Identifying Offensive User Posts in Online Platforms Through Machine Learning

Y. Shakuntala[1], *Mohammed Saif Rehan*[2], Mohammed Aftab Talha[3], *Abdullah Muneeb Siddiqui*[4]
[1]Assistant Professor, Department of CSE (Data Science),
Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.
[2,3,4]UG Students, Department of CSE (Data Science),
Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.

***Abstract*— The increasing use of social media platforms has created new ways for people to communicate, but it has also led to the rise of cyberbullying, which affects users emotionally and psychologically. This project focuses on developing a machine learning-based system to detect cyberbullying in user-generated messages. The system is designed to analyze text posts and classify them as offensive or non-offensive using algorithms such as AdaBoost, Multinomial Naïve Bayes, and Stochastic Gradient Descent (SGD). In addition to detecting abusive content, a Support Vector Machine (SVM) model is used to determine the sentiment of the message as positive or negative. The application provides an interactive platform where users can register, log in, and post messages along with images. Each message is automatically evaluated, and the results are displayed to the user. If offensive language is detected, the system increases the user's offense count, which is visible to the administrator. Based on this count, the admin has the authority to block users who repeatedly violate the rules. The system uses a MySQL database to store user information, posts, and analysis results. This approach reduces the need for manual monitoring and helps in maintaining a safer online environment. The performance of the models is evaluated using accuracy and other metrics, showing that machine learning can effectively identify harmful content. The system serves as a useful tool for managing online interactions responsibly.**

***Keywords*—Cyberbullying detection, social media analysis, machine learning, text classification, Stochastic Gradient Descent (SGD), AdaBoost, Multinomial Naïve Bayes, Support Vector Machine (SVM), sentiment analysis, offensive content detection, data preprocessing, model evaluation metrics, web-based application, MySQL database, user behavior monitoring**

I. INTRODUCTION

Social media has become a major part of everyday life, allowing people to communicate, share ideas, and connect with others easily. Platforms such as messaging applications and networking sites have made communication faster and more accessible. However, along with these benefits, there has been a rise in negative behaviors, especially cyberbullying [6], [10]. Cyberbullying refers to the use of digital platforms to send harmful or abusive messages to others. It can include insulting comments, threats, or spreading false information. Unlike traditional bullying, cyberbullying can happen at any time and can reach a wide audience, making its impact more serious [4]. Victims often feel stress, fear, and emotional pain, which can affect their mental well-being. As the number of users and the volume of online content continue to grow, it becomes difficult to manually monitor all interactions. Human moderators cannot review every message in real time, which allows harmful content to spread quickly. This creates a need for automated systems that can detect and control such behavior. Machine learning provides a practical solution by analyzing patterns in text and identifying abusive language [6]. By using trained models, it is possible to classify messages based on their content. This helps in reducing harmful interactions and improving the safety of online platforms. Therefore, developing an intelligent system for detecting cyberbullying has become an important area of research.

With the advancement of machine learning techniques, analyzing text data has become more accurate and efficient. Earlier methods mainly depended on simple keyword matching, which often failed to understand the actual meaning of a sentence. For example, a message may contain a word that seems offensive but is used in a different

context. Machine learning overcomes this limitation by learning from large datasets and identifying patterns based on context [11], [12]. Algorithms such as Multinomial Naïve Bayes are widely used for text classification because they are simple and perform well with large amounts of data. AdaBoost improves prediction accuracy by combining multiple weak models into a stronger one, while Stochastic Gradient Descent (SGD) helps in optimizing model performance during training. In addition to detecting offensive content, understanding the sentiment of a message is also important. Sentiment analysis helps in identifying whether a message expresses positive or negative emotions. Support Vector Machine (SVM) is commonly used for this purpose due to its ability to classify data effectively. By combining these techniques, a system can perform both cyberbullying detection and sentiment analysis at the same time [5], [7]. This provides a better understanding of user behavior and improves the overall performance of the system. Such an approach reduces manual effort and ensures faster detection of harmful content in social media applications.

The system developed in this project aims to provide a complete solution for detecting cyberbullying through a user-friendly web application. Users can create accounts, upload profile pictures, and post messages, which are then analyzed by the system. Each message is checked using trained machine learning models to determine whether it is offensive or not. At the same time, the sentiment of the message is also identified as positive or negative. The results are displayed immediately, allowing users to understand the nature of their posts. One important feature of the system is the tracking of offensive behavior. Every time a user posts a harmful message, the system increases their offense count. This information is stored in the database and can be viewed by the administrator. If a user repeatedly posts offensive content, the admin can block the account to prevent further misuse. This helps in maintaining discipline and promoting respectful communication. The use of a MySQL database ensures proper management of user data and system records. Although the system performs well, its accuracy depends on the dataset used for training [9]. If new types of abusive language are not included in the dataset, the system may not detect them correctly. Future improvements can focus on using larger and more diverse datasets to enhance performance. Overall, this system provides an effective way to control cyberbullying in online platforms.

II. RELATED WORK

Reynolds et al., (2011) [6] Reynolds et al. explored the use of machine learning techniques to detect cyberbullying in online communication. Their study focused on analyzing textual data from social media platforms to identify harmful content. They applied classification algorithms to distinguish between normal and abusive messages. The research highlighted the challenges involved in detecting cyberbullying due to variations in language and context. Their results showed that machine learning methods can effectively improve detection accuracy compared to traditional approaches. The authors also emphasized the importance of feature selection in improving model performance. This work contributed to the early development of automated cyberbullying detection systems. It provided a base for further research in applying advanced algorithms for online safety. The study demonstrated the practical use of data-driven techniques in identifying harmful behavior.

Zhao et al., (2016) [4] Zhao et al. proposed an approach for automatically detecting cyberbullying on social networks using specific bullying-related features. Their research focused on identifying patterns in user messages that indicate aggressive or harmful intent. They extracted linguistic and behavioral features from social media data to improve classification performance. The study showed that combining multiple features enhances the accuracy of detection systems. The authors also discussed the importance of understanding user interactions in identifying cyberbullying. Their findings indicated that feature-based models can effectively capture hidden patterns in text data. This research contributed to improving automated detection methods by focusing on meaningful attributes. It also highlighted the need for more advanced techniques to handle complex language structures. The study provided valuable insights for building efficient cyberbullying detection systems.

Banerjee et al., (2019) [5] Banerjee et al. introduced a deep learning-based approach for detecting cyberbullying in online platforms. Their study utilized neural network models to analyze textual data and identify abusive content. The authors demonstrated that deep learning methods can capture complex patterns in language more effectively than traditional models. They trained their model on a labeled dataset and evaluated its performance using standard metrics. The results showed improved accuracy in detecting offensive messages. The study also discussed the advantages of using deep neural networks in handling large datasets. Their work highlighted the growing importance of deep learning in text classification tasks. It provided a strong foundation for future research in advanced cyberbullying detection techniques. The research emphasized the need for

scalable and efficient systems in real-world applications.

Waseem and Hovy, (2016) [11] Waseem and Hovy focused on detecting hate speech on social media by analyzing linguistic and user-based features. Their study examined whether offensive content is better identified through language patterns or user behavior. They developed models that classify tweets into categories such as hateful or non-hateful. The research highlighted the importance of feature engineering in improving classification results. Their findings showed that combining different types of features leads to better performance. The authors also discussed challenges such as bias in datasets and annotation issues. This work contributed significantly to understanding hate speech detection in online environments. It also provided insights into the relationship between language and user intent. The study supports the development of more accurate and fair detection systems.

Yadav et al., (2020) [7] Yadav et al. proposed a cyberbullying detection system using a pre-trained BERT model. Their research focused on leveraging transformer-based architectures to improve text classification accuracy. The model was trained on large-scale data and fine-tuned for cyberbullying detection tasks. The authors demonstrated that BERT can understand the context of words more effectively than traditional models. Their results showed significant improvements in detecting abusive and harmful messages. The study also highlighted the importance of contextual understanding in natural language processing. This research represents a shift towards using advanced deep learning techniques for cyberbullying detection. It provided a modern approach to handling complex language patterns. The study contributes to the development of more intelligent and accurate systems for online content moderation.

III. DATASET DETAILS

The dataset used in this project is related to detecting cyberbullying in social media messages and consists of textual data collected from user posts. It mainly includes message content along with corresponding labels that indicate whether the text is offensive or non-offensive and whether the sentiment is positive or negative. These attributes are important for training machine learning models to understand the nature of user-generated content. During the initial stage, the dataset was reviewed to ensure that the text data was meaningful and suitable for analysis. Basic preprocessing steps such as removing unnecessary symbols and handling missing values were considered to maintain data quality. To better understand the dataset, simple analysis was carried out by observing the distribution of offensive and non-offensive messages. This helped in identifying how the data is balanced and how different types of messages are represented. Such analysis is useful for improving the learning process of the models and ensuring better classification results.

The dataset was then prepared for training the machine learning models used in the system. The text data was converted into a numerical format using appropriate techniques so that it could be processed by the algorithms. The input features, which represent the message content, and the output labels, which indicate the classification results, were clearly separated. The dataset was divided into training and testing sets, where most of the data was used to train the models and the remaining portion was used to evaluate their performance. This approach helps in checking how well the model works on new and unseen data. Different machine learning algorithms such as AdaBoost, Multinomial Naïve Bayes, and SGD were trained using this processed dataset. The quality and structure of the dataset play a key role in determining the accuracy of the system. A well-prepared dataset ensures that the model can correctly identify offensive content and provide reliable results for real-time cyberbullying detection.

IV. PROPOSED METHODOLOGY

The proposed system follows a structured machine learning approach to detect cyberbullying in social media content using the Stochastic Gradient Descent (SGD) algorithm. Initially, the dataset containing user messages is loaded and analyzed to understand its structure, class distribution, and overall quality. Preprocessing steps are applied to clean the text data by removing unwanted characters and converting the text into a suitable format for analysis. The input features, which represent the message content, and the target labels, indicating whether a message is offensive or non-offensive, are clearly separated. The processed dataset is then divided into training and testing sets to ensure that the model can be evaluated on unseen data. This step helps in building a reliable system that performs well in real-time scenarios.

The SGD classifier is used as the main model for detecting cyberbullying due to its efficiency in handling large-scale text data and its ability to perform well with linear classification problems.

The model is trained using the prepared dataset to classify messages as offensive or non-offensive. During training, SGD updates model parameters iteratively, which helps in achieving faster convergence and improved performance. The effectiveness of the model is evaluated using performance metrics such as accuracy and other relevant measures. The trained model is then integrated into a web-based application where users can register, log in, and post messages. Each message is analyzed instantly, and the result is displayed to the user. The system also maintains a count of offensive messages for each user, allowing the administrator to monitor user behavior and take appropriate action, such as blocking accounts when necessary. This approach provides a simple and efficient solution for detecting and controlling cyberbullying in online platforms.

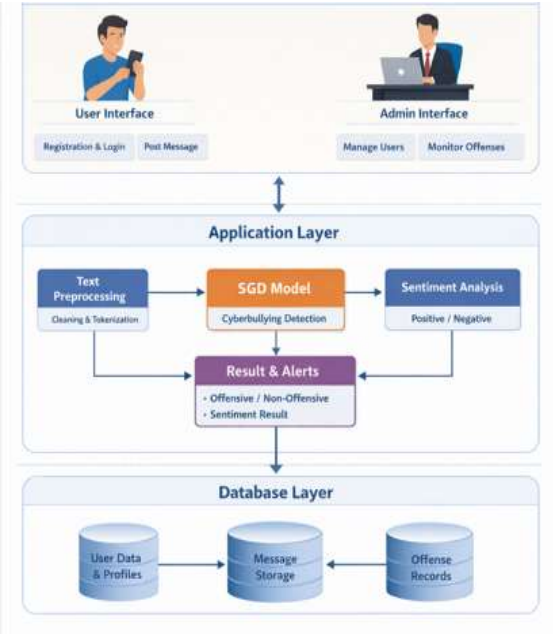

Figure [1]: System Architecture of cyberbullying detection

Figure [1] This diagram shows the working process of the cyberbullying detection system. It begins with users accessing the application through the web interface, where they can register, log in, and post messages. The entered message is then processed and analysed using the Stochastic Gradient Descent (SGD) model to check whether it is offensive or not. The system also identifies the sentiment of the message. All details such as user information, posted messages, and offensive counts are stored in the database. The output is displayed immediately to the user, and the admin can view user activities and take actions like blocking accounts when needed.

V.RESULT AND DISCUSSION

The experimental results show that the machine learning model used in this system is effective in identifying cyberbullying in user messages. The Stochastic Gradient Descent (SGD) algorithm was trained and tested using the prepared dataset, and it produced reliable classification results for both offensive and non-offensive messages. The model was able to correctly identify most of the harmful messages based on the patterns learned during training. The accuracy of the model indicates that it can be used for real-time applications with good performance. Evaluation measures such as accuracy and classification results confirm that the model performs well in detecting abusive content. The system also provides sentiment results, which help in understanding whether the message is positive or negative. From the output images, it can be observed that offensive messages are correctly flagged, and the system updates the offensive count for each user. Non-offensive messages are also correctly classified, showing that the model can distinguish between different types of content.

The results also show that the system works effectively in a real-time environment, where users can post messages and get immediate feedback. The admin module further supports monitoring by displaying users with offensive counts and allowing necessary actions such as blocking. Overall, the results indicate that the proposed system can successfully detect cyberbullying and help in maintaining a safer online platform.

Figure [2]: Cyberbullying Detection System Home Page

Figure [2] shows the home page of the system. It displays the project title and a menu with options like Home, Admin Login, User Login, and

Register. Users can use these options to move to different sections of the application.

Figure [3]: User Registration and Profile Setup Interface

Figure [3] shows the registration page for new users. It contains fields to enter details such as username, password, contact number, email, and address. There is also an option to upload a profile picture before completing the registration.

Algorithm Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AdaBoost	87.060	87.866	78.521	81.590
SGD	97.294	96.632	96.535	96.583
Multinomial naïve Bayes	90.216	86.768	90.532	88.293

Table [1] : Performance Evaluation of cyberbullying

Table [1] The table shows the comparison of different machine learning algorithms used for cyberbullying detection using measures like accuracy, precision, recall, and F1-score. From the results, the SGD algorithm gives the best performance with the highest accuracy of 97.294%. It also shows strong values in precision, recall, and F1-score, which means it can correctly identify both offensive and non-offensive messages effectively.

Figure [4]: User Dashboard and Message Display Screen

Figure [4] shows the user page after login. The user's profile picture is visible along with a welcome message. A table is also shown where posted messages and their details will be displayed.

Figure [5]: Message Posting and Upload Interface

Figure [5] shows the page where a user can post a message. The user can type a message and upload an image. After clicking submit, the message is sent to the system for analysis.

Figure [6]: Offensive Message Detection and Result Output Screen

Figure [6] shows the output after posting a message with offensive words. The system marks the message as offensive and negative. It also shows a warning that the admin may take action. The message details and results are displayed in a table.

Figure [7] : blocking the offensive users

Figure [7] shows the process of blocking a user based on offensive activity. The admin can view the list of users along with their offensive message count. If a user exceeds the allowed limit, the

system enables the block option. Once the admin clicks on it, the user account is blocked and the user will not be able to log in or access the system further.

DISCUSSION

The results shown in the system images explain how the model performs in detecting cyberbullying from user messages. The SGD model gives accurate results in identifying whether a message is offensive or not. From the outputs, it can be seen that harmful messages are correctly marked, while normal messages are also classified properly. This shows that the model is able to handle text data effectively and give correct predictions in most cases. The image also show that proper text preparation helps in improving the results. When the input message is cleaned and processed correctly, the model can better understand the content and give accurate output. Along with detecting offensive content, the system also shows whether the message is positive or negative. This gives more clarity about the nature of the message and helps in better analysis of user behavior. Another important point is the role of the admin in controlling the system. The application keeps a record of how many offensive messages each user has posted. If a user crosses the limit, the admin can block that account. This helps in reducing repeated misuse and keeps the platform safe. Since the system works through a web application, users can easily access it and get instant results. Overall, the system works well in identifying cyberbullying and helps in managing user activity effectively.

VI. CONCLUSION

This project successfully developed an effective system for detecting cyberbullying using machine learning techniques. The SGD algorithm was implemented and evaluated to identify offensive and non-offensive messages from user posts. The model achieved high accuracy, showing its ability to handle text data and provide reliable classification results. Proper preprocessing of the input text played an important role in improving the performance of the model by ensuring clean and meaningful data for training. The system not only detects offensive content but also identifies the sentiment of the message, which helps in better understanding user behavior. Evaluation results confirm that the model performs well in real-time conditions. In addition, the system maintains a record of offensive messages for each user, allowing continuous monitoring of user activity. The integration of the model into a web-based application makes the system easy to use and accessible. Users can post messages and get immediate results, while the admin can monitor users and block accounts if necessary. This helps in maintaining a safe and controlled environment. Overall, the system provides a simple and efficient solution for identifying cyberbullying and supports better management of online interactions.

REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, “Towards the detection of cyberbullying based on social network mining techniques,” Proceedings of the 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), 2017, pp. 1–6, doi: 10.1109/BESC.2017.8256403.

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, “Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying,” International Conference on Intelligent Data Engineering and Automated Learning, 2014, pp. 419–426, doi: 10.1007/978-3-319-01854-6_43.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, “Collaborative detection of cyberbullying behavior in Twitter data,” IEEE International Conference on Electro/Information Technology (EIT), 2015, pp. 611–616, doi: 10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” Proceedings of the 17th International Conference on Distributed Computing and Networking, 2016, pp. 1–6, doi: 10.1145/2833312.2849567.

[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, “Detection of cyberbullying using deep neural network,” IEEE International Conference on Advances in Computing, Communication and Control Systems (ICACCS), 2019, pp. 1–5, doi: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, “Using machine learning to detect cyberbullying,” IEEE 10th International Conference on Machine Learning and Applications (ICMLA), 2011, pp. 241–244, doi: 10.1109/ICMLA.2011.152.

[7] J. Yadav, D. Kumar, and D. Chauhan, “Cyberbullying detection using pre-trained BERT model,” IEEE International Conference on Electrical, Electronics, and Computer Science

(ICECS), 2020, pp. 1–6, doi: 10.1109/ICESC48915.2020.9155700.

[8] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning-based models: A reproducibility study," arXiv preprint arXiv:1804.07154, 2018.

[9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv preprint arXiv:1801.06482, 2018.

[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp. 137–144, doi: 10.1109/ASONAM.2016.7752420.

[11] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," Proceedings of the NAACL Student Research Workshop, 2016, pp. 88–93, doi: 10.18653/v1/N16-2013.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 2017, pp. 512–515.

[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," Proceedings of the 26th International World Wide Web Conference (WWW), 2017, pp. 1391–1399, doi: 10.1145/3038912.3052591.

[14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," Artificial Intelligence Review, vol. 53, no. 6, pp. 4335–4385, 2020, doi: 10.1007/s10462-019-09794-5.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

Babburi, S. Privacy-Preserving Collaborative Framework with Auditable Federated Learning.

Gaddam, S. Integrating Analytics into the Development Process: Bridging the Gap between Data Insights and Design Execution.

Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. Applied Research for Growth, Innovation and Sustainable Impact, 377–384. https://doi.org/10.1201/9781003684657-63

Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.

Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.

Mahimalur, R. K., Vasgam, M., & Manoharan, D. Devops Lifecycle Management And Cloud Migration Assessments: A Security-Driven CICD Perspective.

Viswanathan, V. (2023). AI-Augmented Decision Intelligence for Enterprise Systems: Integrating Cognitive Analytics for Resource and Talent Optimization.

Agrawal, A. M., Gajula, S., Shinde, R. P., Shah, H., & Ghosh, H. (2025, July). Machine Translation for Long Sequences with Enhanced Attention Mechanisms. In 2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-6). IEEE.

Maturi, S. Y. (2021). Blockbond hardening: Securing pooled-hash protocols against traffic tampering, MITM hash-rate hijacking, and template coercion. International Journal of Communication Networks and Information Security, 13(3), 718–728.

Adabala, P. K. (2024). Utilizing predictive analytics to improve efficiency and decision-making in ERP-connected supply chains. International Journal of Intelligent Systems and Applications in Engineering, 12(22s), 2465

Srikanth Kavuri. (2023). Machine Learning Approaches for Security Vulnerability Detection in Software Testing. Computer Fraud and Security. https://doi.org/10.52710/cfs.837

Gajula, S. (2026). Two Pillars of Banking Intelligence: A Comparative Analysis of AI Techniques for Fraud Prevention and Churn Mitigation. 2026 14th International Symposium on Digital Forensics and Security (ISDFS), 1–6. https://doi.org/10.1109/isdfs69419.2026.11458995

Akinapalli, S. (2026). An AI-powered data trust and quality scoring framework for enterprise decision intelligence systems. International Journal

of Data Science and IoT Management System, 5(1), 946–950.

Susarla, R. S., Boyapati, P. K., & Kandula, S. T. R. (2025, July). Cloud-Based Secure Data Storage in Smart Cities Using Central-Smoothing Hypergraph Neural Networks. In 2025 IEEE 4th World Conference on Applied Intelligence and Computing (AIC) (pp. 279-284). IEEE.

Boyapati, P. K. Building a centralized data operations hub for healthcare enterprise integration. IJSAT-Int. J. Sci. Technol. 16 (2). https://doi.org/10.71097/IJSAT.v16.i2.3219