
Toxicity Detection in online games using Reinforcement Learning

Mrs. Ruquia Taha Khatoon, Dantala Shivaghna, Katta Harsha Vardhan

¹Assistant Professor, Department of CSM, CMR College of Engineering & Technology, Medchal, Telangana, India.

²Student, CSE(AI&ML), CMR College of Engineering & Technology, Medchal, Telangana, India.

³Student, CSE(AI&ML), CMR College of Engineering & Technology, Medchal, Telangana, India.

Abstract

Multiplayer online gaming provides entertainment and social interaction but can also encourage toxic chat behavior, including harassment and abuse. This paper presents an AI-based toxicity detection and moderation system integrated into a multiplayer Snake and Ladders game. Machine learning models trained on Kaggle datasets classify chat messages as toxic or non-toxic. A reinforcement learning agent learns appropriate moderation actions, such as issuing warnings or removing players, based on toxicity severity and frequency. The system integrates real-time chat functionality with a dynamic game board to support interactive gameplay. The proposed approach demonstrates how adaptive AI-driven moderation can promote safer, more enjoyable gaming environments while reducing online toxicity and supporting digital well-being.

Keywords. Toxicity Detection, Reinforcement Learning, Natural Language Processing, Real-Time Chat Moderation.

1. INTRODUCTION

People really like playing games on the internet. They play with others from places and ages. They work together. Make friends in these virtual worlds. They can make friends. They can have fun with people they would never meet in life [1]. There is a big problem, with online gaming. It is called toxicity. Online gaming has this problem. It is very serious [2]

Toxic behaviour in games is really bad [3-10]. It can be people harassing each other using hate speech swearing much bullying and being verbally mean [11]. Younger players and people who are already treated unfairly can be hurt the most by behaviour [12]. If we do not stop behaviour online games can become very unpleasant places to be. This can make players quit damage the trust between players and make online gaming not as welcoming as it should be to everyone. Toxic behaviour, in games is a big problem that needs to be addressed. The old way of moderating things, where people have to report problems or where we just filter out words is not working very well. This system is often slow. It does not work the same way every time. It just can't keep up with the number of things that are

happening in real time in modern games. Modern games have a lot of real-time interactions and this old system is not good enough to handle them.

Artificial Intelligence, Machine Learning, Natural Language Processing, and Reinforcement Learning enable intelligent systems to monitor player communication and behavior. These technologies analyze chat messages, identify toxic content, assess severity, and determine appropriate responses. Based on detected toxicity, the system can issue warnings, temporarily restrict communication, or remove players. This adaptive approach promotes respectful interaction, fairness, safety, and a positive multiplayer gaming environment.

2. RELATED WORK

In the gaming industry, there are various major platforms that have adopted AI-powered moderation tools. For example, companies such as Riot Games and Blizzard Entertainment use AI-powered chat filtering and behaviour analysis tools to monitor player interactions and analyze gamer behaviour in real time. However, most of the existing industrial solutions are heavily dependent on supervised learning approaches. Though supervised learning approaches are highly efficient in detecting explicit forms of abuse, they are often ineffective in dealing with slang, sarcasm, contextual insults, and dynamic patterns of toxic behavior. Moreover, most of these approaches are still based on traditional punishment schemes that are not adaptive to player behavior and context.

Recent studies have shown increased interest in applying Reinforcement Learning approaches for improving moderation approaches by enabling adaptive and sequential decision-making approaches.

Table 1 Comparison of Existing Toxicity Detection and Player Behaviour Analysis Techniques

Study	Data used	Key Techniques	Advantages	Limitations
Deep Learning of toxic comment classification	Kaggle Toxic comment dataset(Wikipedia comments)	Logistic Regression, CNN, LSTM	High classification accuracy, captures contextual patterns	Not specially designed for gaming chats, lacks real-time adaptation
Context-Aware Toxicity Detection Using BERT	Social media and gaming chats datasets	BERT, Transformer-based NLP models	Better understanding of context, sarcasm and subtle toxicity	Computationally expensive, requires large training data
Real-time Chat Moderation in online games	n-game chat logs from multiplayer games	Keyword filtering + Supervised ML classifier	Fast detection, scalable for large platforms	Rule-based filtering struggles with evolving slang and coded language

Behaviour - based Toxicity player detection (Riot Games research)	Player chat logs + behavioral reports	NLP Player behaviour scoring models	+ Combines chat analysis with player history, improves fairness	Static punishment policies, limited adaptive decision-making
---	---------------------------------------	-------------------------------------	---	--

3. SYSTEM ARCHITECTURE

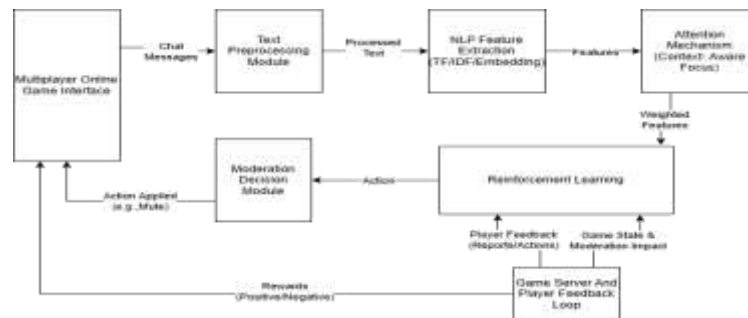


Fig1: System Architecture of Toxicity Detection in Online Games Using Reinforcement Learning

4. FLOWCHART

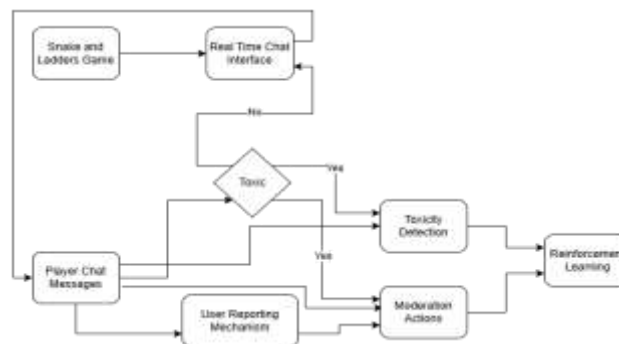


Fig2: Flow Chart of Toxicity detection in Online games using Reinforcement Learning

5. REFERENCES

- [1] Rosenfeld, A., & Richardson, A. (2021). The Role of Artificial Intelligence in Detecting and Preventing Online Toxicity. *IEEE Transactions on Games*.
- [2] Papangelis, A., et al. (2020). Toxicity Detection in Online Games: A Context-Aware NLP Approach. *Proceedings of the IEEE Conference on Games (CoG)*.

- [3] Blackburn, J., Kwak, H., & Lee, S. (2022). STFU Noob!: Predicting Crowdsourced Decisions on Toxic Behavior in Online Games. ACM Conference on World Wide Web.
- [4] S. Gore and C. Puthillate, "Authentication and authorization of users in an information handling system between baseboard management controller and host operating system users," U.S. Patent 11,038,874, 2021.
- [5] Kaggle Dataset. uetchy/sensai – Toxic Comment Classification Dataset. <https://www.kaggle.com/datasets/uetchy/sensai>.
- [6] Gao, Y., et al. (2022). Real-Time Toxicity Detection and Intervention in Online Games Using Reinforcement Learning. Journal of AI Research & Applications.
- [7] V. Antony Asir Daniel, K. Vijayalakshmi, Priyanka Pramod Pawar, Deepak Kumar, A. Bhuvanesh, A. Josephine Christilda, Enhanced affinity propagation clustering with a modified extreme learning machine for segmentation and classification of hyperspectral imaging, e-Prime - Advances in Electrical Engineering, Electronics and Energy, Volume 9, 2024, 100704, ISSN 2772-6711, <https://doi.org/10.1016/j.prime.2024.100704>.
- [8] Y. Yang, J. Li, Z. Wang, and H. Zhang, "Towards detecting contextual real-time toxicity for in-game chat," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023.
- [9] Tereshchenko and M. Hämmäläinen, "Efficient toxicity detection in gaming chats: A comparative study," arXiv Preprint, 2025.
- [10] Cornel, J., et al. (2025). "Comparative Analysis of Two-Class and Multi-Class Toxicity Detection Using Multi-Source Gaming Chat Data."
- [11] T. Fesalbon, L. Perez, and J. Martin, "Fine-tuning pre-trained language models to detect in-game trash talks," arXiv Preprint, 2024.
- [12] U. Naseem, M. Imran, and R. Razzak, "GameTox: A comprehensive dataset and analysis for enhanced toxicity detection in online games," Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2025.

Biographies



Ruquia Taha Khatoon is currently working as an Assistant Professor in the Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning), CMR College of Engineering & Technology, Medchal, Telangana, India. Her areas of interest include Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, and Data Science. She is actively involved in teaching, academic research, and mentoring students in emerging technologies. Her research interests focus on the development and application of intelligent computing techniques to solve real-world problems.



Dantala Shivaghna received her Bachelor of Technology (B.Tech) degree in Computer Science and Engineering (Artificial Intelligence & Machine Learning) from CMR College of Engineering & Technology, Medchal, Telangana, India. Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, and Data Science. She is interested in exploring emerging technologies and their applications in real-world problems. She has been actively involved in academic projects and research activities in the field of Artificial Intelligence and Machine Learning.



Katta Harsha Vardhan received her Bachelor of Technology (B.Tech) degree in Computer Science and Engineering (Artificial Intelligence & Machine Learning) from CMR College of Engineering & Technology, Medchal, Telangana, India. Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, and Data Science. She is interested in exploring emerging technologies and their applications in real-world problems. She has been actively involved in academic projects and research activities in the field of Artificial Intelligence and Machine Learning.