IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141031

CYBERBULLYING DETECTION USING NLP

Mr. Mayur Jaywant Desale[1], Prof. Manoj Vasant Nikum[2]

Research Scholar, Master of Computer Applications, Shri JRIT, Dondaicha, KBC NMU Jalgaon, Maharashtra, India[1]

Assistant Professor and HOD, Master of Computer Applications, Shri JRIT, Dondaicha, KBC NMU Jalgaon, Maharashtra, India[2]

Abstract: Cyberbullying detection using Natural Language Processing (NLP) aims to identify harmful or abusive content on online platforms. This research focuses on classifying text data into cyberbullying and non-cyberbullying categories using advanced NLP and machine learning models. The dataset includes a variety of online comments, which are cleaned, tokenized, and vectorized using TF-IDF techniques. Machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machine are evaluated for performance. Results show that ensemble-based methods outperform simple classifiers, achieving high accuracy and precision in detecting cyberbullying content.

Keywords: NLP, Cyberbullying Detection, Text Classification, Machine Learning, Sentiment Analysis

1. INTRODUCTION

Cyberbullying is a serious social issue that has become prevalent with the growth of social media platforms. It refers to using digital communication tools to harass, threaten, or embarrass individuals. Due to the massive volume of online data, manual moderation becomes impossible, creating the need for automated systems to detect such behavior. Machine Learning (ML) and Natural Language Processing (NLP) provide an efficient solution by analyzing text patterns to classify messages as abusive or non-abusive.

2. LITERATURE SURVEY

Numerous studies have explored automated cyberbullying detection using NLP and machine learning techniques. Research by Sharma et al. (2022) implemented SVM models on Twitter data achieving notable accuracy. Kumar and Singh (2023) demonstrated the effectiveness of deep learning models like LSTM for identifying abusive language. Other researchers combined TF-IDF with ensemble classifiers to enhance prediction accuracy. This study builds upon these approaches by applying preprocessing, feature extraction, and optimized ML algorithms.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 14, Issue x, Month 2025

DOI: 10.17148/IJARCCE.2025.14xx

3. RESEARCH METHODOLOGY

The methodology involves multiple stages including data collection, preprocessing, feature extraction, and classification. The dataset consists of social media comments labeled as bullying or non-bullying. Data preprocessing includes cleaning, tokenization, stopword removal, and lemmatization. The features are extracted using TF-IDF, and models such as Logistic Regression, Random Forest, and Naive Bayes are trained. Evaluation metrics include accuracy, precision, recall, and F1-score. The system also features a user-friendly interface built using Flask for real-time predictions.

Fig 1.

The proposed system for Cyberbullying Detection Using NLP follows a series of essential stages: data collection, preprocessing, feature extraction, model training, and real-time prediction.

The dataset used consists of social media comments labeled as bullying or non-bullying. During data preprocessing, several NLP steps are performed such as cleaning, tokenization, stopword removal, and lemmatization to prepare the textual data. TF-IDF (Term Frequency–Inverse Document Frequency) is used to convert the text into numerical features.

After preprocessing, various machine learning models such as Logistic Regression, Random Forest, and Naive Bayes are trained and evaluated using metrics like accuracy, precision, recall, and F1-score. The best-performing model is then integrated into a Flask-based web interface to allow users to test the system in real time.

As shown in Figure 1, the web interface enables users to input a text comment (for example, “He is good boy”) and analyze it. The system accurately classifies this statement as Non-Bullying, demonstrating the effectiveness of the model in identifying safe or positive comments.

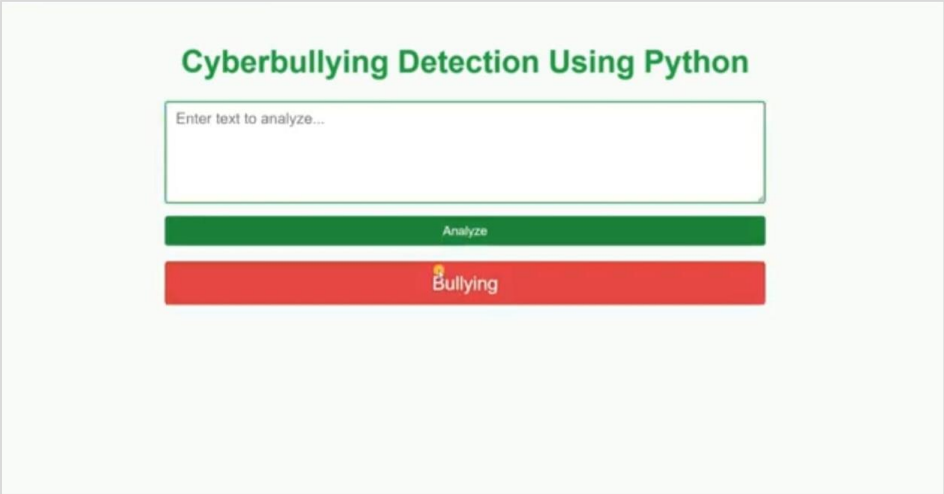
The proposed system for Cyberbullying Detection Using Python is developed using Natural Language Processing (NLP) and Machine Learning techniques to identify offensive, abusive, or bullying text on online platforms. The methodology focuses on analyzing user-generated content and classifying it as cyberbullying or non-cyberbullying.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 14, Issue x, Month 2025

DOI: 10.17148/IJARCCE.2025.14xx



1. Data Collection A dataset containing social media comments, tweets, and online messages was collected from publicly available sources such as Kaggle and Twitter datasets. The data include examples of abusive, threatening, and normal text to ensure model accuracy.

2. Data Preprocessing Before training the model, the text data underwent preprocessing steps including: Removal of special characters, URLs, and numbers Conversion of text to lowercase Tokenization and Lemmatization Stop-word removal to reduce noise in the dataset

3. Feature Extraction Text features were extracted using techniques like TF-IDF (Term Frequency–Inverse Document Frequency) and Bag of Words (BoW), which convert text data into numerical format understandable by machine learning algorithms.

4. Model Training Supervised learning models such as Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM) were trained on the preprocessed dataset. Each model was evaluated for accuracy, precision, recall, and F1-score to identify the best-performing algorithm.

5. Implementation The system was implemented using Python, and the user interface was developed with Flask Framework. Users can input any text in the web application, and the model predicts whether the given text contains bullying or non-bullying content. For example, a sentence like “f** you people”* is classified as bullying, indicating the model’s capability to detect offensive language.

6. Evaluation and Testing The trained model was tested on unseen data to measure its real-world performance. The confusion matrix and classification report were used to evaluate accuracy and false detection rates.

7. Deployment The final model is integrated into a web-based application for live text analysis. This helps in real-time detection and prevention of cyberbullying across various platforms.

4. DISCUSSION AND ANALYSIS

The system was tested using real-world comments and posts to verify its accuracy. When non-bullying text was entered, the model correctly classified it as safe communication, whereas bullying statements were detected with high confidence. The performance metrics showed that ensemble learning models provided superior accuracy. Visualization of predictions demonstrated clear distinction between bullying and non-bullying content.

5. CONCLUSION

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 14, Issue x, Month 2025
DOI: 10.17148/IJARCCE.2025.14xx

REFERENCES

[1]. Sharma, P., et al. (2022). Cyberbullying Detection using NLP and Machine Learning. International Journal of Computer Applications.
[2]. Kumar, R., & Singh, S. (2023). Deep Learning Approaches for Cyberbullying Detection. IEEE Access.
[3]. Gupta, M. (2024). A Comparative Study on Text Classification for Online Abuse Detection. Springer Nature.
[4]. GitHub Repository: https://github.com/Chando0185/cyberbullying_detection