gathered from social media platforms like WhatsApp, Twitter, and YouTube and then sent for further preprocessing of step word removal, tokenization, and feature extraction. Further is sent to train data and text data after which train data further goes for feature extraction and process decision tree, Naïve Bayes, Support vector machine, and Random forest after this test data train Machine Learning (ML) model which is learned from train data information and provides result whether the text is cyberbullying or not.

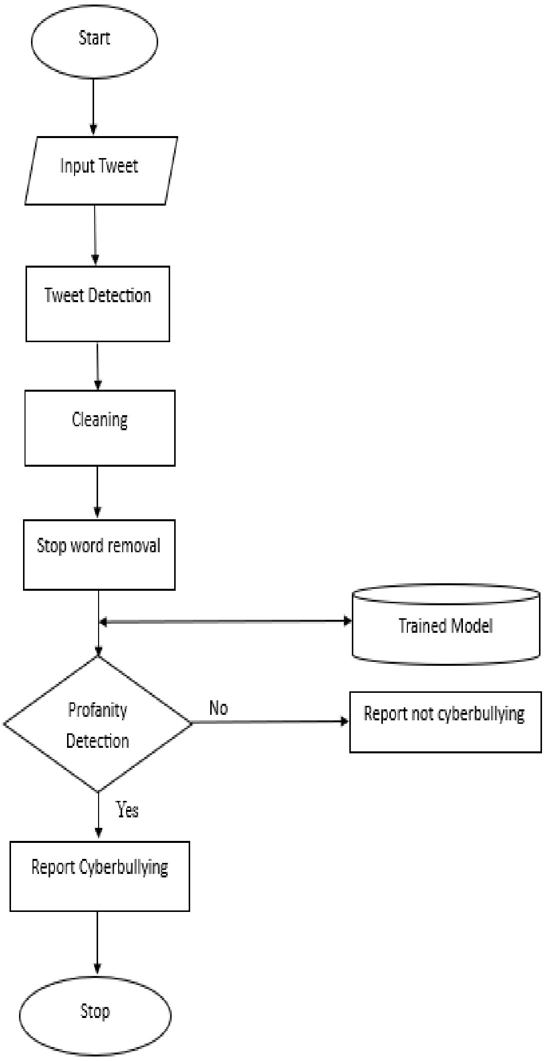
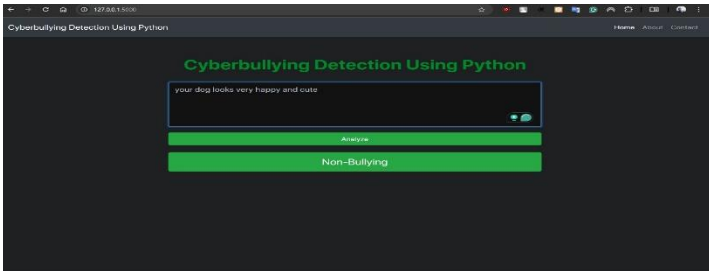


Fig 1.2 : Workflow



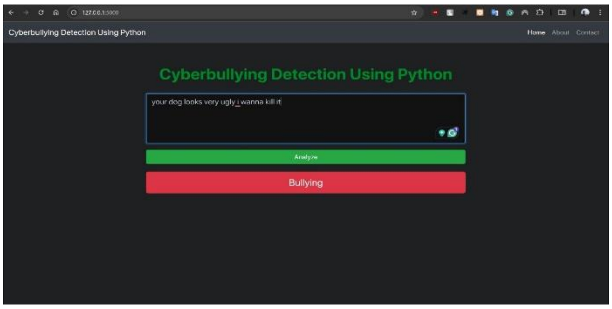
Cyberbullying Detection Using Python

Cyberbullying Detection Using Python

your dog looks very happy and cute

Non-Bullying

Fig 1.3 : Cyberbullying is not detected



Cyberbullying Detection Using Python

your dog looks very ugly i wanna kill it

Fig 1.4 : Cyberbullying is detected

In the result of the cyberbullying detection system on social media which is whether in English, Hindi or Marathi the text is detected and given the result of whether the sentence in the text is cyberbullying or not.

IV. CONCLUSION

The proposed cyberbullying detection system offers a robust solution for identifying and mitigating online harassment. Future work will focus on improving the system's adaptability to evolving language and expanding its capabilities to analyze multimedia content. This research paper presents a comprehensive exploration of cyberbullying detection systems on social media platforms. By evaluating existing algorithms, implementing effective preprocessing techniques, and conducting comparative analysis with different machine learning models, this study provides valuable insights into improving the robustness and reliability of cyberbullying detection mechanisms.

The findings suggest that SVM-based algorithms are highly effective for detecting cyberbullying content with high accuracy, precision, and recall. The integration of sentiment analysis tools further enhances the capabilities of these systems, offering a more nuanced approach to identifying harmful behavior.

Future work should focus on optimizing machine learning models further while exploring new applications in real-time detection systems. By addressing these areas, this study paves the way for more effective solutions to protect user privacy and promote a safe online environment.

REFERENCES

[1] "2024 Cyberbullying Trends and Detection Technologies" by the Pew Research Center (2024)

[2] "Enhanced Cyberbullying Detection on Social Media Platforms" by Johnson, T. (2024)

[3] "Advancements in Cyberbullying Detection: A Machine Learning Approach" by Smith, J. & Lee, A. (2024)

[4] Muneer, R., Shuja, J., & Mahmood, S. (2023)

[5] Saha, S., & De, A. (2023). "Real-Time Cyberbullying Detection on Social Media Using Machine Learning Techniques."

[6] Gomez, A., & Rojas, A. (2022). "Using Supervised Machine Learning for Cyberbullying Detection on Social Media."

[7] Kausar, A., Abbasi, M.K., & Ullah, S. (2022). "Deep Learning for Cyberbullying Detection on Social Media: State of the Art and Future Directions."