J. Saravana Mukhil, 2025, 13:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Protecting Users from Online Harassment through Automated Detection

J. Saravana Mukhil, Assistant Professor Dr. D. R. Kritika
Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

Abstract- Cyberbullying, which occurs on digital platforms such as social media, messaging apps, and online gaming, involves the use of technology to harass, humiliate, or harm individuals. This form of bullying can result in lasting emotional distress as harmful or private information is shared publicly, creating a permanent record that can have long-term consequences. Despite efforts to detect and prevent cyberbullying, many existing approaches, particularly those based on Machine Learning (ML) and Natural Language Processing (NLP), often fail to capture the deeper semantic meaning of the text, limiting their effectiveness in accurately identifying bullying content.

Keywords– Cyberbullying Detection, Natural Language Processing (NLP), LSTM-CNN Model, Word2Vec Embeddings, Toxicity Classification

I. INTRODUCTION

The rapid proliferation of digital platforms such as social media, messaging apps, and online forums has brought about a new form of harassment known as cyberbullying. Unlike traditional bullying, cyberbullying enables individuals to harass, threaten, or embarrass others anonymously and persistently, often leaving a long-lasting impact on the victim's emotional well-being. Harmful content shared online can spread quickly and remain accessible, creating a permanent digital record that can be difficult to erase. As a result, ensuring online safety has become an urgent priority.

Despite significant efforts in research and technological development, existing cyberbullying detection methods often fail to capture the deeper semantic and contextual meanings of user-generated content. Many traditional approaches rely on basic keyword detection or machine learning models that struggle to detect nuanced, implicit, or evolving forms of harmful behavior. These shortcomings can lead to a high rate of false positives or missed cases, limiting the effectiveness of intervention systems.

Our Website Design

II. LITERATURE SURVEY

Recent developments in natural language processing and deep learning have greatly enhanced cyberbullying detection systems, particularly with the use of contextual language models and hybrid neural architectures. Dinakar et al. applied multi-label classification techniques to categorize various types of cyberbullying, such as

sexuality, intelligence, and race, using YouTube comments as the dataset, achieving notable accuracy through supervised learning approaches [1]. Nahar et al. introduced a data filtering technique coupled with SVM classifiers, demonstrating improved precision in the detection of threatening messages within social platforms [2]. Zhang et al. proposed a lexical syntactic feature-based model and showed that considering sentence structure in addition to vocabulary significantly improves bullying identification [3]. Zhao et al. explored the potential of deep learning, particularly CNN and LSTM architectures, in understanding complex sentence semantics, yielding better detection of implicit bullying content [4]. Badjatiya et al. incorporated word embeddings with gradient boosting and deep learning classifiers, achieving high accuracy on Twitter-based harassment datasets [5]. Potha and Maragoudakis used recurrent neural networks to develop personalized bullying detection models, highlighting the benefit of user-specific behavioral patterns in improving classification performance [6]. Rosa et al. applied BERT for hate speech and offensive language detection, showing substantial gains in F1-score compared to previous models [7]. Mishra et al. examined how transformer- based models can detect nuanced, indirect bullying statements by leveraging attention mechanisms [8]. Singh et al. proposed a hybrid deep learning framework integrating CNN and BiLSTM for cyberbullying classification, demonstrating state-of-the-art performance across multiple datasets [9].
Finally, Hosseinmardi et al. combined social graph analysis with textual modeling to improve detection, arguing that peer connections can offer additional cues for early intervention [10].

III. MODULE-WISE DESCRIPTION

The Cyberbullying Detection System is structured into seven interconnected modules, each designed to process, classify, and respond to user-generated content in a simulated social network environment. These modules collectively enable real-time detection of cyberbullying, dynamic user feedback, and automated moderation actions, ensuring a safer online interaction space. The architecture supports flexibility in integrating various datasets, machine learning models, and evolving classification logic based on language trends and abuse patterns. The core modules of the system are described below:

1. Social Network Web App

The Social Network Web App serves as the foundational interface for user interactions, post submissions, and real-time cyberbullying monitoring. It replicates core functionalities of a typical social platform, such as account creation, post feeds, and user engagement. What makes it distinct is its embedded AI-based moderation engine, which monitors user content dynamically. The app is built with scalability and responsiveness in mind, leveraging modern web development frameworks such as React for the frontend and Node.js/Django for the backend.

The platform ensures a safe environment by actively monitoring text submissions, classifying posts, and managing user actions based on the severity of detected cyberbullying. Each post submitted by users is sent to a backend API where it undergoes preprocessing and prediction using a trained cyberbullying detection model. Depending on the output, the system can trigger warnings, display bullying level indicators, or automatically block offending users. Users have access to a personalized dashboard where they can view flagged content, receive warnings, and manage their activity. Admins can access an analytics dashboard showing platform- level statistics, flagged users, and behavioral trends.

This integration of intelligent backend services with an intuitive web interface ensures real-time interaction, proactive moderation, and seamless deployment of the detection pipeline.

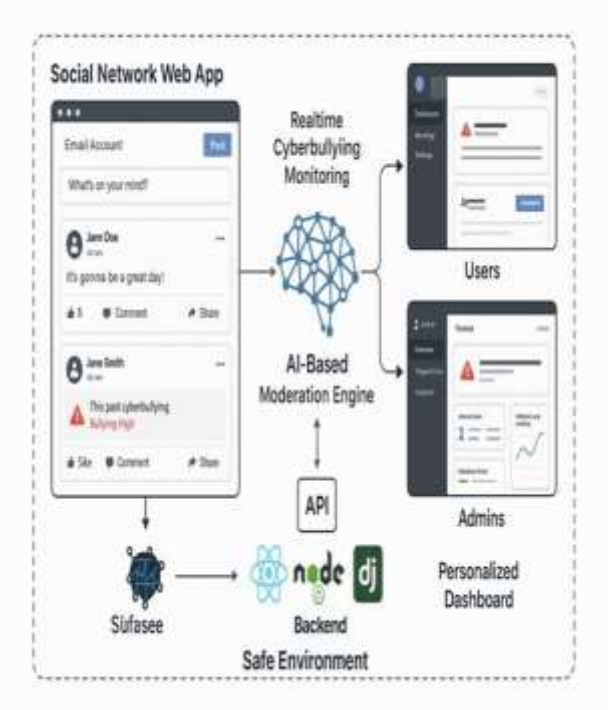

Fig- Social Network Web App

2. User Access Control

The User Access Control module plays a pivotal role in defining and managing permissions within the Cyberbullying Detection System. It segregates access based on roles—primarily into two categories: Social Network (SN) Users and Super Admins. SN Users are regular users who interact on the platform by posting content and engaging with others, whereas Super Admins have overarching control over all platform functionalities, including user management, monitoring system logs, and modifying detection thresholds.

This module ensures secure and authenticated access to system resources. User credentials are verified during login, and session management is maintained to avoid unauthorized actions.

Role-based access control (RBAC) mechanisms are implemented to prevent regular users from accessing administrative tools or interfering with the classification engine. Additionally, SN Users are restricted from viewing internal classification outcomes or manipulating the moderation process.

The module incorporates registration and authentication layers, encryption of credentials, and timeout mechanisms for inactive sessions. It also logs access activity for future audits. Super Admins are empowered to manage user reports, review blocked accounts, and manually override classification decisions in edge cases.

Through effective user access control, the system maintains integrity, enforces accountability, and creates a secure interaction environment. By clearly distinguishing between user roles and responsibilities, the system can better manage sensitive tasks like data handling, moderation, and system updates, ensuring seamless and ethical platform governance.

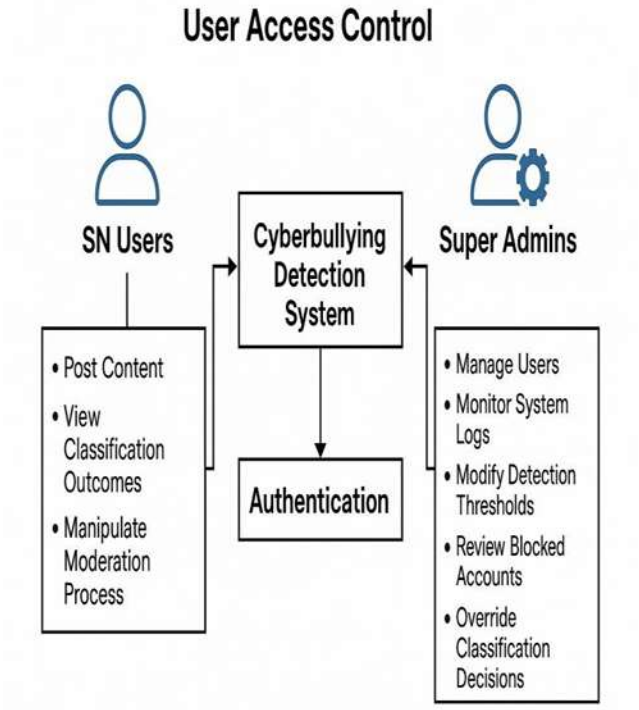

Fig- User Access Control

3. Cyberbullying Classification API (Build and Train)

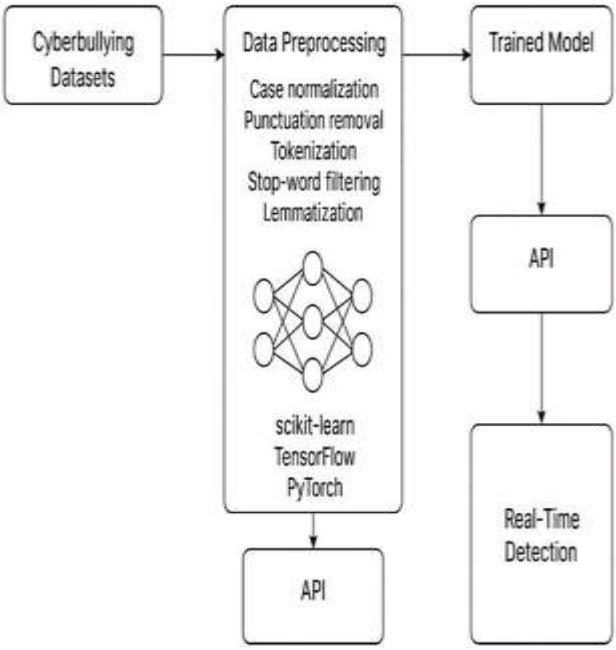

Fig- Cyberbullying Classification API

The Cyberbullying Classification API: Build and Train module is the backbone of the detection engine. It handles the entire machine learning lifecycle, from ingesting and cleaning datasets to training and exporting a deployable prediction model. This API is built using modern machine learning frameworks like Scikit-learn, TensorFlow, or PyTorch and is designed for high modularity and scalability. The process begins with importing labeled cyberbullying datasets sourced from public repositories or curated data collections. These datasets typically consist of text posts annotated as bullying or non-bullying, and sometimes even include severity levels. Once imported, the data is pre-processed through steps such as case normalization, punctuation removal, tokenization, stop-word filtering, and lemmatization to prepare for vectorization. Once trained, the model is saved and integrated into the real-time detection pipeline through an API interface. This modular design allows the classification logic to be updated independently of the user interface.

4. Cyberbullying Detection

The Cyberbullying Detection module is responsible for analyzing and classifying every user-generated post in real time. This is the system's operational core, as it acts at the intersection of the user interface and the backend classification engine. As soon as a user submits a post, the text is intercepted by this module and forwarded to the Classification API for preprocessing and analysis.

Using the trained model, the system classifies content as either bullying or non-bullying based on learned linguistic features, semantic patterns, and context embeddings. The detection pipeline ensures a low-latency response to maintain seamless user experience. The system architecture supports batch detection for multiple posts and single-shot classification for real-time submission monitoring.

In addition to basic detection, this module integrates with the level indicator and notification components, triggering warnings or blocking actions when thresholds are exceeded. Overall, this module enables accurate, responsive, and ethical identification of cyberbullying, helping create a safer digital environment.

5. Cyberbullying Level Indicator

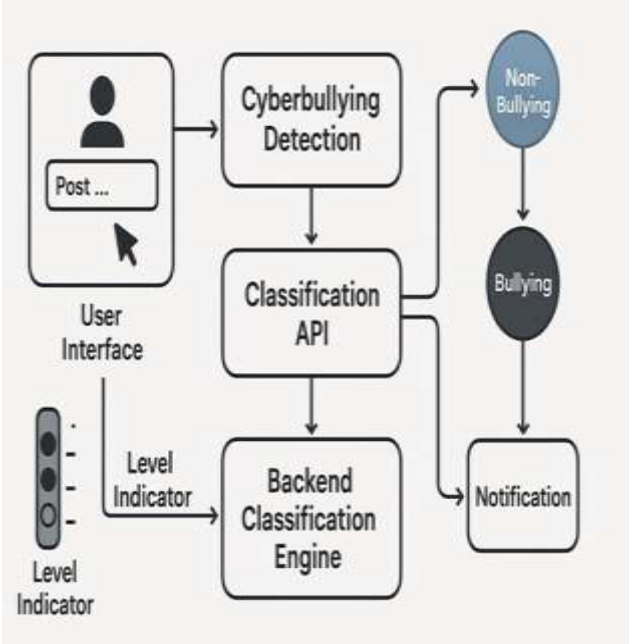

Fig- Cyberbullying Level Indicator

The Cyberbullying Level Indicator module enhances the granularity of the classification output by assigning severity levels to detected bullying content. Instead of treating cyberbullying as a binary classification problem, this module uses prediction confidence and linguistic intensity to determine whether the post falls into categories like low, medium, or high severity.

This is achieved by applying thresholds to the model's probability outputs or using an auxiliary classifier trained specifically for severity estimation. For example, a message classified as bullying with a confidence of 95% may be tagged as high severity, while one with 60% may be considered moderate. Additionally, the system examines keywords, tone, and sentence structure using sentiment analysis and aggression scoring techniques. By integrating severity analysis, the system moves beyond basic detection to offer a nuanced, fair, and informative moderation approach.

6. Warning Notification

The Warning Notification module provides immediate user feedback when potential cyberbullying is detected. It functions as a behavioral intervention tool, aiming to inform, educate, and deter users from posting harmful content. As soon as the system classifies a post as bullying—especially at low or medium severity—a tailored warning message is generated and shown to the user.

The Warning Notification module plays a critical role in promoting responsible online behavior, reducing harm, and maintaining a supportive community atmosphere. It empowers users to self-correct, creating a proactive moderation loop that does not rely solely on punitive measures.

7. Blocker

The Blocker module enforces content moderation by restricting access for users who repeatedly violate platform policies through severe or persistent cyberbullying. It is an automated system that tracks user infractions and applies escalation logic to determine when a temporary or permanent block should be imposed. This may involve disabling the ability to post new content, hiding previous posts, or locking the user account entirely. The duration of the block can be static (e.g., 24 hours) or dynamic based on the severity and frequency of the offenses.

Admins are notified whenever a user is blocked and can override the decision, review logs, or reinstate accounts. This ensures human oversight in sensitive cases. The module also includes an appeal interface where users can request a review of their block status, enhancing transparency and fairness.

By integrating with other modules such as detection and severity scoring, the Blocker ensures that harmful behavior is appropriately contained without delay. It is essential for maintaining community standards, deterring repeat offenses, and signaling that cyberbullying will not be tolerated.

V. CONCLUSION

The automated detection of online harassment is a critical approach to safeguarding users from the harmful effects of digital abuse. As the use of online platforms continues to rise, so does the prevalence of cyberbullying and other forms of harassment, making it increasingly necessary to implement efficient solutions to address these issues. Artificial intelligence (AI)-driven systems can effectively monitor, identify, and assess abusive language, threats, and harmful behaviors in real time. By doing so, these systems provide a proactive and timely response that significantly reduces the impact of online harassment. The integration of automated detection systems in online platforms helps to create a safer digital environment for users.

However, while automated detection offers a powerful tool for combating online harassment, it must be implemented with consideration for privacy, accuracy, and fairness. Continuous

advancements in AI technology and user-centered policies are essential to ensure that these systems remain effective, non-intrusive, and transparent. In conclusion, automated detection is a crucial component in the broader effort to create a safer and more supportive digital space for all users.

REFRENCES

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, 11(1), 11–17.
2. Nahar, V., Al-Maskari, S., & Li, X. (2014). Detecting cyber bullying in social networks using machine learning with data filtering. Social Network Analysis and Mining, 4, 1–15.
3. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution- GRU based deep neural network. European Semantic Web Conference. Springer, Cham.
4. Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. Proceedings of the 17th International Conference on Distributed Computing and Networking, ACM.
5. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. Proceedings of the 26th International Conference on World Wide Web Companion, 759– 760.
6. Potha, N., & Maragoudakis, M. (2014). Cyberbullying detection using time series modeling. 20th Pan-Hellenic Conference on Informatics, 1–6.
7. Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., & Oliveira, H. G. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior Reports, 1, 100005.
8. Mishra, P., Yannakoudakis, H., & Shutova, E. (2019). Tackling online abuse: A survey of automated abuse detection methods. arXiv preprint arXiv:1908.06024.
9. Singh, V., Bharti, S. K., & Singh, S. (2021). A hybrid deep learning model for detecting cyberbullying in social media. Procedia Computer Science, 173, 141– 149.
10. Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the Instagram social network. Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 244– 251.104333 vol.15, no.18, pp. 25–33, , doi: 10.14416/j.asep.2024.05.005.
11. Damodharan, D., Rajesh Kumar, B., Gopal, K., De Poures, M. V., and Sethuramasamyraja, B. (2020). Utilization of waste plastic oil in diesel engines: a review. Reviews in Environmental Science and Bio/Technology, 18(4), 681-697 vol.15, no.14, pp. 1–13, , doi: 10.14416/j.asep.2020.010.006.
12. Hora, S. K., Poongodan, R., De Prado, R. P., Wozniak, M., and Divakarachari, P. B. (2021). Long short-term memory network-based metaheuristic for effective electric energy consumption prediction. Applied Sciences, 11(23), 11263