International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

ChatEclipse360: Intelligent Toxic Concealment System

Chakka Naga Venkata Satya Sai Siri[1], Gonnuri Lalitha Sai Jayamani[2], Paidikondala Sarasvathi Sai Himaja[3], Ms. Ch Naga Padma Latha[4]

[1] CST, Sri Vasavi Engineering College(A), Pedatadepalli, Tadepalligudem – 534101.
[2] CST, Sri Vasavi Engineering College(A), Pedatadepalli, Tadepalligudem – 534101.
[3] CST, Sri Vasavi Engineering College(A), Pedatadepalli, Tadepalligudem –534101.
[4] Assistant Professor, Department CSE, Sri Vasavi Engineering College(A), Pedatadepalli, Tadepalligudem-534101

Abstract - *The rise of online communication has led to an increased prevalence of toxic messages that negatively impact user experiences. Our project focuses on developing an intelligent system to identify and mitigate toxic messages within chat applications. The proposed solution incorporates Natural Language Processing (NLP) and machine learning techniques for detecting and masking toxic content in real-time. When a sender attempts to send a message containing toxic words or phrases, the system immediately intervenes by masking the message. The receiver will see a placeholder indicating the message was identified as toxic, without revealing the specific toxic content. This preserves transparency without propagating harmful language. The sender receives a notification indicating that the message contains inappropriate content, encouraging more respectful communication behavior. This project aims to foster healthier online interactions by preventing the circulation of offensive or harmful language. Our solution is highly scalable, ensuring seamless integration into existing chat applications, while maintaining user privacy and optimizing performance*

***Key Words*: BERT, Concealment System, Content moderation, Real -time Chat Security, ML, NLP, Real-time moderation, Respectful communication, Toxic message detection, User safety, Message filtering, NLP-based moderation, Offensive language detection**

1. INTRODUCTION

In today's digital era, online communication platforms have become an integral part of daily interactions. However, the widespread use of these platforms has led to an increase in toxic and harmful messages, negatively impacting user experiences and fostering an unsafe digital environment.

To address this issue, our project, ChatEclipse360, introduces an Intelligent Toxic Message Concealment System designed to detect and moderate harmful content in real-time. Leveraging the power of Natural Language Processing (NLP) and Machine Learning, this system identifies toxic messages before they are delivered, ensuring that inappropriate content does not propagate further. Unlike traditional moderation tools that either completely block messages or rely on manual intervention, ChatEclipse360 takes a more refined approach by masking harmful messages while still notifying the sender. This feature helps educate users about respectful communication and prevents disruptions in conversation flow.

By implementing this solution, we aim to create a safer and more positive digital space where users can engage in meaningful and respectful conversations without the fear of encountering harmful content.

2. LITERATURE SURVEY

2.1**) Navoneel Chakrabarty.**"**A Machine Learning Approach to Comment Toxicity Classification" (2020)**. AI Research Journal. This study explores the use of machine learning techniques to classify toxic content based on categories such as threats and identity-based abuse obscenity, threats, insults, and identity-based hatred, to filter harmful content.[1]

2.2 **H. Masoorian, M. Ahmadi, N. Mohammadzadeh, and S. M. Ayyoubzadeh.** "**Research of Techniques used in Toxicity Detection" (2021).** International Journal of Artificial Intelligence. They proposed the use of neural networks and machine learning techniques for automatic detection of toxic comments, aiming to protect users from harmful online behavior.[2]

2.3 **P. G. Davange, P. Chaudhari, S. T. Patil, and A. Bhojawala**." **Toxic Chat Detection using Deep Learning**. **(2023)".** They proposed the use of machine learning and deep learning, particularly LSTM with BERT word embeddings, to categorize and filter toxic comments, achieving 94% accuracy.[3]

2.4 **D. Nithya, Nanthine K.S., Thenmozhi S., & Varshini Priya R.** "**Advanced social media Toxic Comments Detection System Using AI" (2024)**. They proposed the development of an automated system for detecting and flagging toxic comments in real-time on social media platforms using NLP and ML techniques.[4]

2.5 **Z Yang, D Tullo, R Rabbany." ToxiSight: Insights Towards Detected Chat Toxicity" (2024).**. They propose an explainability dashboard for in-game chat toxicity detection, integrating XAI techniques like token importance analysis, model visualization, and dataset attribution. The

dashboard provides nearest example insights, word sense analysis for moderators, and free-text explanations to improve interpretability.[5]

2.6 **George Beknazar-Yuzbashev, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment," (2025)** conducted a six-week field experiment using a browser extension to hide toxic content on Facebook, Twitter, and YouTube. The study found that reducing exposure to toxicity led to decreased advertising impressions, time spent on platforms, and other engagement metrics, as well as a reduction in the toxicity of user-generated content. A survey experiment indicated that toxic content triggers curiosity, suggesting a trade-off for platforms between curbing toxicity and maintaining user engagement.[6]

3. EXISTING SYSTEM

1. Limitations of Traditional Moderation

Manual or fixed filters are slow, ineffective in real-time, and may store user data without consent, raising privacy concerns.

2. Conversation Disruption

Some systems completely block messages, interrupting the natural flow of conversations and causing communication breakdowns.

3. Ineffective User Notification

Certain systems fail to notify users properly, making them less effective in promoting responsible and positive communication.

4. PROPOSED SYSTEM

- Uses NLP and Machine Learning to detect and filter toxic messages in real-time.
- Notifies senders for message revision instead of completely blocking messages.
- Replaces harmful messages with placeholders to maintain context.

Ensures seamless integration with chat platforms while maintaining user privacy.

5. ADVANTAGES

1. Real-Time Toxicity Detection:

Uses NLP and Machine Learning to accurately detect and filter toxic messages as they are sent, ensuring a seamless conversation flow.

2. User-Friendly Moderation:

Notifies the sender to revise toxic messages instead of blocking them entirely, promoting positive communication habits.

3. Context-Preserving Filtering:

Replaces harmful messages with placeholders, allowing the receiver to understand the context without exposure to toxicity.

4. Scalable & Privacy-Focused Integration:

Seamlessly integrates with chat platforms while maintaining user privacy, ensuring compliance with data protection policies.

6. METHODS

Methods for Toxicity Detection:

1.Data Collection and Preprocessing:

Input: Gather a dataset of online conversations, chat logs, or social media comments.

Preprocessing: Clean and preprocess the data by removing noise, special characters, and irrelevant content. Tokenize the text and normalize (lowercase, remove stop words, etc.) for better analysis.

2.Feature Extraction:

NLP-Based Feature Extraction: Use NLP techniques like **TF-IDF (Term Frequency-Inverse Document Frequency)** or word embeddings such as **Word2Vec** or **BERT embeddings** to convert textual data into numerical vectors that capture the semantic meaning of the words.

Toxicity Features: Identify specific features related to toxicity, such as sentiment analysis, frequency of offensive words, or context of certain phrases.

3.Model Selection and Training:

Supervised Learning Models: Train machine learning models such as **SVM**, **Random Forests**, or deep learning models like **LSTM** or **CNN** to classify messages as toxic or non-toxic. Models can be trained using labeled datasets of toxic and non-toxic messages.

Pre-trained BERT Model: Fine-tune a **BERT** (Bidirectional Encoder Representations from Transformers) model for the task of toxicity detection. This approach utilizes contextual understanding and semantic knowledge to capture nuances in language.

4.Real-Time Moderation and Masking:

Toxicity Detection: Once a model is trained, use it to classify messages in real-time. If a message is flagged as toxic, mask the message by replacing offensive words with placeholders or symbols.

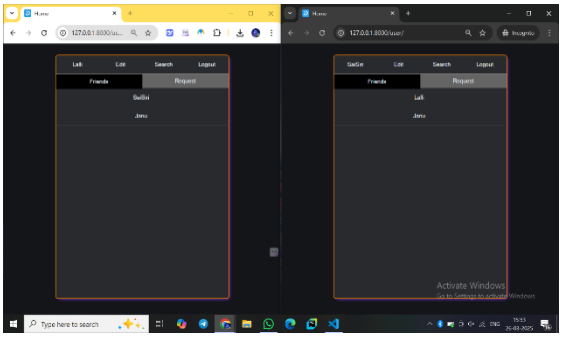
7. RESULT AND DISCUSSION

7.1. Login /SignUp:

Users can create a new account or log in using their credentials to access the chat application. Secure authentication ensures user data privacy and prevents unauthorized access.

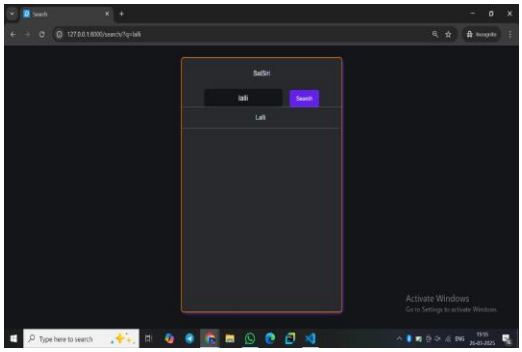
7.2. Home Page:

The home page provides an intuitive interface displaying active chats, user suggestions, and recent conversations. Users can navigate seamlessly to different sections of the chat application.



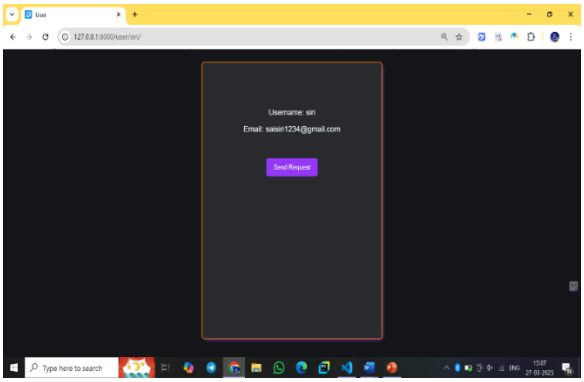
7.3. User Search:

Users can search for other registered members using their username or profile details. The search function helps in quickly finding and initiating conversations with specific users.



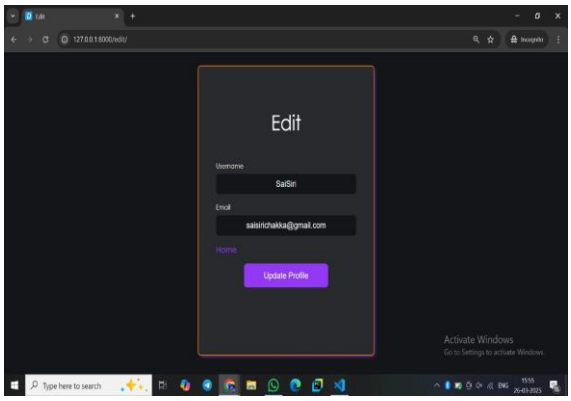
7.4. User Request:

Users can send connection requests to others before initiating a chat. Upon acceptance, both users can engage in seamless communication.



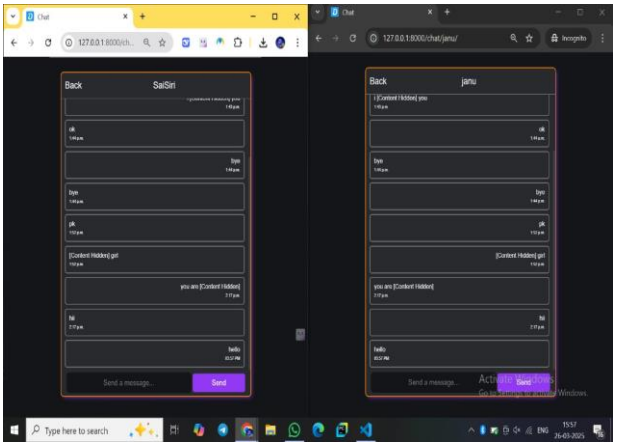
7.5. Update Profile:

Users can edit and update their profile details, such as name, bio, and profile picture. Profile updates ensure personalized interaction and better user engagement.



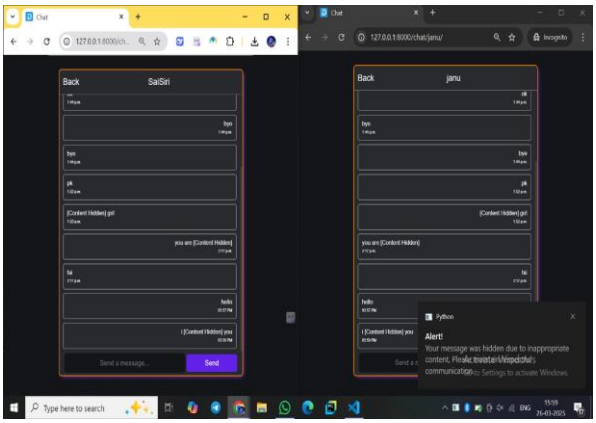
7.6. Chat Interface:

The chat interface supports real-time messaging with text input, emojis, and media sharing. It also integrates message moderation to filter toxic content before displaying it to users.



7.7. Result: (Message Moderation and Toxic Message Notification):

When a user sends a toxic message, the system automatically detects and masks harmful content. The sender receives a notification about inappropriate language, promoting responsible communication.



8. CONCLUSION

ChatEclipse360 is a chat moderation system that detects and filters toxic messages in real time. It notifies users about inappropriate content and replaces harmful words with placeholders instead of blocking messages. The system integrates seamlessly with chat platforms while ensuring user privacy. Its scalable design allows for smooth deployment without disrupting communication. Future improvements will enhance accuracy, support multimedia moderation, and adapt to evolving online interactions.

9. ACKNOWLEDGEMENT

We are really grateful to **Ms Ch Naga Padma Latha, M.Tech, Assistant Professor,** Department of Computer Science and Engineering, for all of her help and assistance during the creation of this project. Her meticulous critiques, insightful observations, and continuous encouragement have been crucial in helping to shape and bring our research to completion.

We also extend our appreciation to Dr.D. Jaya Kumari and Dr.G.V.N.S.R. Ratnakar Rao for cooperation and encouragement.

Lastly, we also acknowledge the unwavering support of the staff of the Department of Computer Science and Engineering played a crucial role in facilitating our project work, and we thank them for their contribution

REFERENCES

[1] N. Chakrabarty, A Machine Learning Approach to Comment Toxicity Classification, arXiv preprint, arXiv:1903.06765, 2019, https://arxiv.org/abs/1903.06765

[2] H. Masoorian, M. Ahmadi, N. Mohammadzadeh, S. M. Ayyoubzadeh, Research of techniques used in toxicity detection, AIP Conference Proceedings, Volume 2705, Issue 1, 2023, Article 030005, ISSN 0094- 243X, https://doi.org/10.1063/5.0143130

[3] P. G. Davange, P. Chaudhari, S. T. Patil, and A. Bhojawala, "Toxic Chat Detection using Deep Learning," International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), vol. 11, no. 1, pp. 7–11, Jan. 2024. https://ijercse.com/article/CII_721563_Pratik%20Gopal%20Davange_IJERCSE%20--%20AL%20--%20%207%20%20to%20%2011%20(1).pdf

[4] D. Nithya, K. S. Nanthine, S. Thenmozhi, R. Varshini Priya, Advanced social media Toxic Comments Detection System Using AI, International Journal for Research in Applied Science and Engineering Technology (IJRASET), Volume 12, Issue IV, April 2024, https://www.ijraset.com/researchpaper/advanced-social-media-toxic-comments-detection-system-using-ai

[5] Z. Yang, D. Tullo, and R. Rabbany, "ToxiSight: Insights Towards Detected Chat Toxicity," Blackbox NLP 2024, 21 Sept 2024, https://openreview.net/forum?id=iL6zxTh2HW

[6] G. Beknazar-Yuzbashev, R. Jiménez-Durán, J. McCrosky, M. Stalinski, Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment, CAGE Working Paper Series, 2025, https://warwick.ac.uk/fac/soc/economics/research/centres/cage/publications/workingpapers/2025/toxic_content_and_user_engagement_on_social_media_evidence_from_a_field_experiment/

BIOGRAPHIES

Chakka Naga Venkata Satya Sai Siri I am pursuing my fourth year in the Department of Computer Science and Technology at Sri Vasavi Engineering College (VSVT), Tadepalligudem, affiliated with JNTU Kakinada, AP. I am a proactive and dedicated individual with a strong drive to continuously improve and adapt to new challenges..

Gonnuri Lalitha Sai Jayamani
Iam Pursuing fourth year in the Department of Computer Science and Technology at **Sri Vasavi Engineering College (VSVT),** Tadepalligudem, affiliated with JNTU Kakinada, AP. I am an enthusiastic and adaptable individual, always eager to take on new opportunities and challenges for personal and professional growth.

Paidikondala Sarasvathi Sai Himaja.
I am Pursuing fourth year in the Department of Computer Science and Technology at **Sri Vasavi Engineering College (VSVT),** Tadepalligudem, affiliated with JNTU Kakinada, AP. I am a committed and results-driven individual, always striving for excellence and ready to embrace challenges that foster continuous learning and growth.

Ch.N.P.Latha ,
completed my Masters in computer Science and Engineering at Sri Sai Aditya College of Science and Technology, Surampalem. Presently working as a Assistant Professor in Computer science and Technology department at Sri Vasavi Engineering College, Pedatadepalli Wes Godavari District. I could deal with different academic projects in Web technology and Machine learning and also train-up the students in R, Python, .NET technologies.