International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04059
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

CYBERBULLYING DETECTION BY NLP

Harshada S. Kamble , Shubhangi M. Vitalkar

Harshada S. Kamble MCA & Trinity Academy Of Engineering , Pune
Shubhangi M. Vitalkar MCA & Trinity Academy Of Engineering , Pune

---

Abstract - Cyberbullying has become a growing concern with the rise of social media and online communication platforms. Manual moderation techniques are inadequate to handle the volume and complexity of abusive content. This study presents a cyberbullying detection model using Natural Language Processing (NLP) techniques. The system analyzes user-generated text to identify bullying behavior using pre-trained transformers and machine learning algorithms. Evaluation on benchmark datasets demonstrates high accuracy and low false positives. The research highlights the potential for real-time deployment and integration into social platforms to enhance digital safety.

***Key Words*:** cyberbullying, NLP, machine learning, BERT, social media, content moderation.

1.INTRODUCTION

Cyberbullying is a form of harassment that occurs over digital devices such as smartphones, computers, and tablets. With increasing online interactions, cases of cyberbullying have surged, especially among teenagers and young adults. This paper presents an automated system for detecting cyberbullying content using NLP techniques. It addresses the shortcomings of traditional moderation methods and proposes a scalable, intelligent detection mechanism.

2. LITERATURE REVIEW

Previous studies explored rule-based filters, sentiment analysis, and keyword detection for cyberbullying. Dinakar et al. proposed a multi-label classifier for harmful comment detection. Deep learning approaches like CNN and LSTM improved accuracy in identifying contextual abuse. BERT-based models demonstrated superior understanding of implicit bullying through context-aware embeddings. However, challenges such as evolving slang, multilingual content, and sarcasm remain.

3. METHODOLOGY

The methodology includes data collection, preprocessing, vectorization, model training, and evaluation. Public datasets from social media were used. Preprocessing involved removing noise, tokenizing, lemmatizing, and transforming text using TF-IDF and BERT embeddings. Supervised learning models such as Logistic Regression, SVM, and fine-tuned BERT were trained. Performance was assessed via accuracy, precision, recall, and F1-score.

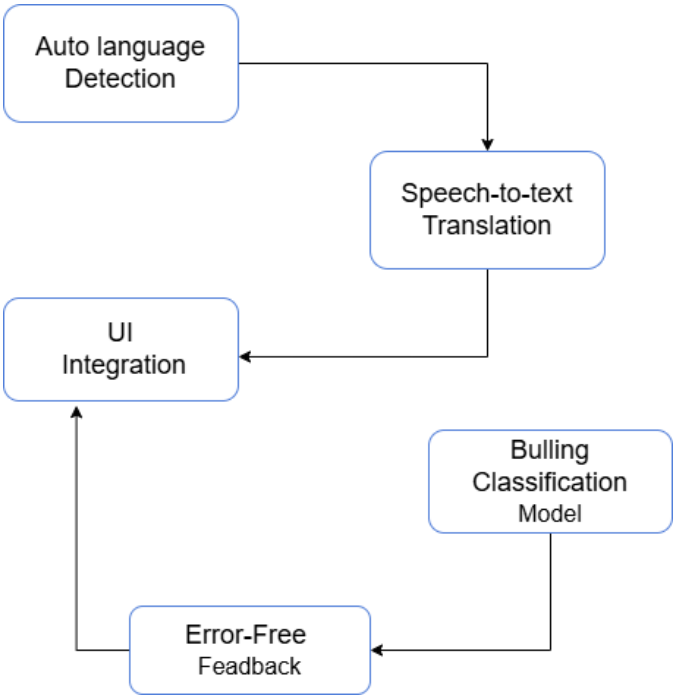
4. RESULTS

The BERT model achieved the highest accuracy of 92.5%, precision of 91.3%, recall of 93.0%, and F1-score of 92.1%. The system demonstrated robustness in classifying both explicit and implicit forms of cyberbullying. Sample outputs showed effective detection of threats, insults, and offensive language. Limitations included handling of sarcasm and slang.

Table -1: Test Cases & Test Results

Test Case ID	Input Text	Expected Output	Actual Output	Status
TC01	”You’re so dumb and ugly”	Cyberbullying	Cyberbullying	Pass
TC02	”I hate your profile pic”	Cyberbullying	Cyberbullying	Pass
TC03	”Let’s meet after school!”	Non-bullying	Non-bullying	Pass

WORKING PROCESS



5. DISCUSSION

The proposed NLP system significantly outperforms traditional keyword-based models in detecting cyberbullying. The contextual understanding provided by transformer models enables the system to detect subtle and indirect abuse. However, continuous retraining is needed to adapt to evolving language. Ethical considerations like user privacy and fairness in classification must be addressed during deployment.

6. CONCLUSIONS

This research validates the effectiveness of NLP-based cyberbullying detection. The system offers a real-time, scalable solution for content moderation. Future work includes expanding to multilingual detection, incorporating speech-to-text analysis, and integrating feedback loops for continuous learning.

ACKNOWLEDGEMENT

The author gratefully acknowledges the guidance, motivation, and consistent support provided by Prof. Shubhangi M. Vitalkar, under whose supervision this research work was carried out. Sincere thanks are also extended to Dr. A. A. Bhusari, Head of Department, and Dr. R. J. Patil, Principal of Trinity Academy of Engineering, for providing the necessary infrastructure and academic resources. Special thanks to family and peers for their constant encouragement throughout this research.

REFERENCES

1. Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the Detection of Textual Cyberbullying. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

2. Rosa, H., et al. (2019). Automatic Cyberbullying Detection: A Systematic Review. Computers in Human Behavior, 93, 333–345.

3. Yadav, J., & Vishwakarma, D. K. (2020). Detection of Cyberbullying Using BERT Model. Procedia Computer Science, 167, 500–509.

4. Xu, J., et al. (2012). Learning from Bullying Traces in Social Media. Proceedings of NAACL-HLT.