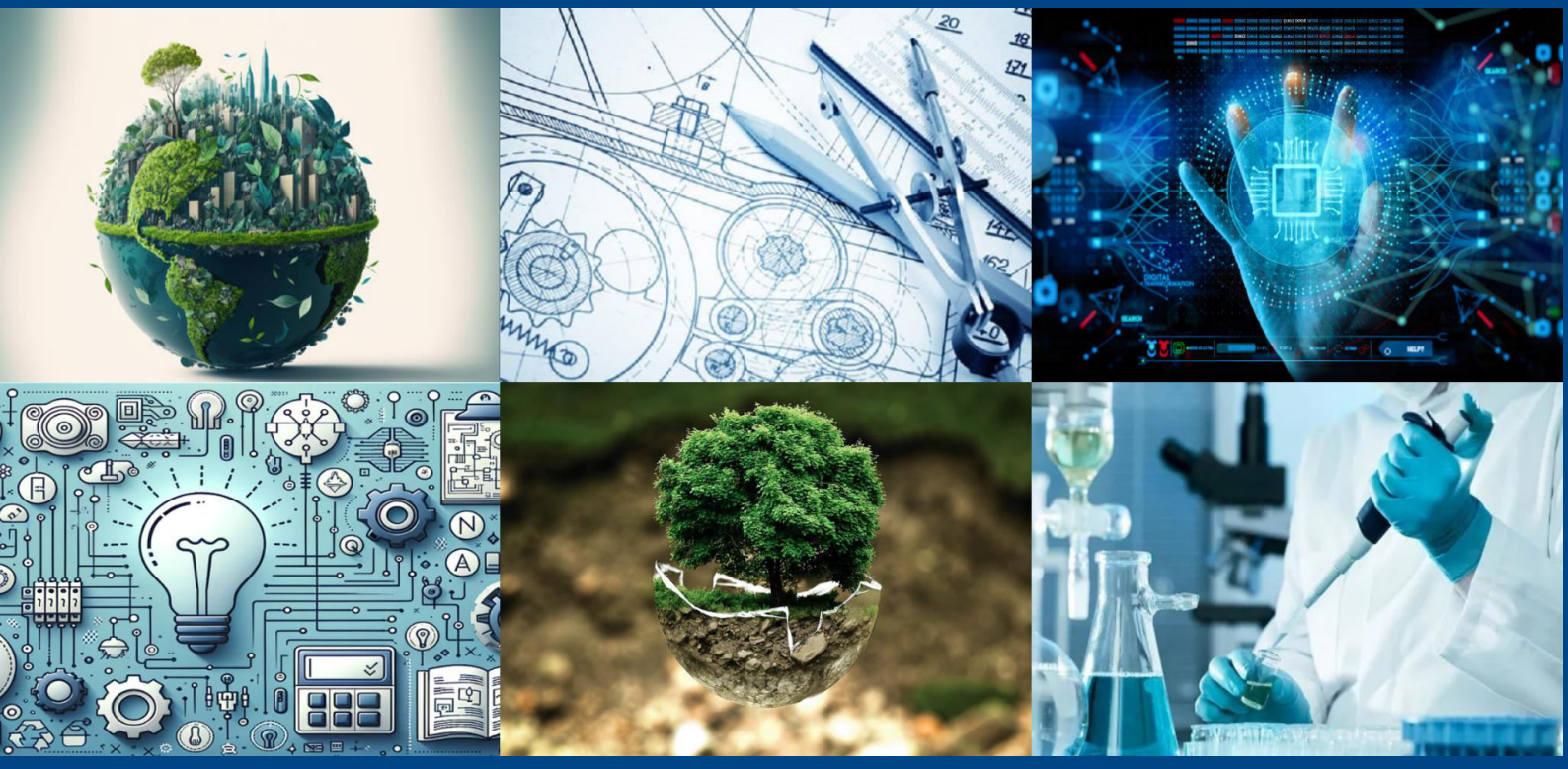
ISSN: 2582-7219

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 5, May 2025**

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI: 10.15680/IJMRSET.2025.0805223

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

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

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

Smart Chat Guardian

Dr.Nisha Dayana, Poongody S

Associate Professor, Department of Computer Science and Information Technology

Vels Institute of Science, Technology and Advanced Studies, Chennai, India

Student, Department of Computer Science and Information Technology

Vels Institute of Science, Technology and Advanced Studies, Chennai, India

ABSTRACT: Abstract – Smart Chat Guardian Smart Chat Guardian is an AI-driven chat moderation system that leverages Machine Learning (ML) and Natural Language Processing (NLP) to ensure a safe and respectful online communication environment. The system automatically detects and filters offensive, toxic, and inappropriate messages in real-time, enforcing progressive penalties such as warnings, temporary mutes, and bans based on user behavior. Additionally, it enhances user interactions by providing smart message suggestions, encouraging polite communication. The system also incorporates emoji-based sentiment analysis to detect aggressive intent and supports multi-language filtering for diverse online communities. Admins can customize chat rules using an AI-powered keyword management system, making moderation flexible and adaptable. By integrating deep learning models for text classification and reinforcement learning for user behavior tracking, Smart Chat Guardian ensures an intelligent, automated, and efficient approach to maintaining a positive and inclusive chat environment across digital platforms

KEYWORD: Smart Chat Guardian is an AI-based chat moderation tool that filters offensive and spam messages.

I. INTRODUCTION

In today's digital era, online communication has become an essential part of our daily lives. However, with the rise of online platforms, toxic behavior, offensive messages, and cyberbullying have also increased significantly. To address this growing concern, we present Smart Chat Guardian, an intelligent chat moderation system powered by Machine Learning (ML) and Natural Language Processing (NLP).

Smart Chat Guardian is designed to detect, filter, and prevent harmful or offensive content in real-time chats. Unlike traditional keyword-based systems, our project uses AI models to understand the context, sentiment, and intent behind messages, including those written in multiple languages or using emojis.

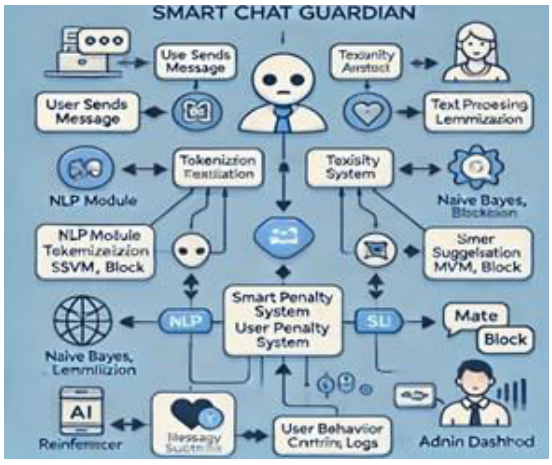


Fig. 1: System Architecture

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI: 10.15680/IJMRSET.2025.0805223

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

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

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

II. LITERATURE REVIEW

Earlier systems used simple keyword filters, but they failed to understand context. Recent works using ML and NLP (like Naive Bayes, SVM, LSTM) improved detection. Our system builds on these by adding emoji analysis, multi-language support, and user behavior tracking for smarter moderation.

III. SYSTEM ANALYSIS PROPOSED SYSTEM

We propose an intelligent, AI-driven chat moderation system that ensures a safe and respectful communication environment in real time.

Key Points:

- NLP Integration: Analyzes text using preprocessing, tokenization, and lemmatization to understand message context.
- ML Classifier: Uses Naive Bayes, SVM, or Deep Learning to classify messages as safe or toxic.
- Emoji Sentiment Analysis: Interprets emotional tone from emojis to enhance understanding.
- Smart Penalty System: Automatically warns, mutes, or blocks users based on message toxicity.
- User Behavior Tracker: Uses reinforcement learning to monitor and adapt to user behavior over time.
- Multi-language Support: Detects harmful content across different languages.
- Admin Dashboard: Allows keyword management, log viewing, and system customization.

Message Suggestions: Recommends friendly alternatives to offensive messages.

IV. EXISTING SYSTEM

Most existing chat moderation systems rely heavily on rule-based approaches or static keyword filtering. These systems simply block predefined offensive words or phrases, offering limited flexibility and poor contextual understanding. Users can easily bypass these filters by using alternate spellings, symbols, or slang variations.

Some platforms have adopted basic Machine Learning models like Naive Bayes or Logistic Regression for toxic message classification. However, these models often lack accuracy due to insufficient training on diverse datasets and fail to handle multi-language inputs or emoji-based sentiments.

Moreover, existing systems rarely track user behavior over time or adapt penalties dynamically. They do not offer features like smart message suggestions, admin keyword customization, or reinforcement learning-based behavior analysis.

Overall, the existing systems are static, limited in scope, and inefficient for real-time, multilingual, and context-aware moderation—highlighting the need for a more robust, intelligent solution like Smart Chat Guardian.

V. IMPLEMENTATION MODEL

The implementation of Smart Chat Guardian is divided into multiple modular components, each responsible for a specific function within the chat moderation system. The overall model ensures real-time message processing, intelligent classification, and adaptive user management.

1. Message Input Module: Captures and processes text, emojis, and multi-language messages.
2. NLP Processing Unit: Cleans and structures input data through tokenization, stopword removal, and lemmatization.
3. Emoji Sentiment Analyzer: Analyzes emotional tone conveyed by emojis, enhancing sentiment accuracy.
4. Machine Learning Classifier: Classifies messages into Safe, Offensive, or Highly Toxic using algorithms like Naive Bayes, SVM, and LSTM.
5. Toxicity Detection Engine: Flags messages that exceed a predefined toxicity threshold.
6. Smart Penalty System: Applies penalties based on message toxicity and past user behavior.
7. User Behavior Tracker: Tracks and adapts to user behavior using reinforcement learning.
8. Message Suggestion Module: Suggests polite alternatives to flagged messages.
9. Admin Dashboard: Provides tools for managing keywords, logs, thresholds, and user behavior insights.

Technologies: Python, TensorFlow, Scikit-learn,NLTK, React.js

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

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

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

VI. CONCLUSION

The Smart Chat Guardian system presents a comprehensive and intelligent solution for real-time chat moderation using advanced AI technologies. By integrating Machine Learning, NLP, deep learning, and reinforcement learning, the system successfully detects, classifies, and manages toxic messages with high accuracy. Its modular design enables scalability and flexibility, while features like emoji sentiment analysis, smart message suggestions, and customizable penalty mechanisms promote respectful and engaging communication. With its real-time performance and adaptability to multilingual inputs, the system addresses the growing need for effective content moderation in digital platforms.

Future Enhancements

To further enhance the system's capabilities, the following developments are proposed:

1. Voice and Video Chat Moderation: Integration of voice-to-text transcription and video content analysis for moderating voice/video chats.
2. Advanced Multilingual NLP Models: Incorporating transformer-based models (e.g., BERT, mBERT, XLM-R) for improved multi-language understanding.
3. User Reputation Scoring: Introducing a dynamic scoring system to track long-term user behavior and adjust moderation sensitivity.
4. Gamification for Positive Behavior: Rewarding users with badges or ranks for consistent positive communication.
5. Integration with Third-party Platforms: Providing APIs and plugins for integration with popular chat services like Discord, Slack, and gaming platforms.
6. Real-time Feedback Loop: Enabling users to flag or correct moderation errors, allowing the model to improve over time.

REFERENCES

1. Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Stanford University.
2. Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. Semantic Web, 10(5), 925–945.
3. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
4. Pedregosa, F., et al. (2011).Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova,

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