Accelerated Innovation (AI): Sustainability for Better Humanity
21–24 May 2025
Bangkok, Thailand
make learn
TIIM
PIConf
Management, Knowledge and Learning
International Conference
Technology, Innovation and Industrial Management
Sustainable Development

A Token-to-Label-to-Pattern Approach for Toxic Chat Detection in Online Games

Raveekiat Singhaphandu
Artificial Intelligence and Computer Engineering Program Department, CMKL University, Thailand.
raveekiat@cmkl.ac.th

Pavinee Rerkjirattikal
Department of Technology and Operations Management, Faculty of Business Administration, Kasetsart University, Thailand.
pavinee.re@ku.th

Eaint Hmue Khin
School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand.
m6722040026@g.siit.tu.ac.th

SangGyu Nam
School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand.
sanggyu@siit.tu.ac.th

Abstract

Toxic behaviour in online multiplayer games, particularly through in-game chat, remains a persistent challenge that negatively affects player experience and community well-being. Existing methods often rely on black-box models that lack interpretability and contextual sensitivity. This paper presents a token-to-label-to-pattern approach for toxic chat detection, offering fine-grained interpretability and domain-specific adaptability. Using the CONDA dataset of annotated Dota 2 chat logs, we apply structured preprocessing and extend token-level slot labelling. We then perform pattern mining on token sequences to identify frequent structures linked to communicative intents, including explicit toxicity, implicit aggression, and gameplay-related actions. Our preliminary findings show that a substantial number of token patterns correlate strongly with toxic behaviour classes, underscoring the importance of structural cues in chat-based toxicity detection.

Keywords: toxic chat detection, slot labelling, natural language processing, online multiplayer games

INTRODUCTION

Toxic behaviours such as verbal abuse, discrimination, harassment, and trolling are widespread and remain a significant challenge in online multiplayer games. Persistent toxicity diminishes gameplay enjoyment, reduce player retention, lead to reputational damage and potential revenue loss for game companies, as it discourages long-term engagement (Kordyaka et al., 2020). Various interventions have been introduced including chat filters, player reporting systems, block/mute functions, and human moderation. However, as Wijkstra et al. (2024) emphasize, the first step in addressing toxic behaviour is accurate detection as it a crucial role in enabling real-time and automated moderation.

While toxic behaviour can manifest in many forms, this study focuses on detecting toxic in-game chat, which is often immediate, emotionally impactful, and visible to both teammates and opponents. In-game chat serves as a primary channel for hostility and verbal aggression. Targeted toxic messages can cause embarrassment, frustration, or anger, prompting players to leave the game prematurely, making detection a critical point of early intervention. However, the volume, real-time nature, and linguistic diversity of chat, including slang and game-specific terms make manual moderation infeasible. As a result, there is a growing need for accurate and automated detection systems.

Previous approaches to toxic chat detection have primarily relied on techniques, such as sentiment analysis, keyword filtering, and deep learning-based classifiers. While these methods have proven successful, they often treat chat messages as unstructured blocks of text and lack interpretability, particularly when dealing with sarcasm, indirect toxicity, or domain-specific slang. This limitation makes it difficult for developers and moderators to understand the reasoning behind flagged messages or to respond appropriately.

To address these challenges, this paper proposes a novel token-to-label-to-pattern approach that leverages fine-grained slot labelling, domain-specific preprocessing, and interpretable pattern mining for toxic chat detection. Rather than classifying entire messages, our approach tags individual tokens like insults, threats, or sarcasm. These token sequences are then analysed to discover frequent and interpretable toxicity patterns, which are used for both real-time message flagging and as features for downstream classifiers. The proposed approach is validated using the chatlog from the game Dota 2.

BACKGROUND

This section explores toxic behaviours and reviews recent approaches to toxic text detection.

Toxic behaviours in multiplayer online games

Multiplayer online games are interactive digital environments where multiple players coexist, collaborate, or compete in real time. Communication occurs via text chat, voice channels, and in-game actions, all are potential channels for toxic behaviour. Toxicity refers to negative or disruptive actions that provoke frustration, anger, or annoyance, and may take the form of abusive language or gameplay behaviours like trolling, griefing, rage quitting, or other unsportsmanlike conduct (Beres et al., 2021).

The nature of toxic behaviour often varies depending on the game genre. In role-playing games (RPGs), it may involve stealing loot from teammates, while in first-person shooters (FPS), it can include intentionally killing teammates. Multiplayer Online Battle Arena (MOBA) games, such as League of Legends (LoL) or Dota 2 feature their own distinct categories of toxicity. Kou (2020) categorizes toxic

behaviours in MOBAs into five types: communicative aggression (e.g., verbal abuse, insults, hate speech), cheating (e.g., scripting, smurfing), hostage holding (e.g., refusing to surrender), mediocritizing (e.g., off-meta play), and sabotaging (e.g., intentional feeding, rage quitting).

Among these forms, text-based chat toxicity stands out as the most prevalent due to low expression barrier. It is often immediate, emotionally impactful, and visible to teammates or opponents in real time. Unlike gameplay-based toxicity, which can be ambiguous or effortful, sending toxic messages is fast, easy, and often impulsive. Moreover, chat messages leave a textual trail that can be logged, analysed, and moderated, offering opportunities for both real-time intervention and post-game accountability. However, chat messages often contain nuanced and implicit expressions, which are difficult to detect with broad classification labels. Sarcasm, passive aggression, platform-specific emotes, and coded language all contribute to subtle forms of toxicity that require contextual understanding and structured analysis. These complexities have made toxic chat detection a focal point in both academic and industry-led efforts to create safer and more inclusive gaming environments.

Toxic text detection methods

The rise of abusive language across forums, live-stream platforms, and online multiplayer games has made toxic text detection an increasingly important area of research. Methods for toxic text detection can be broadly categorized into deep learning-based and non-deep learning-based approaches.

Deep learning-based methods

Recent studies have leveraged a wide range of deep learning techniques to detect toxicity in in-game chats, forums, live streams, and even voice communications. For example, Abbasi et al. (2022) evaluated several deep learning architectures, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Gated Recurrent Units (GRUs) for multilabel classification problem of toxic comments related to religion and ethnicity in Wikipedia.

Vo et al. (2021) explored deep learning and transformer-based models such as TextCNN, GRU, and Toxic-BERT to identify toxic comments on online game forums. Their work addressed key challenges in toxic text detection, including class imbalance, informal language, and creative spelling that often hinder classification accuracy in gaming-related discussions. Kim et al. (2022) developed a bi-directional LSTM model to detect toxic chats on Twitch, where toxicity is often embedded in hybrid messages combining emotes and text. The model successfully detected previously undetected toxicity from visual and textual data. Ismail et al. (2025) proposed an embedding-based valence lexicon GloVe to improve detection of implicit toxicity, including nuanced forms such as sarcasm, gaming slang, and coded expressions commonly found on platforms like Twitch. Expanding beyond text, Yousefi & Emmanouilidou (2021) demonstrated the effectiveness of self-attentive CNNs in detecting toxic speech from audio data in multiplayer games, capturing acoustic cues that text-based models may overlook.

While deep learning methods have demonstrated strong performance in detecting toxic content across online platforms, they are often applied to longer and more structured texts, where sufficient linguistic context supports semantic modelling. In contrast, in-game chat messages are typically short, fragmented, and highly contextual, making them less compatible with conventional deep learning models that rely on longer inputs for meaningful feature extraction. As a result, toxic chat detection in multiplayer games has received comparatively less attention in the deep learning literature.

Moreover, deep learning models often function as black boxes, offering high accuracy without revealing why a particular message is classified as toxic. This lack of interpretability poses challenges especially in moderation contexts where fairness and transparency matter. These models also require large amounts of labelled data, which are scarce in fast-evolving gaming domains where new slang and behaviours frequently emerge. Additionally, models trained on general corpora (e.g., Wikipedia, Twitter) may not generalize well to the unique linguistic patterns in online games. These limitations underscore the need for more structured, interpretable, and domain-adaptive approaches to toxic chat detection.

Non-deep learning-based methods

Several studies have employed rule-based, statistical, or traditional Natural Language Processing (NLP) approaches for toxic text detection. These methods offer greater transparency and lower computational complexity, making them especially suitable in contexts where explainability is crucial.

Martens et al. (2015) implemented a rule-based annotation system to detect toxicity in DotA chat logs. Using n-gram patterns and custom rules, they distinguished general profanity from intentional verbal aggression and highlighted that kill events often trigger toxic language. Weld et al. (2021) introduced a context-aware dual annotation scheme and classifier using dialogue structure and token-level features. Their method emphasized capturing implicit toxicity by segmenting chats into messages or utterances and identifying contextual intent, demonstrating that token-level labels and dialogue history significantly improve classification performance without relying on deep models. Neto & Becker (2018) used statistical and linguistic analysis to explore and cluster in-game LoL chat–logs into thematic categories (e.g., tactics, insults, arguments). They identified a strong correlation between negative topics and match outcome. Ghosh (2021) applied classic NLP techniques (e.g., Bag-of-Words, sentiment analysis, Word2Vec) to social media posts related to online games, successfully identifying racism, sexism, and political toxicity. Though not explicitly targeting in-game chat, the study demonstrated that traditional NLP pipelines can be adapted to capture toxic sentiments from player-generated content.

Beyond text-based features, Canossa et al. (2021) proposed a supervised machine learning framework using Random Forests and Support Vector Machines classify toxic players in For Honor based on behavioural metrics (e.g., performance, disengagement, chat activity), achieving 90.7% accuracy. Together, these studies emphasize the importance of transparent, explainable methods, particularly where moderation decisions must be justified, auditable, or integrated with community-driven feedback.

Despite the progress in both deep learning and traditional methods, research that combines token-level slot labelling with sequential pattern mining remains limited. In addition, most existing works often overlooking the rich semantic structure and flow of conversation. This paper addresses these gaps by proposing a token-to-label-to-pattern approach, then mines interpretable token-label patterns to classify nuanced intent of the text. This approach enhances explainability, captures context sensitivity, and can be used to support real-time flagging, offering a more robust and interpretable solution for toxic chat detection in multiplayer games.

METHODOLOGY

This section describes the dataset, preprocessing pipeline, extended token labelling scheme, message-level classification, and the pattern mining approach**.**

Dataset

We validate the proposed Token-to-Label-to-Pattern toxic chat detection approach using the CONDA dataset (Contextual Offensive and Non-Offensive Dialogue Acts), which contains slot labels in-game chat logs from Dota 2 matches. The dataset includes both toxic and non-toxic utterances, enriched with dialogue context and manually assigned labels as introduced in the work done by Weld et al. (2021). The dataset includes 12,000 utterances drawn from in-game chat logs across 1,921 Dota 2 matches. Each message was classified into one of four categories, following a scheme adapted from Weld et al. (2021):

- **E (Explicit toxicity):** Messages with direct toxic expressions (e.g., profanity, personal attacks, clear verbal abuse).
- **I (Implicit toxicity):** Messages with veiled hostility or passive-aggressive tone, lacking obvious toxic words but carrying harmful intent.
- **A (Action):** Messages indicating player actions, such as pausing, reporting, or quitting.
- **(Others):** Neutral or ambiguous messages that do not fall into the above categories. These may include profanity without a target, self-deprecating jokes, or non-directed expressions.

Data preprocessing

To prepare the chat data for labeling and analysis, we implemented a preprocessing pipeline in Python, using several widely adopted libraries to standardize and clean the text while preserving key contextual information. The first step involved expanding English contractions (e.g., converting "don't" to "do not") with the help of the contractions Python library. We then used the ekphrasis toolkit (Baziotis et al., 2017), which is particularly effective for processing informal and social media text, to normalize gaming slang, elongated words, and emojis. Regular expressions were applied to remove unnecessary punctuation, symbols, and special characters that could introduce noise. Lastly, we used the spaCy library (Honnibal et al., 2020) to lemmatize the text, reducing each word to its base dictionary form for consistency in further analysis. An example of the data preprocessing steps is shown in Table 1.

Table 1: Step-by-step text cleaning and normalization process

Process	Input	Output
Contraction	You aren't carrying your team, n00b ;) gj keep it up!!	You are not carrying your team, n00b ;) gj keep it up!!
Normalizing	You are not carrying your team, n00b ;) gj keep it up!!	You are not carrying your team, noob ;) good job keep it up!!
Punctuation and Symbol Cleaning	You are not carrying your team, noob ;) good job keep it up!!	You are not carrying your team, noob ;) good job keep it up
Lemmatization	You are not carrying your team, noob ;) good job keep it up	You be not carry[†] your team, noob[*] ;) good job keep it up

[†] "Carry" is a slang often refers to high-impact players who significantly contribute to their team.

[*] "Noob" is a slang term for new or inexperienced players, often used as insult.

Labelling and Tokens

To capture both direct and subtle toxic expressions in player communication, we adopted a token-level slot labelling scheme. The dataset includes original slot labels provided by the authors, which cover common forms of toxic and non-toxic expressions. However, these labels are limited in their ability to distinguish between explicit toxicity (e.g., profanity, direct insults) and implicit toxicity (e.g., sarcasm, veiled hostility). To better capture the nuanced and context-sensitive nature of in-game communication, we expanded the token label set by introducing additional tags that reflect subtle linguistic cues frequently observed in player chats. This extended scheme provides a more fine-grained representation of player language and enables detection of both implicit and explicit toxic behaviours.

- **Original slot labels**: T (Toxicity), C (Character), D (Dota-specific), S (Game slang), P (Pronoun), SEPA (Separation of utterance), O (Other)

- **Additional slot labels**: IN (Insult), PF (Profanity), TH (Threats), SA (Sarcasm), CH (Cheating accusations), EN (Encouragement).

Table 2 illustrates how individual tokens in a chat are labelled using the combined slot label set.

Table 2: Example of message labeling process with tags assigned to each word or phrase.

Input	Input	Label (in parentheses)
Labeling	You be not carry your team, noob ;) good job keep it up	you (P), be (O), not (O), carry (D), your (P), team (S), noob (IN), ;) (EN), good job (EN), keep it up (EN)

Pattern mining

To better understand the linguistic structure of user utterances associated with different intent classes, we performed sequential pattern mining on slot label sequences. Our goal was to identify frequently occurring and distinctive patterns that characterize each intent, thereby enhancing interpretability and downstream intent classification. We developed a custom contiguous pattern miner to extract all subsequence of slot labels that are contiguous within each utterance. Our miner enforces strict contiguity, ensuring that identified patterns directly reflect localized linguistic behaviour. For each intent class, we collected all contiguous subsequence and counted their frequency. To evaluate whether a pattern is distinctive to an intent, we compute a class-confidence score:

$$\text{Class confidence} = \frac{\text{Frequency in class}}{\text{Total frequency across all classes}}$$

PRELIMINARY RESULT

Table 3 presents a sample of the token-level patterns identified through pattern mining, along with their associated intent class, confidence score, and representative examples extracted. These patterns demonstrate how recurring token sequences often correlate with certain communicative intents, including explicit or implicit toxicity, gameplay actions, and neutral remarks.

Each pattern is represented as a sequence of slot labels (e.g., P→O→O→O→O), where the labels correspond to annotated token types. The intent class indicates the general meaning or function of messages that contain the pattern (e.g., explicit toxicity or in-game action), and the class confidence reflects how consistently the pattern appears within that intent.

Key observations from the results include:

- Patterns containing IN (Insult) or PF (Profanity) tokens (e.g., IN→O, O→PF) show high confidence for explicitly toxic messages (E class), with confidence scores above 0.85.
- Sarcasm-related patterns such as S or S→S→S tend to indicate implicit toxicity (I class), though with slightly lower confidence values.
- Patterns like S→S→C and C→TH are more commonly found in action-based messages (A class), such as reporting or calling for votes.
- Neutral or ambiguous patterns like P→O→O→O→O appear frequently but are less reliably tied to a specific toxic class (O class, confidence = 0.733), reflecting their more general usage.

Tabel 3: Preliminary pattern mining results: identified token sequences and their associated intent

Pattern	Intent Class	Class Confidence	Example
P→O→O→O→O	O	0.733	- This (P) could (O) not (O) be (O) won (O) - We (P) play (O) 5x4 (O) 22 (O) min (O)
SA	O	0.872	- HAHA (SA) - rofl (SA) indeed (O)
IN→O	E	0.947	- Noob (IN) team (O) - report (S) doom (C) this (P) idiot (IN) has (O) never (O) fought (O) with (O) us (P)
O→PF	E	0.852	- You (P) are (O) asshole (PF) - good (O) game (O) will (O) fucked (PF)
S→S→S	I	0.796	- ez (S) ez (S) ez (S) - Ez (S) top (S) 2 (O) vs (O) 1 (O) still (O) EZ (S) EZ (S) Carry (S) Ez (S) Blame (O)
S→S	I	0.74	- SO (O) EZ (S) BOTTOM (S) - Farm (S) for (O) nothing (O) ez (S) mid (S) ez (S) win (O)
C→TH	A	0.63	- LUNA (C) AFK (TH) REPORT (S) THX (O) - tiny (C) afk (TH) wait (O) pls (O)
S→S→C	A	0.9	- gg (S) report (S) puck (C) plz (O) - :P (EN) gg (S) report (S) qop (C) ty (O)

These findings support the idea that intent can often be inferred from structural token patterns, even when surface wording varies. The extracted patterns serve as the foundation for higher-level analysis, including message-level classification and real-time toxicity detection in future work.

CONCLUSION

This paper proposes a structured token-to-label-to-pattern approach for toxic chat detection, combining token-level labelling with pattern mining. Using the CONDA dataset, advanced preprocessing, and extended labelling, we uncover distinctive token patterns aligned with various communicative intents, including explicit toxicity, implicit aggression, and gameplay-related actions. These patterns enhance interpretability and offer promising directions for real-time toxicity moderation in online multiplayer games. As the current focus is on token-level labelling and pattern discovery rather than full message-level classification, we have not yet applied standard evaluation methods for detection accuracy.

Future work will include validating our approach on separate validation data to assess robustness and consistency across different data subsets. We also plan to broaden the scope of analysis to include other forms of toxicity, such as voice-based harassment and gameplay sabotage, which often involve more complex and variable behavioural patterns. Real-world integration with active interventions, such as automated muting or adaptive matchmaking, also present important directions for further research.

ACKNOWLEDGEMENT

The 3rd author acknowledges the Excellent Foreign Student (EFS) scholarship awarded by the Sirindhorn International Institute of Technology, Thammasat University.

REFERENCES

Abbasi, A., Javed, A. R., Iqbal, F., Kryvinska, N., & Jalil, Z. (2022). Deep learning for religious and continent-based toxic content detection and classification. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-22523-3

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 747–754. https://doi.org/10.18653/v1/S17-2126

Beres, N. A., Frommel, J., Reid, E., Mandryk, R. L., & Klarkowski, M. (2021). Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming. N CHI Conference on Human Factors in Computing Systems (CHI '21), 1–15. https://doi.org/10.1145/3411764.3445157

Canossa, A., Salimov, D., Azadvar, A., Harteveld, C., & Yannakakis, G. (2021). For Honor, for Toxicity: Detecting Toxic Behaviour through Gameplay. Proceedings of the ACM on Human-Computer Interaction, 5(CHIPLAY). https://doi.org/10.1145/3474680

Ghosh, A. (2021). Analysing Toxicity in Online Gaming Communities. In Turkish Journal of Computer and Mathematics Education (Vol. 12, Issue 10).

Honnibal, M., Montani, I., Landeghem, S. Van, & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. Zenodo.

Ismail, H., Khalil, A., & Jasmy, A. (2025). Enhancing online toxicity detection on gaming networks: a novel embeddings-based valence lexicon approach. International Journal of Data Science and Analytics. https://doi.org/10.1007/s41060-025-00730-1

Kim, J., Wohn, D. Y., & Cha, M. (2022). Understanding and identifying the use of emotes in toxic chat on Twitch. Online Social Networks and Media, 27. https://doi.org/10.1016/j.osnem.2021.100180

Kordyaka, B., Jahn, K., & Niehaves, B. (2020). Towards a unified theory of toxic behaviour in video games. Internet Research, 30(4), 1081–1102. https://doi.org/10.1108/INTR-08-2019-0343

Kou, Y. (2020). Toxic Behaviours in Team-Based Competitive Gaming: The Case of League of Legends. CHI PLAY 2020 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play, 81–92. https://doi.org/10.1145/3410404.3414243

Martens, M., Shen, S., Iosup, A., & Kuipers, F. (2015). Toxicity detection in multiplayer online games. 2015 International Workshop on Network and Systems Support for Games (NetGames), 1–6. https://doi.org/10.1109/NetGames.2015.7382991

Neto, J. A. de M., & Becker, K. (2018). Relating conversational topics and toxic behaviour effects in a MOBA game. Entertainment Computing, 26, 10–29. https://doi.org/10.1016/j.entcom.2017.12.004

Vo, H. H. P., Trung Tran, H., & Luu, S. T. (2021). Automatically Detecting Cyberbullying Comments on Online Game Forums. Proceedings - 2021 RIVF International Conference on Computing and Communication Technologies, RIVF 2021. https://doi.org/10.1109/RIVF51545.2021.9642116

Weld, H., Huang, G., Lee, J., Zhang, T., Wang, K., Guo, X., Long, S., Soyeon, J. P., & Han, C. (2021). CONDA: a CONtextual Dual-Annotated dataset for in-game toxicity understanding and detection. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2406–2416. https://doi.org/10.18653/v1/2021.findings-acl.213

Wijkstra, M., Rogers, K., Mandryk, R. L., Veltkamp, R. C., & Frommel, J. (2024). How To Tame a Toxic Player? A Systematic Literature Review on Intervention Systems for Toxic Behaviours in Online Video Games. Proceedings of the ACM on Human-Computer Interaction, 8(CHI PLAY), 1–32. https://doi.org/10.1145/3677080

Yousefi, M., & Emmanouilidou, D. (2021). Audio-based Toxic Language Classification using Self-attentive Convolutional Neural Network. 29th European Signal Processing Conference (EUSIPCO 2021): Proceedings.