Volume 1, Issue 9 (2026)
PP 01-12

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL
VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Cyber Bullying Detection and Alert System for School

Vaishali Shimpi[1], Pallavi Mohapatra[2], Mahek Shaikh[3], Rashmi Borse[4], Saloni Salvi[5]

[1,2,3,4](*Department of CSE (Artificial Intelligence and Machine Learning), Viva Institute of Technology, Mumbai University, India)*

Abstract: *Cyberbullying has become a persistent challenge in digitally meditated school environments, where student communication increasingly occurs through online messaging platforms, schoolmanaged communication systems. Unlike traditional bullying, harmful interactions can occur continuously, making manual monitoring and delayed reporting ineffective. This paper presents CyberShield, an AI-driven cyberbullying detection and alert framework specifically designed for school communication ecosystems. The system combines natural language processing with supervised machine-learning techniques to analyze student text messages and flag potentially harmful content in real time. The architecture includes preprocessing, statistical feature extraction, classification, and an alert module that enables human review and administrative intervention. The framework was evaluated on a processed HateXplain dataset containing 19,200 samples. TFIDF features were used with several classical machine-learning models, and a hybrid pipeline combining random forest classification with rule-based severity assessment achieved 94% accuracy, outperforming baseline approaches. The system is structured for deployment within controlled school platforms, supporting explainable alerts, structured reporting, and timely intervention while maintaining human oversight for ethical compliance.* ***Keywords -*** *AI-based monitoring, Artificial Intelligence, Cyberbullying, Machine Learning, Natural Language Processing, Text classification.*

1. INTRODUCTION

Digital communication technologies have become integral to student interaction within modern educational institutions. Platforms such as learning management systems, internal messaging tools, and online discussion spaces facilitate collaboration and information exchange, yet they also create conditions in which harmful interpersonal behaviors can occur. Among these challenges, cyberbullying has emerged as a critical issue due to its sustained presence, rapid dissemination, and disproportionate psychological impact on young users. Cyberbullying extends beyond the limitations of physical environments, allowing harmful content to persist and circulate across time and space. Messages intended to intimate, ridicule, or isolate can be accessed repeatedly and shared widely, amplifying their effect on victims. Prior studies associate prolonged exposure to online harassment with emotional distress, declining academic engagement, and elevated risk of self-harm among adolescents. Despite increasing awareness, many schools continue to rely on reactive, report-based mechanisms that depend heavily on subjective judgment and delayed intervention.

Automated detection of abusive and harmful online content has been extensively studied using natural language processing and machine learning techniques [1],[2]. Harmful intent is often conveyed indirectly through sarcasm, contextual references, or evolving slang rather than explicit abusive languages. Additionally, interpretations of

bullying vary across cultural, institutional, and age-specific contexts. These factors significantly reduce the effectiveness of static, rule-based filtering approaches and motivate the use of adaptive models capable of learning complex linguistic patterns from data. Recent research in machine learning and deep learning has produced effective methods for abusive language detection, particularly within large-scale social media environments. Supervised classifiers, ensemble techniques, and neural architectures incorporating attention mechanisms have shown measurable performance gains. However, such models are typically trained on data from open platforms, where communications dynamics and moderation objectives differ substantially from those of school systems. Deploying these models in educational settings without contextual adaptation may result in reduced accuracy, bias amplification, and ethical concerns.

This paper proposes CyberShield, an AI-driven cyberbullying detection and alert framework specifically designed for school-based digital communication environments. A key contribution of the proposed system is its ability to support multilingual and code-mixed student communication through text normalization and Unicode aware preprocessing. While the current evaluation is conducted on a publicly available hate speech dataset, the framework is designed for future extension to regional language specific models for improved performance in Indian school environments. The primary goal of cybershield is not limited to labeling content as abusive but to facilitate early identification and responsible intervention within a controlled educational setting. To achieve this, the system is implemented within a custom built communication platform named ShieldChat, which serves as a secure medium for student interactions. The platform incorporates privacy aware monitoring and administrative alert mechanisms that enable policy compliant analysis of potentially harmful messages while maintaining user data protection. CyberShield analyzes messages in real time, flags risks for administrator review, and includes a support chatbot for guidance and reporting. By combining multilingual detection with human oversight, it offers a practical alternative to purely automated moderation systems.

2. LITERATURE REVIEW

Cyberbullying detection has become a significant research focus due to the expansion of digital communication among students and adolescents. Harmful interactions occurring through online messaging systems and educational platforms can negatively affect mental health, academic engagement, and social development. Early approaches relied on manual moderation and simple keyword filtering, which proved ineffective in identifying implicit or context-dependent harassment. Consequently, researchers have increasingly explored automated detection systems using natural language processing and machine learning to identify harmful content in digital communication environments [3], [4].

Supervised machine learning techniques have formed the foundation of many cyberbullying detection systems. These methods rely on annotated datasets to train classifiers capable of distinguishing between harmful and nonharmful messages. Vidgen and colleagues demonstrated that supervised models trained on large-scale abusive language datasets can detect offensive patterns with reasonable accuracy, though challenges remain in understanding sarcasm, slang, and contextual nuance [5]. Their work highlights the importance of structured data and feature representation in improving classification reliability.

Feature extraction plays a central role in text classification systems. Term Frequency–Inverse Document Frequency (TF-IDF) remains one of the most widely used techniques for representing textual data numerically. TF-IDF helps identify important terms within messages while maintaining computational efficiency, making it suitable for real-time monitoring applications. Studies have shown that TF-IDF representations combined with classical classifiers perform effectively on short and informal communication typical of student messaging platforms [6].

Classical machine learning models such as Logistic Regression and Support Vector Machines have been widely adopted in abusive language detection research. These models perform well on high-dimensional sparse text data generated from feature extraction methods like TF-IDF. Research has demonstrated that linear classifiers provide strong baseline performance while maintaining interpretability and low computational cost, making them suitable for real-time institutional deployment [7], [8]. Their transparency also supports responsible decision-making in educational environments.

Ensemble learning techniques have been explored to improve classification robustness and accuracy. Agrawal and Awekar proposed a cyberbullying detection system combining multiple classifiers to enhance prediction performance across social media datasets. Ensemble approaches can capture complex feature interactions and

improve generalization, though many implementations focus primarily on open social networks rather than controlled school-based platforms [9]. This highlights the need for adaptable systems designed for institutional contexts.

Hybrid detection strategies combining machine learning with rule-based filtering have gained attention for their ability to detect both explicit and implicit forms of harassment. Rosa and colleagues introduced a context-aware cyberbullying detection framework integrating linguistic rules with statistical classification to improve sensitivity to subtle bullying behaviors. Hybrid models have been shown to improve reliability in identifying high-severity harmful expressions while maintaining interpretability and efficiency [10], [11].

Some studies have explored incorporating contextual or user-based information into detection models. Dadvar et al. examined the use of user behavior and profile features to improve cyberbullying classification accuracy. Although this approach enhances predictive performance, reliance on personal user data introduces privacy concerns and limits applicability in school-based monitoring systems where data protection is essential [12]. As a result, text-focused detection methods remain more suitable for educational platforms.

Human oversight is increasingly recognized as a critical component of automated moderation systems. Humanin-the-loop frameworks allow machine learning models to flag potentially harmful content while leaving final decisions to administrators or moderators. Research indicates that combining automated detection with human review improves reliability, reduces false positives, and supports ethical deployment in safety-critical environments such as schools [13], [14].

Recent studies emphasize the need for detection systems tailored specifically to institutional communication environments. Many existing models are trained on open social media datasets, which differ significantly from structured school platforms in language use, moderation requirements, and privacy considerations. Prior research suggests that lightweight machine learning models combined with hybrid detection and administrative alert mechanisms provide a practical balance between accuracy, transparency, and deployment feasibility in educational settings. These findings motivate the development of school-oriented cyberbullying detection frameworks designed for early intervention and responsible monitoring [15],[25].

3. METHODOLOGY

3.1.1. Proposed Hybrid Model

The final CyberShield framework follows a hybrid detection approach that combines supervised machine learning with rule-based filtering to identify harmful student messages in school communication platforms. The objective is to detect both indirect bullying patterns present in everyday conversation and explicit abusive expressions that may require immediate attention. Incoming text messages are first processed and transformed into structured numerical representations using TF-IDF, enabling the system to capture important words and short phrase patterns commonly associated with bullying behaviour. These feature vectors are then analyzed using a trained Random Forest classifier, which predicts whether a message belongs to the bullying or non-bullying category based on patterns learned from labeled training data. This data-driven classification enables consistent detection while maintaining efficiency suitable for real-time school monitoring environments.

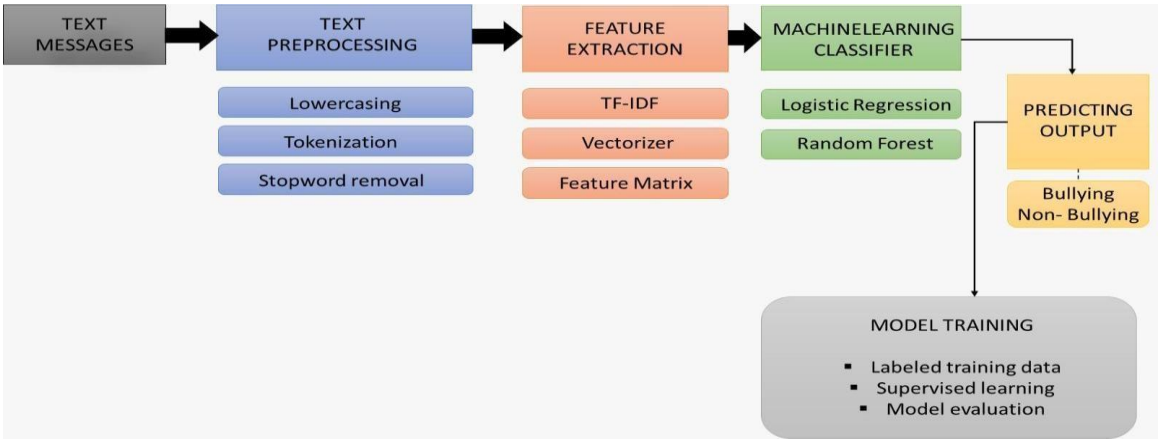

FIGURE 1: DETAILED ARCHITECTURE OF THE CYBERSHIELD SYSTEM

To improve reliability in high-risk scenarios, the framework includes an additional rule-based severity module that scans messages for strongly offensive or sensitive terms. Messages flagged by the classifier are further evaluated using predefined linguistic rules and a confidence-based decision mechanism to determine whether an administrative alert should be generated. This layered design allows the system to capture both subtle and explicit forms of harmful communication while reducing false alarms. By combining probabilistic machine learning predictions with deterministic filtering checks, the hybrid architecture supports faster processing, clearer interpretation of results, and practical deployment within school-managed digital platforms where timely intervention and transparency are essential.

3.1.2. Alert And Response Mechanism

CyberShield includes an alert generation module that activates whenever a message is identified as potentially harmful. The system flags the message and sends a notification to designated school administrators or moderators along with relevant context so that the interaction can be reviewed carefully. This ensures that automated detection supports human decision-making rather than replacing it. Administrative staff evaluate flagged messages and determine appropriate action, and no disciplinary measures are applied automatically based solely on model output.

3.1.3. Ethical Considerations and Human Oversight

Because student communication monitoring is sensitive, the system is designed with strong human oversight and ethical safeguards. Automated detection models may occasionally misinterpret slang, cultural expressions, or multilingual text, so CyberShield functions as a decision-support tool rather than an autonomous enforcement system. All final decisions remain with authorized school personnel, helping maintain fairness, transparency, and responsible intervention while protecting student privacy.

3.1.4. Design Scope and Justification

The methodology focuses on lightweight, interpretable machine learning techniques instead of large-scale deep learning or external social media integrations. Complex neural models often require extensive computational resources and large datasets and may introduce concerns related to bias and explainability. By using a traditional machine learning framework tailored to school communication environments, the system prioritizes reliability, clarity of predictions, and practical deployment within institutional platforms.

3.2. Multilingual Dataset Description

The dataset used for this study is based on a publicly available collection of annotated social media comments containing examples of harmful and non-harmful language. Each text sample is manually labeled according to whether it reflects abusive, offensive, or normal communication, allowing the system to learn patterns associated with bullying-related behaviour. For the purpose of this work, a curated subset of approximately 19,200 text messages was used to simulate communication patterns similar to those found in student messaging environments. To support practical deployment in school platforms, the original multi-class labels were simplified into two categories: toxic and non-toxic. This binary structure allows the system to identify potentially harmful messages more quickly and supports efficient real-time monitoring.

Table 1: Dataset Statistics and Class Distribution

Category	**Count (Approx.)**
Total Messages	12,000
Bullying samples	5,600
Non-Bullying samples	6,400
Avg.message length	12-25 words
Code-mixed samples	~30%
Languages	11

Table 2: Languages Included in the Multilingual Cyberbullying Dataset

Language ID	**Language**	**Script/Form**
L1	English	Latin
L2	Hindi	Devanagari
L3	Hindi-English(Hinglish)	Roman(Code-mixed)
L4	Marathi	Devanagari
L5	Gujrati	Gujrati
L6	Bengali	Bengali
L7	Tamil	Tamil
L8	Telugu	Telugu
L9	Kannada	Kannada
L10	Malayalam	Malayalam
L11	Punjabi	Gurmukhi/Roman

Before training, the dataset underwent a structured preprocessing pipeline to improve consistency and reduce noise in the text. All messages were converted to lowercase, and unnecessary elements such as URLs, special characters, and extra symbols were removed. Common stop words were eliminated, and Unicode normalization was applied to ensure consistent handling of multilingual and code-mixed text frequently seen in student communication. The processed dataset was then divided into training and testing sets using an 80:20 split, where the larger portion was

used for model learning and the remaining portion for evaluation. To ensure stable performance and reduce bias from a single data split, a five-fold cross-validation process was applied during training. This approach allowed the model to be tested across multiple data partitions, resulting in more reliable and consistent performance estimates.

TABLE 3: NLP Pipeline Overview

Stage	**Technique used**	**Purpose**
Text Cleaning	Regex,Unicode normalization	Noise Reduction
Tokenization	Language-aware tokenizers	Word separation
Feature Extraction	TF-IDF	Text-Numeric
Classification	Supervised ML	Bullying detection
Alerting	Rule + threshold	Admin notification

3.3 Cyberbullying-Aware Chatbot Module

The CyberShield system includes an integrated chatbot that provides students with guidance, awareness, and confidential reporting support related to online harassment. The chatbot allows users to seek help, understand responsible digital behaviour, and report bullying incidents without immediate exposure. It operates alongside the detection system so that when harmful messages are identified, students can receive instant assistance while administrators are notified separately. This ensures both early support and structured monitoring within the platform.

Technically, the chatbot follows a modular design with an intent recognition pipeline that interprets student queries and retrieves appropriate responses from a structured knowledge base. It maintains conversation context through session tracking and can shift into a supportive mode if signs of distress are detected, offering focused guidance and safety resources. The chatbot is embedded directly into the platform interface, enabling smooth integration into school systems while maintaining reliability and controlled responses.

4. Evaluation Matrix

TF-IDF Formula:

$$TF - IDF\ (t, d) = TF\ (t, d) \times \log\ (N / DF\ (t))$$

Where:

TF (t, d) = term frequency

DF (t) = document frequency

N = total number of documents

Table 4: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1
Logistic Regression	86%	0.85	0.84	0.84
SVM	89%	0.88	0.87	0.87

Random Forest	92%	0.91	0.90	0.90
Proposed Hybrid	94%	0.93	0.92	0.92

The proposed hybrid model achieved the highest performance by combining statistical machine learning classification with rule based severity detection. The model demonstrated consistent performance across validation folds and improved identification of high-risk bullying phrases compared to individual classifiers.

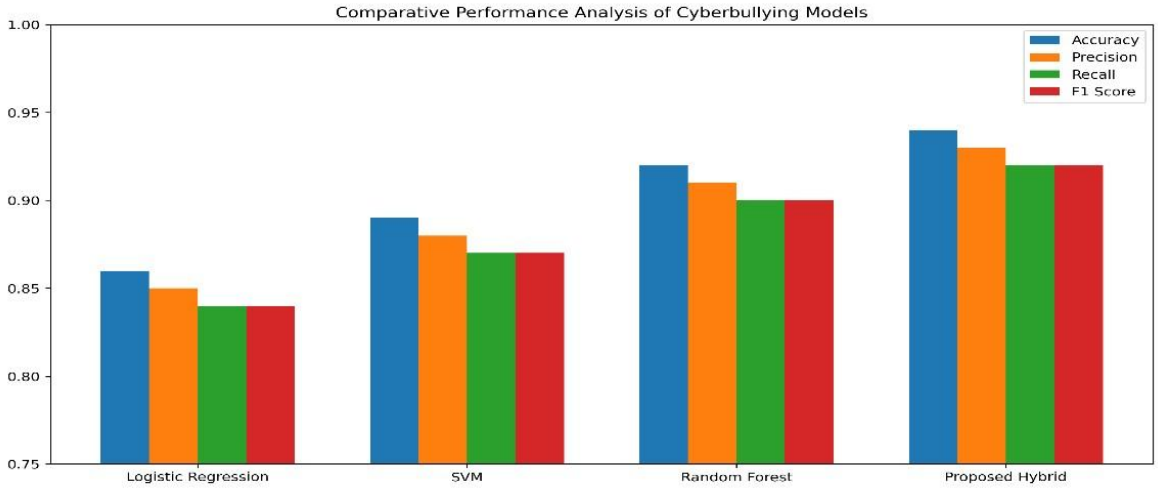

FIGURE 2: COMPARATIVE PERFORMANCE ANALYSIS OF CYBERBULLYING MODELS **5.**

RESULTS:

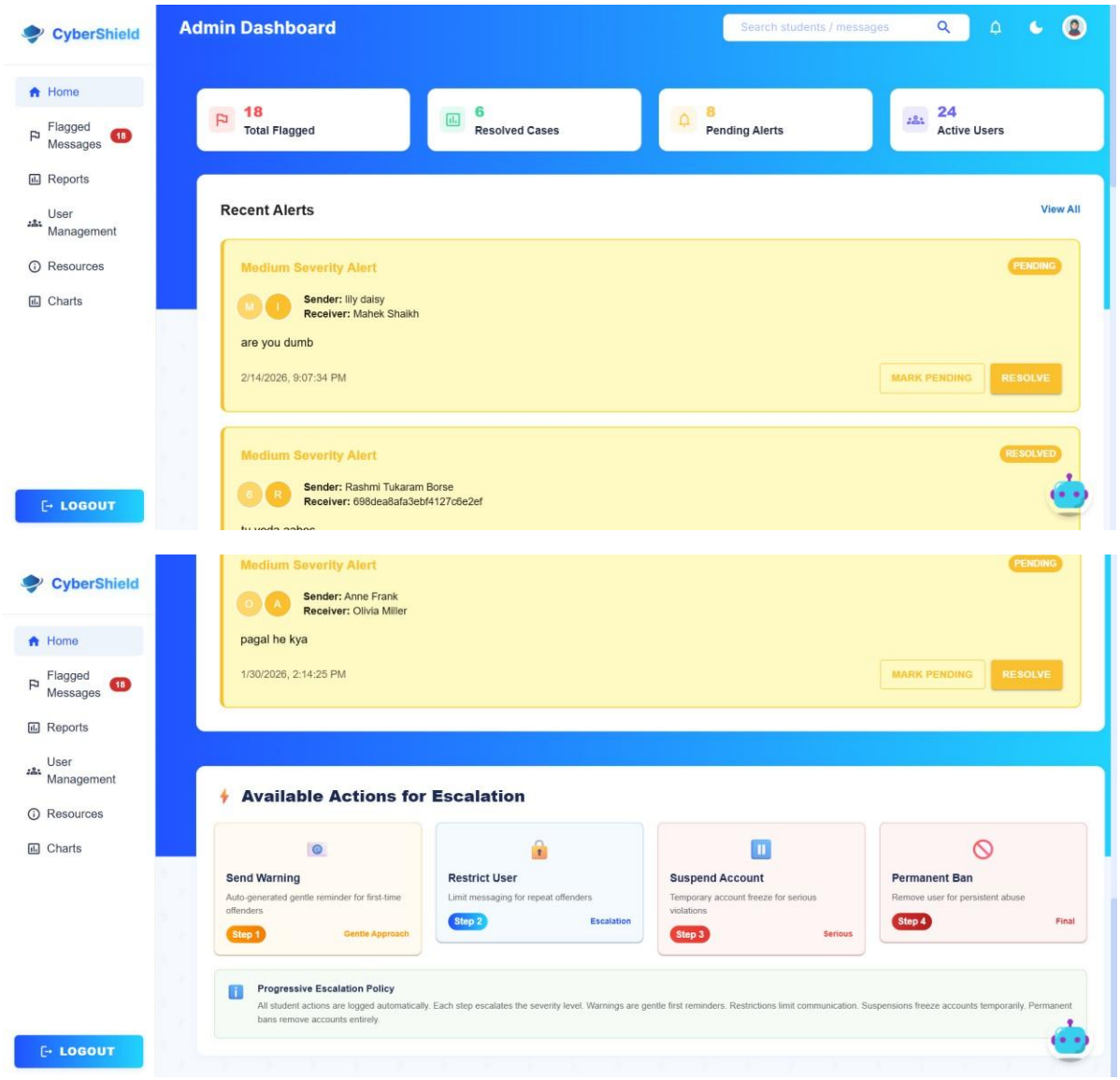

Figure 3 : Admin Dashboard Interface

The figure presents the Admin Dashboard of the CyberShield system, which enables school authorities to monitor and manage cyberbullying incidents efficiently. It provides a centralized overview of key system data, including flagged messages, resolved cases, pending alerts, and active users. The dashboard also features structured navigation modules and an escalation policy with four progressive actions- warning, restriction, temporary suspension, and permanent ban ensuring fair, consistent, and safety-focused intervention.

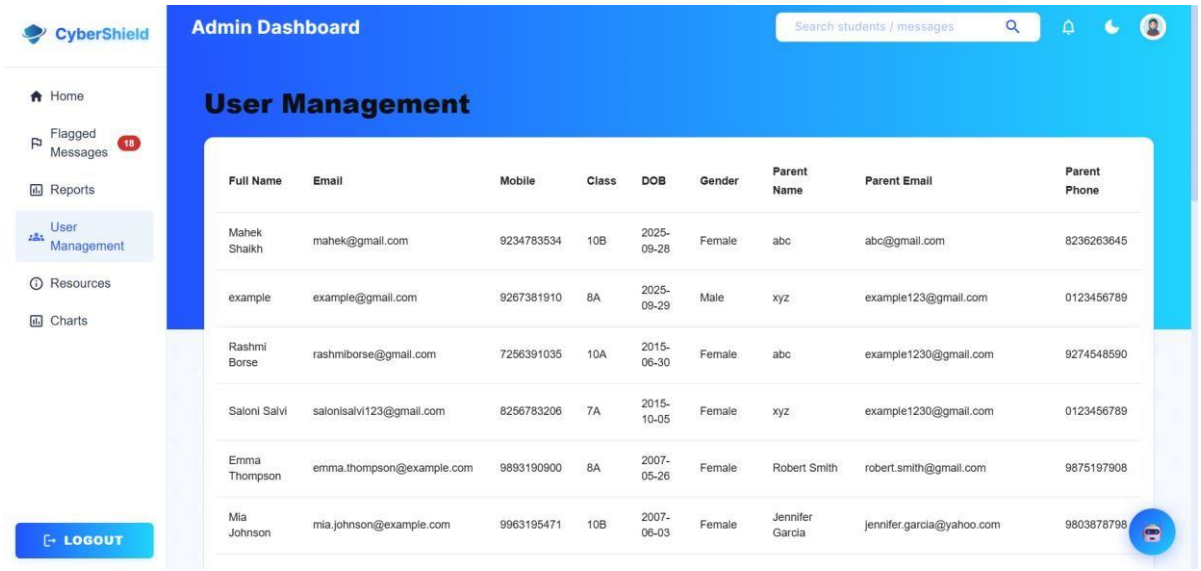

Figure 4 : User Management Data from Admin Dashboard

This figure displays the User Management module, which provides administrators with access to detailed student records, including personal information, class details, and parent contact information. The centralized table format enables efficient monitoring, verification, and communication, ensuring transparency and effective coordination with guardians when necessary.

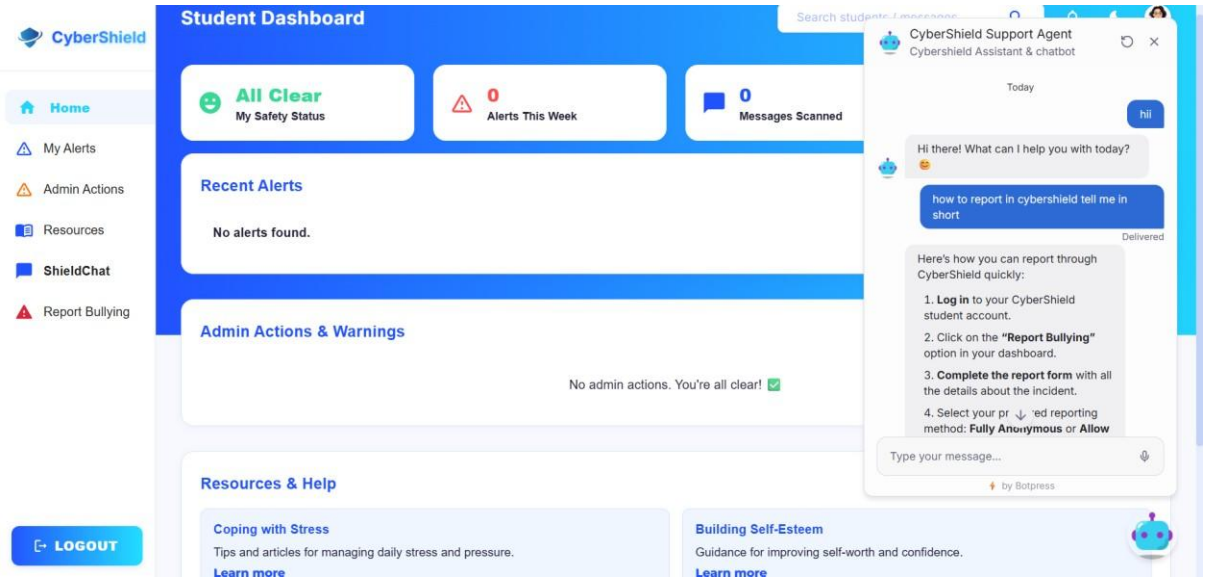

Figure 5 : CyberShield Support Agent - Student Assistance Chat Interface

This figure shows the CyberShield Support Agent integrated into the system. The chatbot acts as an AI-powered assistant that helps students quickly find information and guidance. It provides instant responses to questions about reporting bullying, account usage, safety features, and platform navigation. In the example shown, the student asks how to report bullying, and the chatbot gives clear step-by-step instructions, including logging in, selecting "Report Bullying", completing the form, and choosing a reporting method.

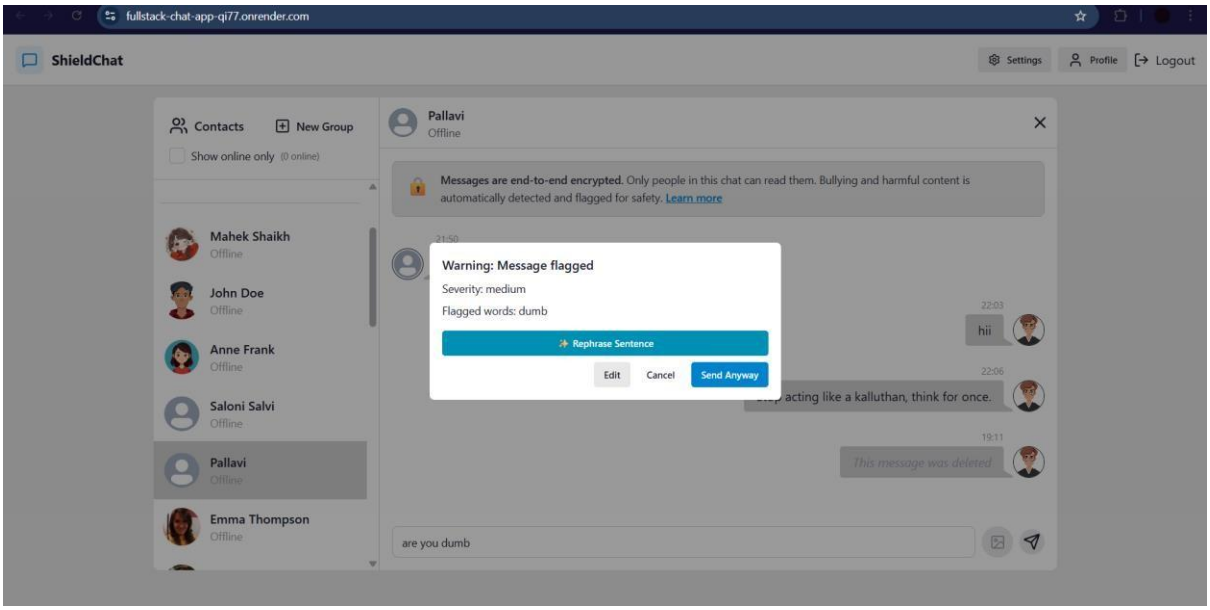

Figure 6 : Real- time message warning for flagged content in ShieldChat

This figure illustrates the student chat interface integrated with the cyberbullying detection mechanisms. When a harmful message is sent, the system automatically flags it and displays a warning popup indicating the severity level and detected offensive words. The user is given options to edit, rephrase, or send the message anyway. This proactive harmful communication before it is delivered, prompting responsible digital behavior.

6. CONCLUSION AND FUTURE WORK

This study presented a hybrid cyberbullying detection framework that combines statistical machine learning with rule-based severity analysis to support reliable real-time monitoring in school communication platforms. By integrating TF-IDF feature representation, a Random Forest classifier, keyword-based severity scoring, and ruledriven filtering, the system improves both detection performance and interpretability. The hybrid structure allows the model to identify subtle patterns of bullying as well as explicit harmful expressions while maintaining low computational cost and clear decision logic. Experimental evaluation showed that the combined approach performs better than individual baseline classifiers across standard performance metrics. The framework is designed for controlled educational environments where transparency, efficiency, and timely administrative intervention are essential, making it suitable for responsible deployment within school-managed digital systems.

Future work can focus on enhancing early-risk detection by analyzing behavioural patterns over time rather than relying only on individual messages. Incorporating sequence-level analysis and temporal pattern monitoring could enable earlier intervention before harmful behaviour escalates. Additional improvements may include refining severity scoring using insights from behavioural and educational research, as well as strengthening explainability features to help administrators better understand model decisions. Long-term deployment may also benefit from continuous monitoring mechanisms that adapt to changes in student language usage and communication trends, ensuring stable performance and sustained reliability in evolving digital environments.

Acknowledgements

The authors gratefully acknowledge the support and guidance provided by their project guide throughout the development of this work. The authors also recognize the collective efforts of all group members, whose consistent collaboration and shared responsibility were central to completing the project. The support and the academic environment provided by the department faculty are also sincerely appreciated.

REFERENCES

[1] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, 2018.

[2] M. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. SocialNLP*, Paris, France, 2017, pp. 1–10.

[3] B. Vidgen, S. Hale, E. Guest, H. Margetts, and L. Derczynski, "Detecting abusive language using supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 74, pp. 1–34, 2022.

[4] H. Rosa, P. Pereira, R. Ribeiro, et al., "Context-aware cyberbullying detection using hybrid NLP techniques," *Expert Systems with Applications*, vol. 209, 2023.

[5] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Proc. 35th European Conference on Information Retrieval (ECIR)*, Moscow, Russia, 2013, pp. 693–696.

[6] P. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. 39th European Conference on Information Retrieval (ECIR)*, Aberdeen, UK, 2018, pp. 141–153.

[7] B. Vidgen and L. Derczynski, "Directions in abusive language training data: A systematic review," in *Proc. ACL*, 2021.

[8] B. Mathew, P. Saha, H. Tharad, et al., "Thou shalt not hate: Countering online hate speech," in *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, 2021.

[9] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Modeling textual cyberbullying," in *Proc. ICWSM*, 2011.

[10] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. ICWSM*, Montreal, Canada, 2017, pp. 512–515.

[11] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. NAACL-HLT*, 2019.

[12] T. Mandl, S. Modha, P. Majumder, et al., "Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification," in *Proc. FIRE*, 2019.

[13] B. Vidgen, A. Yasseri, and H. Margetts, "HateXplain: A benchmark dataset for explainable hate speech detection," in *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[14] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research," *Psychological Bulletin*, vol. 140, no. 4, pp. 1073–1137, 2014.

[15] J. W. Patchin and S. Hinduja, "Cyberbullying and self-harm," *Journal of School Health*, vol. 80, no. 12, pp. 614–621, 2010.

[16] A. Bohra, D. Vijay, V. Singh, S. Akhtar, and M. Shrivastava, "A dataset of Hindi-English codemixed social media text for hate speech detection," in *Proc. EMNLP Workshop*, 2018.

[17] B. Vidgen and L. Derczynski, "Challenges and opportunities in abusive language detection," *Computational Linguistics*, 2021.

[18] A. Ganesh et al., "Multilingual cyberbullying detection using hybrid deep learning approaches," *IEEE Access*, 2024.

[19] S. García-Méndez and J. Arriba-Pérez, “Explainable cyberbullying detection using large language models,” *Expert Systems with Applications*, 2024.

[20] T. Brown et al., “Language models are few-shot learners,” in *Proc. NeurIPS*, 2020.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019.

[22] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Proc. NeurIPS*, 2020.

[23] Z. Waseem and D. Hovy, “Predictive features for hate speech detection on Twitter,” in *Proc. NAACL Student Research Workshop*, 2016, pp. 88–93.

[24] S. Badjatiya, M. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in *Proc. WWW Companion*, 2017, pp. 759–760.

[25] R. Saha, S. Sharma, and A. Saxena, “Cyberbullying detection: Challenges and future directions,” *IEEE Access*, vol. 9, pp. 10123–10135, 2021.