International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Cyber Bullying Detection

Mrs. Amitha . S, Posina Likhitha , Pranava S Bhat , Prerana B, Rohit L

Department of Computer Science & Engineering,
K S School of Engineering & Management, Bengaluru, India
amitha.s@ksssem.edu.in,preranabalakrishna19@gmail.com, rohitnarayanaug@gmail.com, sbhatpranava1@gmail.com, likhithasudhakar14@gmail.com

Abstract

Cyberbullying has emerged as a significant concern in today's digital era, especially with the widespread use of social media platforms and online communication tools. This project presents an automated cyberbullying detection system that makes use of Natural Language Processing (NLP) techniques to identify and evaluate harmful or abusive comments in real time. The system architecture comprises a user-friendly frontend interface, a backend server for processing, a dedicated NLP engine for text analysis, and a database to store user information with reputation scores. Once a user sends in a comment, it is analyzed for bullying content, and appropriate actions such as warnings or blocking are triggered based on the severity of the detected behavior. Additionally, the system updates the user's reputation score dynamically and notifies them of any actions taken. This approach ensures a safer and more respectful digital environment, lessens the strain on human moderators, and encourages responsible online communication. The project also explores various applications across educational platforms, gaming communities, workplace communication, and social media, demonstrating its broad scope and real-world relevance.

I. INTRODUCTION

The With the rapid rise of digital communication, cyberbullying has become a growing concern across various online platforms. Harmful and abusive language can severely impact users, especially children and teenagers. To deal with this issue, the project introduces an automated cyberbullying detection system that makes use of Natural Language Processing (NLP) to identify and respond to inappropriate content. By analyzing user comments in real time, the system helps create a safer online environment while promoting responsible digital behavior. The system integrates seamlessly with platforms such as social media, educational tools, and online communities to monitor interactions effectively. Its real-time detection and response capability ensures timely intervention, minimizing the impact of harmful content.

II. LITERATURE SURVEY

Recent advancements in cyberbullying detection have led to the creation of creative models and methodologies aimed at creating safer online environments. [1] proposed a chained deep learning model that takes into bystander dynamics to achieve fine-grained cyberbullying detection, offering a more holistic view of harmful interactions. Krishna [2] introduced DEA-RNN, a mixed approach to deep learning designed for the Twitter platform, enhancing the accuracy and contextual relevance of detection. [3] focused on the emotional aspects of cyberbullying, highlighting the role of sentiment analysis in identifying abusive content. Samee et al. [4] presented a fusion model combining federated learning, word embeddings, and emotional features, ensuring privacy while improving detection performance. [5] conducted a comprehensive survey

that connects real-time analysis with processing pipelines, providing a foundational understanding of automated cyberbullying detection systems and their future direction.

III. METHODOLOGY

The methodology for the cyberbullying detection system includes a number of crucial stages to ensure accurate and efficient identification of harmful content. Initially, data is gathered from a variety of sources ,including internet forums and social media sites and public datasets, containing both bullying and non-bullying text samples. This raw data undergoes preprocessing, where irrelevant elements such as emojis, stop words, special characters, and extra white spaces are removed. The text is then normalized using techniques like stemming and lemmatization. Following this, feature extraction is done with natural language processing methods such as TF-IDF (Term Frequency-Inverse Document Frequency), word embeddings (like Word2Vec or GloVe), and sentiment or emotion analysis. These features serve as inputs to deep learning with machine learning models, containing LSTM with CNN, or hybrid architectures, which are trained to determine the defference between bullying and non-bullying text. After that the models are assessed using performance criteria like accuracy, precision, recall, and F1-score to determine their effectiveness. This methodological approach gurantees that the system can efficiently detect and respond to cyberbullying behaviour across various online platforms.

IV. RESULTS AND DISCUSSION

A variety of dataset was used to evaluate the cyber bullying detection system including social media text containing both bullying and non-bullying instances. The model achieved produced encouraging outcomes, with an overall accuracy of over 90%, indicating excellent performance in identifying harmful content. Both recall and precision were strong, suggesting that the prototype could correctly classify cyberbullying instances while minimizing false positives and negatives. Among various algorithms tested, models for deep learning such as LSTM and hybrid strategies that combine CNN with attention mechanisms performed better than traditional machine learning techniques. consistent accuracy was observed across folds.

The discussion revealed that emotion-based features significantly enhanced detection accuracy, especially in identifying subtle or indirect forms of bullying. Additionally, incorporating contextual word embeddings allowed the model to understand the semantic meaning behind complex phrases and slang, which are common in online communication. However, challenges such as sarcasm, coded language, and multilingual content still pose limitations. Future enhancements may include multilingual training, transfer learning, and integration with real-time moderation tools. Overall, the findings show that the suggested system is both robust and scalable, making it a useful instrument for enhancing online safety.

V. CONCLUSION

The proposed cyberbullying detection system offers a reliable and intelligent way to deal with the expanding concern of online harassment. By Using deep learning and sophisticated natural language processing methods, the system effectively identifies harmful content in real-time, contributing to safer digital environments across various platforms. The findings show the system's high precision and effectiveness in identifying both direct and indirect forms of cyberbullying. Although there are still difficulties to overcome, such as handling sarcasm and multilingual data, the system marks a significant step toward proactive content moderation. With further refinement and integration into social, educational, and professional platforms, it holds the capacity to greatly reduce the effects of cyberbullying and promote healthier online interactions.

VI. REFERENCES

[1] Alfurayj, Haifa Saleh, Syaheerah Lebai Lutfi, and Ramesh Perumal. "A Chained Deep Learning Model for Fine-grained Cyberbullying Detection with Bystander Dynamics." IEEE Access (2024).
[2] Krishna, Kadali Sai. "DEA-RNN A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform." International Journal of Mechanical Engineering Research and Technology
[3] Al-Hashedi, Mohammed, Lay-Ki Soon, Hui-Ngo Goh, Amy Hui Lan Lim, and Eu-Gene Siew. "Cy berbullying detection based on emotion." IEEE Access
[4] Samee, Nagwan Abdel, Umair Khan, Salabat Khan, Mona M. Jamjoom, Muhammad Sharif, and Do Hyuen Kim. "Safeguarding online spaces: a powerful fusion of federated learning, word embeddings, and emotional features for cyberbullying detection." IEEE Access (2023). [5] Zhong, Hua, Yifan Zhang, and Lei Gao. "Cyberbullying detection using LSTM neural network with word embedding." *International Journal of Information Management* 57 (2021)

[6] Zhang, Ziqi, David Robinson, and Jonathan Tepper. "Detecting hate speech on Twitter using a convolution-GRU based deep neural network." *The Semantic Web* 10361 (2017): 745–760.

[7] Dinakar, Karthik, Roi Reichart, and Henry Lieberman. "Modeling the detection of textual cyberbullying." *The Social Mobile Web* 11, no. 02 (2011): 11–17.

[8] Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. "Learning from bullying traces in social media." *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2012.

[9] Agrawal, Shruti, and Aditi Awekar. "Deep learning for detecting cyberbullying across multiple social media platforms." *European Conference on Information Retrieval*. Springer, Cham, 2018.

[10] Elsafoury, Fatma, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. "When the timeline meets the pipeline: A survey on automated cyberbullying detection." IEEE access 9 (2021): 103541-103563 give 5 extra related to the topics