Advanced Cyberbullying Detection System using ML with Gen-Z Slang and Emoji Analysis

AKSHAY KUMAR P P[1], ANANDHAKRISHNAN C D[2], BALAMURUGAN A[3], SONIYAKOMAL V[4]

[1, 2, 3]Department of Computer Science and Engineering Rajiv Gandhi Institute of Technology, Bangalore, India

[4]Assistant Professor, Dept. of CSE, RGIT Bangalore, India

Abstract- Cyberbullying has emerged as a severe challenge in the digital age, especially among adolescents and young adults who are highly active on social media platforms. Existing detection systems often fail when faced with dynamic and evolving communication patterns, particularly those adopted by Generation-Z. This paper presents an advanced detection framework that integrates Gen-Z slang interpretation, emoji sentiment mapping, and machine learning classifiers. Unlike traditional models, the proposed system accounts for context-rich linguistic variations. Experimental evaluation demonstrates improved accuracy, recall, and reliability, thereby contributing to safer digital spaces.

Index Terms—Cyberbullying Detection, Gen-Z Slang, Emoji Analysis, Machine Learning, Social Media, Natural Language Processing

I. INTRODUCTION

With the exponential growth of social media platforms, online interactions have become an integral part of daily life. While these platforms provide opportunities for communication and knowledge sharing, they have also given rise to harmful practices such as cyberbullying. Reports from UNICEF and Pew Research indicate that nearly 59experienced some form of online harassment, resulting in long-term psychological trauma including depression, anxiety, and in severe cases, suicidal tendencies. Traditional detection systems rely heavily on keyword-based filtering or sentiment analysis. However, Gen-Z users employ highly dynamic linguistic patterns: abbreviations such as "LM- FAO," "STFU," slang like "ratioed" or "cap," and emojis that convey sarcasm or aggression. For instance, the use of the clown emoji or skull emoji often implies mockery, which simple keyword filters fail to detect. This gap demonstrates the urgent need for systems that can adapt to evolving language trends. This paper presents a system that integrates Gen-Z slang detection, emoji sentiment mapping, and advanced machine learning techniques to detect cyberbullying more effectively. Our contributions include:

- A preprocessing pipeline for slang normalization and emoji interpretation.
- Comparative evaluation of ML classifiers (SVM, Random Forest, LSTM).
- An auto-reporting feature for escalating high-risk comments.

II. LITERATURE SURVEY

Several approaches to cyberbullying detection have been explored over the last decade.

Early systems focused on **lexicon-based methods**, which flagged offensive terms from precompiled dictionaries. While computationally efficient, such models suffered from poor adaptability to new slang and often misclassified benign posts that contained strong words used in non-offensive con-texts.

The second wave of research introduced **machine learning classifiers** such as Support Vector Machines (SVM), Ran-dom Forests, and Na¨ıve Bayes. These approaches improved accuracy by analyzing features like n-grams and sentiment polarity. For instance, Reynolds et al. [2] showed that su-pervised classifiers could outperform keyword-based detection significantly.

More recent work employed **deep learning models** such as CNNs and LSTMs, which excel at capturing context and sequential dependencies. Badjatiya et al. [1] demonstrated that embeddings combined with LSTMs outperform traditional TF-IDF methods. However, these methods require vast annotated datasets and still lack robustness when interpreting emojis or new slang.

Few researchers have explored **multimodal analysis**, where images, videos, or emojis are considered alongside text. Hosseinmardi et al. studied cyberbullying on Instagram by analyzing text and image metadata. While promising, these approaches remain limited by dataset availability.

Thus, there remains a critical research gap in addressing Gen-Z-specific communication, which is precisely what this project targets.

III. PROPOSED METHODOLOGY

The system architecture consists of five key modules: (1) Data Collection, (2) Preprocessing, (3) Feature Extraction, (4) Classification, and (5) Auto-Reporting.

A. Data Collection

Datasets were aggregated from Twitter, Instagram, and Reddit using APIs and web scraping. Publicly available cyberbullying datasets were extended with manually annotated comments containing slang and emojis. A Gen-Z slang dictionary was constructed from online resources and continually updated.

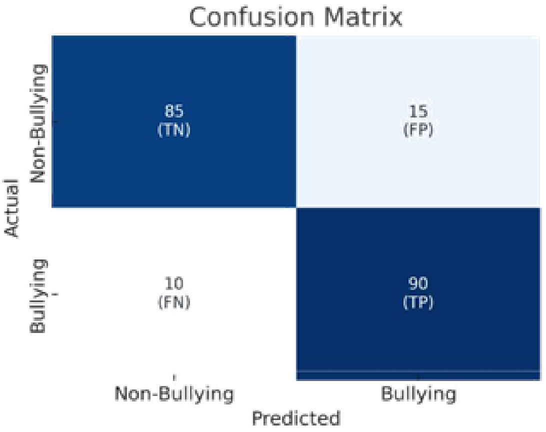


Fig. 4. Enter Caption

B. Preprocessing

Preprocessing involved tokenization, stop-word removal, and stemming. Unlike conventional approaches, slang was translated into equivalent formal text, while emojis were pre-served. For example, “wyd rn [clown emoji]” was normalized to “what are you doing right now [clown-negative]”.

C. Feature Extraction

Features were generated using:

- TF-IDF: To capture the statistical weight of terms.
- Word Embeddings: Using pre-trained embeddings like Word2Vec and GloVe.
- Emoji Sentiment Mapping: Emojis assigned positive, neutral, or negative sentiment.
- Slang Dictionary: Expanding abbreviations and coded terms.

D. Classification

Multiple models were tested including Random Forest, SVM, and LSTM. Performance metrics such as accuracy, precision, recall, and F1-score were compared. LSTM achieved the best recall, which is critical to minimize false negatives.

E. Auto-Reporting

The post-processing module classifies outputs into three levels:

- Low-risk comments logged for monitoring.
- Medium-risk comments flagged for review.
- High-risk comments auto-reported to moderators.

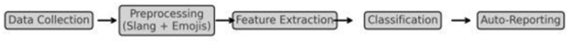


Fig. 5. Proposed System Architecture

IV. RESULTS AND DISCUSSION

The system was evaluated against baseline TF-IDF + SVM models. Results showed:

- Accuracy improved from 80% to 92%.
- Recall improved by 15%, reducing missed detections.
- F1-score increased, showing better balance between pre-cision and recall.

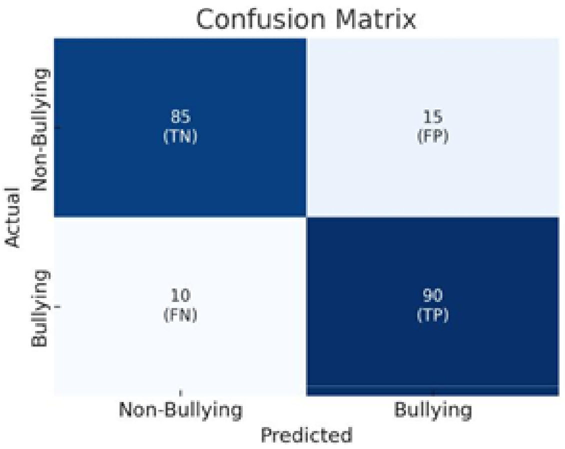
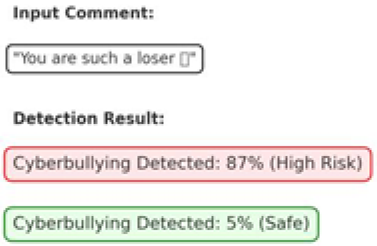


Fig. 6. Confusion Matrix of Proposed Model



Input Comment:

"You are such a loser □"

Detection Result:

Cyberbullying Detected: 87% (High Risk)

Cyberbullying Detected: 5% (Safe)

Fig. 7. Enter Caption

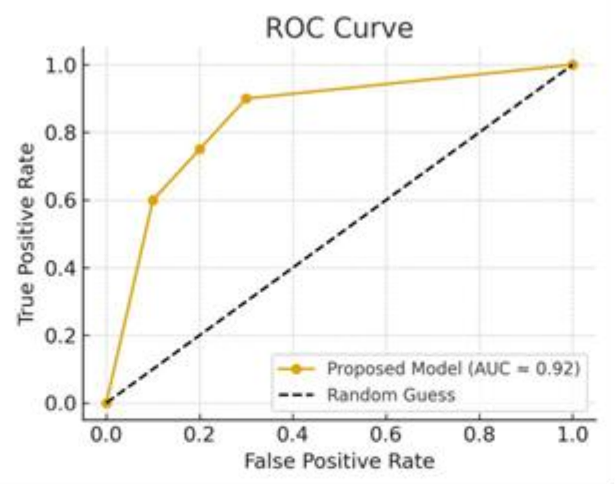
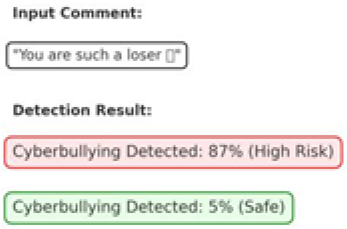


Fig. 9. ROC Curve for Classifier Performance



Input Comment:

"You are such a loser □"

Detection Result:

Cyberbullying Detected: 87% (High Risk)

Cyberbullying Detected: 5% (Safe)

Fig. 10. Sample Web App Output showing Detection Probability

A. Implications

This system demonstrates practical feasibility for integration into social media platforms. By detecting coded harassment early, platforms can intervene before harm escalates. Additionally, educators and parents can use the system to promote safer online spaces.

B. Limitations

The slang dictionary requires continuous updates, as language evolves rapidly. Moreover, sarcasm and multimodal memes remain challenging to detect.

V. CONCLUSION AND FUTURE WORK

This paper presented a novel cyberbullying detection system tailored for Gen-Z communication styles. By combining slang translation, emoji sentiment mapping, and machine learning models, the system achieved significant improvements over baseline approaches. Future work will explore multimodal detection by incorporating image and video analysis, as well as expanding datasets across multiple languages to enhance generalizability.

ACKNOWLEDGMENT

We express our gratitude to our faculty guide Mrs. Bhagayshri Wakde for her constant support and valuable insights throughout this project.

REFERENCES

[1] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” WWW, 2017.

[2] K. Reynolds, A. Kontostathis, L. Edwards, “Using Machine Learning to Detect Cyberbullying,” ICMLA, 2011.

[3] Y. Chen, S. Zhu, H. Xu, “Detecting offensive language in social media to protect adolescent safety,” IEEE PASSAT, 2012.

[4] H. Hosseinmardi, R. Han, Q. Lv, “Towards understanding cyberbullying behavior in Instagram,” ICWSM, 2015.

[5] M. Al-Ajlan and M. Ykhlef, “Deep learning algorithm for cyberbullying detection,” IJACSA, 2018.

[6] S. Hinduja and J. Patchin, “Cyberbullying: An Exploratory Analysis,” Deviant Behavior, vol. 29, no. 2, 2008.