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

Toxic Chat Detection in Online Games Using Hybrid BERT and Character-level CNN

JaeHong Lee
School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand.
m6522041026@g.siit.tu.ac.th

Pavinee Rerkjirattikal
Department of Technology and Operations Management, Faculty of Business Administration, Kasetsart University, Thailand.
pavinee.re@ku.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 chats in online games remain a persistent challenge, disrupting fair play, harming player experience, and undermining community integrity. Traditional mitigation systems often rely on blacklist-based filtering, which struggles to detect intentionally obscured language, such as creative spellings, abbreviations, and evolving slang. These methods require constant manual updates, making them impractical in fast-paced gaming environments. This study proposes a hybrid model that combines BERT and a Character-level Convolutional Neural Network (CharCNN) to address these limitations. Using a real-world dataset from the popular game Dota 2, the model is evaluated against three natural language processing (NLP) approaches: (1) TF-IDF with logistic regression, (2) BERT, and (3) CharCNN. Results show that the hybrid model outperforms all others in classification accuracy and F1 score. By integrating contextual understanding with character-level pattern recognition, the model effectively detects both explicit and subtle toxic expressions. It offers a promising solution for scalable, real-time moderation in dynamic online gaming communities.

Keywords: toxic chat detection, TF-IDF, BERT, Character-level CNN, natural language processing, natural language processing, online multiplayer games

INTRODUCTION

Toxic behaviour in online multiplayer games, including harassment, hate speech, griefing and trolling remains a persistent challenge that threatens harmony of gameplay and game communities. Its continued presence not only diminishes the overall quality of gameplay but also impacts long-term player retention and game reputation, posing considerable commercial risks for developers and publishers (Blackburn & Kwak, 2014; Chandrasekharan et al., 2017). In response, game companies have implemented a range of moderation strategies, such as profanity filters, player reporting systems, mute/block functions, and automated moderators. While these strategies have shown partial, their effectiveness depends on the ability to detect toxic content accurately and in real time, especially in dynamic, fast-paced environments like in-game chat.

Unlike forum posts or post-game reviews, in-game chat occurs instantaneously and often reflects players' immediate reactions. As a primary in-game communication channel, chats become common outlets for impulsive aggression and verbal abuse. Toxic messages targeting teammates or opponents can trigger emotional distress, early match abandonment, and unpleasant gameplay experience. Although numerous toxic chat detection methods exist, challenges remain in capturing and identifying fast evolving and obscured nature of the language used in chats, including slang, abbreviations, or sarcasm, which pose difficulties for both manual and automated moderation efforts.

Traditional automated toxic text detection typically relies on keyword-based filtering, sentiment analysis, or general-purpose machine learning models. However, these approaches often fall short in detecting subtle or context-dependent toxicity and players can easily counteract these approaches using creative spellings, abbreviations, or emojis. A further limitation lies in the lack of transparency and interpretability in many of these models, which limits their implementation in real-world moderation workflows where interpretability and trust are essential (Pavlopoulos et al., 2020).

To address these challenges, this study proposes a hybrid model for toxic chat detection that combines the contextual understanding of BERT with the fine-grained linguistic sensitivity of a character-level Convolutional Neural Network (CNN). Using a real-world dataset from the popular multiplayer game Dota 2, the hybrid model is evaluated against three established natural language processing (NLP) baselines: TF-IDF with logistic regression (TF-IDF+LR), fine-tuned BERT, and character-level CNN. Experimental results demonstrate that the hybrid model consistently outperforms these baselines in terms of accuracy and F1 score. These findings highlight the effectiveness of integrating contextual and character-level features, offering a promising foundation for scalable, real-time moderation systems capable of detecting nuanced and evasive toxic behaviours in online gaming environments.

BACKGROUND

Detecting toxic language in online communication has become an essential task in maintaining healthy digital environments, particularly in real-time interactions in multiplayer games. In-game chat is often spontaneous, informal, and dynamic, making it a common ground for harmful language. Early approaches to toxic chat moderation primarily relied on banned word lists and simple rule-based systems. However, these methods quickly proved inadequate, as users adopted creative misspellings, special character substitutions, and other evasion techniques to bypass detection. To overcome these limitations, researchers began exploring more advanced NLP techniques capable of capturing the nuances and variability of human language.

Traditional machine learning methods employed a range of text representation techniques, including word frequency models, TF-IDF vectorisation, and n-gram analysis. However, these approaches demonstrated limitations in detecting context-dependent expressions and deliberately modified harmful language (Nobata et al., 2016; Waseem & Hovy, 2016).

The emergence of deep learning led to significant advances in toxic content detection. CNNs and Long Short-Term Memory (LSTM) networks became widely adopted for their ability to capture local patterns and sequential dependencies in text. CNNs are effective at identifying specific patterns or phrases that indicate toxicity, while LSTMs are better at understanding the overall meaning of longer or more complex sentences by capturing word order and relationships (Park & Fung, 2017; Yin et al., 2017).

As users continued to modify text to avoid detection using techniques like misspelling or replacing characters, researchers began to explore character-level models. These models process text at the level of individual characters rather than whole words. Character-level CNNs (CharCNN) emerged as an effective text classification methodology applied directly to raw character data without relying on word-level representations or linguistic feature engineering (Zhang et al., 2015). CharCNN models have demonstrated particularly strong performance in online communication environments with non-standard text. Studies have shown that character-level models often outperform word-level models in detecting harmful language containing indirect spelling and character substitutions on social media platforms (Founta et al., 2018).

More recently, transformer-based models such as BERT have further advanced toxic language detection by capturing the full context of a sentence. BERT's bidirectional learning approach effectively captures complex contextual relationships between words in sentences (Devlin et al., 2019). This characteristic has proven useful for identifying explicit profanity and subtle forms of toxic expression that vary in interpretation depending on context.

However, despite its strong performance, BERT and other word-based models can still struggle with the misspellings, abbreviations, and non-standard grammar commonly found in gaming chats. To overcome these limitations, recent studies have focused on hybrid approaches that combine word-level semantic understanding with character-level pattern recognition capabilities (Hakimov & Ewerth, 2021). These models aim to capture the meaning of a message while also recognising unusual patterns or altered spellings. In this study, we propose a hybrid model that integrates BERT's contextual understanding with CharCNN's ability to identify character-level features. We evaluate the model on toxic chat messages from the game Dota 2, comparing its performance against standalone BERT and CharCNN models to assess its ability to detect both explicit and subtle forms of toxicity in gaming environments.

METHODOLOGY

This section outlines methodology, including dataset and preprocessing step and toxic chat detection models of toxic chat detection. To comply with ethical and editorial standards, offensive words in chat examples have been sanitized while preserving linguistic patterns necessary for model demonstration.

Dataset and Preprocessing

We used an open-source dataset from Kaggle provided by Gosu.AI, which contains English-language in-game chat messages from Valve's Dota 2 matches (Fesalbon, 2022). The dataset consists of 3,267

chronologically ordered chat messages exchanged between players during gameplay. On average, each message comprises approximately 4.1 words and 20 characters, reflecting the concise and fast-paced nature of in-game communication. The original dataset includes three toxicity labels: normal, aggressive, and strong toxic, which were consolidated into a binary classification: 0 (normal) and 1 (toxic). It resulted in 1,800 normal messages (55.1%) and 1,467 toxic messages (44.9%).

In-game chat is often characterised by informal language, abbreviations, misspellings, and intentional deviations from standard spelling or grammar. To account for these characteristics, we applied a normalisation process which involved converting all text to lowercase, removing HTML tags, URLs, and email addresses, eliminating special characters and emojis (while preserving specific punctuation marks such as full stops, exclamation points, and question marks), and reducing redundant whitespace.

To enhance the effectiveness of the BERT model, we constructed an expanded text sequence (*context_text*) that incorporates conversational context by including messages immediately before and after each target message. These were separated using special tokens: [BEFORE], [CURRENT], and [AFTER]. For instance, given the following chat exchange within a game session:

To effectively train the BERT model, we created expanded text (context_text) that includes the context before and after each message. This was distinguished using special tokens [BEFORE], [CURRENT], and [AFTER]. Each content_text has maximum 2 special tokens. For example, assuming the following conversation occurred sequentially in a game session:

- User 1: “hi”
- User 2: “hello”
- User 3: “good game”

The *context_text* for User 2’s message “hello” would be structured as:

- [BEFORE] hi [CURRENT] hello [AFTER] good game

Table 1 illustrates examples of pre-processed chat messages in their original chronological order, taken from a single game match. These examples illustrate how contextual information was incorporated for the BERT-based model, which received the *context_text* as an input. While CharCNN and TF-IDF+LR models used only the original message text.

Table 1: Data examples used for BERT models.

Match ID	text	context_text	Label
6	why axe is destroying me top	[CURRENT] why axe is destroying me top [AFTER] f**king reported axe [AFTER] sorry nex	Normal
6	fucking reported axe	[BEFORE] why axe is destroying me top [CURRENT] f**king reported axe [AFTER] sorry nex [AFTER] what is the best soup?	Toxic
6	sorry nex	[BEFORE] why axe is destroying me top [BEFORE] f**king reported axe [CURRENT] sorry nex [AFTER] what is the best soup?	Toxic
6	what is the best soup?	[BEFORE] f**king reported axe [BEFORE] sorry nex [CURRENT] what is the best soup?	Normal

Toxic Chat Detection Models

This study implements and evaluates four models for toxic chat detection in gaming environments: (1) TF-IDF combined with Logistic Regression, (2) BERT, (3) Character-level Convolutional Neural Network (CharCNN), and (4) a hybrid BERT+CharCNN model.

The TF-IDF model converts text into weighted word frequency vectors and feeds them into a logistic regression classifier, capturing important keywords and n-gram patterns. BERT is a pre-trained transformer model that understands text by encoding rich bidirectional context from entire sentences and is fine-tuned on game chat data with added special tokens for conversational cues. CharCNN processes text at the character level using convolutional filters to capture subword patterns and spelling variations, making it effective for informal or distorted text.

Our proposed approach leverages the strengths of CharCNN and BERT. The hybrid model integrates BERT's contextual understanding with CharCNN's sensitivity to character-level signals, enhancing robustness to informal and evasive language. The model architecture consists of the following components (see Figure 1):

- Parallel Processing: The input is processed simultaneously along two paths. BERT receives the full contextualised message (*context_text*) embedded with special tokens, while CharCNN processes only the current message (*text*) at the character level.
- Feature Extraction: BERT extracts primary contextual features through the hidden state of the [CLS] token. Additionally, we leverage our introduced special tokens ([BEFORE], [CURRENT], [AFTER]) to differentiate various parts of conversation flow, aiding in contextual differentiation. Meanwhile, CharCNN extracts character-level patterns through character embeddings and convolutional filters of various sizes.
- Feature Combination: Outputs from both models are concatenated.

The two models offer complementary strengths, with CharCNN excelling at detecting spelling variations and evasive expressions using special characters, while BERT captures context-dependent meaning changes crucial for determining intent behind potentially abusive language. The training process involves separate inputs for each component where BERT receives text enriched with contextual markers while CharCNN processes only the current message. We optimize the model using the AdamW optimizer (Loshchilov & Hutter, 2017), with a linear learning rate scheduler and cross-entropy loss function.

EXPERIMENTAL RESULTS

For the experiments, the dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling. In this study, we evaluated model performance using F1 score and accuracy. The performances of the models are compared in Table 2.

The BERT+CharCNN hybrid model achieved the highest F1 score (0.9205), demonstrating improved performance by leveraging both contextual information and character-level features. The BERT and CharCNN models showed comparable results, while the traditional TF-IDF with logistic regression model performed the worst, with an F1 score of 0.8678.

Figure 1: Schematic diagram of the BERT+CharCNN

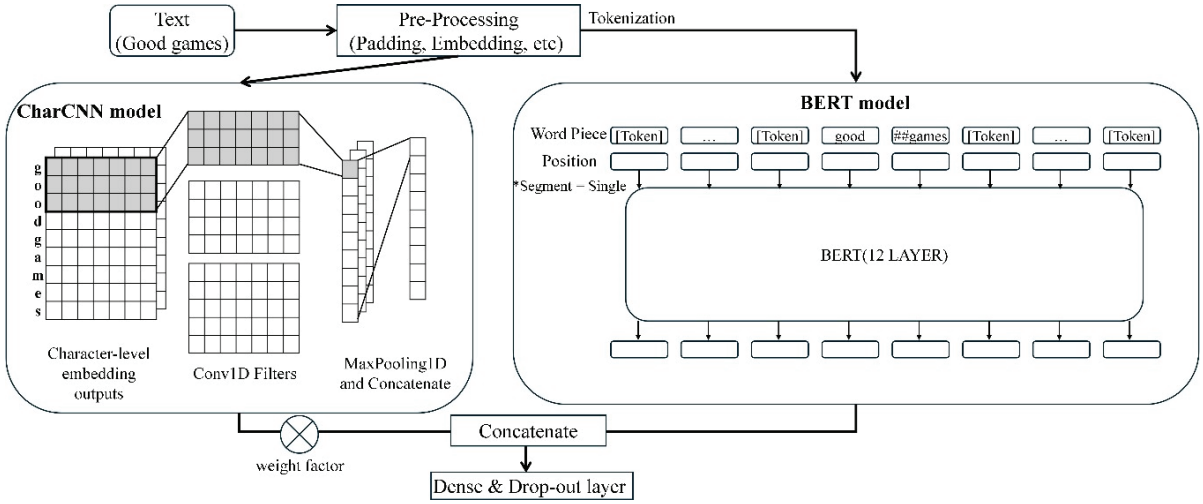

Table 2: Overall performance comparison of models.

Model	Accuracy	F1 score
TF-IDF+LR	0.8696	0.8678
BERT	0.8920	0.8920
CharCNN	0.8961	0.8963
BERT+CharCNN	0.9205	0.9205

Additionally, we analyzed each model's response to four special test cases (disguised toxic chat, compliments mixed with profanity, etc.) that reflect scenarios commonly found in real gaming environments, to assess practical performance. Details of the test cases and each model's evaluation are presented in Tables 3 and 4.

Table 3: Description of special test cases used for practical evaluation.

No	Case	text
1	Praise mixed with profanity, but not toxic (without context)	[BEFORE]: “” [CURRENT]: “what the f**king good skill” [AFTER]: “”
2	Praise mixed with profanity, but not toxic (with context)	[BEFORE]: “good bro” [CURRENT]: “what the f**king good skill” [AFTER]: “thank you”
3	Disguised Toxic Chat (without context)	[BEFORE]: “” [CURRENT]: “ff!@3ckk you” [AFTER]: “”
4	Disguised Toxic Chat (with context)	[BEFORE]: "Stop feeding" [CURRENT]: “ff!@3ckk you” [AFTER]: “”

Table 4: Model predictions on special test cases (Confidence: 0 to 1, probability).

Case No.	Case	TF-IDF+LR	BERT	CharCNN	BERT+ CharCNN
1	Praise mixed with profanity, but not toxic (without context)	Toxic (0.8547)	Toxic (0.9756)	Toxic (1.0000)	Toxic (0.9972)
2	Praise mixed with profanity, but not toxic (with context)	Toxic (0.8547)	Normal (0.9508)	Toxic (1.0000)	Normal (0.9738)
3	Disguised Toxic Chat (without context)	Normal (0.5166)	Normal (0.9385)	Normal (0.6451)	Toxic (0.9266)
4	Disguised Toxic Chat (with context)	Normal (0.5166)	Normal (0.6981)	Normal (0.6451)	Toxic (0.9794)

Key patterns and insights derived from the experimental results are as follows:

- Impact of Contextual Information: Performance differences based on the presence of contextual information were significant. Comparing test cases 1 and 2, the BERT model classified "Praise mixed with profanity, but not toxic" as normal (0.9508) when context was provided, but as toxic (0.9756) without context. This indicates that BERT-based models can effectively utilize contextual information.
- Effect of Character-Level Features: The CharCNN model demonstrated high confidence in toxic chat detection but showed limitations in contextual judgment. This is evidenced by test case 2, where it classified "praise mixed with profanity" as toxic despite the provided context.
- Difficulty in Detecting Disguised toxic chat: As seen in test cases 3 and 4, BERT and TF-IDF+LR models struggled to effectively detect profanity containing special characters or in modified forms. This suggests limitations in word-based approaches for recognizing character-level patterns in disguised profanity.
- Superiority of the Hybrid Model: The BERT+CharCNN model maintained high confidence and accurate classification across most cases. Notably, in test cases 3 and 4, it accurately detected disguised toxic chat ("ff!@3ckk you") with high confidence (0.9266, 0.9794), while other models failed. This can be attributed to the combination of character-level pattern recognition and contextual understanding capabilities.

In conclusion, the BERT+CharCNN hybrid model proved effective in detecting profanity across various types of game chat by integrating contextual information and character-level patterns. It demonstrated high accuracy particularly in boundary cases such as disguised profanity and context-dependent expressions. These findings validate the efficacy of multi-layered feature integration for toxic content detection in dynamic digital communication environments like game chat.

CONCLUSION

In this study, we implemented and compared four models (TF_IDF+LR, BERT, CharCNN, and BERT+CharCNN hybrid) for detecting toxic language in game chat. The experimental results showed that the BERT+CharCNN hybrid model, which combines the contextual information capabilities of BERT and the character-level pattern recognition of CharCNN, achieved the best performance (F1 score of 0.9205).

For future work, we plan to focus on improving model interpretability. Specifically, we intend to apply traditional approaches such as word contribution analysis along with attention weight visualization and

LIME (Linear Interpretable Model-agnostic Explanation). Through these methods, we aim to gain deeper insights into which text patterns the model focuses on when determining toxicity, contributing to the development of more effective toxic detection systems and healthy communication guidelines in gaming environments.

REFERENCES

Blackburn, J., & Kwak, H. (2014). STFU NOOB!: predicting crowdsourced decisions on toxic behaviour in online games. *Proceedings of the 23rd International Conference on World Wide Web*, 877–888. https://doi.org/10.1145/2566486.2567987

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–22. https://doi.org/10.1145/3134666

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Fesalbon, D. (2022). *GOSU AI English Dota 2 Game Chats*. Kaggle. https://www.kaggle.com/datasets/danielfesalbon/gosu-ai-english-dota-chat

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behaviour. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.14991

Hakimov, S., & Ewerth, R. (2021). *Combining Textual Features for the Detection of Hateful and Offensive Language*. https://github.com/sherzod-hakimov/HASOC-2021---Hate-Speech-Detection

Loshchilov, I., & Hutter, F. (2017). *Decoupled Weight Decay Regularization*. http://arxiv.org/abs/1711.05101

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*, 145–153. https://doi.org/10.1145/2872427.2883062

Park, J. H., & Fung, P. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter. *Proceedings of the First Workshop on Abusive Language Online*, 41–45. https://doi.org/10.18653/v1/W17-3006

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). Toxicity Detection: Does Context Really Matter? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4296–4305. https://doi.org/10.18653/v1/2020.acl-main.396

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). *Comparative Study of CNN and RNN for Natural Language Processing*.

Zhang, X., Zhao, J., & Yann, L. (2015). Character-level Convolutional Networks for Text Classification. *Proceedings of the 29th International Conference on Neural Information Processing Systems-Volume 1*, 649–657. http://arxiv.org/abs/1502.01710