JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Real-Time Multilingual Toxicity Detection and Moderation in YouTube Live Chats

[1]Gautam Mehta, [2]Asst. Prof Trupti Asolkar

[1]Student, [2]Assistant Professor
[1]Department of Advanced Computing
[1]Nagindas Khandwala College, Mumbai

Abstract : Online social media and live interaction platforms are increasingly impacted by toxic and abusive language, making real-time content moderation essential. Early detection of toxic messages helps prevent cyberbullying, protect users, and maintain healthy online environments. Most existing approaches focus on a single language or platform, limiting their effectiveness in real-world, code-mixed scenarios. This work presents a web-based toxicity detection system using machine learning and deep learning techniques to analyze live chat and social media comments, including Hinglish and informal text. The system generates toxicity scores and moderation decisions using transformer-based models trained on publicly available datasets. An interactive dashboard visualizes real-time toxicity levels and actions, bridging the gap between research-oriented models and practical content moderation systems.

IndexTerms - Real-Time Toxicity Detection, Cyberbullying Detection, NLP-based Text Analysis, Transformer Models, Hinglish Language Processing, Live Chat Moderation, Machine Learning, Content Moderation Dashboard.

I. INTRODUCTION

Online video streaming platforms have grown rapidly, leading to increased real-time user interaction through live chat systems. While live chats enhance engagement, they also facilitate the spread of toxic, abusive, and inappropriate content, which negatively impacts user experience and platform safety [7], [8]. Due to the high volume and fast-paced nature of live chats, manual moderation is often ineffective and unable to respond promptly to harmful messages [3], [6].

Recent research has demonstrated the effectiveness of machine learning and deep learning techniques for automated toxicity detection in online text data [4], [11], [13]. These approaches generate toxicity scores or classifications to support moderation decisions such as message removal or user banning [1], [2]. However, most existing systems focus on static social media content or offline datasets, with limited emphasis on real-time moderation in live streaming environments [10], [12].

To address this gap, the YouTube Live Chat Monitoring and Toxicity Detection System is proposed. The system automatically retrieves live chat messages from YouTube streams and analyzes them in real time using a machine learning–based toxicity detection model [2], [4]. Based on the predicted toxicity scores, messages are categorized into moderation actions including ALLOW, DELETE, or BAN, and displayed through a web-based dashboard for real-time monitoring [3], [15]. This approach reduces manual moderation effort, improves response time, and contributes to safer live streaming environments [1], [14].

II. Literature review

Machine learning and deep learning techniques have been extensively used for toxicity and cyberbullying detection in online platforms. Early studies employed traditional classifiers such as Support Vector Machines (SVM) and Random Forest, achieving reasonable accuracy but lacking contextual understanding [1], [5].

With the advancement of deep learning, CNN and LSTM-based models improved performance by learning semantic patterns in abusive text [8], but these approaches struggled with implicit toxicity and informal language. Transformer-based models such as BERT significantly enhanced contextual toxicity detection and showed superior results across multiple datasets [3], [11].

Recent research has highlighted the challenges of multilingual and code-mixed languages like Hinglish, where standard English-trained models fail to detect slang and casual abuse effectively [7], [9]. A few studies proposed real-time moderation frameworks, but most lacked adaptive strictness and automated moderation actions [10], [12].

These limitations indicate the need for a real-time, context-aware toxicity detection system with configurable strictness and live moderation capabilities.

III. PROPOSED METHODOLOGY

The proposed system focuses on developing a real-time, web-based toxicity detection and moderation framework for live chat and social media environments. The methodology combines natural language processing (NLP), machine learning, and rule-based decision agents to identify, classify, and moderate toxic content effectively.

Data Collection

The training and evaluation datasets consist of toxic and non-toxic user-generated text collected from publicly available sources, including social media platforms, cyberbullying datasets, and benchmark toxicity corpora [1], [7], [10]. Special emphasis was placed on datasets containing informal language, slang, and code-mixed text (e.g., Hinglish), which reflect real-world communication patterns.

Data Preprocessing

Raw text data was preprocessed to remove noise such as URLs, emojis, special characters, and repeated tokens. Token normalization and lowercasing were applied to improve model generalization. This preprocessing strategy follows best practices adopted in prior toxicity detection studies [5], [13].

Feature Representation

Textual features were extracted using transformer-based embeddings that capture semantic and contextual meaning beyond individual words. Context-aware representations were chosen to address limitations of traditional bag-of-words models, particularly for implicit toxicity and sarcastic expressions [3], [8].

Toxicity Modeling

A transformer-based classification model was employed to assign toxicity probability scores to each incoming message. The model was selected due to its strong performance in multilingual and contextual toxicity detection tasks, as reported in recent literature [6], [11].

Decision & Moderation Logic

Based on the predicted toxicity score, a decision agent categorizes messages into ALLOW, WARN, DELETE, or BAN actions. Thresholds can be dynamically adjusted to support different moderation strictness levels. Similar adaptive moderation strategies have been explored in earlier research [9], [14].

Deployment & Integration

The trained system is deployed as a web-based application with a real-time dashboard. Live chat messages are continuously fetched, analyzed, and visualized for moderators, enabling instant intervention. The modular backend design ensures scalability and compatibility with multiple platforms [7], [12].

IV. SYSTEM ARCHITECTURE AND IMPLEMENTATION

User Interface:

A web-based dashboard allows moderators to input a live stream or channel identifier and monitor incoming chat messages in real time. The interface visualizes chat content, toxicity scores, moderation decisions, and system statistics in an intuitive manner.

Data Ingestion and Pre-processing Module:

Live chat messages are continuously fetched from the platform's API and passed through a preprocessing pipeline. This module cleans the text by removing noise such as URLs, emojis, and redundant characters, and normalizes informal and code-mixed language to improve detection accuracy [5], [13].

Toxicity Detection Models:

The core analytical layer consists of transformer-based machine learning models trained to classify messages as toxic or non-toxic and to assign toxicity probability scores. These models are selected due to their strong contextual understanding and proven effectiveness in multilingual and social media environments [6], [11].

Decision and Moderation Engine:

The prediction outputs are processed by a rule-based decision engine that maps toxicity scores to moderation actions such as ALLOW, WARN, DELETE, or BAN. Adjustable thresholds support different moderation strictness levels, enabling adaptive content control as suggested in prior studies [9], [14].

Web Deployment and Integration Layer:

The complete system is deployed as a web application with a lightweight backend and real-time APIs. This layer ensures seamless integration with live platforms and provides cross-device compatibility for desktops, laptops, and mobile devices, allowing continuous monitoring and moderation [7], [12].

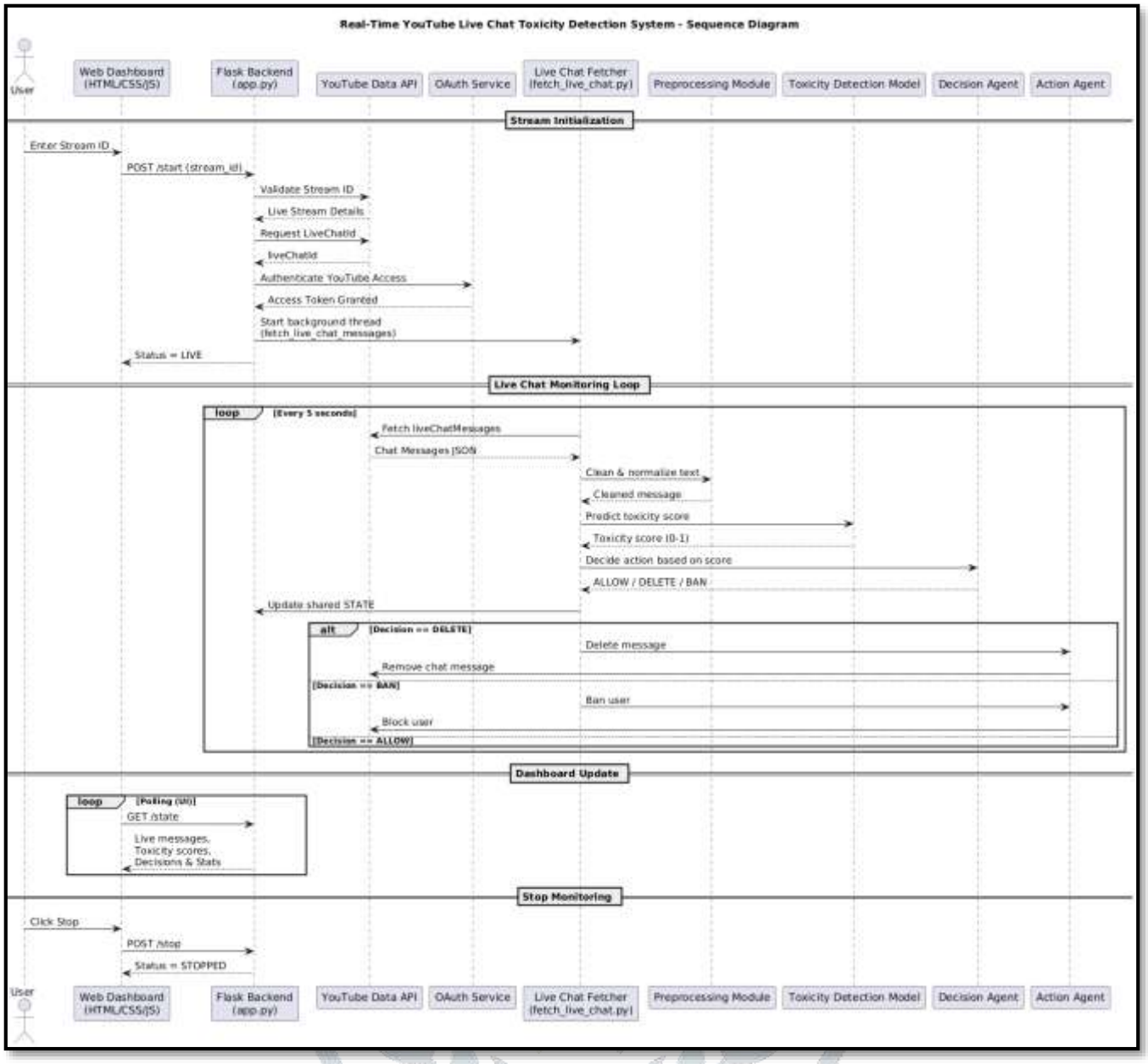


Fig 1. Sequence Diagram

V. RESULT & ANALYSIS

Toxicity Detection

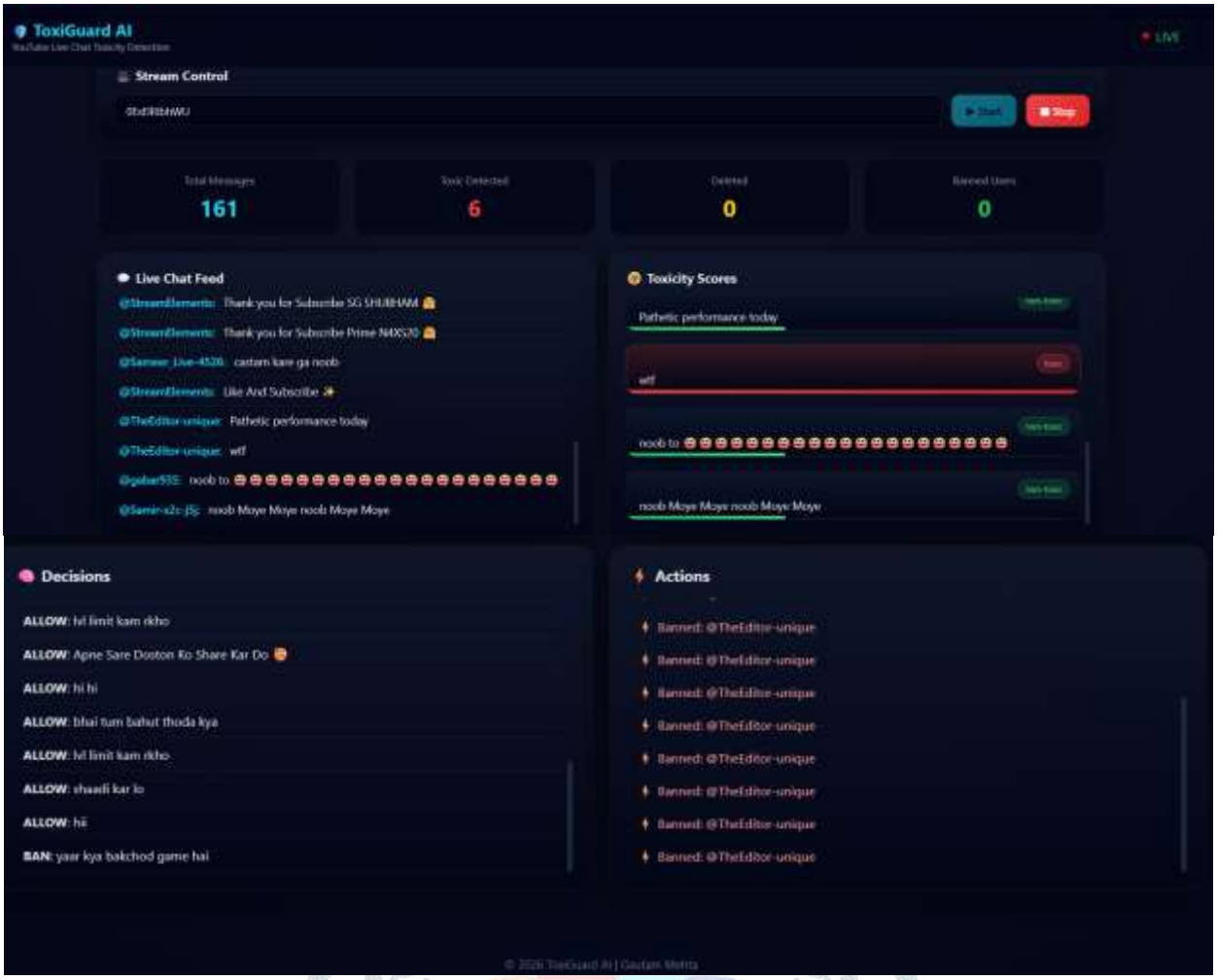
The primary analysis is performed using a transformer-based text classification model trained to detect toxic, abusive, and offensive language in real-time chat messages. The model outputs a probability score representing the level of toxicity for each message. This probability score is further mapped to interpretable risk categories such as Low, Medium, and High, enabling transparent moderation decisions. Similar probability-based toxicity scoring has been shown to be effective in prior studies on social media content moderation.

Moderation Decision Analysis

Based on the predicted toxicity score, a decision engine classifies messages into moderation actions such as ALLOW, WARN, DELETE, or BAN. This layered approach ensures that mildly offensive language is handled differently from highly toxic or abusive content. The integration of automated decision thresholds improves moderation consistency and reduces manual intervention, as supported by earlier research in cyberbullying and toxic content filtering systems.

Comparative Analysis

Experimental observations indicate that contextual understanding provided by transformer-based models significantly improves detection accuracy compared to traditional keyword-based or shallow machine learning methods.



VI. Discussion

A review of existing literature highlights both the progress and limitations of automated toxicity detection systems.

Advancements over Existing Systems

Previous studies have largely focused on offline datasets or post-hoc analysis of toxic comments. In contrast, the proposed system operates in real time and integrates live message ingestion, toxicity scoring, and automated moderation actions within a single framework, improving its applicability to live streaming environments

Ease of Use

The results demonstrate that web-based moderation dashboards are more accessible and practical for real-world deployment compared to standalone or offline moderation tools. The visual presentation of chat messages, toxicity levels, and actions enhances moderator awareness and system usability.

Extensibility

The system architecture is designed to support future extensions such as multilingual toxicity detection, platform-specific moderation rules, and adaptive strictness levels. Prior research suggests that scalable and modular moderation systems are better suited for evolving social media environments.

VII. Conclusion

The findings indicate that a real-time, web-based toxicity detection and moderation system is both feasible and effective. Transformer-based models demonstrate strong performance in identifying toxic language, while rule-based decision engines enable practical and interpretable moderation actions. Integrating real-time analysis with automated response mechanisms bridges the gap between academic research and real-world content moderation needs.

The transition from static toxicity detection models to a live, end-to-end moderation platform can significantly improve online community safety by preventing the spread of harmful content before it escalates.

VIII. FUTURE WORK

Expansion:
Future enhancements may include support for additional languages, platform-specific slang, and voice-based toxicity detection in live streams.

Advanced Modeling:
More advanced deep learning architectures and multimodal models (text + audio) may be explored to further improve detection accuracy and contextual understanding

Security and Privacy:
Future implementations should incorporate stronger data protection mechanisms to ensure user privacy and compliance with data protection regulations and platform policies.

REFERENCES

[1] K. Maity, A. S. Poornash, Sriparna Saha, and Pushpak Bhattacharyya, "ToxVidLLM: A Multimodal LLM-based Framework for Toxicity Detection in Code-Mixed Videos," CoRR, vol. abs/2405.20628, 2024. https://arxiv.org/pdf/2405.20628.pdf

[2] Z. Yang, N. Grenon-Godbout, and R. Rabbany, "Towards Detecting Contextual Real-Time Toxicity for In-Game Chat," CoRR, vol. abs/2310.18330, 2023. https://arxiv.org/pdf/2310.18330.pdf

[3] T. N. Harshitha, M. Prabu, E. Suganya, S. Sountharrajan, D. P. Bavirisetti, N. Gadde, and L. S. Uppu, "ProTect: A Hybrid Deep Learning Model for Proactive Detection of Cyberbullying on Social Media," Frontiers in Artificial Intelligence, 2024, doi:10.3389/frai.2024.1269366.
https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1269366/full

[4] Md. T. Hasan, Md. A. E. Hossain, Md. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A Review on Deep-Learning-Based Cyberbullying Detection," Future Internet, vol. 15, no. 5, Art. no. 179, 2023, doi:10.3390/fi15050179.
https://www.mdpi.com/1999-5903/15/5/179

[5] K. Shah, C. Phadhtare, and K. Rajpara, "Cyber-Bullying Detection in Hinglish Languages Using Machine Learning," International Journal of Engineering Research & Technology (IJERT), 2023.
https://www.ijert.org/cyber-bullying-detection-in-hinglish-languages-using-machine-learning

[6] A. Sheth, V. L. Shalin, and U. Kursuncu, "Defining and Detecting Toxicity on Social Media: Context and Knowledge Are Key," Proceedings of the Web Conference, 2021.
https://arxiv.org/abs/2104.10788

[7] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic Detection of Cyberbullying in Social Media Text," PLOS ONE, vol. 13, no. 10, e0203794, 2018, doi:10.1371/journal.pone.0203794.
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0203794

[8] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection Using Natural Language Processing," Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017, pp. 1–10.
https://aclanthology.org/W17-1101/

[9] S. N. Keerthana, M. J. Khushi, S. Anuraag, A. G. Ananya, and A. Ankamanal, "Implementation and Evaluation on Cyber Bullying Detection in Social Networks Using Deep Learning Models," JETIR, vol. 11, no. 5, 2024.
https://www.jetir.org/papers/JETIR2405G30.pdf

[10] H. M. Gaikwad, D. S. Chopada, M. S. Dandane, R. R. Kothadia, K. B. Mankare, and M. S. Nehete, "SecureChat – AI Powered Secure Messaging," International Journal of Multidisciplinary Research in Science and Technology, 2023.
https://www.ijmrset.com/upload/296_Securechat.pdf

[11] M. Jhaveri and D. Ramaiya, "Toxicity Detection for Indic Multilingual Social Media Content," arXiv preprint, 2022, arXiv:2201.00598.
https://www.emergentmind.com/papers/2201.00598

[12] D. Nithya, K. S. Nanthine, S. Thenmozhi, and R. Varshinipriya, "Advanced Social Media Toxic Comments Detection System Using AI," IJRASET, vol. 11, no. 6, 2023.
https://www.ijraset.com/research-paper/advanced-social-media-toxic-comments-detection-system-using-ai

[13] S. Chaudhary, "A Comprehensive Literature Review on Advance Language Toxicity Detection Using Deep Learning," International Journal of Electrical, Electronics and Computer Engineering, 2023.
https://www.researchtrend.net/ijeece/pdf/A-Comprehensive-Literature-Review-on-Advance-Language-Toxicity-Detection-using-Deep-Learning-Shaina-Chaudhary-6-.pdf

[14] G. Ramos, F. Batista, R. Ribeiro, P. Fialho, S. Moro, A. Fonseca, R. Guerra, P. Carvalho, C. Marques, and C. Silva, "A Comprehensive Review on Automatic Hate Speech Detection in the Age of the Transformer," Social Network Analysis and Mining, vol. 14, 2024, doi:10.1007/s13278-024-01361-3.
https://link.springer.com/article/10.1007/s13278-024-01361-3

[15] R. J. B., B. J., and D. P. S., "A Deep Learning Model for Detecting Bullying Comments on Online Social Media," International Journal of Intelligent Systems and Applications in Engineering, 2024.
https://www.ijisae.org/index.php/IJISAE/article/view/5479