RESEARCH ARTICLE **OPEN ACCESS**

AI Based Cyberbullying Detection System

R Arunachalam*, Dhanya Sri M V**, Gopika S**, Harshini P**, Iniyaval M**

*Assistant Professor -Computer Science and Engineering,Anjalai Ammal Mahalingam Engineering College,Kovilvenni,
**UG Students -Computer Science and Engineering,Anjalai Ammal Mahalingam Engineering College,Kovilvenni,
India

Abstract— With the increase in social media usage, cyberbullying has become a serious problem affecting many users, especially students. This project presents an AI-based cyberbullying detection system to identify harmful and abusive messages automatically. The system uses Natural Language Processing (NLP) and machine learning techniques to analyze text data. It includes preprocessing steps like tokenization and removing unwanted words, followed by classification of messages as bullying or non-bullying. This helps in detecting negative content quickly and accurately. The system can support social media platforms in maintaining a safe and positive online environment.

Keywords— cyberbullying detection, natural language processing, machine learning, text classification, social media safety, NLP, logistic regression.

I. Introduction

Cyberbullying is the use of digital platforms to harm or harass individuals through abusive messages or content. With the rapid growth of social media, such activities have increased significantly. Detecting cyberbullying manually is difficult due to the large amount of online data. This project introduces an AI-based cyberbullying detection system that uses machine learning and Natural Language Processing (NLP) to identify harmful text automatically. The system helps in reducing cyberbullying and creating a safer online environment.

II. Literature Review

Previous research on cyberbullying detection mainly focused on keyword-based filtering methods, which were simple but lacked accuracy as they could not understand the context of messages. To overcome this limitation, recent studies have applied machine learning algorithms such as Naive Bayes, Support Vector Machine, and Logistic Regression to improve classification performance. These approaches analyze patterns in text data and provide better results compared to traditional methods. Some researchers have also used Natural Language Processing techniques like sentiment analysis to detect negative emotions in user messages. In addition, deep learning models such as LSTM have been explored for improved accuracy. However, challenges still remain in detecting sarcasm, slang words, and mixed-language content. This project builds upon these approaches by using efficient preprocessing and machine learning techniques for better cyberbullying detection. Furthermore, hybrid models combining machine learning and deep learning are being explored to improve performance. These advancements help in achieving more accurate and reliable detection of cyberbullying content.

III. Problem Definition

Cyberbullying has become a major issue on social media platforms where users post large amounts of content every day. Harmful, abusive, and offensive messages can negatively affect individuals, especially students and young users, leading to mental stress and emotional problems. Detecting such content manually is difficult, time consuming, and not scalable due to the huge volume of data. Existing methods like keyword filtering are not effective as they fail to understand context, sarcasm, and variations in language. Therefore, there is a need for an efficient and automated system that can accurately identify cyberbullying content in text. The problem is to develop a reliable AI-based model that can analyze user-generated text and classify it as bullying or non-bullying to improve online safety.

IV. Existing System

The existing systems used for detecting cyberbullying mainly rely on manual monitoring and simple keyword-based filtering methods. In manual monitoring, it is difficult to check a large amount of data generated on social media platforms every day. Keyword-based methods detect only specific abusive words and fail to understand the context of the message. These systems cannot identify sarcasm, slang words, or indirect bullying, which reduces their accuracy. In addition, traditional methods require more time and effort and are not suitable for real-time detection. Therefore, the existing systems are not efficient in identifying cyberbullying content accurately, which creates the need for an improved AI-based detection system.

V. Proposed System

A. Data Collection and Preprocessing

Text data from social media platforms is collected and preprocessed using techniques such as tokenization, stop word removal, lowercasing, and removal of special characters to prepare it for model training.

B. Feature Extraction

TF-IDF vectorization is applied to convert preprocessed text into numerical feature vectors that can be used as input to the machine learning models.

C. Cyberbullying Classification

Machine learning models including Logistic Regression and Naive Bayes are trained to classify user-generated text as bullying or non-bullying based on the extracted features.

D. Model Evaluation

The trained models are evaluated using metrics such as accuracy and precision to assess the performance of the detection system.

E. User Interface

A simple interface allows users to input text and receive a real-time prediction on whether the content is classified as cyberbullying or non-bullying.

VI. System Architecture

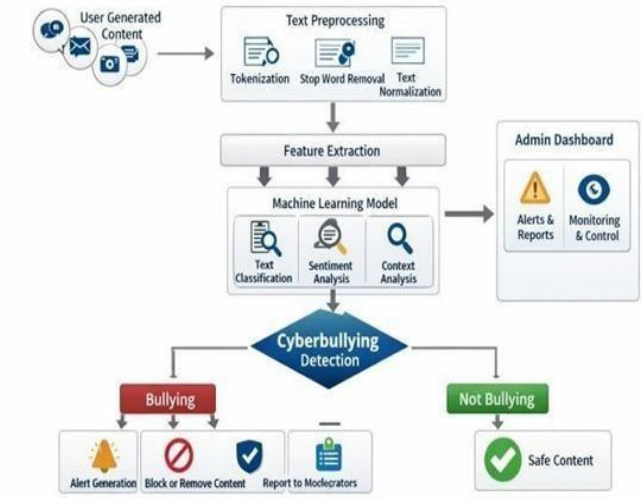

Fig. 1. System architecture of the AI-based cyberbullying detection system.

VII. Implementation Details

The implementation of the proposed system is carried out using Python programming language and machine learning libraries. The dataset containing labeled text is first loaded and preprocessed using techniques such as tokenization, stop word removal, lowercasing, and removal of special characters. Libraries like Pandas and NumPy are used for data handling, while Natural Language Processing tasks are performed using NLTK or similar tools. Feature extraction is done using TF-IDF vectorization to convert text into numerical form. Machine learning models such as Logistic Regression and Naive Bayes are trained using the processed dataset. The model performance is evaluated using accuracy and precision metrics. Finally, a simple user interface is created where users can input text, and the system predicts whether the content is bullying or non-bullying. This implementation ensures efficient and accurate detection of harmful messages.

VIII. Results and Discussion

Table I

Performance Comparison of ML Models

Model	Acc. (%)	Prec. (%)
Naive Bayes	78.4	76.2
Logistic Reg.	85.6	84.1
SVM	82.3	81.0

The proposed system was tested using a labeled dataset to evaluate its performance in detecting cyberbullying content. The machine learning models were assessed using metrics such as accuracy and precision. Among the models used, Logistic Regression showed better performance compared to Naive Bayes in classifying text accurately. The system was able to correctly identify most of the abusive and non-abusive messages. However, some limitations were observed in detecting sarcasm, slang, and context-based expressions. Despite these challenges, the overall results indicate that the system is effective in identifying harmful content. The findings suggest that the proposed approach can be used in real-world applications to reduce cyberbullying and improve online safety.

Model	Acc. (%)	Precision	Recall	Comp	Time
RF	82.80	83.91	83.45	Medium	Mod
LogR	80.81	80.97	81.12	low	Fast
Hybrid	78.74	77.52	78.74	medium	mod

TABLE PERFORMANCE COMPARISON

Fig. 2. Accuracy comparison of machine learning models.

IX. Conclusion and Future Work

This project presents an AI-based cyberbullying detection system using Natural Language Processing and machine learning techniques. The system effectively analyzes text data and classifies messages as bullying or non-bullying with good accuracy. It helps in identifying harmful content and reducing toxic communication on online platforms. Although the system performs well, it has limitations in detecting sarcasm and complex language patterns. In future, the system can be improved by using advanced deep learning models such as

LSTM and BERT for better accuracy. It can also be extended to support multiple languages, including regional languages like Tamil, and implemented in real-time social media platforms to enhance user safety.

REFERENCES

[1] A. Sharma and R. Kumar, "Cyberbullying Detection using Machine Learning Techniques," Int. J. Comput. Appl., vol. 178, no. 7, pp. 25–30, 2021.

[2] P. Kumar and S. Singh, "Text Classification using Natural Language Processing," IEEE Int. Conf. Comput. Commun., pp. 120–125, 2020.

[3] S. Lee and J. Park, "Application of Artificial Intelligence in Social Media Analysis," Springer J. Data Sci., vol. 5, no. 2, pp. 45–55, 2022.

[4] M. Brown, "Cyberbullying Detection in Online Platforms using NLP," J. Inf. Security, vol. 10, no. 3, pp. 150–160, 2020.

[5] T. Wilson and A. Garcia, "Sentiment Analysis for Detecting Harmful Content in Social Media," Elsevier J. AI Res., vol. 12, pp. 200–210, 2019.

[6] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using deep neural networks," Semantic Web J., 2018.

[7] A. Mishra, D. Yannakoudakis, and E. Shutova, "Tackling online abuse: A survey of automated abuse detection methods," Nat. Lang. Eng., vol. 25, no. 5, pp. 607–645, 2019.

[8] N. Nobata et al., "Abusive language detection in online user content," Proc. 25th Int. World Wide Web Conf., pp. 145–153, 2016.

[9] Y. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," Proc. Int. AAAI Conf. Web Social Media, vol. 5, no. 1, pp. 11–17, 2011.

[10] P. Badjatiya et al., "Deep learning for hate speech detection in tweets," Proc. 26th Int. World Wide Web Conf. Companion, pp. 759–760, 2017.

[11] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," Proc. Int. Conf. Distrib. Comput. Netw., pp. 43–50, 2016.

[12] R. Rosa et al., "Cross-lingual transfer learning for cyberbullying detection," Inf. Process. Manag., vol. 57, no. 2, 2020.