GameTox: A Comprehensive Dataset and Analysis for Enhanced Toxicity Detection in Online Gaming Communities

Usman Naseem[1], Shuvam Shiwakoti[2], Siddhant Bikram Shah[3], Surendrabikram Thapa[2], Qi Zhang[4]
[1]Macquarie University, [2]Virginia Tech,
[3]Northeastern University, [4]Tongji University

Abstract

The prevalence of toxic behavior in online gaming communities necessitates robust detection methods to ensure user safety. We introduce GameTox, a novel dataset comprising 53K game chat utterances annotated for toxicity detection through intent classification and slot filling. This dataset captures the complex relationship between user intent and specific linguistic features that contribute to toxic interactions. We extensively analyze the dataset to uncover key insights into the nature of toxic speech in gaming environments. Furthermore, we establish baseline performance metrics using state-of-the-art natural language processing and large language models, demonstrating the dataset's contribution towards enhancing the detection of toxic behavior and revealing the limitations of contemporary models. Our results indicate that leveraging both intent detection and slot filling provides a significantly more granular and context-aware understanding of harmful messages. This dataset serves as a valuable resource to train advanced models that can effectively mitigate toxicity in online gaming and foster healthier digital spaces. Our dataset is publicly available at: https://github.com/shucoll/GameTox.

1 Introduction

The rapid expansion of online gaming has revolutionized entertainment, creating dynamic and engaging experiences for players worldwide. However, with this growth arises the challenge of maintaining a safe environment amidst a backdrop of increasingly toxic behavior (da Silva et al., 2020). Toxic behavior refers to negative actions by players that harm the gaming experience for others, such as harassment, griefing, or aggressive communication (Blackburn and Kwak, 2014), which can significantly detract from the user experience and lead to psychological harm (Kwak et al., 2015).

Several techniques have been used to manage toxic speech in online games and promote a positive online environment. These include word censorship, shadow banning users, and restricting their ability to communicate (Maher, 2016). While efforts have been made to develop frameworks and curate datasets to advance automated toxicity detection in online games, current datasets focus only on utterance-level annotation (Märtens et al., 2015; Blackburn and Kwak, 2014; Stoop et al., 2019). While utterance-level annotation of samples is intuitively reasonable for intent classification, using only one label for long sequences can lead to ambiguity and misclassification (Mielke et al., 2021), especially in online interactions which typically use a large amount of metaphors and slang (Do Dinh and Gurevych, 2016).

Slot filling, or the annotation of each word in a sentence, has emerged as a promising method in Natural Language Processing (NLP) as it offers an abundance of labels for data-hungry deep learning models. Further, slot filling facilitates the extraction of semantic concepts from text sequences, which improves the generalization ability of language models (Chen et al., 2019). The addition of token-level labels enhances the performance of models for tasks such as utterance-level classification (Weld et al., 2022). However, despite the benefits of joint task datasets spanning both intent classification and slot filling, data resources in this field remain limited.

To address these gaps, we propose GameTox, a toxicity detection dataset consisting of 53,000 online game chats from the game World of Tanks (WoT) collected through the WoT-record[1] database. The data comprises manual annotations for 6 classes at the utterance level (intent classification) and automated lexicon-based annotations for 4 classes at the token level (slot filling). With GameTox, we aim to facilitate the development of robust

[1] https://wot-record.com/

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 440–447
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

and granular toxicity detection models, ultimately contributing to safer online gaming communities.

2 Related Works

2.1 Toxicity detection in online games

Researchers have proposed various frameworks and datasets for automated toxicity detection in online games. Blackburn and Kwak (2014) utilized crowdsourced in-game user reports from League of Legends (LoL) for toxic behavior detection by extracting 534 features from in-game performance, user reports, and chat logs and employed the Random Forest Classifier for toxicity detection. Stoop et al. (2019) used a similar approach for data collection and introduced the RNN-based HaRe framework that tracked toxicity estimates for each user individually, updated the estimate with every new utterance, concatenated all of the utterances of each user, and classified the combined text. Märtens et al. (2015) proposed a novel lexicon-based annotation strategy for game chat toxicity detection to devise the DotAlicious dataset consisting of chat replays from 12,923 Defense of the Ancients (DOTA) matches.

2.2 Other Toxicity and Hate speech datasets

Detection of hate speech and toxicity in online environments has seen significant progress in recent years. Qian et al. (2019) introduced two labeled hate speech datasets collected from Reddit (22k comments) and Gab (33k comments) containing manually-written intervention responses. Wijesiriwardene et al. (2020) focused on toxic behaviors among youngsters and introduced ALONE, a dataset for toxic behavior detection among adolescents on Twitter, consisting of 16,901 tweets in 688 interactions and labeled for toxic vs non-toxic classes. Founta et al. (2018) analyzed abusive behavior on Twitter by releasing a dataset of 80,000 tweets annotated for seven labels: offensive, abusive, hate speech labels, aggressive, cyberbullying, spam, and normal. Mathew et al. (2021) introduced HateXplain, a dataset for explainable hate speech detection, consisting of 20,148 posts collected from Twitter and Gab annotated for three classes: hate, offensive, and normal, alongside target communities within hate. They further annotated the sections of the post that guide the labeling rationale. Zampieri et al. (2019) released an offensive language detection dataset comprising 14,100 tweets categorizing offensive language and its targets, consisting of offensiveness detection with three target classes: Individual, Group, and Other. To discern multiple aspects within cyberbullying, Salawu et al. (2021) curated an extensive dataset for cyberbullying detection comprising 62,587 tweets annotated for multiple aspects including Bullying, Profanity, Sarcasm, Threat, and Spam. Table 1 provides a summary of related literature in the domain.

3 Dataset

3.1 Data Collection and Pre-processing

We collected 53,000 utterances from the WoT-Record database, which stores chat recordings from the game World Of Tanks. Among these utterances, 42,963 samples contained only English text, and the rest were in other languages or a code-mixed format. The 42,963 English utterances were annotated for intent, and all samples were annotated for slot filling by converting the code-mixed samples to English by using Google Translate [2]. We converted all text to lowercase to ensure uniformity. We removed all duplicated text from the corpus, which may otherwise create biases. Further, we removed all user identifiers such as usernames and gamer tags to preserve the privacy of players.

3.2 Annotations

3.2.1 Slot Annotations

An automatic keyword-based slot labeling procedure was implemented for slot filling. We defined a set of 4 slot types - **T** (Toxic), **G** (Game Slang), **V** (Verb), **O** (Other). A corpus of labeled words was used to label each token in the dataset. To ensure correct labels for contemporary slang, we developed game toxicity labels by incorporating supplemental materials from Palomino et al. (2021), Märtens et al. (2015), and ElSherief et al. (2018). We also utilized Google's list of profanity[3] words and toxic utterances to expand the toxic word list. The final toxic word list consisted of 21,094 entries. Furthermore, among the slot annotation labels, all non-Latin script words and those from less common languages were grouped under the *other* category.

3.2.2 Intent Annotations

A two-step annotation process was followed for intent annotations. Large Language Models (LLMs)

[2] https://translate.google.com

[3] https://github.com/coffee-and-fun/google-profanity-words

Work	Data Source	Utt. lv.	T lv.	Labels
(Blackburn and Kwak, 2014)	LoL(Game)	✓	✗	toxic, non-toxic
(Märtens et al., 2015)	DOTA(Game)	✓	✗	toxic, non-toxic
(Founta et al., 2018)	Twitter	✓	✗	offensive, abusive, hateful speech aggressive, cyberbullying, spam, normal
(Stoop et al., 2019)	LoL(Game)	✓	✗	toxic, non-toxic
(Zampieri et al., 2019)	Twitter	✓	✗	offensive, non offensive targets - individual, group, others
(Qian et al., 2019)	Reddit and Gab	✓	✗	hate, no-hate
(Wijesiriwardene et al., 2020)	Twitter	✓	✗	toxic, non-toxic
(Mathew et al., 2021)	Twitter& Gab	✓	✗	hate, offensive, normal, target communities
(Salawu et al., 2021)	Twitter	✓	✗	insult, bullying, profanity, sarcasm, threat, exclusion, porn and spam
GameTox (Ours)	**WoT(Game)**	✓	✓	**Intents - Hate and Harassment, Threats, Extremism, Insults and Flaming, Other Offensive Texts, and Non-Toxic. Slots - Game Slang, Toxic, Verb, Other**

Table 1: Summary of datasets used in the literature. Utt. lv. and T lv. represent Utterance level and Token level respectively.

exhibit stellar reasoning capabilities in NLP tasks and hold promise as annotators that can label samples much faster than humans. However, they are prone to misannotating samples due to insufficient context or inherent biases. To overcome these challenges, we adopt a human-LLM collaborative annotation system similar to Wang et al. (2024). For efficiency, we initially create pseudo-labels by using ChatGPT, which are then verified by human annotators. All human labels take precedence over LLM labels. For manual annotations, five experienced annotators were employed for manual intent annotations with all the utterances being equally divided among the annotators to annotate. Each utterance was classified into either *Non-toxic* or one of the five toxicity labels: *Hate and Harassment*, *Threats*, *Extremism*, *Insults and Flaming*, and *Other Offensive Texts*.

Accurate and consistent annotations are essential for the reliability and validity of any analysis or model developed using labeled data. To achieve precise intent annotations, we implemented a three-phase annotation process. Further, the annotators followed comprehensive guidelines to maintain consistency and reliability in their work.

We used Fleiss' Kappa (κ) (Falotico and Quatto, 2015) as a statistical measure to assess the inter-annotator agreement. The κ for intent annotation was 0.78 and 0.91 in the pilot and consolidation phases respectively. This increase in κ reflects the effectiveness of the 3-phase annotation schema.

3.2.3 3-phase Annotation Schema

Pilot Run. In the first phase, a pilot run with 500 utterances was conducted to ensure that all annotators understood the annotation instructions. Since labeling text can be challenging, it was crucial to establish a shared understanding of the varieties and constituents of toxicity. During this phase, some confusion arose among the annotators, prompting revisions to the instructions to clarify ambiguities.

Revision Phase. In the second phase, all five annotators labeled 1500 utterances to ensure the clarity of the revised instructions from the first stage. The annotators used these updated guidelines to annotate the utterances, confirming that the revised instructions were clear and that they could consistently identify the presence of toxicity and its type.

Consolidation Phase. In the third phase, the annotators participated in a group discussion to address conflicts identified during the second phase of annotation while annotating 500 utterances after revising the instructions. This consensus-building process facilitated a thorough review of the annotations and ensured a shared understanding of the final guidelines. Occasional ambiguities were resolved through regular meetings and consultations with annotation experts, including academic professors. This phase was crucial for resolving disagreements and ensuring consistent labeling of all utterances, thereby enhancing the overall quality of the dataset.

3.2.4 Annotation Guidelines

Each utterance was labeled to one of 6 labels: *Non-toxic* if toxicity was not present and one of the five toxicity labels if toxicity was present. Annotation guidelines for each label are mentioned below.

Hate and Harassment. Utterances with the presence of identity-based hate or harassment (e.g., racism, sexism, homophobia) like *jap, greek***, pozor Ukraine, shut up homo, u guys play like fckng russians, asian monkey go away, fgt, poofer.*

Threats. Utterances with threats of violence, physical harm to another player, employee, or property, terrorism, or releasing a player's real-world personal information (e.g., doxing). like *I will kill u, go die, your family die in fire*

Extremism. Utterances with extremist views (e.g., white supremacy), attempts to groom or recruit for an extremist group, or repeated sharing of political or religious beliefs like *nazis, muslim*.

Insults and Flaming. Insults or attacks on another player or team (not based on player or team's real or perceived identity) like *fcking morons, delete this game idiots, noobs, idiots, bots*.

Other Offensive Texts. Any message not covered in the aforementioned categories that is offensive or harms a player's reasonable enjoyment of the game. Examples - *Easy lose, ok lose, another rigged game, Give up, FFS*.

Non-Toxic. Utterances without any toxicity.

3.3 Data Analysis

Label	#Samples	%
Non-Toxic	34679	80.71
Insults and Flaming	6049	14.07
Other Offensive Texts	1885	4.38
Hate and Harassment	274	0.63
Threats	53	0.12
Extremism	23	0.053

Table 2: Label distribution for intent classification.

Token	%
Other	67.17
Verb	15.51
Game Slang	7.72
Toxic	9.59

Table 3: Token distribution for slot classification.

Intent and Slot Distribution. Table 2 provides the class distribution of intent across the 42,963 English utterances, and Table 3 provides the slot filling distribution across all utterances. Most utterances are non-toxic in nature and a notable data imbalance is present. However, this is in line with real-world data distributions, where extremely toxic labels such as Hate and Harassment, Threats, and Extremism are often moderated or automatically suppressed. Figure 1 illustrates the word cloud for all intent labels.

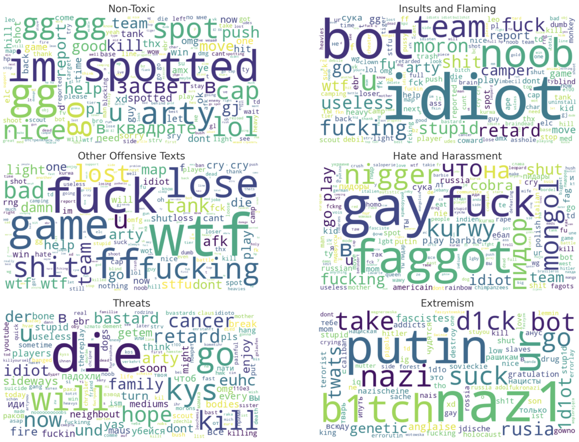

Figure 1: Wordcloud of words in each intent label.

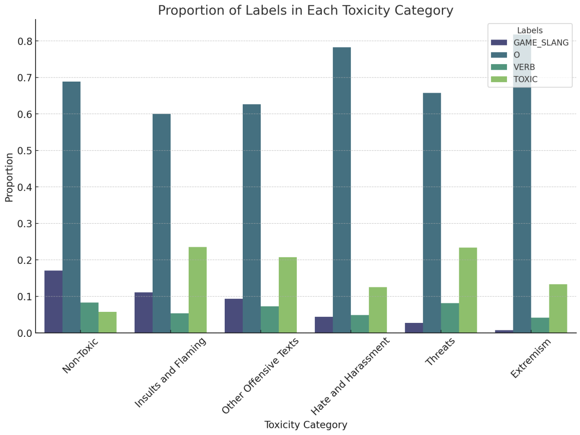

Figure 2: Slot token proportions in each intent label.

Intent-Slot Correlation. We analyze the relationship of each intent label with the slot tokens. Figure 2 provides the proportion of the tokens in each intent class. We find that toxic words have a high concentration within *Insults and Flaming*, *Other Offensive Texts*, and *Hate and Harassment* labels, and are less frequent in *Non-Toxic* utterances, but remain non-negligible. Game slangs have a high proportion within *Non-toxic* and *Insults and Flaming* labels, and are less frequent in *Extremism* and *Threats*, whereas verb tokens are more uniform across all labels. To further probe the relationship between intent labels and slot tokens, we obtain the most frequent slot tokens for 'Game Slang' and 'Toxic' tokens within each intent label, and Table 5 provides the top 5 Game Slang and Toxic tokens within each intent label.

4 Baselines and Analysis

We conduct classification experiments for the entire dataset (53,000 samples) and English-only (42,963 samples) utterances by using 12 baseline models. Appendix A.2 describes the models used. Table 4 presents the baseline results for intent and slot classification in GameTox's English-only and all language subsets. All the models perform better in slot classification over intent classification, indicating that identifying intent in human utterances poses a

Model	English					All				
	JSA	JAF	I-F1	S-F1	ICA	JSA	JAF	I-F1	S-F1	ICA
ToXCL (Hoang et al., 2024)	-	-	0.87	-	0.88	-	-	0.85	-	0.85
Mistral-7B (Jiang et al., 2023)	-	-	0.69	-	0.71	-	-	0.60	-	0.60
Llama-2-7B (Touvron et al., 2023)	-	-	0.65	-	0.68	-	-	0.59	-	0.62
Flan-T5-XL (Chung et al., 2024)	-	-	0.68	-	0.71	-	-	0.53	-	0.53
Gemma-7B (Team et al., 2024)	-	-	0.74	-	0.74	-	-	0.66	-	0.69
RNN-NLU (Liu and Lane, 2016)	0.78	0.89	0.84	0.93	0.85	0.76	0.88	0.84	0.91	0.85
Slot-gated (Goo et al., 2018)	0.85	0.93	0.87	0.98	0.87	0.73	0.88	0.87	0.88	0.87
Capsule NN (Zhang et al., 2018)	0.81	0.88	0.77	0.98	0.84	0.81	0.87	0.77	0.97	0.84
Inter-BiLSTM (Wang et al., 2018)	0.81	0.91	0.87	0.94	0.88	0.83	0.92	0.85	**0.98**	0.85
Inter-BiLSTM (Attn.) (Wang et al., 2018)	0.78	0.9	0.87	0.92	0.87	0.83	0.92	0.86	0.97	0.86
Joint mBERT (Chen et al., 2019)	0.86	0.93	0.88	0.98	0.88	**0.86**	0.93	**0.89**	0.97	**0.89**
Joint BERT (Chen et al., 2019)	**0.88**	**0.94**	**0.89**	**0.99**	**0.89**	0.85	**0.94**	**0.89**	**0.98**	**0.89**

Table 4: Classification performance along Intent and Slot levels. Joint Semantic Accuracy (JSA) gives comprehensive accuracy across intent and slot classification, where an utterance is considered accurately analyzed only when the intent and all slot labels, are correctly identified. Joint Average F1 (JAF) gives the joint Macro F1-score across both intent and slot classification. Intent-F1 (I-F1) and Slot-F1 (S-F1) give the Macro F1 score across all intent classes and slot types respectively. Intent Classification Accuracy (ICA) gives the intent-level accuracy of the models.

Extremism		Hate and Harassment	
Game Slang	Toxic	Game Slang	Toxic
xd	destroy	cap	battle
	crying	dps	faggots
	suck	heavy	nie
	b11ch	game	pussy
	nazi	skoda	wtf

Insults and Flaming		Threats	
Game Slang	Toxic	Game Slang	Toxic
cap	nie	strv	die
wn8	spammer	t100	cancer
t43	reta	omg	kill
mod	pussy	arty	fire
arty	kills	maus	retard

Other Offensive		Non-Toxic	
Game Slang	Toxic	Game Slang	Toxic
cap	broken	cap	hullu
wn8	battle	wn8	nie
arty	dirty	t43	blah
lit	nie	mod	kills
lmao	injuries	glhf	pussy

Table 5: Top 5 slot Game Slang and Toxic tokens across all intent labels

larger challenge to the models, leaving more room for improvement. The transformer models outperform the traditional neural architectures across all tasks. Amongst all the experiments, the Joint BERT models perform significantly better than the other models as they benefit from the extensive linguistic supervision provided by both types of labels during pre-training. The smaller transformer, mBERT, is surpassed by the bigger model BERT across almost all the metrics, which may indicate that larger models are better suited to utilize the large amounts of labeled data provided by the GameTox dataset. The ToXCL framework (Hoang et al., 2024) and LLM models result in subpar performance despite having large and complex model sizes, indicating the benefits of implementing slot-filling labels in supporting methods.

5 Conclusion

In this work, we introduce GameTox, a dataset for toxicity intent detection and slot filling in gaming environments. Our dataset is unique in its dual focus, capturing both the intentions behind toxic utterances and the specific components of speech that contribute to toxicity. We conducted baseline classification experiments using state-of-the-art NLP models, validating the dataset's utility in both intent detection and slot-filling tasks. Our experiments provide a benchmark for future research, highlighting the dataset's potential to enhance the precision and depth of toxicity detection methods. With GameTox, we aim to foster further innovation in the development of sophisticated, context-aware toxicity detection systems. Future work can focus on expanding the dataset, refining these models, and exploring their applications across diverse online platforms to mitigate toxic interactions and promote healthier online communities.

Ethical Statement

Privacy and Anonymity. The data utilized in this study originates from publicly available game chat logs. Further, all chat utterances included in the dataset have been anonymized to protect the privacy of the individuals involved. We adhered to strict data handling protocols to ensure that the privacy of all users is maintained.

Potential Risks. GameTox includes utterances that target specific individuals, communities, ethnic groups, and other entities with hate/toxicity. Although our intention in releasing this dataset is to strengthen chat moderation in online games and create safer online environments, there is a risk that it could be misused to propagate hate and discrimination. Further, we urge researchers to be mindful of the inherent biases within the dataset, as these may adversely affect the development of toxicity detection and moderation techniques.

Annotations. We hired 5 annotators with at least an undergraduate degree to annotate samples for GameTox. The annotators were either native English speakers or had taken the English language test (either TOEFL, PTE, or IELTS) ensuring accurate and reliable annotations. They were compensated appropriately according to the standard local rate.

Bias and Fairness. In the developmental phase of our dataset, we took measures to address and minimize potential biases. We implemented a rigorous annotation process to ensure that the labeling of toxic behavior was fair and consistent across different contexts. Additionally, we regularly reviewed and updated our guidelines to reflect the shared understanding of toxic behavior and its impact on individuals.

Limitations

While GameTox provides a comprehensive dataset for toxicity detection in online gaming, it has several limitations. Firstly, the dataset is sourced from WoT game chat logs, which may not fully represent the diversity of language and toxic behavior across different gaming communities. Additionally, the dataset may inherit inherent biases from the annotators' subjective interpretations of toxicity, despite rigorous annotation protocols. Moreover, the models trained on GameTox may exhibit overfitting on the specific patterns of toxicity present in the dataset, potentially reducing their generalizability.

References

Jeremy Blackburn and Haewoon Kwak. 2014. Stfu noob! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*, pages 877–888.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Bruno Mendes da Silva, Mirian Tavares, Filipa Cerol, Susana Mendes da Silva, Paulo Falcão, and Beatriz Isca Alves. 2020. Playing against hate speech–how teens see hate speech in video games and online gaming communities. *Journal of Digital Media and Interaction*, 3:34–52.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Nhat M Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. 2024. Toxcl: A unified framework for toxic speech detection and explanation. *arXiv preprint arXiv:2403.16685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3739–3748.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Brendan Maher. 2016. Can a video game company tame toxic behaviour? *Nature*, 531(7596):568–572.

Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6. IEEE.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Marco Palomino, Dawid Grad, and James Bedwell. 2021. Goldenwind at semeval-2021 task 5: Orthrus-an ensemble approach to identify toxicity. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Semiu Salawu, Jo Lumsden, and Yulan He. 2021. A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 146–156, Online. Association for Computational Linguistics.

Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 19–24.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. *arXiv preprint arXiv:1812.10235*.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8):1–38.

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 427–439. Springer.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S Yu. 2018. Joint slot filling and intent detection via capsule neural networks. *arXiv preprint arXiv:1812.09471*.

A Appendix

A.1 GPT prompt

To generate the initial pseudo-labels for intent classification we used the following prompt:

"*Categories: Hate and Harassment: Identity-based hate or harassment (e.g., racism, sexism, homophobia). Threats: Threats of violence, physical safety to another player, employee or property, terrorism, or releasing a player's real-world personal information (e.g., doxxing).Extremism: Extremist views (e.g., white supremacy), attempts to groom or* *recruit for an extremist group or repeated sharing of political, religious, or social beliefs. Insults and Flaming: Insults or attacks on another player or team (not based on player or team's real or perceived identity) Other Offensive Texts: Any other message not covered in the above categories that is offensive and/or harms a player's reasonable enjoyment of the game. Given the following messages, Classify each one according to the categories listed above. Must Only return the category. {chat}.* " Here, "{chat}" is replaced by one dataset sample.

A.2 Baseline Models

ToXCL (Hoang et al., 2024): ToXCL is a unified framework tackling implicit toxic speech detection and explanation, leveraging a target group generator, encoder-decoder, and knowledge distillation.

Mistral-7B (Jiang et al., 2023): A 7B-parameter LLM employing a transformer-based architecture with multi-head self-attention.

Llama-2-7B (Touvron et al., 2023): A 7B-parameter variant of the Llama-2 family of LLMs that leverages a transformer backbone with scaled multi-head attention.

Flan-T5-XL (Chung et al., 2024): A T5-based LLM with 3B parameters that undergoes instruction-focused fine-tuning via the FLAN methodology. It leverages a unified sequence-to-sequence framework.

Gemma-7B (Team et al., 2024): A 7B-parameter LLM built on a transformer foundation with specialized gating mechanisms.

RNN-NLU (Liu and Lane, 2016): An attention-based bi-directional recurrent neural network model that simultaneously predicts the current slot and intent at each time step, utilizing shared hidden states and attention mechanisms.

Slot-gated (Goo et al., 2018): An attention-based BiLSTM model that constructs distinct attended contexts for slot filling and intent classification. It explicitly incorporates the intent context into the slot-filling process through a gating mechanism.

Capsule NN (Zhang et al., 2018): A capsule-based neural network designed to explicitly capture the semantic hierarchical relationships among words, slots, and intents using a dynamic routing-by-agreement mechanism.

Inter-BiLSTM (Wang et al., 2018): A model that integrates two interconnected BiLSTMs that perform slot filling and intent classification respectively. Information is exchanged between the two tasks by sharing hidden states at each time step, facilitating the decoding process on both sides.

Inter-BiLSTM (Attn.) (Wang et al., 2018): We combined the Inter-BiLSTM model with the default attention mechanism (Vaswani et al., 2017).

Joint mBERT (Chen et al., 2019): The multilingual model mBERT is used for joint intent classification and slot filling in code-mixed data.

Joint BERT (Chen et al., 2019): leverages the strengths of pre-trained BERT by performing joint prediction intent and slot prediction using the [CLS] token embedding for intent classification and token embeddings for slot filling.