Journal of Embedded Intelligence and Vision Systems

May 2025, Volume 2, Issue 5

Research on Detecting of Cyberbullying Using Machine Learning Networking

[1] Prof. P.M.Walchale, [2] Mr. Akash Santosh Paraskar, [3] Mr. Vishal Rajaram Wakudkar [4]. Vrushali Pramodrao Raut.

[1] Assistant Professor, Dept. Computer Science And Engineering, Sanmati Engineering College Washim, Maharashtra. , India

[2.3,4] UG Scholar, Computer Science And Engineering, Sanmati Engineering College Washim, Maharashtra. , India

Copyright © Journal of Embedded Intelligence and Vision Systems by Chief Editor is licensed under CC BY-NC-SA 4.0

Abstract- *In the digital communication era, the use of social media platforms has skyrocketed, providing both opportunities for connection and threats like cyberbullying and hate speech. "Detecting Cyberbullying Using Machine Learning" is a paper aimed at developing a Flask-based social media platform integrated with intelligent content moderation features. The system allows users to register, post text and images, and interact with others while ensuring that all content undergoes real-time scanning for abusive language and inappropriate visuals using machine learning techniques. By employing automated abusive word filtering, NSFW image detection, a community reporting mechanism, and an admin review system, the paper ensures that users experience a clean, safe, and respectful digital environment. This solution demonstrates how machine learning can be applied to tackle real-world issues and promote safer online communities.*

Index-Cyberbullying Detection, Machine Learning, FlaskFramework.

I. INTRODUCTION

The expansion of social media platforms has enabled millions of users worldwide to share ideas, communicate, and build communities. However, this technological growth has also brought challenges, most notably the rise of cyberbullying and the spread of abusive content. Unchecked, such issues can lead to psychological harm, distress, and unsafe digital spaces. Therefore, creating a proactive solution to monitor and moderate content becomes crucial.

In this paper, a Flask-based social media web application is developed with integrated machine learning models to detect and filter abusive and harmful content automatically. The platform provides a user-friendly interface where users can log in, post, like, comment, and interact securely. Simultaneously, it uses abusive word filters, hate speech classification models, and NSFW image detection to ensure that the communication remains respectful and free of toxicity. Through user reporting features and an admin panel for moderation, the paper ensures a dynamic, constantly updated safe space for digital interactions.

II. LITERATURE REVIEW

Islam, Md Manowarul, et al. designed and developed an effective technique to detect online abusive and bullying messages by merging natural language processing with machine learning. The study utilized two distinct features—Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF)—to evaluate the accuracy of four different machine learning algorithms in identifying cyberbullying content. The comparative analysis demonstrated the strengths of these feature extraction methods in enhancing classifier performance.

Neelakandan, S., M. Sridevi, et al. proposed the FSSDL-CBDC technique, which integrates preprocessing, feature selection, and classification to detect cyberbullying. Their method employs a Binary Coyote Optimization-based Feature Subset Selection (BCO-FSS) approach to select the most relevant features, enhancing classification accuracy. Additionally, they introduced a hybrid model combining the Salp Swarm Algorithm with a Deep Belief Network (SSA-DBN), effectively identifying and characterizing cyberbullying activities in online environments. The originality and superior classification performance of their proposed framework were validated through extensive simulations.

Hani, John, et al. presented a supervised machine learning framework for the detection and prevention of cyberbullying. Multiple classifiers were trained on a cyberbullying dataset, revealing that Neural Networks (NN) achieved the highest performance with an accuracy of 92.8%, followed by Support Vector Machines (SVM) at 90.3%. Their results indicate that NN outperforms other classifiers used in similar studies on the same dataset.

Muneer, Amgad, et al. compiled a comprehensive global dataset comprising 37,373 unique tweets from Twitter to analyze cyberbullying. They employed seven machine learning algorithms—Logistic Regression (LR), LightGBM, Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB), and SVM—and evaluated them using accuracy, precision, recall, and F1 score. Among these, Logistic Regression demonstrated the best overall performance with a median accuracy of 90.57% and the highest F1 score of 0.928. SGD achieved the best precision, and SVM excelled in recall.

Kargutkar, Saloni Mahesh, et al. proposed a unique approach using Convolutional Neural Networks (CNN) to perform dual characterization of cyberbullying content. Unlike previous naïve classification techniques, their method showed improved accuracy and precision by leveraging CNNs for deeper content analysis. Their experimental evaluation using a public dataset revealed significant performance improvements over existing models.

Azeez, Nureni Ayofe, et al. reviewed a broad spectrum of classification algorithms and introduced an ensemble model for cyberbullying detection using Twitter datasets. Their experimentation involved classifiers such as Naive Bayes, K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Linear SVC, AdaBoost, SGD, and Bagging. Performance was evaluated through four key metrics: accuracy, precision, recall, and F1 score. Their comparative analysis provided insights into the strengths and weaknesses of each classifier in detecting online abuse.

Sultan, Daniyar, et al. conducted a comprehensive literature review on cyberbullying detection, focusing on machine learning and natural language processing frameworks. Their study outlined the typical workflow in cyberbullying detection: data collection, preprocessing, feature selection and extraction, and classification. This review offers a consolidated perspective on current techniques and serves as a guide for developing more practical and consistent future solutions.

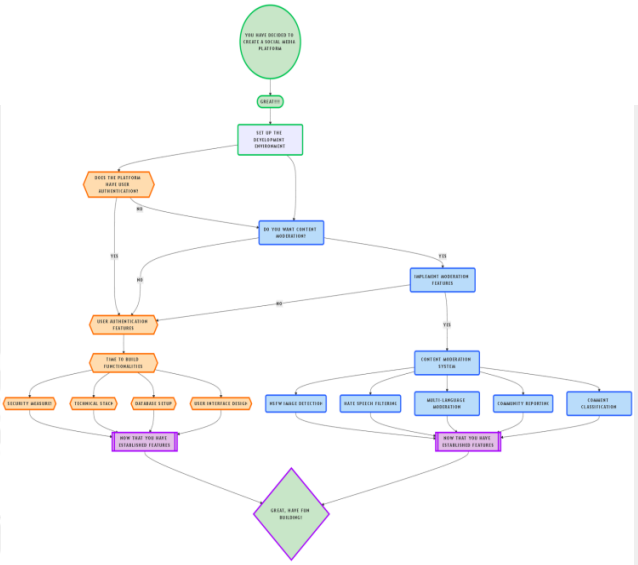
Iwendi, Celestine, et al. carried out an empirical analysis to evaluate the effectiveness of deep learning models in detecting offensive content in social commentary. The models tested included BLSTM, GRU, LSTM, and RNN. They followed thorough preprocessing steps, including cleaning, tokenization, stemming, lemmatization, and stop word removal. Their findings demonstrated the superior performance of deep learning algorithms in accurately identifying cyberbullying content.

Raj, Mitushi, et al. introduced a deep learning-based framework capable of analyzing real-time Twitter tweets and social media posts to detect cyberbullying content. Their model proved more effective than traditional approaches, especially in handling multilingual content such as English, Hindi, and Hinglish. This work highlights the growing importance and success of deep learning techniques in tackling the complexities of cyberbullying detection.

III. RESEARCH METHODOLOGY

The proposed system is a comprehensive Flask-based social media application that integrates powerful machine learning-driven content moderation to combat cyberbullying effectively. The system will begin by allowing users to register and log in securely using bcrypt encryption for password protection and Flask-Login for session management. After successful login, users are taken to a dynamic dashboard where they can create and view posts, similar to any social media platform. When a user submits a post or blog, the system automatically scans the text using a predefined abusive word filter along with a machine learning comment classifier to detect hate speech or bullying language. Similarly, images uploaded by users are processed through a TensorFlow-based NSFW image detection model to check for inappropriate content. If any abusive or harmful content is detected either in text or images, the post is either blocked automatically or flagged for admin review. The system also empowers users to report posts that they find offensive, triggering an alert to the admin panel for further action. An admin panel is available to manage moderation rules, update blocked words, and monitor reported content. Thus, the entire platform works step-by-step from secure user login, to safe content posting, to intelligent and proactive content moderation, ensuring a clean, positive, and respectful digital environment.

FLOW CHART



WORKING

The system starts when a new user accesses the web application and registers with a username, email, and password, all securely handled through bcrypt encryption. Once logged in, users are redirected to a social-media-like dashboard where they can create posts comprising text and images. As soon as a post is submitted, the backend instantly analyzes the content: text inputs are passed through a multi-layered abusive word detection mechanism and a machine learning-based comment classification model to identify any potential cyberbullying or hate speech, while uploaded images are automatically checked by a TensorFlow-powered NSFW classifier to detect and block inappropriate visuals. If a post is identified as harmful, the system either removes it or hides it from public view while simultaneously notifying the admin. Users also have the ability to manually report any posts they find offensive, ensuring community participation in content moderation. An administrator can review reported posts through a separate admin interface and modify moderation policies as needed. Throughout the process, user sessions, post data, and reports are securely stored and managed using SQLAlchemy ORM with SQLite, maintaining performance, accuracy, and a secure, respectful user environment.

IV. SYSTEM REQUIREMENT

1. **Frontend:** HTML5, CSS3, JavaScript, Bootstrap
2. **Backend:** Python (Flask Framework), SQLAlchemy, Flask-Login, TensorFlow, Bcrypt
3. **Database:** SQLite

V. EXPERIMENTAL SETUP & RESULT

IMPLEMENTATION

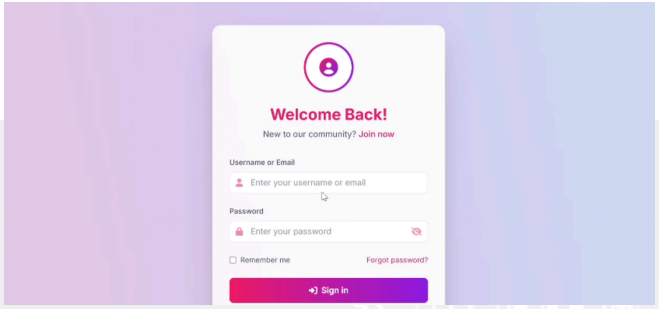


Fig. Shows User Login Interface

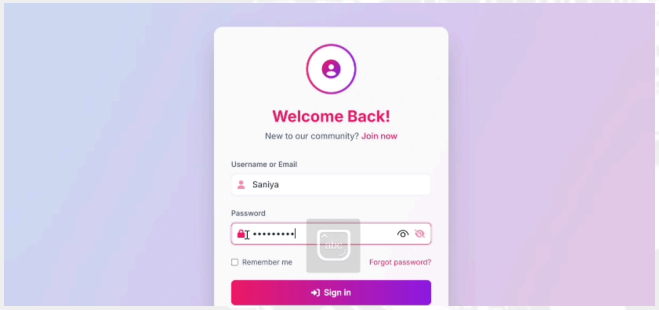


Fig. Shows Users Can Log In Using their Username or Email and Password

Step 1: User Dashboard and Post Creation Interface

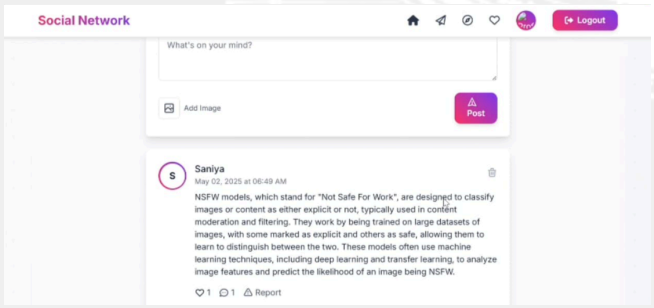


Fig. Shows User Dashboard Displaying the Post Creation Area

This figure shows the user's dashboard after successful login. At the top, a text input field labeled "What's on your mind?" allows users to create new posts. Below this, an "Add Image" button enables users to upload images along with their text content. The prominent "Post" button facilitates the submission of user-generated content, which will then be processed by the system's cyberbullying detection mechanisms.

Step 2: Content Display and User Interaction

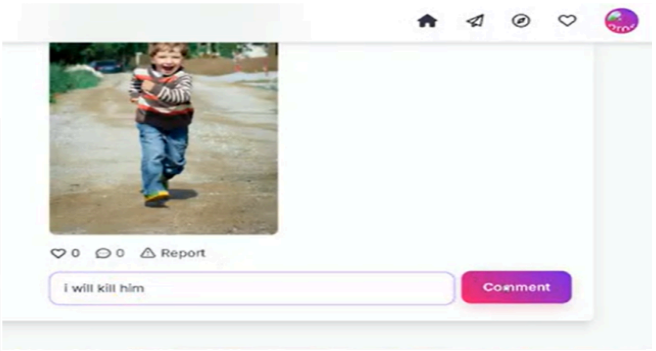


Fig. Shows Displayed Post With User Comment and Reporting Option

This figure showcases a user-generated post containing an image and a comment. Below the post, users can see options for liking, commenting, and reporting the content. In this instance, a potentially harmful comment, "I will kill him," has been posted.

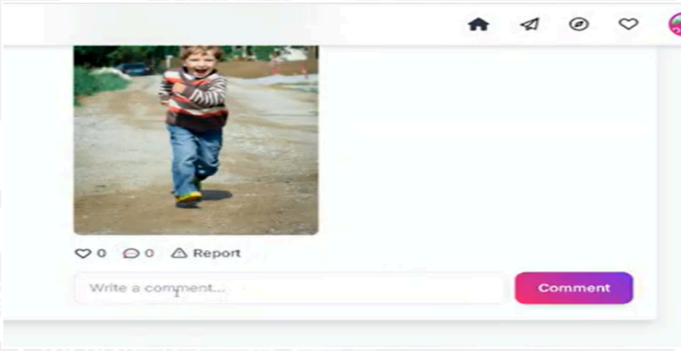


Fig. Shows Post View After the Removal of the Offensive Comment

This figure displays the same post after the offensive comment ("I will kill him") has been removed. This action is a result of the proposed system's automatic content moderation capabilities or following administrative action on a user report. This step demonstrates the importance of identifying and promptly removing harmful content to maintain a safe and respectful online environment.

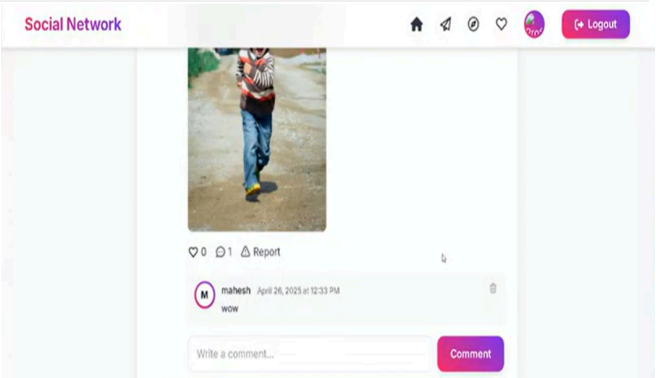


Fig. Shows Display of a User Post with a Positive Comment

This figure showcases a user-generated post with a positive comment, "wow." The comment is displayed below the image, indicating that it has passed the system's content moderation filters. This demonstrates the system's ability to allow positive and appropriate interactions, fostering a healthy and engaging environment for users. The display of the comment alongside the like and report options allows other users to interact with the content and ensures that even seemingly innocuous comments can be reported if deemed necessary.

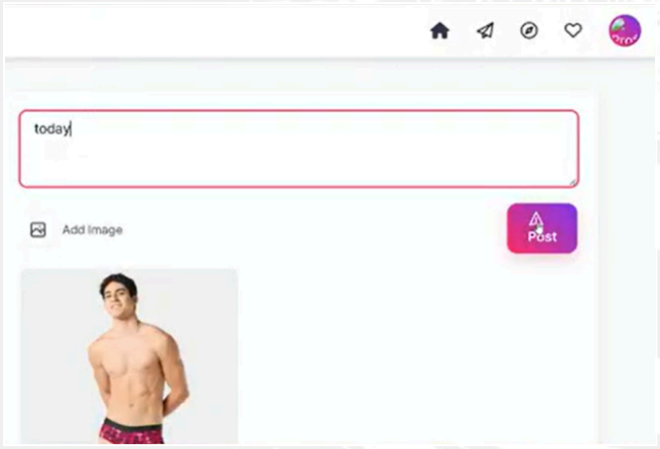


Fig. Shows an Attempt to Post Inappropriate Image Content

During the content posting process, if the uploaded image is identified as inappropriate by the NSFW detection model, or if the accompanying text triggers the abusive word filter or the machine learning classifier, the system will take immediate action. This action involves either automatically blocking the post from being published, thus preventing the inappropriate content from becoming visible to other users, or flagging the post for administrative review. Furthermore, the system allows for user-driven moderation, enabling users to report content they deem offensive. Such reports generate an alert within the administrator panel, prompting further investigation and appropriate moderation actions to maintain a safe and respectful platform environment.

RESULT

The result of this paper demonstrates the effectiveness of using machine learning and Flask-based web development to create a secure and proactive platform for detecting and preventing cyberbullying in an online social media environment. The system successfully integrates user authentication, content creation, and advanced moderation features, such as real-time text analysis for abusive language and NSFW image detection. The machine learning models, powered by TensorFlow, efficiently flag harmful content, ensuring that posts are automatically screened before being published or flagged for admin review. Additionally, users are empowered to report offensive content, contributing to a community-driven approach to moderation. Ultimately, the paper shows how combining machine learning, secure backend technologies, and intuitive frontend design can provide a cleaner, safer, and more respectful digital space for users, effectively combating cyberbullying while promoting positive interactions.

VI. CONCLUSION

The "Detecting of Cyberbullying Using Machine Learning Networking or Cyber Security" paper offers a robust and intelligent solution to combat the ever-growing problem of online abuse. By harnessing the power of machine learning models, Flask-based development, and user-driven moderation features, this platform successfully creates a safer online environment where users can express themselves freely while being protected against harmful content. Through automated moderation, multi-language support, and dynamic updating capabilities, this paper not only addresses today's challenges but also lays the groundwork for future innovations in digital safety and community engagement.

VII. REFERENCE

[1] Kumar, Rupesh, Shreyas Parakh, and CNS Vinoth Kumar. "Detection of cyberbullying using machine learning." Turkish Journal of Computer and Mathematics Education 12, no. 9 (2021): 656-660.

[2] Islam, Md Manowarul, Md Ashraf Uddin, Linta Islam, Arnisha Akter, Selina Sharmin, and Uzzal Kumar Acharjee. "Cyberbullying detection on social networks using machine learning approaches." In 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1-6. IEEE, 2020.

[3] Neelakandan, S., M. Sridevi, Saravanan Chandrasekaran, K. Murugeswari, Aditya Kumar Singh Pundir, R. Sridevi, and T. Bheema Lingaiah. "Deep learning approaches for cyberbullying detection and classification on social media." Computational Intelligence and Neuroscience 2022 (2022): 2163458.

[4] Hani, John, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, and Ammar Mohammed. "Social media cyberbullying detection using machine learning." International Journal of Advanced Computer Science and Applications 10, no. 5 (2019): 703-707.

[5] Muneer, Amgad, and Suliman Mohamed Fati. "A comparative analysis of machine learning techniques for cyberbullying detection on twitter." Future Internet 12, no. 11 (2020): 187.

[6] Kargutkar, Saloni Mahesh, and Vidya Chitre. "A study of cyberbullying detection using machine learning techniques." In 2020 fourth international conference on computing methodologies and communication (ICCMC), pp. 734-739. IEEE, 2020.

[7] Azeez, Nureni Ayofe, Sunday O. Idiakose, Chinazo Juliet Onyema, and Charles Van Der Vyver. "Cyberbullying detection in social networks: Artificial intelligence approach." Journal of Cyber Security and Mobility (2021): 745-774.

[8] Sultan, Daniyar, Batyrkhan Omarov, Zhazira Kozhamkulova, Gulnur Kazbekova, Laura Alimzhanova, Aigul Dautbayeva, Yernar Zholdassov, and Rustam Abdrakhmanov. "A Review of Machine Learning Techniques in Cyberbullying Detection." Computers, Materials & Continua 74, no. 3 (2023).

[9] Iwendi, Celestine, Gautam Srivastava, Suleman Khan, and Praveen Kumar Reddy Maddikunta. "Cyberbullying detection solutions based on deep learning architectures." Multimedia Systems 29, no. 3 (2023): 1839-1852.

[10] Raj, Mitushi, Samridhi Singh, Kanishka Solanki, and Ramani Selvanambi. "An application to detect cyberbullying using machine learning and deep learning techniques." SN computer science 3, no. 5 (2022): 401.