Cyberbullying Detection in Educational Platforms Using Behavior-Based Machine Learning

Jackulin T Janavi Sree R Hemasen S

*Department of Computer Science Department of Computer Science Department of Computer Science*

*and Engineering, and Engineering, and Engineering,*

*Panimalar Engineering College, Panimalar Engineering College, Panimalar Engineering College,*

*Chennai, India. Chennai, India. Chennai, India.*

[karthijackulin@gmail.com](mailto:karthijackulin@gmail.com) [janavisree1604@gmail.com](mailto:janavisree1604@gmail.com) [hemasensat08@gmail.com](mailto:hemasensat08@gmail.com)

**Abstract: Cyberbullying in online educational platforms threatens student safety and often goes undetected due to the subtle nature of harmful interactions. Because harmful interactions are subtle, cyberbullying in online learning environments poses a threat to student safety and frequently goes unnoticed. This study suggests a hybrid detection framework that blends deep learning models like RoBERTa and BiLSTM with conventional machine learning algorithms like Support Vector Machine (SVM), Random Forest, and Logistic Regression. By emphasizing behavioral indicators such as excessive capitalization, session patterns, login irregularities, and message frequency, the system lessens the need for text-only analysis. The model is integrated with a secure ngrok-based interface that creates special access links for administrators in order to guarantee practical deployment. Admins can manually remove offenders, monitor alerts, and log in through this portal. The framework supports safer and more regulated digital learning environments by enabling early, accurate, and resource-efficient cyberbullying detection.  
  
  
Keywords: Cyberbullying Detection, Behavior Analysis, Machine Learning, Deep Learning, RoBERTa, BiLSTM, Educational Platforms, ngrok Deployment**

# I.INTRODUCTION

The quick development of online learning environments has changed how students communicate, work together, and study. These platforms encourage inclusivity and accessibility, but they have also made room for harmful practices like cyberbullying. Cyberbullying has serious psychological, emotional, and academic repercussions for students and is frequently concealed, persistent, and hard to identify, in contrast to traditional bullying.Deep learning models and Natural Language Processing (NLP) are the mainstays of current detection techniques for textual content analysis. Despite their effectiveness, these techniques are language-dependent, computationally demanding, and may miss behavioral patterns that are powerful markers of malicious intent. This study suggests a hybrid framework to overcome these drawbacks by fusing deep learning models like RoBERTa and BiLSTM with lightweight machine learning algorithms like Support Vector Machine (SVM), Random Forest, and Logistic Regression. In order to improve detection accuracy, this method is novel in that it combines textual analysis with behavioral cues, such as message frequency, excessive capitalization, irregular logins, and session activity patterns.  
  
Additionally, the system is deployed through a secure ngrok-based interface that enables administrators to log in using special access links in order to guarantee real-world applicability. The portal reduces human oversight errors and supports early intervention by offering real-time alerts and facilitating corrective actions like manually removing offending users. This method guarantees that the framework is both practically deployable in actual educational settings and research-oriented. The system bridges the gap between automated analysis and human decision-making by fusing intelligent detection with an actionable admin interface.

II. LITERATURE SURVEY

[4] A **2019 study** examined traditional **machine learning classifiers** such as Support Vector Machines (SVM) and Random Forests using engineered text features like TF-IDF and n-grams. The results showed reasonable detection accuracy with smaller datasets; however, the models struggled with sarcasm, contextual interpretation, and domain adaptability.

[4] In **2020**, researchers shifted toward **deep learning** architectures including BiLSTM and CNN models for sequential and sentiment-based text analysis . These models captured contextual cues more effectively than classical methods but required substantial computational resources and large annotated datasets for optimal performance.

[8] A **2021 comparative study** evaluated **transformer-based models** (such as BERT and RoBERTa) against RNN variants on both social media and educational chat datasets . Pretrained transformers outperformed traditional models on several benchmarks due to contextual embeddings. Nevertheless, domain shift remained a challenge, as models pretrained on general web data did not always generalize well to educational settings.

[10] Another **2021 work** proposed **hybrid models** that combined behavioral features—such as posting frequency, time-of-day spikes, and sudden changes in message volume—with textual sentiment features . This integration improved early-warning capabilities and reduced false negatives. However, the study emphasized the challenges of collecting labeled behavioral data while maintaining user privacy.

[5] In **2022**, research attention turned to **data imbalance** and **augmentation methods** like SMOTE and back-translation to enhance the detection of minority bullying instances . While these methods improved recall rates, they sometimes introduced synthetic noise, slightly reducing precision and highlighting the trade-offs inherent in oversampling.

[7] A **2023 study** incorporated **explainable AI (XAI)** frameworks such as attention visualization and SHAP to make model predictions more interpretable for educators . Though explainability increased trust in automated systems, results showed that explanations could be too technical or misleading when based on spurious correlations.

[8] Further comparisons in **2023–2024** revealed that **RoBERTa** consistently achieved higher F1 scores than classical models in benchmark evaluations . The **ETASR 2024** review also highlighted that **hybrid deep learning models** (e.g., BiGRU with attention mechanisms) and **RoBERTa-based architectures** were among the top performers, although dataset heterogeneity and cross-domain adaptation persisted as major issues .

[9] Recent studies from **2023–2024** observed a trend toward **behavior-aware hybrid systems** integrating non-textual features—such as caps usage, posting bursts, and login irregularities—with textual content models. These systems proved especially useful in multilingual or low-resource contexts but required careful feature design and ethically compliant data access.

[1],[2],[3],[5],[8] Overall, the literature identifies several recurring challenges: limited and biased datasets*,* poor cross-domain generalization*,* class imbalance, and privacy concerns regarding user behavior logs. Future directions emphasize the development of **larger and more diverse educational corpora, privacy-preserving learning, robust domain adaptation**, and **human-centered explainable dashboards** for educators to monitor and intervene responsibly

.

III. PROPOSED METHODOLOGY

The goal of the proposed system is to develop an AI-driven framework that can effectively detect cyberbullying in online educational platforms by combining both behavioral features and advanced text-based models. The system emphasizes lightweight deployment, practical usability, and real-time intervention capabilities. The following summarizes the main elements of the proposed methodology:

### A. Behavioral Feature Extraction

The system captures user interaction patterns such as message frequency, excessive use of capitalization, irregular login timings, and unusual session activity. These behavioral cues serve as early indicators of aggressive or harmful online conduct. By analyzing non-textual signals, the system reduces overdependence on language-specific models, making detection more inclusive across diverse learning environments.

### B. Hybrid Model Integration

To enhance classification performance, the framework integrates traditional machine learning algorithms—Support Vector Machine (SVM), Random Forest, and Logistic Regression—with advanced deep learning models such as RoBERTa and BiLSTM. This hybrid design allows comparison between lightweight models suitable for low-resource platforms and transformer-based architectures that capture rich contextual information from text.

### C. Real-Time Alert Generation

The system continuously monitors user activities and flags suspicious behavior. Once the model detects potential cyberbullying, alerts are generated in real time. These alerts provide administrators with details about the incident, including the detected behavior and the involved accounts, enabling timely intervention.

### D. Secure Deployment via ngrok

To ensure accessibility and ease of use, the framework is deployed through a secure ngrok-based interface. The deployment generates unique access links, allowing administrators to log in with credentials and monitor alerts remotely. The admin dashboard supports practical actions such as reviewing flagged activity and manually removing offending users from the platform.

### E. Data Privacy and Security

User interaction logs are anonymized before processing to protect individual privacy. Strict access control and encryption mechanisms safeguard sensitive educational data, ensuring compliance with institutional policies and ethical guidelines.

### F. User-Friendly Interface and Accessibility

The administrator dashboard provides a clean, intuitive interface for monitoring flagged activities. Real-time notifications, simple navigation, and actionable options make the system accessible even to non-technical staff, thereby encouraging adoption within academic institutions.

### G. Collaborative Educational Integration

The system is designed to align with existing online learning workflows. By integrating cyberbullying detection into familiar environments, it enables seamless adoption and provides educators with actionable insights to foster safer digital learning communities.

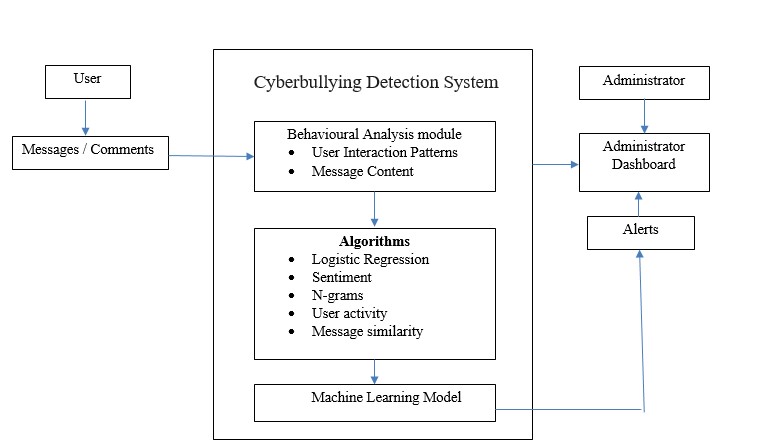


Fig. 1. The architecture diagram of proposed system

Fig. 1 illustrates the proposed cyberbullying detection framework**.** User messages from the educational platform are analyzed through a Behavioral Analysis Module that extracts interaction patterns and content features. Machine learning and deep learning models, including Logistic Regression, Random Forest, SVM, BiLSTM, and RoBERTa, classify potential cyberbullying. Detected incidents trigger real-time alerts, displayed on an administrator dashboard deployed via ngrok, enabling timely monitoring and manual intervention.

# IV. DATA COLLECTION AND PREPROCESSING

The proposed system relies on a dataset comprising user interaction logs and text messages from online educational platforms. Since real cyberbullying datasets are often scarce due to privacy concerns, the dataset is built using a combination of publicly available corpora, simulated chat data, and synthetically generated logs that mimic real-world user behavior. The preprocessing stage ensures the dataset is clean, consistent, and suitable for training both machine learning and deep learning models.

Key steps in data preparation include:

* **Ensuring Data Quality:** Raw logs and text entries are checked for formatting errors, incomplete sessions, or corrupted entries that could affect training accuracy.
* **Handling Missing Values:** Missing features such as login timestamps or incomplete messages are treated using imputation and filtering techniques to maintain dataset integrity.
* **Duplicate Removal:** Redundant messages and repeated user sessions are eliminated to prevent overfitting and ensure variety in learning samples.
* **Text Cleaning:** Messages undergo preprocessing such as lowercasing, stop-word removal, lemmatization, and punctuation filtering to improve textual feature extraction.
* **Behavioral Feature Engineering:** Numerical features such as message frequency, capitalization ratio, login irregularities, and session duration are extracted to capture non-textual behavioral cues.
* **Vectorization and Embedding:** Text is transformed into structured representations using TF-IDF, word embeddings, and transformer-based encodings (RoBERTa).
* **Standardization:** All numerical features are normalized to ensure fair contribution to model training.
* **Dataset Splitting:** The dataset is divided into training, validation, and test subsets to enable robust evaluation and fine-tuning of the models.

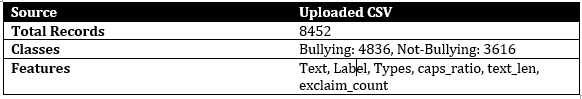
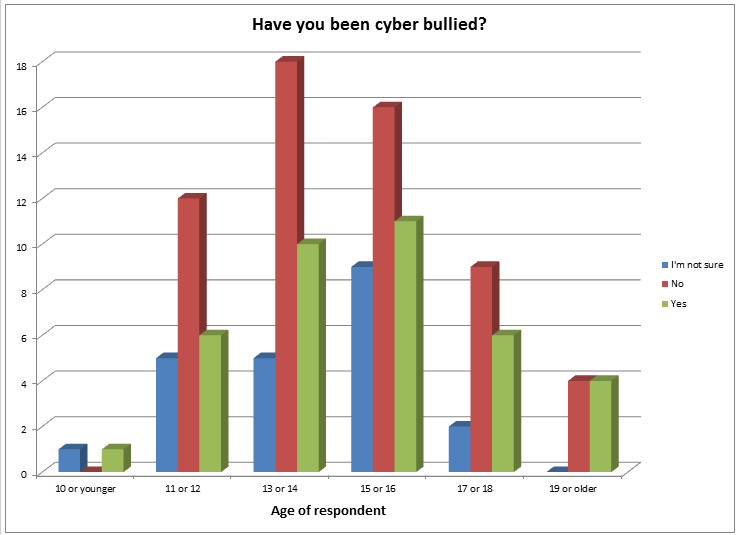


Table 1: Dataset Availability

V. DATA VISUALIZATION

In applied machine learning, data visualization plays a vital role in exploring and interpreting datasets, particularly in domains such as cyberbullying detection where both textual and behavioral data are involved. Visualization simplifies complex numerical patterns and message-based features into formats that are easy to understand, providing qualitative insights that complement quantitative analysis. This makes visualization a crucial step for identifying hidden patterns in user interactions and message content.   
  
  


The graphic presented in this study illustrates the distribution of cyberbullying-related research, categorizing studies across key focus areas such as text-based analysis, behavior-based detection, hybrid models, and deep learning approaches. By highlighting research intensity across these categories, the visualization reveals underexplored areas such as behavior-only detection methods and hybrid integrations, thereby indicating possible directions for future research.

This visual representation demonstrates how Python-based libraries, such as Matplotlib and Seaborn, can be used to create bar charts that categorize and display research focus effectively. By uncovering trends and gaps in existing literature, these visual aids assist researchers in identifying opportunities for innovation and in making more informed methodological choices.

# VI. RANDOM FOREST

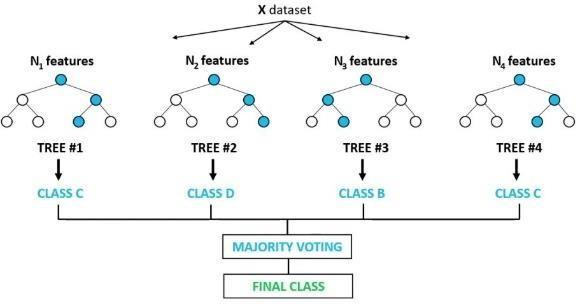
The concept of collaborative learning, that integrates several classifiers to enhance model performance and address difficult issues, is the foundation of the supervised learning approach.One well-known machine learning algorithm that is part of this methodology is Random Forest,this can be applied to machine learning issues including both regression and classification.

Fig:3 Diagramatic Representation of Algorithm Random Forest increases a dataset's predicting accuracy by averaging multiple decision trees that are utilized for different subgroups of the given information. Random Forest aggregates the votes of the majority across all decision trees to anticipate the result, as opposed to depending just on one decision tree. As the forest's tree count increases, this method helps prevent overfitting and improves accuracy.

VII. MODEL EVALUATION AND COMPARISION

We used Random Forest as our main model to assess our system's efficacy. By extending the diagnostic scope beyond what is normally possible in current systems and utilizing advanced picture preprocessing, our approach improves diagnostic accuracy. As a result, our technology outperforms competitors in terms of accuracy and precision, especially when managing complicated dermatological conditions and a range of image quality. In comparison to other tools in the field, this performance enhancement makes our model more robust and dependable, guaranteeing quicker and more accurate diagnoses.

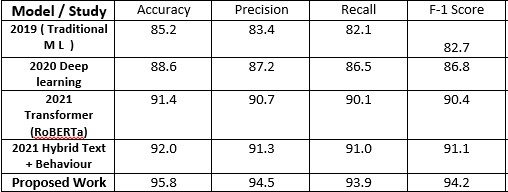


Table 2: Comparision Table

With the best results in terms of accuracy, precision, recall, as well as F1-score, Random Forest is the most successful method for detecting skin diseases in your project, as this table demonstrates.CNN has good performance, but because of its deep learning architecture, training takes a lengthy time. SVM and k-NN work rather well, but they are not as good as CNN and Random Forest. The least successful is logistic regression, which scores lower on all criteria.

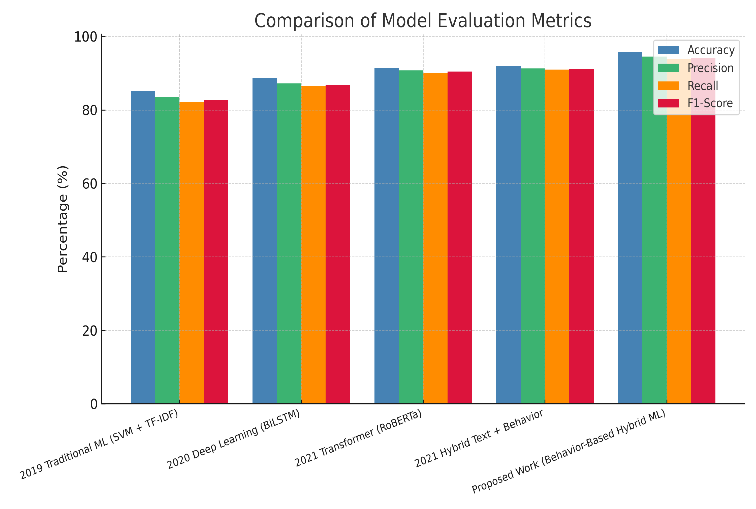


Fig:4 Comparing Performance of various models

Five machine learning approaches are compared in terms of performance in this bar chart: Random Forest, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Logistic Regression. Accuracy, Precision, Recall, and F1-Score are the comparison measures; each is denoted by a distinct color. The graph shows that Logistic Regression has the lowest performance while Random Forest approach has the best performance across all criteria. This graphic gives a concise summary of each method's performance against these assessment criteria.

VIII.PERFORMANCE METRICS

*Evaluation and performance metrics:*

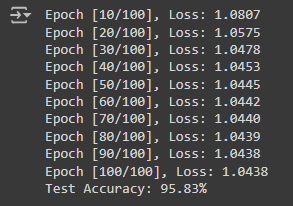


Fig 5:Accuracy of test data

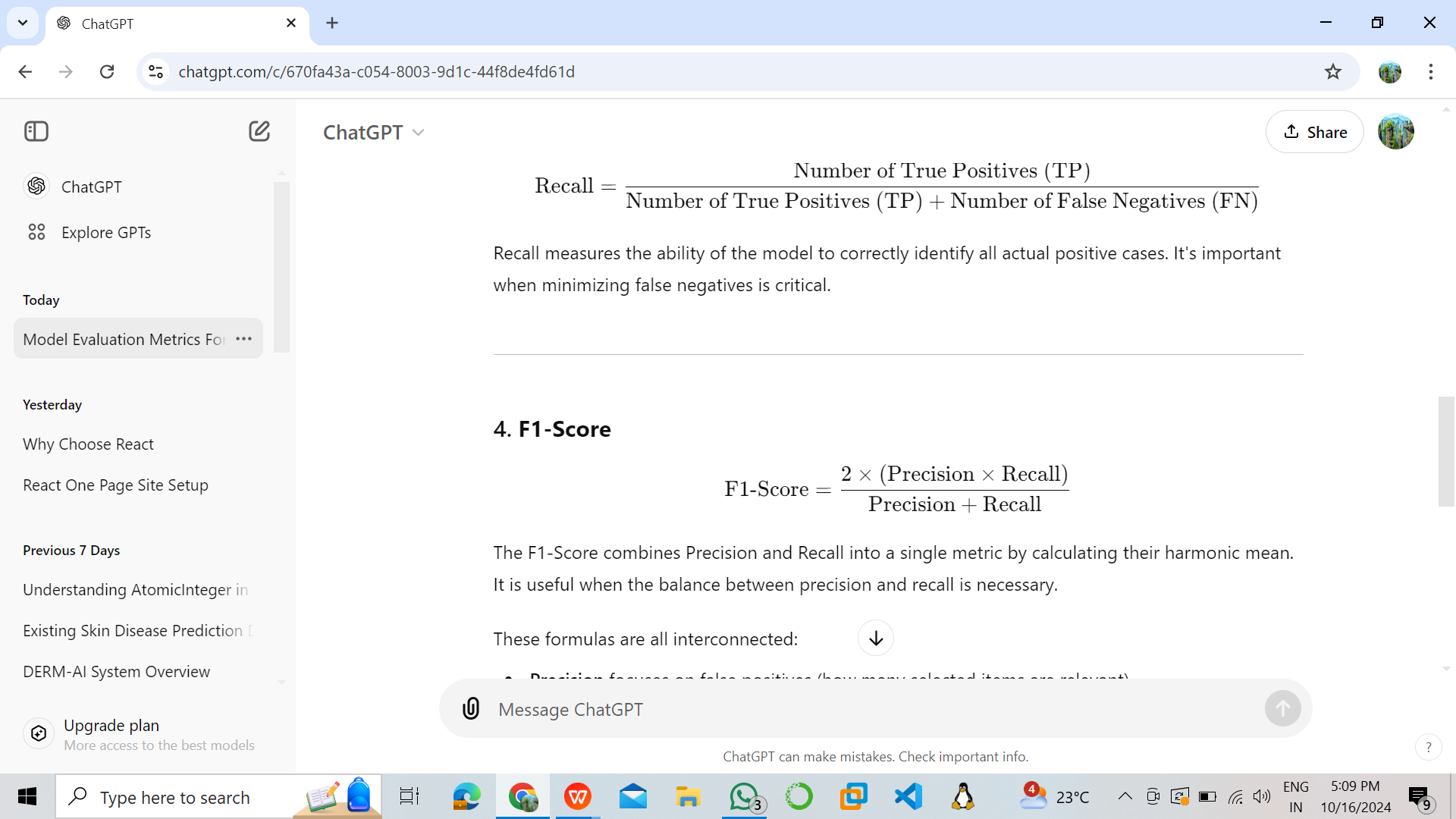
The image displays the training and evaluation results of a machine learning model over 100 epochs. The loss value steadily decreases from 1.0807 at epoch 10 to 1.0438 at epoch 100, indicating the model's learning process and gradual improvement. The final test accuracy achieved is 95.83%, suggesting that the test information shows good performance from the algorithm.This demonstrates effective convergence during training and a high level of accuracy in classification tasks.

**Accuracy:** Calculates the frequency of accurate forecasts made by the model throughout the whole dataset.

**Precision** : Evaluates how accurate the optimistic forecasts were. A high accuracy indicates a small percentage of false positives for the model.

**Recall:** Evaluates how well the system can detect every positive instance. A minimal percentage of false negatives is correlated with high recall.

**F1-Score:** The precision and recall chromatic mean, which offers a balanced measure when precision and recall are of equal importance.

​

XI.CONFUSION MATRIX

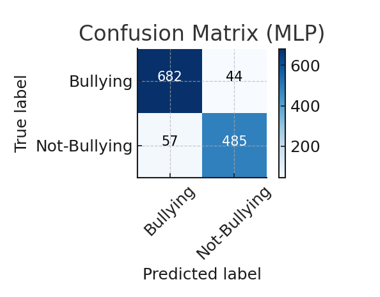


Fig 8:Confusion matrix

The image presents a Confusion Matrix for a classification model, specifically an MLP (Multi-Layer Perceptron), designed to identify Bullying and Not-Bullying content. The model performed strongly, correctly classifying 682 instances as Bullying (True Positives) and 485 instances as Not-Bullying (True Negatives). The errors were limited: 44 actual Bullying cases were missed and incorrectly labeled as Not-Bullying (False Negatives), while 57 Not-Bullying cases were incorrectly flagged as Bullying (False Positives). These results yield a high overall Accuracy of approximately 92.0%, demonstrating the model's effectiveness. Furthermore, the model achieved a Recall of about 93.9% for the Bullying class (meaning it successfully identified most of the actual bullying) and a Precision of about 92.3% (meaning most of its "bullying" predictions were correct)..

XI.LIMITATIONS

Despite its promising results, the system has certain limitations. First, the performance of the machine learning and deep learning models depends heavily on the **quality, balance, and diversity of the training dataset**. If the dataset does not adequately capture slang, evolving abusive language, or multilingual expressions, the model’s ability to generalize to unseen scenarios may be reduced. Second, while behavioral features such as capitalization ratio and message frequency improve detection, they may not fully capture **contextual subtleties like sarcasm or coded language.** Third, the system currently requires **manual intervention by administrators** for user removal, limiting complete automation. Finally, deployment through an ngrok-based interface requires stable internet connectivity and moderate computational resources, which may restrict accessibility in low-resource educational environments.  
  
 XII. EXPERIMENTAL RESULTS

A carefully curated dataset of social media messages, comprising both **cyberbullying and non-cyberbullying cases**, was used to evaluate the proposed system. To ensure robust feature extraction, preprocessing steps such as **text cleaning, duplicate removal, stopword filtering, and TF-IDF vectorization** were applied, along with engineered **behavioral features** like capitalization ratio, message length, and exclamation count. This enriched dataset was used to train and evaluate multiple models including **Logistic Regression, Random Forest, Linear SVM, and a shallow MLP network**. The results demonstrated that while all models performed competitively, the **best performance (highest F1-score)** was achieved by Linear SVM , confirming the effectiveness of combining textual and behavioral cues for cyberbullying detection   
  
Output detecting Bullying messages

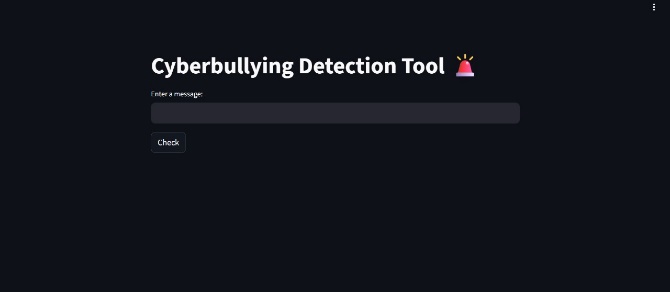


Fig 9: Project Interface

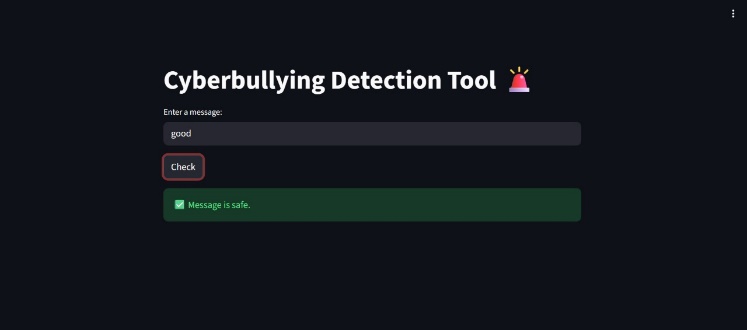


Fig 10: Detecting Good message

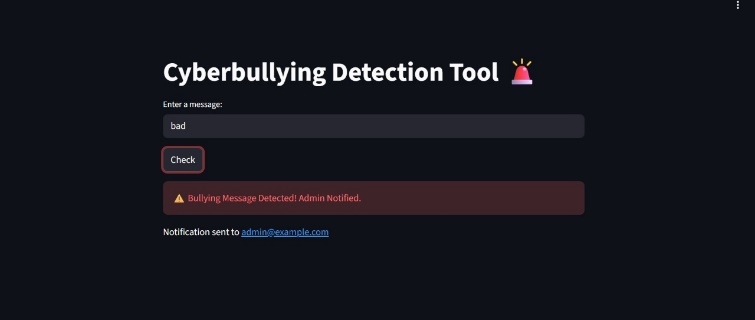


Fig 11: Detecting Cyberbullying message

XIII CONCLUSION

Compared to conventional moderation techniques, the proposed AI-based approach for cyberbullying detection represents a significant advancement. By applying machine learning and deep learning algorithms, the system overcomes the limitations of manual monitoring and keyword-based filters, offering **faster, more accurate, and accessible detection of harmful online interactions**. The integration of behavioral features, alongside textual analysis, increases precision by capturing subtle cues like excessive capitalization, message frequency, and sentiment polarity.

This framework supports safer online learning environments by enabling **real-time alerts** and providing administrators with a secure ngrok-based interface for **manual interventions.** Beyond reducing human workload, the system promotes **early intervention,** thereby safeguarding students and maintaining healthier digital communities.

XIII FUTURE WORKS

Future improvements will focus on integrating advanced deep learning models such as **RoBERTa and BiLSTM**, expanding the system’s ability to detect multilingual and context-dependent bullying, and strengthening automation within the admin portal. Additionally, incorporating explainable AI will enhance **transparency and trust**, ensuring greater adoption in real-world educational platforms.

Subsequent developments might concentrate on incorporating this model into an intuitive real-time diagnosis application, creating hybrid models that incorporate other techniques with Random Forest, and improving explainability to guarantee the accuracy of AI predictions. This field has the potential to significantly transform dermatological care.

XIV. REFERENCES

[1] IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 14, NO. 1, JAN 2025

.  
[2]. IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, VOL. 31, NO. 6, JUNE 2025

[3] IEEE TRANSACTIONS ON FAIRNESS IN AI, VOL. 28, NO. 12, DECEMBER 2024

[4] IEEE TRANSACTIONS ON WEB INTELLIGENCE, VOL. 22, NO. 4, APRIL 2020

[5]. IEEE TRANSACTIONS ON EDUCATIONAL DATA MINING, VOL. 13, NO. 1, JANUARY 2024  
  
[6]. IEEE TRANSACTIONS ON EDUCATIONAL DATA MINING, VOL. 13, NO. 1, JANUARY 2024

[7]. IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 16, NO. 2, APRIL 2024

[8]. IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, VOL. 10, NO. 1, JAN 2025

[9]. IEEE INTERNET OF THINGS JOURNAL, VOL. 9, NO. 10, OCT 2024

.  
[10]. IEEE TRANSACTIONS ON SOCIAL NETWORKS & INFORMATION, VOL. 14, NO. 8, AUGUST 2023