Cyberbullying Detection on Text and Images Using Machine Learning Algorithms

Manohar Singh
Computer Science and Engineering
Graphhic Era Hill University
Dehradun, Uttarakhand, India
burathimannu@gmail.com

Manish Bailwal
Computer Science and Engineering
Graphic Era Hill University
Dehradun, Uttarakhand, India
manishbailwal02@gmail.com

Piyush Bagla
Computer Science and Engineering
Graphic Era Hill University
Dehradun, Uttarakhand, India
Pbagla@gehu.ac.in

Kavindra Singh
Computer Science and Engineering
Graphhic Era Hill University
Dehradun, Uttarakhand, India
Kavindrakapkoti5@gmail.com

Abstract **- In today's world, the rise of social media has provided everyone with a platform to share their thoughts and communicate with each other. However, it has also led to harmful behaviors such as cyberbullying. This research addresses the challenge of detecting cyberbullying in both textual and image-based content using advanced machine learning algorithms. We propose a comprehensive system that preprocesses and analyzes data to train multiple models and evaluate their effectiveness. Our approach integrates text preprocessing techniques such as normalization and tokenization, alongside image preprocessing techniques including optical character recognition (OCR) for text extraction. We evaluate a range of machine learning models including Random Forest, SVM, Logistic Regression, Naive Bayes, and Gradient Boosting, aiming to identify the best-performing model for multimodal content analysis. The system is implemented using a Streamlit application, facilitating user interaction for input and detection. Results demonstrate the efficacy of combining text and image analysis, underscoring the importance of multimodal approaches in improving cyberbullying and hate speech detection accuracy. Our proposed models yield the best performance of 91% when we applied Random Forest Classifier.**

Index Terms - Cyberbullying Detection , Machine learning algorithms, Data preprocessing, Text analysis, Image analysis.

I. INTRODUCTION

Cyberbullying is a significant public health issue that affects millions of individuals worldwide, causing emotional distress, psychological harm, and, in severe cases, leading to tragic consequences. As the prevalence of social media platforms continues to grow, so does the incidence of cyberbullying, making it a pressing concern for individuals, families and socitey. Social media platforms like Facebook, Twitter, Instagram, and Snapchat have become central to modern communication, allowing users to connect and share content effortlessly. However, this ease of access and anonymity has also facilitated the spread of harmful behavior, including cyberbullying and hate speech. Cyberbullying involves using technology to harass, threaten, embarrass, or target another person. Online threats, as well as mean, aggressive, or rude texts, tweets, posts, or messages, all count as forms of cyberbullying. These behaviors can lead to severe psychological consequences for victims, necessitating effective detection and intervention strategies.

Detecting cyberbullying and hate speech is a complex task. Traditional methods mainly focus on analyzing text using natural language processing (NLP) techniques. While these methods can be effective, they often miss some abusive content because language changes and depends on context. Additionally, with more images being shared on social media, it's important to also analyze images to find visual signs of cyberbullying and hate speech.

Many research studies in this area have used various machine learning and deep learning techniques to achieve significant results in detecting and preventing cyberbullying. However, most of these studies have focused mainly on textual data. In this research, we will develop a comprehensive preprocessing pipeline, evaluate multiple machine learning models, create an integrated system for text and image analysis, and conduct a comparative analysis of model effectiveness.

II. LITERATURE SURVEY

Cyberbullying has grown in popularity as a result of technical advancement and increased accessibility. In addition to the emotional, social, and intellectual consequences of cyberbullying and online harassment, many victims of cyberbullying have attempted suicide as a result of their psychological trauma [1] .

Cyberbullying has been recognised as a growing global issue, with preventative measures being considered for prospective victims. Consequently, research studies should look at detecting cyberbullying and devising preventive methods [2,3].
Traditional approaches are, however, difficult to scale and assess in this situation. These strategies are often based on human patterns and social networking sites with comparatively small data sampling [4].

Most research papers have extracted data from a single source and conducted comparative studies on various machine learning techniques combined with different word vectors or feature extraction techniques to determine the best combination. Only a few studies have focused on optimizing detection models by either building ensemble machine learning models or layering different feature pre processing techniques. Most works in this area have primarily focused on textual data, neglecting images that may contain cyberbullying. Nowadays, images containing cyberbullying are widely circulated on Twitter and other social media platforms. Therefore, we are also focusing on analysing image content to detect cyberbullying.

In 2019, Emmery et al. presented an extensive survey that reviews existing annotation schemes and datasets specific to cyberbullying. This survey, published in a peer-reviewed journal, offers a detailed overview of the methodologies used to annotate cyberbullying data, the types of cyberbullying covered, and the sources of the data. It serves as a crucial reference for researchers looking to understand the current landscape of cyberbullying datasets and to identify gaps that need to be addressed. The survey emphasizes the importance of standardized annotation schemes to ensure consistency and reliability in data labeling, which is essential for developing robust and generalizable machine learning models [5] .

In 2020 a study was done by Amgad Muneer and Suliman Mohamed Fati to detect cyberbullying without involving victims by analyzing 37,373 tweets using seven machine learning classifiers: Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB), and Support Vector Machine (SVM). Evaluated on accuracy, precision, recall, and F1 score, LR showed the highest performance with a median accuracy of 90.57%. Additionally, LR achieved the best F1 score (0.928), SGD the best precision (0.968), and SVM the best recall (1.0) [6].

In 2020, R. R. Dalvi, S. Baliram Chavan, and A. Halbe presented a machine learning model aimed at detecting and preventing cyberbullying on Twitter. The study explored the application of various machine learning techniques to accurately identify instances of cyberbullying on the social media platform. The researchers utilized a combination of classifiers and evaluated their effectiveness in identifying true positive scenarios [7].

In 2021 a research was done by Kazi Saeed Alam ,Shovan Bhowmik, and Kundu Prosun who developed an ensemble-based machine learning approach for detecting offensive texts with high accuracy[8] .

In 2021 a study was done by Stefano Menini, Alessio Palmero Aprosio and Sara Tonelli and presented a novel dataset of images and comments in Italian, created with teenagers to raise awareness of cyberbullying. They collected potentially offensive comments for over 1,000 images and categorized them semantically. This analysis reveals that the presence and gender of human subjects in images trigger different types of comments, providing new insights into the connection between social media images and offensive messages. They also compared their dataset with a similar one from WhatsApp, highlighting differences between image-based and text-only interactions [9].

Detecting abusive language and cyberbullying on social media has become increasingly critical to prevent future damage. As a result, various studies have been conducted over the past decade to address this issue. For instance, Sadiq et al. utilized machine learning techniques to identify vandalism on Wikipedia. They applied methods from text analysis and communication theory to detect cyber-violence from sexual abusers on the internet [10].

In 2021 E. Sarac Essiz and M. Oturakci proposed an innovative approach using an Artificial Bee Colony (ABC)-based feature selection algorithm to enhance the detection of cyberbullying. That study highlights the efficiency of ABC in selecting the most relevant features from a dataset, thereby improving the performance of machine learning models in identifying cyberbullying. The ABC algorithm mimics the foraging behavior of bees to explore and exploit the search space effectively. By applying this technique, the authors were able to reduce the dimensionality of the data, which not only speeds up the computation process but also increases the accuracy of the cyberbullying detection models. Their results demonstrate significant improvements in model performance, suggesting that ABC-based feature selection can be a powerful tool in the fight against online abuse [11].

In 2022 C. E. Gomez, M. O. Sztainberg, and R. E. Trana conducted a study to curate cyberbullying datasets using a collaborative method that combines human expertise and AI. This study proposed a framework where AI algorithms pre-label potential instances of cyberbullying, which are then verified by human experts. This approach ensures accurate and comprehensive datasets, enhancing the training of

effective machine learning models for cyberbullying detection [12].

A study done by Rezvani et al. presented an intelligent cyberbullying identification systemwhich: (i) extracts attributes from the image, image metadata and textual items; (ii) contextualizes the drawn out attributes by creating a crowdsourced feedback loop; (iii) merges the attributes with a neural network to classify and develop potentially handy attributes. Their proposed approach has been able to substantially upgrade most metrics with the incorporation of contextual attributes [13].

Bandeh Ali Talpur and Declan O'Sullivan developed a feature□based system to detect the magnitude of cyberbullying in a tweet [19]. This architecture utilizes tweet-content features in order to create an ML classifier for identifying tweets as non-cyberbullied, and moderate, medium and extreme bullied tweets [14].

A model for detecting tweets that contain racist text was proposed by performing the sentiment analysis of tweets. A stacked ensemble deep learning model is assembled by combining GRU, CNN and RNN, called Gated Convolutional Recurrent- Neural Networks (GCR□NN) [15]. The performance of the model is optimized by setting different structures in terms of the number of layers, loss function, optimizer and number of neurons, etc. Proposed model showed substantially better performance than those of machine learning models.

Embedding word representations and deep-learning approaches were recognize and classify toxic speech. A Twitter corpora was used to conduct binary and multi-class classification, and two main approaches were investigated: extracting word embeddings accompanied by utilizing a DNN classifier and fine-tuning the pre-trained BERT classifier. BERT fine-tuning was found to be substantially more effective [16].

III. METHODOLOGY

This section outlines the methodology for detecting cyberbullying and hate speech in text and images using machine learning algorithms. The approach includes data collection, preprocessing, model training, and system implementation. Coming to the implemented approach that is shown in Figure 1. It depicts how cyberbullying was detected in the acquired dataset.

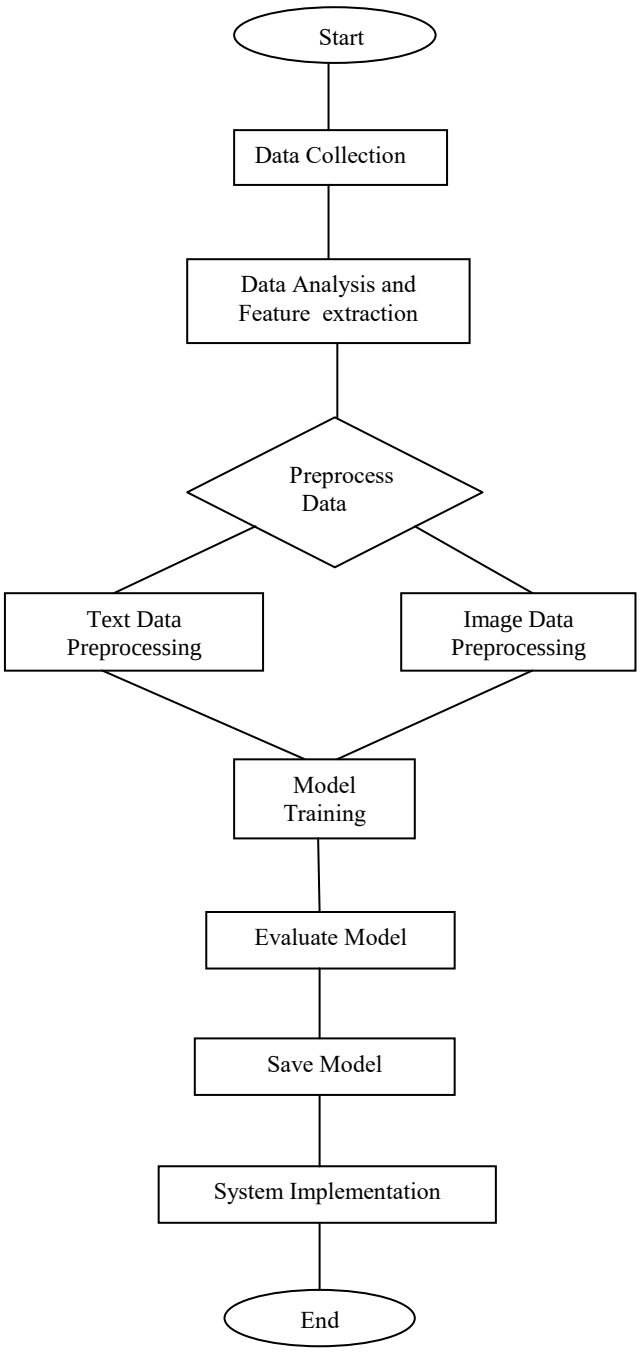

Figure 1 : Flowchart of implemented approach

A. Data Collection : The dataset for this study includes both text data from social media platforms and images containing textual content. The dataset for this study includes both text data from social media platforms and images containing textual content. The preprocessing involves cleaning and preparing both text and image data to ensure optimal performance of the machine learning models. The text dataset contains 74,682

entries, each labeled with a sentiment . For image dataset we have various images but in limited amount.

B. Text Preprocessing *:* Text data undergoes several preprocessing steps, including converting all text to lowercase for consistency, eliminating URLs, HTML tags, special characters, and retweet tags to clean the text, splitting the text into individual words or tokens, reducing words to their base or root form through lemmatization, and removing common words (stopwords) that do not contribute to the text's meaning, such as "and," "the," and "is."

C. Image Preprocessing : In order to reduce computational complexity, images are preprocessed by converting them to grayscale, standardizing image sizes to a uniform dimension, using optical character recognition (OCR) to extract text embedded in images, and then preprocessing the extracted text using the same methods as textual data.

D. Feature Engineering: Feature engineering is the process of changing raw data into features that machine learning algorithms can use to improve performance. This includes turning cleaned text data into numerical representations using approaches such as TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, as well as integrating extracted text from images with additional attributes such as text length and sentiment scores.

E. Model Selection : Model selection entails selecting the optimal machine learning algorithms to train on our processed data. The performance of various classifiers is examined, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, AdaBoost, Logistic Regression, Naive Bayes, and Gradient Boosting .

F. Model Training: Model training entails utilizing the chosen algorithms to extract patterns from training data. The procedure consists of training each classifier on the vectorized text data from the training set, adjusting hyperparameters for each model to optimize performance, and saving the trained models and vectorizers for later use.

G. Model Evaluation : It involves analyzing the performance of trained models on testing data to ensure that they generalize effectively to new, previously unknown data. This includes assessing the accuracy of each model using metrics such as accuracy score and determining the model with the highest accuracy or other relevant performance indicators.

H. Model Deployment : Model deployment involves integrating the trained model into an application that can be used to make predictions on new data. This includes developing a user-friendly interface using a framework like Streamlit to allow users to input text or upload images, loading the saved models and vectorizers into the application, using the trained model to make predictions based on user input and displaying the results, and continuously monitoring the deployed model's performance and updating it as necessary to maintain accuracy.

IV. RESULT

The performance of various machine learning algorithms for detecting cyberbullying from text and images was evaluated. The algorithms included Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, AdaBoost, Logistic Regression, Naive Bayes, and Gradient Boosting. Each model was trained using preprocessed text data vectorized with TF-IDF and was evaluated based on accuracy. The results are summarized in Figure 2.

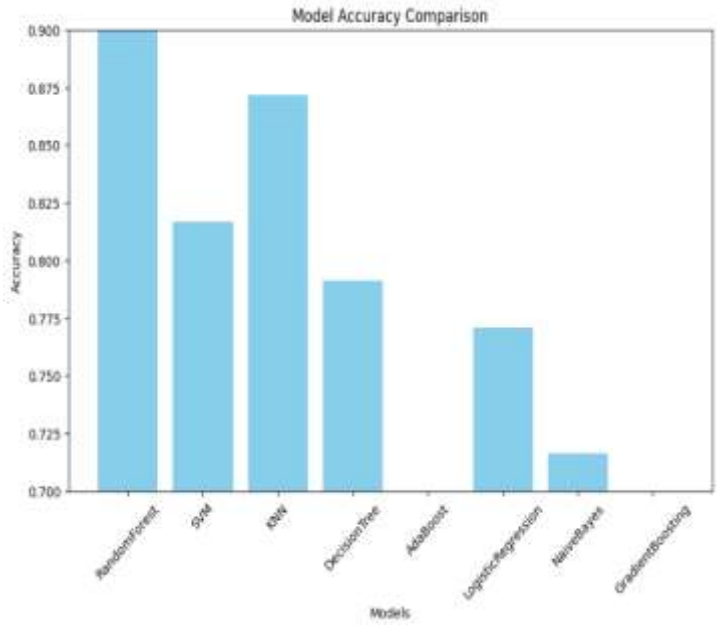

Figure 2 : Model Performance Comparison

Key Findings :

1. Highest Accuracy**:** The Random Forest Classifier achieved the highest accuracy of

91%, making it the best-performing model for this study.

2. Multimodal Improvement: The integration of text and image analysis improved the overall detection performance, demonstrating the effectiveness of a multimodal approach.

Combining text and image data provides a more comprehensive understanding of online content, enabling more accurate detection of cyberbullying and hate speech. The success of the Random Forest Classifier highlights the effectiveness of ensemble methods in handling the variability in text and image data. The use of OCR for image text extraction and its analysis underscores the importance of context in accurately detecting harmful content.

The findings suggest that multimodal approaches have significant potential in real-world applications, such as social media monitoring and content moderation, to reduce harmful content and improve user safety. The results can inform policymakers and platform developers about effective strategies for detecting and mitigating cyberbullying and hate speech. Implementing such detection systems on social media platforms can significantly reduce users' exposure to harmful content, enhancing overall user safety and experience.that ensemble methods are effective in handling the variability in text and image data.

The study is limited by the availability of labeled data for training, particularly for image-based content, which can impact model generalizability. The system's performance may vary across different social media platforms due to variations in user behavior, content styles, and platform-specific nuances. Additionally, while ensemble methods perform well, they are often less interpretable than simpler models, limiting the understanding of the decision-making process.

V. Conclusion & Future works

Cyberbullying messages and posts over social media are continuously affecting individuals particularly teenagers and society and often lead to a series of consequences even suicidal thoughts among the victims. In this research, we have built a comprehensive system for detecting cyberbullying and hate speech in both text and images using machine learning algorithms. By preprocessing data, training multiple models, and integrating a multimodal approach, the study achieved high detection accuracy. The findings emphasize the importance of combining text and image analysis for more effective content moderation. Future work should focus on enhancing the dataset and exploring advanced deep learning models to further improve the system's performance.

Future research should focus on expanding the dataset to include more diverse examples of cyberbullying and hate speech, covering different languages, cultures, and platforms. Exploring advanced deep learning techniques, such as multimodal transformers and transfer learning, could further enhance detection performance and robustness. Incorporating user feedback mechanisms can help refine the model by adapting to new trends and emerging types of harmful content. Conducting cross-platform analysis can provide insights into the model's generalizability and help create more adaptable and versatile detection systems.

Additionally, collaborating with social media platforms to implement real- time detection systems can integrate real-time detection technologies can greatly improve user safety. When implementing such systems, it is imperative to take privacy protection and ethical considerations into account. By using explainable AI strategies to improve the interpretability of complicated models, consumers' trust will be increased and decision-making processes will become easier to grasp. Furthermore, combining sociological and psychological knowledge can aid in the creation of more complex models that take the intensity and context of cyberbullying occurrences into account. Having conversations with legislators, parents, and educators can help comprehensive plans to stop cyberbullying and make the internet a safer place for all users to use become more widely adopted.

References

[1] D. Hall, Y. Silva, Y. Wheeler, L. Cheng and K. Baumel, “Harnessing the power of interdisciplinary research with psychology-informed cyberbullying detection models,” International Journal of Bullying Prevention, vol. 4, no. 1, pp. 47–54, 2021.

[2] K. Arce-Ruelas, “Automatic cyberbullying detection: A Mexican case in high school and Higher Education Students,” IEEE Latin America Transactions, vol. 20, no. 5, pp. 770–779, 2022.

[3] T. Ahmed, M. Rahman, S. Nur, A. Islam and D. Das, “Natural language processing and machine learning based cyberbullying detection for Bangla and romanized bangla texts,” TELKOMNIKA (Telecommunica tion Computing Electronics and Control), vol. 20, no. 1, pp. 89–97, 2021.

[4] C. Theng, N. Othman, R. Abdullah, S. Anawar, Z. Ayop et al., “Cyberbullying detection in twitter using sentiment analysis,” International Journal of Computer Science & Network Security, vol. 21, no. 11, pp. 1–10, 2021.

[5] Emmery, C., Verhoeven, B., Daelemans, W., & Jacobs, G. (2019). Current Limitations in Cyberbullying Detection: On Evaluation and Data Collection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6080-6090). ACL.

[6] Muneer, A., & Fati, S. M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter.

Department of Computer and Information Sciences, Universiti Teknologi PETRONAS; Information Systems Department, College of Computer and Information Sciences, Prince Sultan University.

[7] R. R. Dalvi, S. Baliram Chavan and A. Halbe, ”Detecting A Twitter Cyberbullying Using Machine Learning,” 2020 4th International Confer ence on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893

[8] Alam, K. S., Bhowmik, S., & Kundu Prosun, P. R. (2024). Cyberbullying detection: An ensemble based machine learning approach. In Proceedings of the 2024 International Conference on Machine Learning and Data Engineering (MLDE). Khulna University of Engineering & Technology, Bangladesh Army International University of Science and Technology.

[9] Menini, S., Palmero Aprosio, A., & Tonelli, S. (2024). A multimodal dataset of images and text to study abusive language. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Fondazione Bruno Kessler, Trento, Italy..

[10] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. Choi et al., “Aggression detection through deep neural model on twitter,” Future Generation Computer Systems, vol. 114, no. 1, pp. 120–129, 2021.

[11] E. Sarac Essiz and M. Oturakci, “Artificial bee colony-based feature selection algorithm for cyberbullying,” The Computer Journal, vol. 64, no. 3, pp. 305–313, 2021.

[12] C. E. Gomez, M. O. Sztainberg and R. E. Trana, “Curating cyberbullying datasets: A human-AI collabo□rative approach,” International Journal of Bullying Prevention, vol. 4, no. 1, pp. 35–46, 2022.

[13] Nabi Rezvani and Alireza Tabebordbar. ”Linking Textual and Contextual Features for Intelligent Cyberbullying Detection in Social Media.”

[14] Bandeh Ali Talpur and Declan O’Sullivan. ”Multi-Class Imbalance in Text Classification: A Feature Engineering Approach to Detect Cyber□bullying in Twitter.” Informatics. Vol. 7. No. 4. Multidisciplinary Digital Publishing Institute, 2020.

[15] Lee E, Rustam F, Washington PB, El Barakaz F, Aljedaani W, Ashraf I. Racism detection by analyzing diferential opinions through sentiment analysis of tweets using stacked ensemble gcr nn model. IEEE Access. 2022;10:9717–28.

[16] d’Sa, A.G., Illina, I., Fohr, D.: Bert and fasttext embeddings for automatic detection of toxic speech. In: 2020 International Multi-Conference on:“Organization of Knowledge and Advanced Technologies”(OCTA), pp. 1–5 (2020). IEEE