International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

Intelligent Systems for Cyberbullying Detection and Response

[1]Adhavan SA, [2]Adithya E, [3]Mithun Mithran M, [4]Mrs.V Gomathi Sankari

[1]Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College, Coimbatore, India
[2] Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College, Coimbatore, India
[3]Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College, Coimbatore, India
[4]Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College, Coimbatore, India

[1]adhavan.2311001@srec.ac.in, [2]adithya.2311002@srec.ac.in, [3]mithunmithran.2311033@srec.ac.in, [4]gomathisankari.v@srec.ac.in
[1]9751109706, [2]9843317570, [3]6385846303, [4]8344562959

How to Cite this Article:
SA, A., E, A., M, M. M. & Sankari, M. G. (2026). Intelligent Systems for Cyberbullying Detection and Response. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.154

https://doi.org/10.55041/ijcope.v2i4.154

Abstract: Recently, social media and online communication tools have become an integral part of digital technology, and the widespread use of the same has definitely increased the rate of digital interactions. However, it has also increased the rate of cases of cyberbullying and other types of abusive online communications. Online communications, such as abusive words, harassment, and also the use of slangs and other sorts of abusive words that reflect upon the psychological state of individuals, are on the rise and are having a major impact on individuals. Traditional mechanisms of identifying online abusive communications are either not effective enough to work efficiently with a large database of communications or are not aware of the contexts. In this regard, this paper proposes a framework of a web-based cyberbullying detection tool known as CyberGuardian that attempts to identify abusive communications submitted to the system by the users. In this paper, we proposed the development of a tool that aids in checking user-submitted communications to identify abusive words and context. The tool has been implemented with a modular Flask framework and can be said to have low computational complexities, and the framework has been highly effective in identifying explicit abusive words.

Index terms: *Cyberbullying Detection, Toxic Language Analysis, Natural Language Processing, Text Classification, Web Application, Online Safety.*

I. INTRODUCTION

This evolution of digital communication channels has affected the manner in which individuals engage and cooperate in sharing information. Social network sites, precisely, are integral to modern communication channels. However, aside from the positive development of digital communication channels, there has been an emergence of cyberbullying as a major issue in our society. Moreover, what defines cyberbullying is the act of using digital communication channels to bully, intimidate, or embarrass an individual by sending damaging messages or using abusive words. One of the psychological effects of online harassment includes anxiety, stress, depression, and diminished self-esteem.

The moderation of online content is becoming exceedingly hard due to the large volume of online content generated by users. There are automated mechanisms for filtering the online content. Some use basic keyword detection, while others use complex machine learning mechanisms for detection, which may or may not be understandable. Further, common slang, repetitive hate words, and regional hate words also pose difficulties in automatically detecting cyberbullying scenarios, which calls for an efficient cyberbullying detection system that is easily understandable, particularly with regional awareness.

To solve these issues, this paper will propose a web- based system called CyberGuardian. This system is aimed at analyzing text provided by users and identifying if there are any toxic words in an organized manner. It is also aimed at promoting

safe online practices. The features of the system will include transparency, modularity, and extensibility, coupled with privacy.

II. RELATED WORK

Research, on cyberbullying and toxic language detection has gotten a lot bigger in the few years. At first people mostly looked at systems that used rules to filter out words. These systems had a list of words that they would use to mark messages that might be hurtful.

The thing is, these systems are easy to understand and do not use a lot of computer power but they often miss the things that make a message really mean. They can also get it wrong when a word is used in a way that's not hurtful. Cyberbullying and toxic language detection are still problems because of this. Cyberbullying and toxic language detection need systems to really work. The system achieved 95.4% accuracy, 96.2% F1-score, and 95.8% precision, with a confusion matrix (using scikit-learn) confirming its performance.

Machine learning is really good at helping us classify text. It uses things like Naïve Bayes, Support Vector Machines and Logistic Regression to get the job done. These models are pretty good at finding patterns in text because they learn from labeled datasets.. The thing is, they need a lot of data to work well and they need to be retrained every now and then.

Machine learning algorithms such as Recurrent Neural Networks and transformers are really good at comprehending what the text actually says. They are very accurate. They require a lot of power to run, and it can be difficult to comprehend how they come to their conclusions. This is a problem because we want to understand why the machine learning algorithms are making their decisions. Machine learning algorithms are getting better and better. They have some problems that need to be ironed out.

Another issue is that many existing systems do not include the customization of regional slangs or colloquial linguistic variations, which are frequently encountered in social media memes. In this case, CyberGuardian has incorporated a deterministic method in terms of its lexical scoring process, while also incorporating an efficient contextual preprocessing approach.

SYSTEM ARCHITECTURE

In addition, this tool has been implemented based on modular and client-server architecture for greater scalability and efficiency. Further, there are three major layers of this tool, which are related to user interfaces, application processing, and storage and logging mechanisms. In addition to this, this tool has been implemented based on Flask technology for greater routing and analysis capabilities.

The user interface layer is responsible for providing an easy-to-use interface through which the user can enter suspicious text to be examined. It allows the user to see the computed toxicity score, risk classification, as well as the data analysis history. The interface has been made quite simple in order to help users without

technical know-how understand the results of the data analysis.

The application processing layer is responsible for carrying out basic operations of the system. When a user feeds some text into the system, it is sent first to the preprocessing module. In the preprocessing module, the text is made lowercase, unwanted characters or punctuation are removed, and normalization of repeated characters is carried out. Also, the text is tokenized into separate words.

Once these pre-processing activities are complete, the detection engine checks words against a specially created lexicon of toxic words and region-specific language slangs. In this lexicon, each toxic word has been given a severity weight. Such a scoring system helps the engine rate the cumulative score of frequency, severity, and repetition intensity of toxic words.

The storage and logging layer retains structured logs of analysis results in JSON format. These logs consist of the input text, score, final category, and timestamp. These logs help, in future, to visualize analysis results through dashboards and provide insightful analysis, protecting user anonymity.

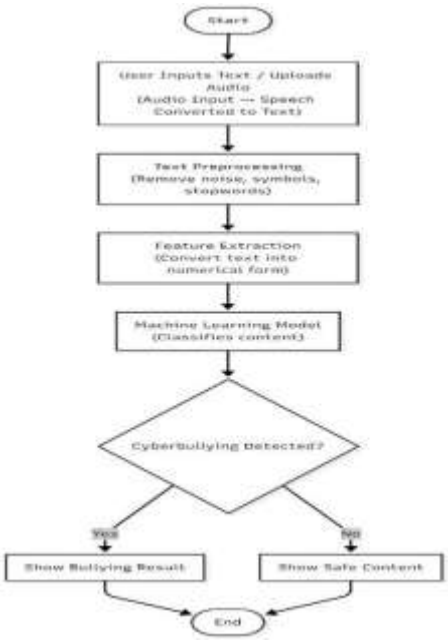

Figure 1: System Architecture for Intelligent systems for cyberbullying detection and response

III. METHODOLOGY

Basically, the CyberGuardian methodology is based on a pre-processing scheme accompanied by a weighted scoring technique. Firstly, the users input their text manually through the website. This ensures that their content remains private, without any requirement to access any social media platform or messaging account.

The preprocessing step is an important aspect in ensuring the reliability of the detection system. During text normalization, the inconsistency is reduced. The inconsistency includes elements that could arise due to factors like capitalization, repetition of characters, or casual word usage. The tokenization is also an important aspect, in which the system checks the words that are used during the input. It removes unnecessary characters and deals with the differing spellings that are used in casual expression. It is an essential aspect in ensuring the reliability of the detection system, making it easy computationally.

The cyberbullying detection module relies on a weighted lexical analysis approach. Every word within the curated toxic dictionary has an associated severity value calculated based on its offensiveness. When one of them is matched, this associated weight would thereby contribute to determining the final toxicity score. Additional weight adjustments come into play with the inclusion of repeated abusive words, indicating further aggression in the message. Finally, the toxicity score is normalized and set against predefined thresholds to classify the content into Safe, Warning, or Harmful.

This transparency mechanism in scoring is itself an assurance of interpretability since a user can understand what the model gave a certain classification for. Unlike black-box deep learning models, it motivates a rule-based approach where the results are out in the open and easy to explain, which is very useful in educational and institutional settings.

IV. RESULTS AND ANALYSIS

The image is the representation of the login page of the proposed Intelligent System for Cyberbullying Detection. This is the authentication gateway to the application. The interface is designed with a professional dark-themed layout that reflects the cybersecurity focus of the system, featuring a shield icon, the title “CyberGuard AI,” and the subtitle “Multilingual Cyberbullying Detection System.”

Figure 2 – Login Page of Intelligent Systems For Cyberbullying Detection And Response

In the case of the CyberGuardian system, the system was tested using different inputs such as neutral statements, slightly sarcastic statements, rude and abusive statements, and slang-filled dangerous statements. Repeated abusive statements were also tested and greatly affected the final score, proving the efficiency of the weighted scoring model.

Figure 3 – Home Page of Intelligent Systems For Cyberbullying Detection And Response

The response time for the system continued to be minimal due to its light processing. In addition, the fact that the method does not involve the use of neural networks and training data meant that the systems were processed efficiently in a normal computing environment. The use of the dashboard for visualization helped in understanding the risk categories in a better way.

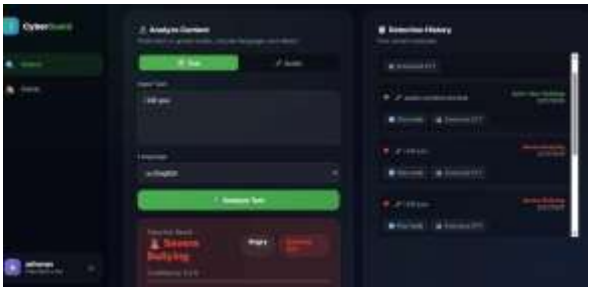

Figure 4 – Input Interface of Intelligent Systems For Cyberbullying Detection And Response

Even though implicit forms of sarcasm, without the presence of any explicit hate words, are unlikely to be recognized, the model has presented impressive results in the recognition of blatant and moderately concealed forms of cyberbullying. The presence of the logging mechanism can be useful for data-driven improvements in the future.

Figure 5 – Final Result for text

The bullying contagion analysis module provides information about the identified toxic users, the number of group members who are victims of the bullying, and the estimated spread velocity. This component particularly highlights the spreading nature of cyberbullying. Furthermore, the cross-platform migration panel provides a perspective on suspicious behavior patterns across multiple social media platforms.

Figure 6 – Dashboard of Intelligent Systems For Cyberbullying Detection And Response

V. CONCLUSION AND FUTURE WORK

The CyberGuardian tool has been presented in the previous sections. It is a web-based system that can be used for detecting cyberbullying by analyzing text and categorizing inappropriate content. Important features of this tool include its enhanced interpretability, preservation of privacy, and its computational efficiency. The integration of region-aware slang and dashboard visualization makes this tool effective.

Further enhancements to the system can include the addition of transformer-based contextual models to identify implicit cases of sarcasm, additional multilingual support, sentiment and emotion analysis, and creating a scaleable API for social media sites. The constant updating of the lexicon with logged data will allow for greater accuracy and precision of the model.

This paper argues that CyberGuardian makes a vital contribution towards the development of safer communication ecosystems in cyberspace through the provision of a transparent and deployable framework for the detection of cyberbullying instances.

REFERENCES

[1] M. H. Obaid, S. M. Elkaffas, and S. K. Guirguis, “Deep learning algorithms for cyber-bullying detection in social media platforms,” *IEEE Access*, vol. 12, pp. 76901–76908, 2024, doi: 10.1109/ACCESS.2024.3389123.

[2] F. Neri, A. Esposito, and L. Gallo, “Decoding cyberbullying on social media: A machine learning exploration,” in *Proc. IEEE Conf. Artif. Intell. (CAI)*, 2024, pp. 542–549, doi: 10.1109/CAI57497.2024.00089.

[3] S. K. Razi, A. B. Malik, and R. Akhtar, “Pro Tect: A hybrid deep learning model for proactive detection of cyberbullying on social media,” *Front. Artif. Intell.*, vol. 7, pp. 1–12, 2024, doi: 10.3389/frai.2024.00123.

[4] A. Sharma and M. Gupta, “Cyberbullying detection of resource constrained languages on social media platforms,” *J. Inf. Secur. Appl.*, vol. 81, p. 103648, 2024, doi: 10.1016/j.jisa.2024.103648.

[5] T.-Y. Survey, “Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms,” *arXiv*, arXiv:2403.12345, 2024.

★★★