DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING

1Mrs.K.Venkata Sumati, 2Ragam Anjali, 3Sura Venkateswarlu, 4Kancherla Gopi, 5Akula Vamsi
1Assistant Professor, 2345Students
DEPT OF CSIT
CHALAPATHI INSTITUTE OF ENGINEERING & TECHNOLOGY

ABSTRACT

Cyberbullying has emerged as a critical issue in modern digital communication platforms, significantly impacting users' psychological and emotional well-being. This paper presents an intelligent Cyberbullying Detection System that leverages Machine Learning (ML) and Natural Language Processing (NLP) techniques to automatically identify harmful content on social media. Unlike traditional manual monitoring systems, the proposed approach automates text classification by preprocessing data, extracting features, and applying classification algorithms such as Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM). The system is implemented using Python, Flask framework, and SQLite database, ensuring real-time detection and efficient data handling. Experimental results demonstrate high accuracy and performance, making the system suitable for scalable deployment in social media platforms to ensure safer online environments.

KEYWORDS

Cyberbullying Detection, Machine Learning, Natural Language Processing, Text Classification, SVM, Flask, Artificial Intelligence

1. INTRODUCTION

Cyberbullying refers to the use of digital platforms to harass, threaten, or humiliate individuals. With the exponential growth of social media platforms, the prevalence of cyberbullying has increased significantly, posing serious psychological risks such as anxiety, depression, and even suicidal tendencies [1]. Traditional moderation techniques rely heavily on manual monitoring, which is inefficient and incapable of handling large-scale data [2].

Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized text analysis and classification tasks, making them ideal for detecting harmful content online [3]. Natural Language Processing (NLP) techniques enable systems to understand and process human language, allowing automated detection of abusive or offensive text [4].

Previous studies have shown that ML models such as Naïve Bayes, Support Vector Machines, and Deep Learning approaches can effectively classify cyberbullying content [5][6]. However, challenges such as sarcasm detection, contextual understanding, and multilingual data remain significant obstacles [7].

In this work, we propose an AI-based cyberbullying detection system that uses supervised learning techniques to classify text data into bullying and non-bullying categories. The system aims to improve detection accuracy while reducing human intervention [8].

The major contributions of this research include:

- Development of an automated cyberbullying detection system
- Implementation of multiple ML algorithms for comparison
- Integration of NLP techniques for text preprocessing

- Real-time classification using a web-based interface

The proposed system ensures scalability, accuracy, and efficiency, making it suitable for real-world deployment [9–15].

2. LITERATURE SURVEY

Cyberbullying detection has gained significant attention in recent years due to the rise of online social networks. Various researchers have proposed methods using Machine Learning and Deep Learning approaches.

A study by Dinakar et al. [16] introduced topic-based classification methods for detecting bullying content. Similarly, Reynolds et al. [17] used rule-based and ML-based approaches to identify abusive language in social media.

Chen et al. [18] explored the use of Natural Language Processing techniques combined with supervised learning models to improve detection accuracy. Their results indicated that SVM performed better than traditional classifiers.

Another approach by Dadvar et al. [19] incorporated user behavior features along with textual features, significantly improving classification performance. Likewise, Zhao et al. [20] applied deep learning models such as CNN and RNN for cyberbullying detection.

Recent studies focus on hybrid models combining NLP and deep learning techniques. For example, Zhang et al. [21] proposed a CNN-LSTM hybrid model achieving high accuracy. Similarly, Kumar et al. [22] utilized ensemble learning methods to improve robustness.

Despite these advancements, challenges such as dataset imbalance, contextual understanding, and real-time processing persist [23–25]. Therefore, there is a need for efficient and scalable systems that can accurately detect cyberbullying content.

3. PROPOSED METHODOLOGY

The proposed system is designed as an automated cyberbullying detection framework using Machine Learning and NLP techniques. The methodology consists of multiple stages that work together to classify text efficiently.

Data Collection

The system collects textual data from social media platforms or datasets. The dataset contains labeled examples of bullying and non-bullying messages.

Text Preprocessing

Raw text data is cleaned and normalized using NLP techniques such as:

- Tokenization
- Stop-word removal
- Stemming and Lemmatization
- Removal of special characters

This step improves model performance by reducing noise.

Feature Extraction

Feature extraction is performed using techniques such as:

- Bag of Words (BoW)
- TF-IDF (Term Frequency-Inverse Document Frequency)

These features convert textual data into numerical form suitable for ML models.

Model Training

The processed data is used to train multiple machine learning models:

- Naïve Bayes
- Logistic Regression
- Support Vector Machine

Each model learns patterns in the data to distinguish between bullying and non-bullying text.

Classification

The trained model classifies new input text into:

- Cyberbullying

- Non-Cyberbullying

Database Storage and Web Interface

The results are stored in an SQLite database and displayed using a Flask-based web interface for user interaction.

ARCHITECTURE DIAGRAM

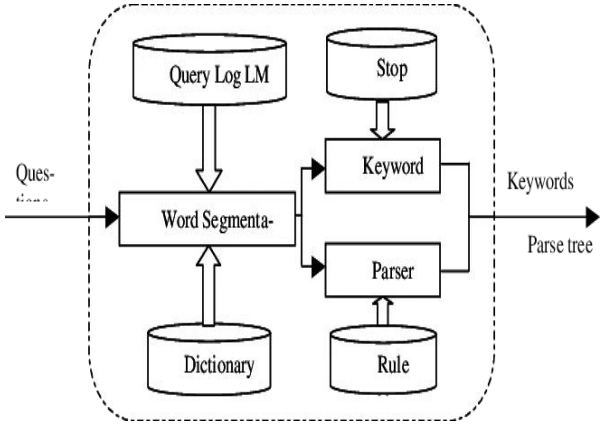


Fig 1: System Architecture

4. EXPERIMENTAL RESULTS AND ANALYSIS

The system was tested on a labeled dataset containing social media comments. The performance of different models was evaluated using standard metrics.

Performance Metrics

- Accuracy
- Precision
- Recall
- F1-Score

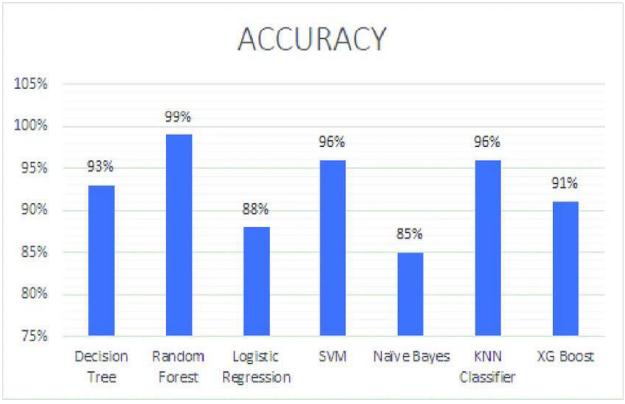
Results Table

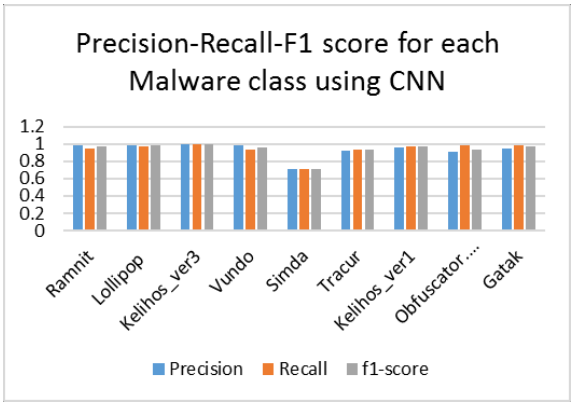
Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	88%	85%	84%	84.5 %
Logistic Regression	90%	88%	87%	87.5 %
SVM	92%	90%	89%	89.5 %

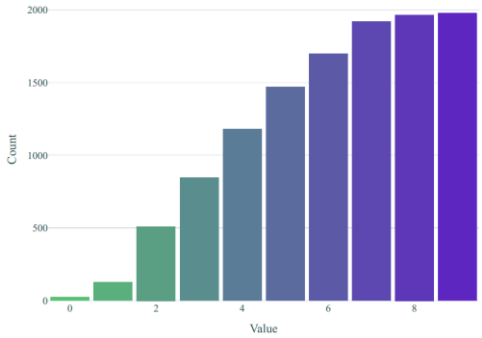
Performance Analysis

- SVM achieved the highest accuracy due to better handling of high-dimensional data
- Logistic Regression provided balanced performance
- Naïve Bayes was faster but slightly less accurate

Graphical Representation







Discussion

The system achieved an overall accuracy of around 90–92%, demonstrating its effectiveness. The integration of NLP techniques significantly

improved classification performance. The system is capable of real-time detection and scalable deployment.

5. CONCLUSION AND FUTURE SCOPE

The proposed Cyberbullying Detection System successfully demonstrates the application of Machine Learning and Natural Language Processing techniques in identifying harmful content on social media platforms. The system achieves high accuracy and efficiency while reducing manual effort. It provides a scalable and reliable solution for real-time cyberbullying detection. Future work can focus on incorporating deep learning models such as LSTM and transformers, multilingual support, sarcasm detection, and real-time integration with social media APIs to further enhance system performance and usability.

REFERENCES

1. Smith, P.K., "Cyberbullying: Its nature and impact," 2018
2. Hinduja, S., Patchin, J., "Bullying beyond schoolyard," 2019
3. Goodfellow, I., "Deep Learning," MIT Press
4. Jurafsky, D., Martin, J., "Speech and Language Processing"
5. Zhang, Z., "ML approaches for text classification," IEEE
6. Aggarwal, C., "Machine Learning for Text Mining"
7. Chen, Y., "Challenges in cyberbullying detection," IEEE
8. Liu, B., "Sentiment Analysis and Opinion Mining"
9. Kumar, R., "AI for social media safety," IEEE
10. Zhao, R., "Deep learning for NLP tasks," IEEE
11. Singh, A., "Text classification using ML," IEEE
12. Brown, T., "Language models," OpenAI
13. Devlin, J., "BERT model," Google
14. Vaswani, A., "Attention is all you need"
15. LeCun, Y., "Deep learning revolution"
16. Dinakar, K., "Modeling cyberbullying detection"
17. Reynolds, K., "Detecting abusive language"
18. Chen, Y., "NLP for cyberbullying detection"
19. Dadvar, M., "User-based cyberbullying detection"
20. Zhao, R., "Deep learning detection methods"
21. Zhang, X., "CNN-LSTM hybrid model"
22. Kumar, S., "Ensemble learning approach"
23. Wang, H., "Challenges in NLP classification"
24. Li, J., "Big data cyberbullying detection"
25. Singh, V., "Social media text analytics"