# Classifying Toxicity in Video Games

Brianna Ta

Columbia University

## 1 Introduction

In recent years, multiplayer video games have become more popular, offering opportunities to connect and be social. While these games can create a fun online community, in-game toxicity is rampant and affects players' mental health and interactions. Toxicity can be characterized by abusive communication, such as harassment or verbal abuse, and disruptive behaviors, such as spamming. It has been shown to negatively impact player's mental health and heighten loneliness.

Toxicity has become increasingly normalized, especially in competitive games such as League of Legends. Perceived toxicity discourages social interaction and can reduce the typical social and entertainment benefits that video games provide. Understanding and addressing toxicity is important to keeping online gaming communities positive.

In this paper, natural language processing techniques will be employed to classify toxicity in text from common gaming social media platforms. Specifically, pre-trained transformer models, GPT-2 and BERT, will assess whether the text extracted from social media contains toxic language. Automating toxicity detection can provide insights into the patterns of toxic language and help censor messages, mitigating the impact of toxicity in the gaming community (Frommel et al., 2023).

## 2 Related Work

Wessel Stoop (2021) explored methods for detecting toxic behavior in online conversations, particularly in gaming communities like League of Legends. While the author noticed that you could identify commonly used offensive terms and block those out, these approaches struggle with sarcasm and finding the false positives and negatives. The author developed a neural network architecture, using embedding layers, bidirectional GRUs, and dense layers to identify toxic language. It was trained on 5,000 conversations and detected patterns in volume and repetition of phrases that helped the model achieve better performance in detecting toxicity with less false positives. This project is building the groundwork for more accurate toxicity detection and improving the gaming community as a whole.

In a similar vein to this project, an NLP Group at the University of Sydney created CONDA, a dataset for toxic language detection in games. They compiled 45,000 utterances from 12,000 conversations from the chat logs of 1,900 Dota 2 matches. Each of these messages can be categorized into slot and intent types. The slot types include

T (Toxicity), C (Character), D (Dota-specific), S (Game slang), P (Pronoun), and O (Other). The intent types are E (Explicit toxicity), I (Implicit toxicity), A (Action), and O (Other). Existing toxicity datasets typically limit their approach to the utterance level and the individual message, but CONDA looks at the entire conversation and can identify implicit toxicity based on previous messages. In their findings, they found that "gg" appeared frequently in all four intent type categories, portraying that gamers often use slang that don't appear in the general toxicity domain. This dataset contains useful data samples that could be used to expand on these results in future iterations (Weld et al., 2021).

# 3 Dataset

Five UC Berkeley Stuents partnered with GGWP, the first AI-powered proactive content moderation tool, to provide game developers with the tools to identify toxic behavior communities on public platforms. They collected data about League of Legends and Player Unknown Battlegrounds via data scraping for three major social platforms: Twitter, Twitch, and Discord. From Twitter, they extracted tweets that contained a certain keyword related to the games. From Twitch, they extracted chat logs from popular channels that streamed the games they were interested in investigating. And for Discord, they extracted conversations from large discord channels that discussed the games. They categorized different types of toxicity: toxic, severe toxic, obscene, threat, insult, and identity hate. This project uses the manually labeled data from Twitter, Twitch, and Discord, which their toxic classification model was trained on. They compared Logistic Regression, Support Vector Machine, Random Forest, LSTM, and BERT models. In constraint, this project compares the GPT-2 and BERT pre-trained models with and without fine-tuning (Lai, 2020).

# 4 Methods

This analysis aims to classify text messages as toxic (1) or non-toxic (0) using pre-trained GPT-2 and BERT models. It evaluates which model does better on a limited dataset and how they perform with different techniques of fine-tuning.

## 4.1 Data Preprocessing

The dataset contains combined data from Twitter, Twitch, and Discord on text related to the two games (League of Legends and Player Unknown Battlegrounds).

The dataset contains two columns: text and their corresponding toxicity label:

- toxic (1): text exhibiting aggressive or harmful language
- non-toxic (0): clean text with no observed toxicity

Due to the imbalance of toxic and non-toxic labels, additional preprocessing steps were performed to improve model performance. All of the text messages were cleaned, removing all punctuation, numbers, and excessive whitespace and converted to lower case. For GPT-2, Byte Pair Encoding (BPE) was applied using the GPT-2 tokenizer from the Hugging Face transformers library. For BERT, WordPiece tokenization was

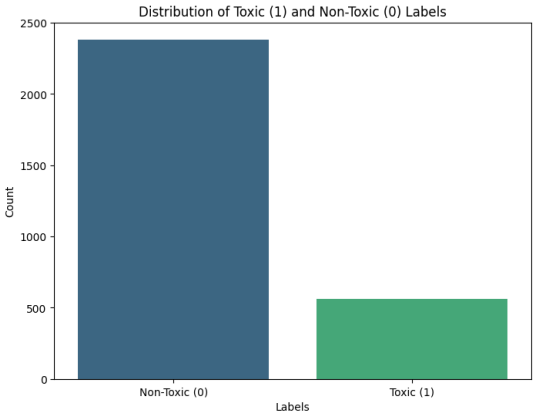



Figure 1: Distribution of Toxic and Non-Toxic Labels

applied using the pre-trained BERT tokenizer. WordPiece tokenization varies from the GPT-2 tokenizer as it breaks words into smaller subwords, which helps handle out-of-vocabulary (OOV) words. The dataset was split into 80% training and 20% validation sets using a stratified test-train split. To address the class imbalance, ROC-AUC (Area Under the Receiver Operating Characteristic Curve) was included in the evaluation metrics.

## 4.2 Models

This analysis uses GPT-2 and BERT models to perform text classification. These models were chosen because they are state-of-the-art and perform well on natural language processing tasks. While the GGWP project used various models, this project studies the differences between these two popular transformer models.

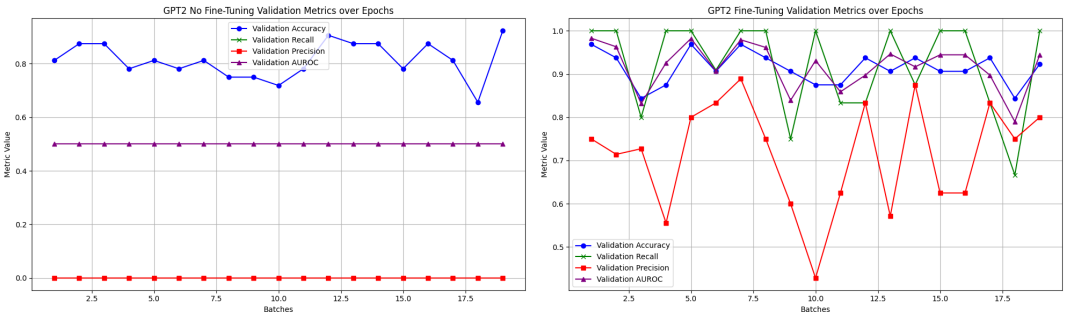
### 4.2.1 GPT-2

GPT-2, developed by OpenAI, is an unsupervised language model that was built on transformer architecture that uses 1.5 billion parameters. It was trained on an extremely large and diverse dataset called WebText, which allowed it to perform well on several natural language processing tasks, such as text generation, machine translation, question answering, and summarization. This model seemed like an ideal candidate to evaluate how it adapts to gaming jargon and slang. (Radford et al., 2019)

Because of time constraints, this analysis used the smallest pre-trained GPT-2 model, which has only 120 million parameters. The pre-trained model with no-fine tuning had poor results, with the validation precision and recall being nearly 0. This shows how pre-trained models have limited performance capabilities when applied to specific tasks without any additional fine-tuning. After fine-tuning, the model's performance improved drastically.

### 4.2.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) is another groundbreaking transformer model developed by Google AI. Unlike traditional models, BERT

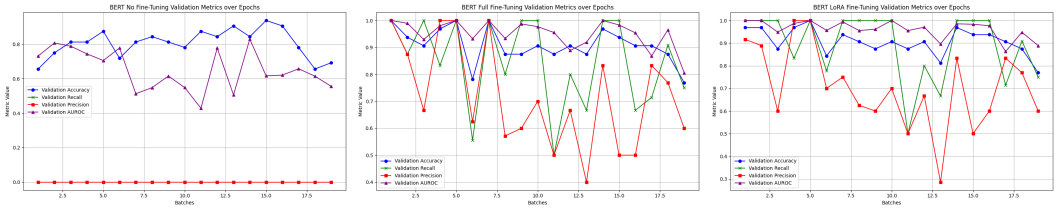



(a) GPT-2 No Fine-Tuning (b) GPT-2 Full Fine-Tuning

Figure 2: GPT-2 Performance

processes language bidirectionally to improve its understanding of words based on their surroundings. This feature helps it perform well on tasks, such as sentiment analysis, question answering, and text summarization. This model was trained on a large dataset as well, using Wikipedia and Google's BookCorpus (Muller, 2022).

Similarly to the GPT-2 model, the BERT model with no fine-tuning performed poorly in precision and recall and significantly improved with fine-tuning techniques. The performance of the full fine-tuning and LoRA fine-tuning models were comparable, with only the slightest of drops in performance for LoRA fine-tuning.



(a) BERT No Fine-Tuning (b) BERT Full Fine-Tuning (c) BERT LoRA Fine-Tuning

Figure 3: BERT Performance

## 5 Performance

Unsurprisingly, both models without any fine-tuning performed poorly. Fine-tuning helps adapt these models to more specific tasks, such as video game toxicity detection in this project.

GPT-2's validation accuracy increased from 0.7923 to 0.9260 with fine-tuning, while the precision and recall improved substantially, increasing from 0.0526 to 0.76714 and from 0.0088 to 0.9103, respectively.

| Fine Tuning | Train Loss | Accuracy | Precision | Recall | AUROC |
|---|---|---|---|---|---|
| None | 0.6099 | 0.7923 | 0.0526 | 0.0088 | 0.4966 |
| Full | 0.2291 | 0.9260 | 0.7614 | 0.9103 | 0.9199 |

Table 1: Performance of GPT-2

BERT also showed substantial improvements with fine-tuning. Similarly to the GPT-2 model, the pre-trained BERT model with no fine-tuning performed poorly with a

validation precision and recall of 0. The difference between the performance of the full fine-tuning and LoRA fine-tuning was negligible, portraying that LoRA would be the better option for projects with time constraints and limited computing resources. LoRA fine-tuning has a great performance, with less parameters.

BERT's validation accuracy increased from 0.8062 with no fine-tuning to 0.9089 and 0.9073 for full and LoRA fine-tuning, respectively.

| Fine Tuning | Train Loss | Accuracy | Precision | Recall | AUROC |
|---|---|---|---|---|---|
| None | 0.4895 | 0.8062 | 0.0000 | 0.0000 | 0.6519 |
| Full | 0.1981 | 0.9089 | 0.7179 | 0.8458 | 0.9511 |
| LoRA | 0.1944 | 0.9073 | 0.7036 | 0.8922 | 0.9616 |

Table 2: Performance of BERT

Overall, the GPT-2 model typically performed better in accuracy, precision, and recall, while the BERT model performed better in reducing training loss and optimizing AUROC.

## 6 Conclusion

This project analyzed the effectiveness of pre-trained transformer models, GPT-2 and BERT, for classifying toxicity about video games posted on social media platforms, like Twitter, Twitch, and Discord. The findings suggest that both models improved significantly with fine-tuning, showing that fine-tuning is needed for specific tasks.

Both the GPT-2 and BERT models performed poorly without fine-tuning, with low validation precision and recall. However, GPT -2 improved more in accuracy, precision, and recall after full fine-tuning than BERT. This suggests that GPT-2 might be the better option for toxicity detection.

In contrast, BERT was better in reducing training loss and optimizing the AUROC, which is an important metric for balancing model performance since this was an imbalanced dataset. With more time, it would be interesting to explore how these metrics change with different batch sizes and learning rates. Implementing other models could give further insight into what architectures are optimal for detecting toxicity in text. Incorporating other datasets, such as chat messages from video games instead of social media posts, could be an interesting avenue as well.

LoRA fine-tuning yielded a similar performance to full fine-tuning with fewer parameters. Testing various methods of fine-tuning could help improve upon the model's performance with limited resources. Future research on other fine-tuning methods, such as Parameter efficient fine-tuning (PEFT), prompt tuning, and adapters might give new insights into more efficient ways to utilize pre-trained models for specific tasks.

This project faced several limitations, such as a relatively small and imbalanced dataset. The size and imbalance could have influenced the toxicity detection, especially considering the slang used in gaming conversations. Using larger and more diverse datasets would improve the understanding of toxicity detection and the generalizability of the models.

## References


Julian Frommel, Daniel Johnson, and Regan L. Mandryk. How perceived toxicity of gaming communities is associated with social capital, satisfaction of relatedness, and loneliness. *Computers in Human Behavior Reports*, 10:100302, 2023. ISSN 2451-9588. doi: https://doi.org/10.1016/j.chbr.2023.100302. URL https://www.sciencedirect.com/science/article/pii/S2451958823000350.

Chris Lai. Ggwp - analyzing game toxicity through public data, 2020. URL https://github.com/pl2599/GGWP-Toxic-Behavior/.

Britney Muller. Bert 101 - state of the art nlp model explained, 2022. URL https://huggingface.co/blog/bert-101.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Henry Weld, Guanghao Huang, Jean Lee, Tongshu Zhang, Kunze Wang, Xinghong Guo, Siqu Long, Josiah Poon, and Caren Han. CONDA: a CONtextual dual-annotated dataset for in-game toxicity understanding and detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2406–2416, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.213. URL https://aclanthology.org/2021.findings-acl.213.

Florian Kunneman Wessel Stoop. Catching cyberbullies with neural networks, Dec 2021. URL https://thegradient.pub/catching-cyberbullies-with-neural-networks/.