ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.206 **Volume 8, Issue 3, March 2025**

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803296

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Securechat – AI Powered Secure Messaging

Mr.H.M.Gaikwad[1] , Mrs. D.S. Chopada[2], Mayank Satish Dandane[3], Rachit Ritesh Kothadia[4], Khushbu Bhushan Mankare[5], Mansi Suhas Nehete[6]

Head, Department of Artificial Intelligence and Machine Learning, K.K Wagh Polytechnic, Nashik, India[1]

Lecturer, Department of Artificial Intelligence and Machine Learning, K.K Wagh Polytechnic, Nashik, India[2]

Students, Department of Artificial Intelligence and Machine Learning, K.K Wagh Polytechnic, Nashik, India[2,3,4,5]

ABSTRACT: The objectives of this system are to enhance personal safety by providing real-time detection and mitigation of harmful content in digital communication platforms. The system aims to protect users from exposure to abusive language, threats, cyberbullying, and other harmful content, with a particular focus on supporting vulnerable populations such as minors. By leveraging machine learning algorithms like Logistic Regression, Linear Support Vector Machines (SVM), and TF-IDF, the system is capable of detecting harmful content in both text and emoji-based communication. These algorithms are trained to recognize patterns associated with abusive behavior, bullying, and harmful interactions, allowing for the swift identification and flagging of potentially dangerous messages. A key feature of this system is the real-time detection and response mechanism, which allows for immediate action when harmful content is detected, such as providing warnings, notifying moderators, or even blocking offensive messages before they are fully delivered. This not only helps to mitigate the negative impact of abusive interactions but also fosters a safer, more respectful environment in which users can communicate freely without fear of harassment or abuse.Applications of the Personal Safety Assistant extend across a range of digital environments, including social media platforms, messaging apps, forums, and online gaming communities. By integrating this system into such platforms, the system contributes to creating a healthier online ecosystem, where abusive interactions are discouraged and users are encouraged to engage in more positive, respectful communication. The system can also be adapted for use in educational settings, workplaces, and public forums to ensure a broader societal impact, protecting users and promoting a culture of safety and inclusion in online spaces. The system ultimately addresses the growing concern around online harassment and harmful digital behavior, making it a vital tool for enhancing personal safety and promoting a healthier, more secure digital communication environment.

KEYWORDS: Personal Safety,Harmful Content Detection,TF-IDF, Linear (Support Vector Machine), Real-time Detection

I. INTRODUCTION

The rapid growth of digital communication platforms—ranging from social media to instant messaging apps—has brought significant benefits, enabling effortless global connectivity and information sharing. However, this expansion has also led to a surge in harmful online behaviors such as cyberbullying, harassment, abusive language, and direct threats. These behaviors pose severe psychological risks, particularly for minors and vulnerable users, resulting in anxiety, depression, and social isolation.

Despite various efforts to address these issues, the prevalence of harmful content remains a critical challenge. Current systems often rely on post-event analysis, where harmful content is detected after being sent, providing limited protection for users during live interactions. This system focuses on developing a Personal Safety Assistant capable of detecting and mitigating harmful content in real- time, specifically targeting abusive language and harmful emojis within chat platforms.

By integrating text and emoji analysis, this system enhances detection accuracy, addressing subtle forms of abuse such as emoji misuse for intimidation. Unlike traditional safety measures, this system emphasizes real-time functionality to prevent escalation and provide immediate interventions, bridging the gap in user protection during sensitive conversations.

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803296

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The central problem addressed is the lack of systems capable of detecting and mitigating harmful content in real-time. This system develops a real-time solution that identifies harmful messages, including abusive language, harmful emojis, and threats, as they are sent, ensuring a safer and more supportive digital communication environment.

II. LITERATURE REVIEW

The increasing prevalence of harmful content, such as hate speech, cyberbullying, and online harassment, on social media and digital communication platforms poses significant challenges to creating safe and respectful online environments. Existing solutions often fall short in detecting and mitigating such harmful behavior in real time, particularly across diverse platforms where communication styles vary. While various approaches, including supervised learning and deep learning techniques, have been explored for detecting offensive language, hate speech, and cyberbullying, there remains a gap in the ability to accurately detect harmful content at scale and in real-time, especially when accounting for nuanced expressions such as emojis, slang, and context- specific communication. This necessitates the development of more robust models capable of efficiently analyzing and mitigating harmful content in dynamic digital environments.

The proliferation of harmful content—including hate speech, cyberbullying, and online harassment— on digital platforms has highlighted the need for real-time, scalable solutions.

While significant research has been conducted to detect offensive language and harmful behaviors using supervised and deep learning techniques, existing methods face challenges in real-time detection, particularly for nuanced expressions like emojis and context-specific communication.

Several studies shaped this system: Anguita et al. [1] offered preprocessing insights for large-scale data, aiding our detection approach. Waseem and Hovy [2] and Yin et al. [5] used text-based features and machine learning for hate speech and harassment detection, informing our feature extraction.
Dinakar et al. [3] and Agrawal and Awekar [6] applied supervised and deep learning to cyberbullying, supporting our TF-IDF use. Zhang et al. [4] emphasized real-time feedback, aligning with our focus. Davidson et al. [7] backed our Logistic Regression and SVM approach, while Fortuna and Nunes [8] guided model selection with a hate speech review.

III. SYSTEM ARCHITECTURE

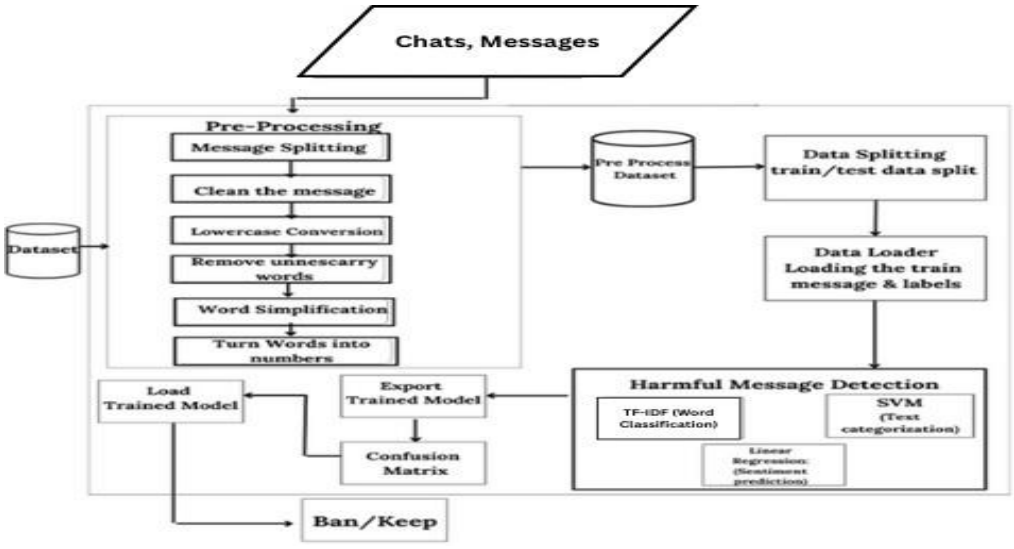

Figure 1. System Architecture

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803296

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In this system, we have trained our harmful content detection model using TF-IDF with a balanced dataset that contains both harmful and benign messages. This ensures that the model does not develop a bias towards either type of content. The system architecture is shown in the diagram. During the development phase, we collected a dataset consisting of real-world chat messages, pre-processed it by tokenizing words and emojis, and created a new structured dataset that captures both textual and emoji-based interactions. The system processes each incoming message for real-time classification, flagging harmful content as it is detected.

IV. METHODOLOGY

Online chat platforms often struggle with harmful content due to slow, inconsistent manual moderation, leaving users vulnerable. SecureChat introduces an automated, real-time detection system to address this issue effectively.

i. Solution to the Problem

The system analyzes each message as it is sent, using a pre-trained machine learning model to detect offensive or harmful content. Upon detection, it flags the message, issues warnings to the sender, tracks violations, and bans repeat offenders, while logging incidents for admin review. This ensures efficient moderation without disrupting genuine conversations.

ii. Data Preprocessing

Incoming messages undergo preprocessing to enhance detection accuracy. This includes tokenizing text into words, decoding emojis into textual descriptions for semantic analysis, and removing stop-words to focus on critical content. These steps standardize the data for consistent model performance.

Figure 2. Data Preprocessing

iii. Feature Extraction

A fine-tuned transformer model, such as BERT or GPT variants, processes preprocessed messages to extract contextual embeddings. These embeddings capture linguistic patterns, sentiment, and relational context, enabling accurate classification of harmful content.

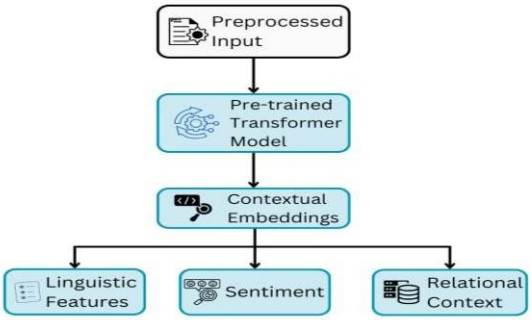

Figure 3. Feature Extraction

i. **Banning Mechanism**

If a user sends a harmful message, the system follows a 5-stage banning process based on the number of violations:

- Stage 1 (Flagged 1): 5-minute ban (User receives a warning).

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803296

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Stage 2 (Flagged 2): 1-hour ban (User is warned again).
- Stage 3 (Flagged 3): 5-hour ban (User is notified of stricter actions).
- Stage 4 (Flagged 4): 24-hour ban (Final warning before permanent action).
- Stage 5 (Flagged 5): Permanent ban (User is permanently blocked from the platform). Each time a user is flagged, they receive a notification about their current status and ban duration.

V. RESULT AND DISCUSSION

The effectiveness of the Personal Safety Assistant for Harmful Content Detection is evaluated based on its real-time detection capabilities, user interface efficiency, and overall system performance. The system integrates a deep learning model trained on diverse datasets to classify and flag harmful interactions in online chats. This section presents the outcomes of the implementation, showcasing the user interface, detection results, and system workflow.

Each figure below provides insights into various components of the system, including the graphical interface, classification results, performance metrics, automated reporting mechanisms, and system architecture. These results demonstrate the system's capability to accurately identify harmful content, ensure user safety, and facilitate proactive moderation.

i) User Interface:

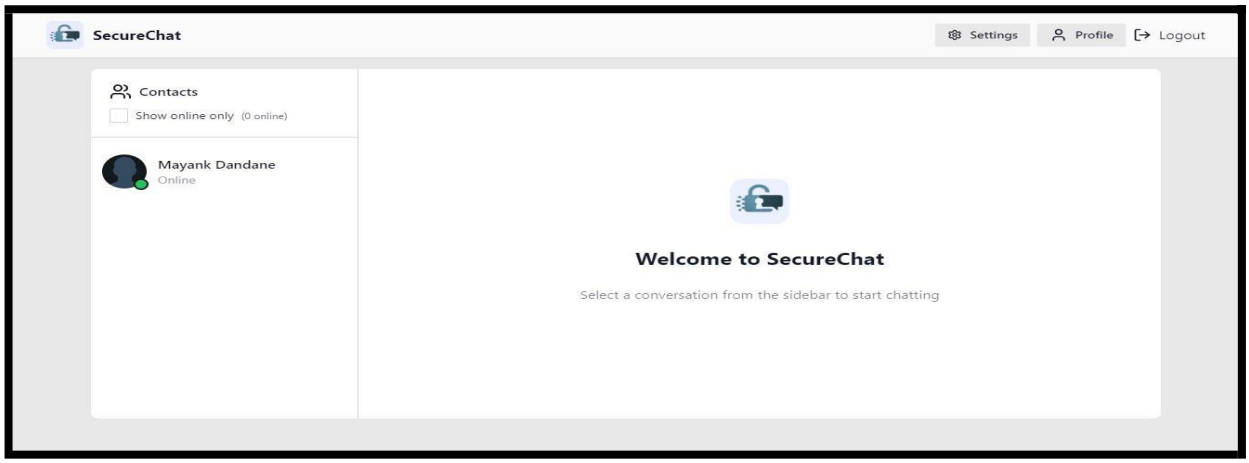

Fig 6. 1 User Interface

User Interface Overview Key Sections

a. Navigation Bar

- **i.** A top bar featuring the SecureChat logo and name on the left.
- **ii.** Buttons for Settings, Profile, and Logout on the right for quick access to user management options.

b. Contacts Panel (Left Sidebar)

- **i.** Contacts List: Displays available users for chatting.
- **ii.** Online Status Filter: A toggle button labeled "Show online only", allowing users to filter contacts.
- **iii.** Empty State Message: Displays "No online users" when no contacts are available.

c. Main Chat Window (Center Panel)

- **i.** Welcome Message: Displays a SecureChat logo along with the text "Welcome to SecureChat".
- **ii.** User Instruction: A message prompting users to "Select a conversation from the sidebar to start chatting."

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803296

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Design & Functionality Considerations

- **Light Theme UI**: A visually appealing interface with white-accented text/icons for contrast.
- **Minimalist Layout**: Focused on usability and easy navigation.
- **Smooth Transitions**: Ensures a seamless user experience when navigating between chats.
- **Responsive Design**: Optimized for different screen sizes and devices.

ii) Sender Using Abusive Language Across Multiple Languages in Chat:

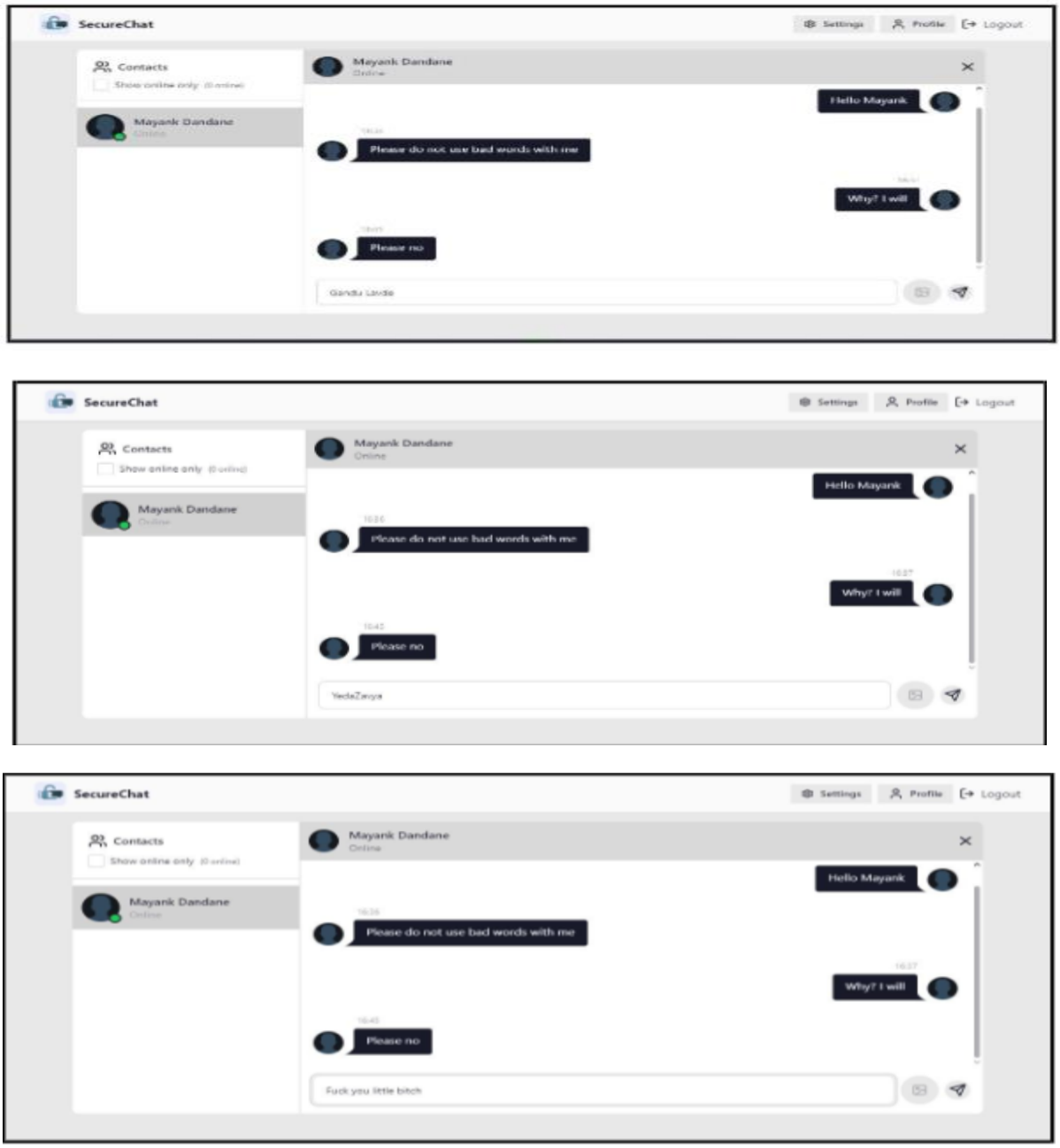

Fig 6. 2 Sender Using Abusive Language Across Multiple Languages in Chat

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803296

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Overview:-

Context: The images depict a chat interface where a sender engages in harmful communication using abusive language across three different languages—Hindi, English, and Marathi. Each message contains offensive words, emphasizing the challenges of moderating multilingual conversations in digital spaces. The system in place effectively detects these violations, ensuring that harmful interactions are promptly addressed.

Detection Mechanism: The chat application is equipped with an intelligent multilingual abusive language detection system, capable of identifying offensive words in real time. This detection process relies on advanced natural language processing (NLP) techniques and a predefined database of harmful terms across multiple languages. The system analyzes each incoming message, flags inappropriate content, and applies the necessary moderation rules. By supporting multiple languages, it demonstrates its robustness in safeguarding digital conversations regardless of linguistic diversity.

Action & Response: Upon detecting abusive language, the system takes immediate action based on predefined moderation policies. Depending on the severity and frequency of violations, users may face temporary suspensions or permanent bans. This ensures that repeat offenders are deterred from continuing harmful behavior. The system also logs flagged messages for further analysis, helping administrators refine detection algorithms and enhance moderation policies over time.

Significance: The ability to detect and moderate abusive language in real time is crucial for maintaining a safe, respectful, and inclusive online communication environment. As digital interactions increasingly span multiple languages, having an efficient and automated content moderation system prevents harassment, cyberbullying, and toxic behavior. By ensuring that conversations remain civil across linguistic barriers, this system fosters a more positive and secure digital space for all users.

iii) User Suspension & Ban Alert:

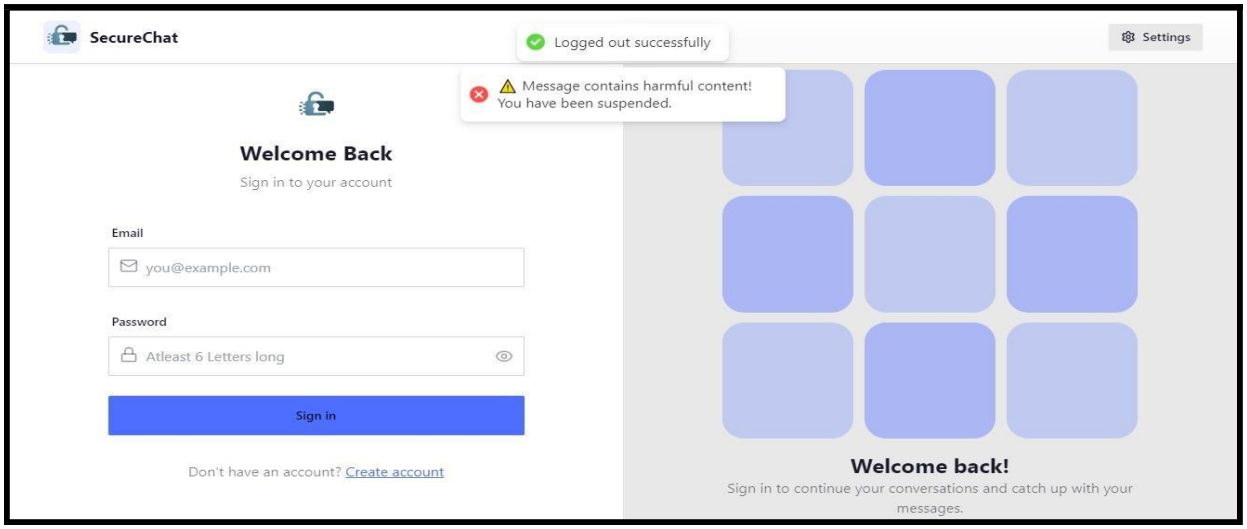

Fig 6. 7 User Suspension & Ban Alert

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803296

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This section covers suspension and banning mechanisms in SecureChat due to harmful content detection.

Key Features:

a. Real-Time Notifications:

i. Users receive instant alerts upon logout if their messages contain harmful content.

ii. Suspension alert appears as a pop-up:

- ⚠**"Message contains harmful content! You have been suspended."**
- If the suspension is temporary, users can log in once the suspension period ends.

b. Ban Implementation:

i. If a user repeatedly violates guidelines, they are permanently banned.

ii. A red alert message appears:

+ "The selected user has been banned due to the implementation of abusive words."

- Banned users cannot log in again.

VI. CONCLUSION

In summary, the Personal Safety Assistant for Harmful Content Detection is a crucial initiative aimed at ensuring safer digital communication by identifying and mitigating harmful content in real-time. Through the development of advanced machine learning and deep learning models, this system enhances online safety, particularly for vulnerable users, by preventing exposure to abusive language, threats, and harmful emojis. By integrating these tools across platforms, it fosters positive interactions and contributes to building a healthier and more respectful online community. This system is vital in adapting to the growing challenges of digital abuse and ensuring user safety in an increasingly connected world.

ACKNOWLEDGEMENT

We express our deepest gratitude to all those who guided and supported us throughout the journey of our system selection, design, and development. Their valuable insights, encouragement, and support have been instrumental in the successful completion of this system.

First and foremost, we extend our sincere thanks to **Prof. P. T. Kadave**, Principal, K. K. Wagh Polytechnic, Nashik, for granting us the permission and opportunity to undertake this system.

We are also profoundly grateful to **Prof. H. M. Gaikwad**, Head of the Artificial Intelligence & Machine Learning Department, for his continuous guidance, valuable suggestions, and timely feedback, which helped us navigate through challenges and improve our work.

Our heartfelt appreciation goes to our Internal Faculty Guide, **Mr. H. M. Gaikwad**, and the entire staff and technical team of the Artificial Intelligence & Machine Learning Department for their unwavering support, technical assistance, and encouragement throughout the system development.

We would like to extend a special note of gratitude to **Mr. Sachin Thete**, Founder of SachiTech, our system sponsor, whose sponsorship and belief in our system idea provided the necessary resources and motivation to bring this system to life. His support has been invaluable, and we are sincerely thankful for his contributions.

We are also deeply thankful to our system mentor, **Mrs. D. S. Chopada**, for her expert guidance and mentorship, which helped shape our system and enabled us to deliver a well-structured and impactful solution.
Lastly, our deepest appreciation goes to our parents for their unwavering support and belief in us, and to all our friends and well-wishers who contributed, directly or indirectly, to the successful completion of this system.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

1. R. P. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition Using Smartphones," Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 437-442, 2013.
2. Waseem, Z. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," Proceedings of the NAACL Student Research Workshop, pp. 88- 93, 2016.
3. K. Dinakar, R. Reichart, H. Lieberman, "Modeling the Detection of Textual Cyberbullying," Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 11- 17, 2011.
4. Z. Zhang, D. Luo, D. Li, M. Zhang, "Real-Time Deep Learning Model for Cyberbullying Detection in Online Chats," IEEE International Conference on Smart City and Social Media, pp. 243- 248, 2018.
5. D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, L. Edwards, "Detection of Harassment on Web 2.0," Proceedings of the Content Analysis in the WEB 2.0 Workshop, pp. 1-7, 2009.
6. S. Agrawal, A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," Proceedings of the European Conference on Information Retrieval (ECIR), pp. 141-153, 2018.
7. T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM), pp. 512-515, 2017.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com