Unveiling Identity Biases in Toxicity Detection : A Game-Focused Dataset and Reactivity Analysis Approach

Josiane Van Dorpe
Ubisoft La Forge
UQAM
van_dorpe.josiane@courrier.uqam.ca

Zachary Yang
Ubisoft La Forge
McGill University, Mila
zachary.yang@mail.mcgill.ca

Nicolas Grenon-Godbout
Ubisoft UDO
nicolas.grenon-godbout@ubisoft.com

Grégoire Winterstein
UQAM
winterstein.gregoire@uqam.ca

Abstract

Identity biases arise commonly from annotated datasets, can be propagated in language models and can cause further harm to marginal groups. Existing bias benchmarking datasets are mainly focused on gender or racial biases and are made to pinpoint which class the model is biased towards. They also are not designed for the gaming industry, a concern for models built for toxicity detection in videogames' chat. We propose a dataset and a method to highlight oversensitive terms using reactivity analysis and the model's performance. We test our dataset against ToxBuster, a language model developed by Ubisoft fine-tuned for toxicity detection on multiplayer videogame's written chat, and Perspective API. We find that these toxicity models often automatically tag terms related to a community's identity as toxic, which prevents members of already marginalized groups to make their presence known or have a mature / normal conversation. Through this process, we have generated an interesting list of terms that trigger the models to varying degrees, along with insights on establishing a baseline through human annotations.

1 Introduction

Online spaces are valuable for exchanging ideas and discussing common interests globally. However, these interactions are often marred by toxic comments and content, evident on platforms like Facebook (Ciftci et al., 2017), Twitter (Watanabe et al., 2018), and Reddit (Mohan et al., 2017). The videogame industry is also not immune to harm and harassment, as evidenced by the rising toxicity in written communications among players(ADL, 2022). This high level of toxicity not only affects gaming choices but also the personal lives of players involved. Consequently, platforms (Hanu and Unitary team, 2020; Muralikumar et al., 2023) and the videogame industry (Miller, 2019; Shi, 2019; Unity, 2021) have turned towards language models for toxicity detection and content moderation due to their excellent performance and contextual understanding.

Although these models can effectively capture toxic content, they can perpetuate and amplify social biases present in their training datasets (Angwin et al., 2016; Caliskan et al., 2017; Dixon et al., 2018; Savoldi et al., 2021). Biases can emerge during dataset creation when practitioners sample data, annotators label data based on personal understanding, culture, and experiences, and practitioners aggregate labels. In this study, we specifically focus on the issue of models overestimating the toxicity of terms associated with certain concepts, leading to problematic false positives and even false negatives (Dixon et al., 2018; Kiritchenko and Mohammad, 2018; Prabhakaran et al., 2019; Sap et al., 2019; Garg et al., 2022). Existing research has primarily focused on biases in toxicity detection without considering the specific use-case of in-game chat, despite the widespread presence of toxicity in that particular context.

To fill this research gap, we begin by meticulously constructing a dataset that aims to uncover identity biases present in language models. This dataset encompasses biases commonly observed within the English-speaking gaming community in North America. Next, we introduce a novel approach that combines reactivity analysis and the model's performance to identify highly sensitive terms. We apply this method to assess the effectiveness of ToxBuster (Yang et al., 2023), a model specifically trained for in-game chat, as well as Perspective API. Through our evaluation, we demonstrate the efficacy of our approach and its potential for evaluating various forms of biases in toxicity detection models. Both the dataset and the method are described in this paper as a proof of concept, as they do not cover every possible bias. Additional iterations are possible and strongly encouraged.

In summary, our contributions are two-fold, presenting a prototype of the following:

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 263–274

1. An identity bias benchmark dataset for toxicity detection models,
2. A novel method that combines reactivity analysis and model performance to identify sensitive terms possibly conveying biases.

2 Related Work

Toxicity detection is inherently complex and subjective, with different definitions and interpretations among researchers (Garg et al., 2022; Kowert, 2020). Biases also vary across communities, influenced by culture, origin and socio-political context. In this study, we define biases as "prejudice in favour of or against one thing, person, or group compared with another usually in a way that's considered to be unfair" (University of California). Natural language processing encompasses a wide range of types of biases, categorized by their sources or the type of harm they cause (Sap et al., 2019; Garg et al., 2022). Our focus lies specifically on lexical identity biases, which refer to biases conveyed by terms related to one's identity or characteristic (Zhou et al., 2021).

Initially, we address the questions posed by Blodgett et al. (2020), more precisely "what kinds of system behaviours are harmful, in what ways, to whom, and why". The harmful behaviour we examine are false positives and false negatives concerning identity biases. In other words, we aim to identify terms that the model consistently tags as toxic or non-toxic, even when they are used in the opposite manner. A false positive resulting from an identity bias prevents marginalized and possibly minority communities from discussing and engaging with members of their own group (Zhou et al., 2021) thereby erasing proper representation of that social group (Dev et al., 2021a; Blodgett et al., 2022). Conversely, a false negative neglects to flag a sentence containing a term that should be identified as toxic, usually associated with an oppressing or majority community.

Previous work on bias detection has focused on creating evaluation metrics (Prabhakaran et al., 2019) or corpora (Kiritchenko and Mohammad, 2018) that consider binary dimensions, such as gender (male, female) or race (black, white), to determine the model's leanings. However, online spaces such as videogames encompass diverse communities that cannot be easily classified into such binary labels, rendering this analysis insufficient. Kiritchenko and Mohammad (2018) developed a dataset of sentence templates and human names to identify biases in a model, however these longer sentences and names are not typically found in videogame chat.

3 Methodology

Here, we present our method for identifying biases and oversensitive terms toxicity models flag in four sections. Section 3.1 discusses the generation of a synthetic dataset using sentence templates and a set of manually identified terms associated with potential biases. Section 3.2 outlines the annotation process for this dataset. Section 3.3 focuses on using reactivity analysis to highlight terms and biases. Lastly, in section 3.4, we describe the two toxicity detection models that we measure.

3.1 Dataset Creation

Following Kiritchenko and Mohammad (2018), we create a set of sentence templates and a corresponding set of terms. Initially, a list of terms was generated by identifying categories commonly assessed in identity bias research, such as race, gender, sexual orientation, and age (Dev et al., 2021b; Garg et al., 2022). To ensure templates and terms relevance to the gaming domain, we sample thousands of chat lines containing this first list of terms to identify sentence formulations frequently used within the chat of an online competitive multiplayer game where communication is permitted among teammates as well as with enemy teams. Each team consists of 5 players.

In order to ensure inclusivity in our template selection and bias categories, we sought input from volunteer members of various employee resource groups within our organization. These groups consist of employees with diverse backgrounds and perspectives. The four employees consulted provided valuable feedback on the chosen terms, sentences, bias dimensions, and proposed additional ones.

Through this process, we compile a final list consisting of 46 terms categorized into 12 categories, as shown in Appendix A. We prioritize terms that exhibited ambiguity in their offensive nature, as they can be used in either a toxic or non-toxic manner - e.g. *you look gay* would be toxic, while *they discussed gay marriage* is not (Wang et al., 2014; Zhou et al., 2021). Terms considered to be insults or slurs without any ambiguity (e.g. ret*rd) regarding their offensive nature in North American

Template	Tag replacement
you must be **<Attribute>**	black, trans, old...
i don't like **<Groups>**	whites, black women, young people...
you play like a **<Person>**	white person, fat gay, mexican artist...

Table 1: Examples of sentence templates and tags replacements

English were not included in this list. We note that the term *fairy* is included in the sexual orientation category, as it can be used to refer to a homosexual man.

We carefully select 22 grammatically correct sentence templates presented in Table 5 of Appendix A to reflect sentences found within in-game chat lines. These templates are also chosen to ensure variation in the syntactic function of the evaluated terms, allowing them to serve as either subjects or objects within sentences. Each sentence includes one or two "tags", representing variables to be replaced by a single term or noun phrases consisting of two terms. The **<Attribute>** tag must be a single-term adjective, such as *black, white, trans, old* or *young*. The **<Person>** tag represents a singular noun (*female, male, artist, developer*...), either alone or preceded by an **<Attribute>** if it is not already included in the sentence (e.g., *black female, black male, black artist*). Similarly, the **<Groups>** tag functions like **<Person>**,but with the noun in its plural form (e.g., *black females, black males, black artists*). Note that a term belonging to **<Attribute>** in its singular form and **<Groups>** in its plural form (*black woman*, but also *trans blacks*) cannot be combined with itself (*black blacks*).

Using the sentence templates and list of terms, a total of 16,008 synthetic chat lines were generated. Examples are seen in Table 1. The dataset is diverse and encompasses categories of biases that are not limited to binary options; its main purpose is to evaluate the model's reactivity to each term irrespective of its category. We emphasize that the dataset serves as a prototype, indicating the potential use of a broad range of terms and templates to subsequently expand the dataset.

3.2 Dataset Annotation

Our dataset is annotated through a two-step process. Firstly, a sample of the dataset is manually annotated. Secondly, a random forest model is trained to propagate these annotations for the entire dataset.

We circumvent making assumptions about the toxicity or lack thereof in a term, sentence template, or their combination by including various categories for terms and sentiment polarity for templates. This is to avoid rejecting terms and templates that may seem inoffensive in themselves or combined, yet might be evaluated as toxic by a human. Human annotations will decide which sentence is toxic or non-toxic. Our method may then reveal unexpected biases, while unbiased terms and templates will manifest their neutrality.

3.2.1 Manual Annotation

We obtain a set of ground truth labels from four participants. These participants were recruited from within the game company that developed ToxBuster (Yang et al., 2023), which is further described in Section 3.4.1.

A total of 1,363 lines, a subset of the complete dataset, was annotated. The decision to annotate only a fraction of the dataset is due to both the exploratory nature of this research and the limitation of resources to annotate a large dataset. We further discuss the decision and motivation in the limitations. Annotations guidelines and details are provided in Appendix B. The process ultimately allowed us to obtain a binary label for each line of the subset.

3.2.2 Annotation Propagation

We propagate the manual annotations to the full dataset by training a Random Forest. In particular, we perform a 5-fold CV over 6 mtry parameters (see section C) with a 20-80 train-test split. The best performing parameters are ntree=500 and mtry=15, with a F1 score of 90.4% on the test set.

Ideally, this step would not be needed as the whole dataset would be manually annotated.

3.3 Reactivity Analysis

We will now elaborate on the process of identifying biases and models reactivity to certain terms in the dataset.

Our objective is to compare the analysis results of each toxicity detection model with the ground truth annotations from the annotated dataset.

We conduct a reactivity analysis by calculating the average predictive difference in the probability of toxicity for each lemmatized term. In other words, we measure the difference of each sentence's toxic probability when the specific term

is present or absent (Gelman and Hill, 2006), providing insights on the influence of the presence of the term on toxicity.

We calculate the reactivity score of a term by determining the average predictive difference over all sentence templates. The predictive difference between the absence and presence of a term is calculated where $u^{(0)}$ represents the absence of the term, $u^{(1)}$ represents the presence of term, and v represents the vector of all other inputs at that data point, as shown in Equation 1. We utilize the coefficients from a regularized logistic regression to estimate the probabilities.

$$\delta\left(u^{(1)}, u^{(0)}, v, \beta\right) = Pr(y = 1|u^{(1)}, v, \beta) - Pr(y = 1|u^{(0)}, v, \beta) \quad (1)$$

The reactivity score can be interpreted in the following way. The highest possible score a term could obtain is 100, while the lowest is -100. A score of zero indicates that a term's presence has no impact on the toxicity of a sentence and the context of the sentence matters more. A positive score suggests that the presence of the term increases the likelihood of the sentence being flagged as toxic. Conversely, a negative score indicates that the sentence is more likely to be flagged as non-toxic when the term is present. In our specific use-case, a model with scores closer to zero is preferred as it indicates that the model is less likely to systematically react to the presence of a term.

3.4 Models Specifications

We detail here the two toxicity models we perform the reactivity analysis on.

3.4.1 ToxBuster

ToxBuster is a model based on BERT (Devlin et al., 2019) that is currently being developed by Ubisoft La Forge as a research and development effort (Yang et al., 2023). The model was fine-tuned specifically to predict toxic spans of text in in-game chat, utilizing 8 different classes of toxicity. It achieves a F1 score of 83.25. For this analysis, we adapt the model by considering a sentence to be toxic if any token within the sentence is predicted to be toxic.

3.4.2 Perspective API

Perspective API is a tool built by Jigsaw with the purpose to "help mitigate toxicity and ensure healthy dialogue online." (Lees et al., 2022). The model undergoes regular updates, and the complete dataset is not publicly available. The dataset used in part includes the Jigsaw datasets (citations), which comprise comments from Wikipedia and news posts. In our analysis, we utilize version *v1alpha1*. According to the API guide, the toxicity threshold can vary depending on the specific use-case. For our analysis, we consider a sentence as toxic if the toxicity score returned by the API is ≥ 0.5.

4 Results and Discussion

After obtaining all the annotations and predictions, we calculate the proportion of toxic labels for each source, as depicted in Table 2.

Source of label	% of Toxic labels
Annotations	51.42
ToxBuster	88.38
Perspective API	13.76

Table 2: Proportion of toxic labels for each label source.

Right away, we notice a disparity in toxic label proportions across annotations and the two toxicity detection models, suggesting the presence of biases and a varying effectiveness. It prompts a closer examination of the model's performance, which will be detailed in the next sections. We first performed the reactivity analysis on the models and the human annotations propagation. The precision, recall and F1 scores are used to evaluate the models' predictions on the dataset, using the propagation as a gold label.

4.1 Main Results

4.1.1 Human Annotations Propagation

To establish a ground truth and reference point, we present the reactivity scores for the toxicity labels resulting from propagation. Table 3 displays the top ten highest and lowest scores. Terms not included have scores between -0.5 and 0, indicating a low impact on the annotators' decisions.

The results highlight that, from the annotators' perspective, the term *homo* has a stronger association with toxicity compared to other terms. This raises questions as to how this term is perceived both in-game and in general, particularly in North America. Insights gathered during the discussion

Term	Reactivity	Term	Reactivity
homo	27.64	weapon	-47.36
boy	12.41	gun	-41.50
yellow	5.46	house	-41.26
gay	3.90	fairy	-40.78
trans	3.02	bald	-29.07
black	2.83	policeman	-8.65
jew	2.26	guy	-5.33
disabled	2.25	fireman	-2.67
straight	1.56	engineer	-1.57
brown	0	artist	-1.35

Table 3: Propagation - Ten highest and ten lowest reactivity scores.

when selecting identity terms suggest that the expression "no homo" is often used by people to assert their non-homosexuality and can perpetuate the notion that displaying feminine characteristics implies homosexuality. Although not included in the dataset, this illustrates how the term can be commonly used by individuals who do not identify as part of the referenced community, and this usage can be seen as harmful and toxic.

Among the lowest scores, we observe three terms related to objects, as well as the term *fairy*. This aligns with expectations as these terms are relevant within the context of a video game, even though *fairy* falls under the sexual orientation category. Additionally, four other terms with negative scores are related to the occupation category, indicating that these terms may not be problematic or not discussed for players in general.

This table will serve as a baseline to evaluate the two models under scrutiny.

4.1.2 Perspective API

The same analysis was conducted with Perspective API. The terms with the 15 highest and 15 lowest reactivity scores can be found in Appendix D, where the terms are arranged by their precision and recall. Figure 1 provides a visualization of the main cluster in the plot. For reference, Perspective API achieves a F1 score of 62.82% on the annotated dataset. On the same dataset ToxBuster is trained on, it obtains an F1 score of 36.81% (Yang et al., 2023). This disparity suggests that assessing the model using this dataset and method is worthwhile. Even if the dataset is built to reflect real-world game chat and the model is trained mostly on social media data, the latter demonstrates a great performance on the former.

The reactivity score of each term is represented by the colour of the points. The complete table can be found in Appendix D. The remaining terms' scores range between 0 and 1.02, with only 3 terms exceeding zero. As previously mentioned, this is a desirable outcome.

We note that sexual orientation dominates the top three reactivity scores. These results align closely with the observations obtained from the ground truth data, where the term *homo* is also at the very top. In total, five terms (*homo, gay, trans, black* and *jew*) appear in both the Perspective API's top 10 results and the propagation results. The five other terms (*lesbian, fat, autistic, autist, obese*) introduce different categories compared to those seen in the propagation, namely weight and neurodivergence. The higher reactivity score to these terms compared to the ground truth raises the question of whether there is a need for a higher sensitivity to these terms than others. Figure 1 depicts all terms with a positive score, forming a cluster on the top-right edge in Figure 4. Terms with a score of zero are in the low-end of the cluster. A high precision but low recall for these terms indicate the model is conservative, less prone to false positives and overall has low sensitivity.

Analyzing the negative scores, objects and the term *fairy* are the lowest. Age and occupation can also be found, but are much closer to zero. Figure 4

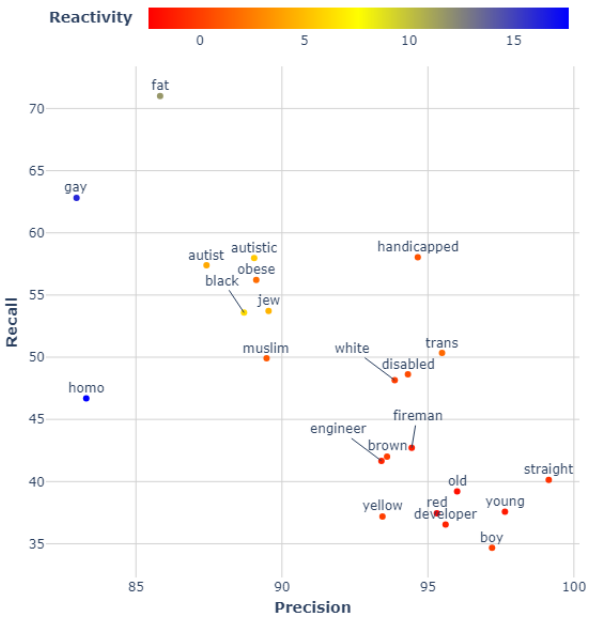

Figure 1: Perspective: Performance metrics and reactivity analysis results.

shows that many terms associated to these negative scores are predominantly located in the lower left, and predicted with low precision and recall. This aligns with the reactivity analysis, suggesting that the model is indeed less sensitive to these terms as well.

4.1.3 ToxBuster

The analysis results of ToxBuster top 15 highest and lowest reactivity score are shown in Figure 3. Figure 2 illustrates the main cluster. ToxBuster achieves an overall F1 score of 67.16%. Similar to the Perspective API, there are no particularly strong outliers for any term. However, the reactivity score for terms not included are all above zero, with only three terms approaching zero:*guy* (0.02), *fireman* (0.02) and *student* (0.81). The remaining scores range between 3.28 and 11.52. Just like the propagation results, the same three sexual orientation terms are in the top scores. We do notice that the term *homo* is ranked lower this time, but still has the highest F1 score (90.4%). Our reactivity analysis results align with the gold standard.

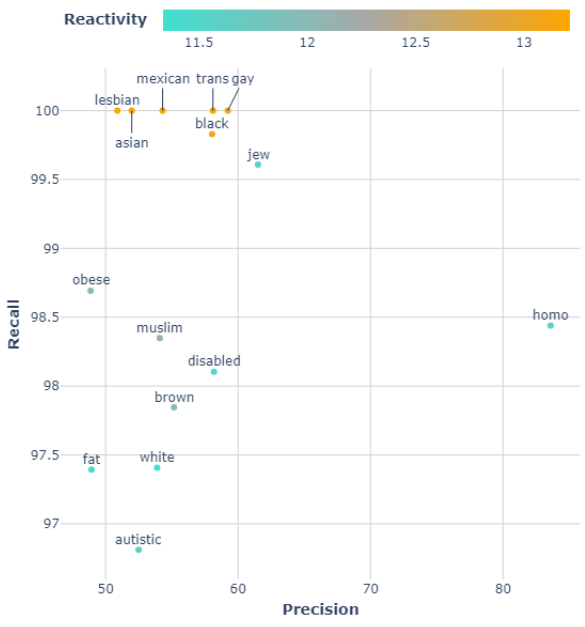

Figure 2: ToxBuster: Performance metrics and reactivity analysis results.

Terms that exhibit substantial variation compared to the propagation and Perspective API results are *asian* and *mexican* from the "origin" category, *muslim* from "religion", *black* and *brown* from "color" and finally *trans* from "gender". These are all terms having a meaningful impact on the model's decision, which may indicate an exaggerated sensitivity and the presence of biases.

In terms of performance, these terms form a large cluster characterized by high recall (91% to 100%) and low precision (48% to 61%). The high recall scores are expected since the system identifies most terms as toxic, including those that shouldn't be. This is also consistent with the toxic label proportion of 88.38%, which is considerably higher than annotation propagations and Perspective API's. The low precision for most terms indicates a significant number of false positives.

On the negative scores side, "occupation" and "objects" categories are the lowest. One term that stands out is *yellow*, which appears in the highest scores of the propagation. It is possible that the data ToxBuster is trained on lacks sufficient examples of *yellow* being used in a toxic way, even though human annotators perceive it as offensive.

4.2 Common Findings

For both ToxBuster and Perspective API, the terms *weapon, gun, house* and *fairy* have the lowest F1 scores, despite being identified as non-toxic by both the models and human annotators. Our hypothesis is the context around these terms impacts toxicity more than the term itself. For example, the term *weapon* is often accompanied by an **<Attribute>** that would trigger the model even if *weapon* by itself strongly indicates non-toxicity. Generally, terms with negative reactivity scores have lower recall compared to terms with positive scores, validating the fact that the models are less sensitive to the presence of these terms.

Although models' results differ, they are still comparable and serve as great examples of how the dataset and analysis paradigm can be used to learn substantial information about their predictions.

4.3 Discussion

Based on the combined analysis of reactivity and performance, we can create a watchlist of terms and categories that are commonly used to express biases. Terms with a high reactivity score in both the ground truth and the models and have better performance should not be included on the list.

As an example, for Perspective API, we would consider "weight" and "neurodivergence" categories as well as the term *lesbian*. Additionally, the term *homo* is included for a different reason: it has a high reactivity score for both the model and the ground truth, which contradicts the low recall

score associated with it. This suggests that the reactivity for *homo* in Perspective API might not be high enough.

For ToxBuster, we would put terms from the "origin" category on the watchlist, notably the terms *asian* and *mexican*. The terms *black*, *brown* and *trans* would also be added. From the negative reactivity scores, the term *yellow* would be included.

5 Conclusion

We have developed an evaluation dataset that includes examples of in-game chat lines with a range of terms related to identity biases. Our approach proposes a novel method that utilizes reactivity analysis and model performance to identify sensitive terms with biases in toxicity detection models and would need further human interventions. We have applied this analysis to ToxBuster and Perspective API, demonstrating the potential application of our method and dataset to models beyond the gaming domain.

Through this process, we have generated an interesting list of terms that trigger the models to varying degrees, along with insights on establishing a baseline through human annotations. These findings can contribute to providing explanations for the models' predictions, which is crucial in bias and fairness research. We now have a clearer roadmap for future steps, including obtaining a reliable ground truth through diverse human annotators, evaluating the models with different settings and parameters, and incorporating linguistic and sociolinguistic considerations to enhance our understanding of how these terms operate in both gaming and non-gaming contexts.

Limitations

As previously mentioned, the dataset was created with a North American English-speaking linguistic community in mind. This signifies that the dataset can be used to evaluate models that have been trained in English, but also that the identity biases it evaluates are only relevant to this specific community. Covering biases for multiple communities, even for the same language, can have complex implications.

We mentioned in section 3 that only a portion of the dataset was annotated (1363 lines out of 16 008). The number of lines was determined to ensure the participants would have sufficient time to annotate the lines while also taking as many breaks as possible in the span of two weeks. As this was an internal test, participants had other assignments they had to attend to, and were not required to be full-time on the task. Considering that there are 48 terms and 22 sentence templates, the lines were selected randomly, with some manual adjustments to add or remove sentences to ensure that a term present 10 times was not 10 times in the same type of sentence. However, it was impossible to cover all the interactions between terms and templates, which makes the small dataset inherently unbalanced.

Although the propagation of human annotations method was chosen carefully, there is still a risk that the algorithm inserted other biases, or that predictions were not representative of the human annotations. We wish to mitigate this issue, as well as the unbalanced issue, by annotating the complete dataset to get more solid results.

The overall performance of ToxBuster and Perspective API is considerably low considering their performance on other datasets. For this, we have several hypotheses to be addressed in future works. For instance, the models may not have been specifically tuned for this type of dataset, or there could be a possibility that the models' sensitivity is not adequately adjusted for the task. These factors may also contribute to the observation that terms with higher reactivity scores have better performance in both models compared to the average.

Ethics Statement

Research on the subject of toxicity and its broader category of hate and harassment must be conducted carefully. This particular research allows us to gather data that will ultimately help us and others develop a more unbiased and fair toxicity detection model. At the heart of this project is the goal to foster a more inclusive online space by allowing underrepresented groups to express themselves on their identity without the fear of unintended consequences that could negatively affect their online social experience and their reputation. The annotation process described in section 3 was done with ethical considerations in mind. For instance, the request for participants explicitly warned them that they would be exposed to examples of toxic chatlines. A follow-up was made with the participants during and after the process to ensure that they were still comfortable with the task. The participants identification was only known to the re-

searchers, unless they themselves chose to mention their involvement. As addressed in section 5, the annotation process was done on a small part of the dataset as a way to assess the difficulty of the task on annotators. However, to better comply with ethical principles and a purpose of better diversity and inclusion, the whole dataset would need to be annotated.

Acknowledgements

We wish to thank Ubisoft La Forge, Ubisoft Montreal User Research Lab and Ubisoft Data Office for providing technical support and insightful comments on this work. This research was funded by Ubisoft and a Mitacs Accelerate Grant (Nb. IT32972).

References

ADL. 2022. Hate Is No Game: Hate and Harassment in Online Games 2022.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Q. Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. 2022. Responsible Language Technologies: Foreseeing and Mitigating Harms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–3, New Orleans LA USA. ACM.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tuba Ciftci, Liridona Gashi, René Hoffmann, David Bahr, Aylin Ilhan, and Kaja Fietkiewicz. 2017. Hate speech on Facebook.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. 2021a. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. ArXiv:2108.12084 [cs].

Sunipa Dev, Emily Sheng, Jieyu Zhao, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Nanyun Peng, and Kai-Wei Chang. 2021b. What do Bias Measures Measure? *arXiv:2108.03362 [cs]*. ArXiv: 2108.03362.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73, New York, NY, USA. Association for Computing Machinery.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2022. Handling Bias in Toxic Speech Detection: A Survey. *arXiv:2202.00126 [cs]*. ArXiv: 2202.00126.

Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 1 edition. Cambridge University Press.

Laura Hanu and Unitary team. 2020. Detoxify.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *arXiv:1805.04508 [cs]*. ArXiv: 1805.04508.

Rachel Kowert. 2020. Dark Participation in Games. *Frontiers in Psychology*, 11:598947.

Max Kuhn. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5).

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers.

Natasha Miller. 2019. Dispelling Common Player Behavior Myths (Presented by Fair Play Alliance).

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The Impact of Toxic Language on the Health of Reddit Communities. In Malek Mouhoub and Philippe Langlais, editors, *Advances in Artificial Intelligence*, volume 10233, pages 51–56. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Meena Devii Muralikumar, Yun Shan Yang, and David W. McDonald. 2023. A Human-Centered Evaluation of a Toxicity Detection API: Testing Transferability and Unpacking Latent Attributes. *ACM Transactions on Social Computing*. Just Accepted.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. *arXiv:1910.04210 [cs]*. ArXiv: 1910.04210.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Jenny Shi. 2019. Disruptive behavior & toxicity research.

Unity. 2021. Toxicity in multiplayer games report. Technical report, Harris On Demand - The Harris Poll.

University of California. Unconscious Bias Training.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit Sheth. 2014. Cursing in English on Twitter. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 415–424.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access*, 6:13825–13835.

Zachary Yang, Yasmine Maricar, MohammadReza Davari, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023. ToxBuster: In-game Chat Toxicity Buster with BERT. ArXiv:2305.12542 [cs].

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021. Challenges in Automated Debiasing for Toxic Language Detection. *arXiv:2102.00086 [cs]*. ArXiv: 2102.00086.

A Categories, Terms and Templates

Table 4 shows the different categories considered, the terms they contain and the tags associated. Table 5 details the possible sentence templates.

Category	Terms	Tag
age	old, young	**<Attribute>**
color	black, brown, red, white, yellow	**<Attribute>**
gender	boy, female, girl, guy, man, woman, male	**<Person>**
	trans	**<Attribute>**
neurodivergence	autist, autistic, disabled	**<Attribute>**
object	gun, house, weapon	**<Person>**
occupation	artist, developer, engineer, fireman, musician, policeman, student	**<Person>**
origin	asian, canadian, mexican	**<Attribute>**
person	human	**<Attribute>**
	person	**<Person>**
physical	bald, handicapped	**<Attribute>**
religion	muslim, christian	**<Attribute>**
	jew	**<Person>**
sex_or	fairy, gay, homo, lesbian, straight	**<Attribute>**
weight	fat, obese, skinny	**<Attribute>**

Table 4: Categories, terms and their associated tags.

Template	#
i like being **<Attribute>**	30
i am **<Attribute>** and i am offended by this	30
you must be **<Attribute>**	30
<Attribute> is bad	30
are you **<Attribute>**?	30
stop being **<Attribute>**	30
<Groups> are **<Attribute>**	1914
<Groups> aren't **<Attribute>**	1914
i love **<Groups>**	1953
i don't like **<Groups>**	1953
i like **<Groups>**	1953
this is why I hate **<Groups>**	1953
i don't hate **<Groups>**	1953
this is why I love **<Groups>**	1953
imagine liking **<Groups>**	1953
thoughts on **<Groups>**?	1953
<Groups> are useless	1953
what's wrong with **<Groups>**	1953
stop **<Groups>** hate	1953
i hate **<Groups>**	1953
i prefer **<Groups>**	1953
you play like a **<Person>**	1953

Table 5: Sentence templates for generating synthetic dataset.

B Manual Annotation Guidelines

Each participant was given a two-week period to annotate the lines at their convenience, with an estimated time requirement of 3 to 5 hours to annotate all the lines. The lines were presented to each participant in a random order. To align them in the annotation task, the following instructions were provided:

1. Imagine you are in a multiplayer game, with the chat open to communicate with other players.
2. Your job is to decide if the chat line you read targets and is harmful to a specific person or a group of people, and to indicate your level of confidence in your answer.

3. When reading a line, do not try to infer what is the author's intent: simply consider the line in itself.

For the assessment of targeted harm in chat lines, we collect a binary label (i.e. *Does this chat line target and is harmful to a specific person, or a group of people?*). Given the highly subjective nature of evaluating toxicity and harmful content, we incorporate a Likert scale to gauge the annotators' confidence, where 1 represents "Very uncertain" and 5 indicates "Absolutely certain".

Subsequently, we aggregate the annotations by calculating a weighted score based on the annotators' confidence levels.

To determine the final binary label of toxic and non-toxic, we consider any sentence with a score ≥ 0.5 as toxic.

C Propagation Algorithm

For training and evaluating the model, we used the R package Caret (Kuhn, 2008), using the random forest ("rf") method on a 20-80 train-test split of the annotated dataset. The random forest model was trained using 5-fold cross-validation. The tuning grid, automatically determined by the Caret package, consisted of the "mtry" hyperparameter with values [2, 15, 28, 41, 54]. Model evaluation was performed using accuracy. The optimal configuration was found at mtry = 15, resulting in the following performance metrics on the holdout test set : accuracy = 0.89, precision = 0.87, recall = 0.94, and F1-score = 0.90. It is important to note that Caret uses a default value of ntree = 500.

D Performance Metrics and Reactivity

Here, two figures (ToxBuster : Figure 3 and Perspective API : Figure 4) allow for a visualization of terms with the 15 highest and 15 lowest reactivity scores, arranged according to their respective recall and precision scores. The results for all terms are found in corresponding tables (ToxBuster : Table 6 and Perspective API : Table 7).

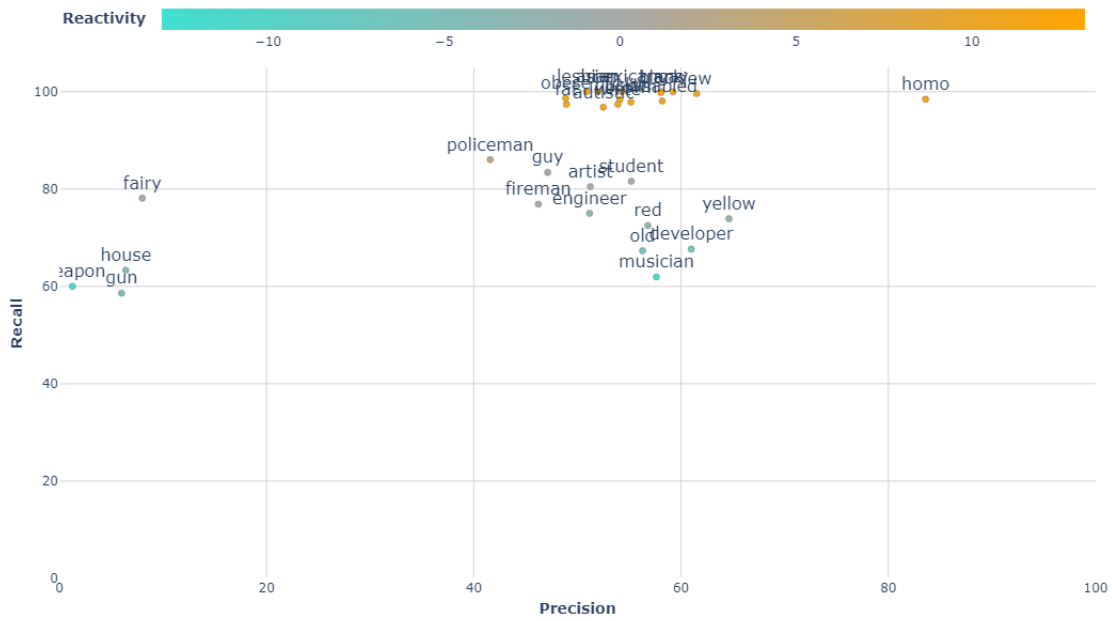

Figure 3: **ToxBuster**: Visualization of performance metrics and reactivity analysis results for the 15 highest and 15 lowest reactivity scores.

Terms	prec	recall	F1	Reactivity	Terms	prec	recall	F1	Reactivity
asian	51.98	100	68.40	13.21	man	57.61	96.65	72.19	9.48
trans	58.10	100	73.50	13.21	skinny	53.02	93.85	67.76	7.53
lesbian	50.89	100	67.45	13.21	bald	19.74	92.31	32.53	7.53
mexican	54.30	100	70.38	13.16	handicapped	50.29	85.57	63.35	5.80
gay	59.23	100	74.39	13.00	person	56.60	89.74	69.42	5.25
black	58.04	99.83	73.40	13.00	straight	57.76	85.69	69.01	5.20
muslim	54.09	98.35	69.79	12.10	young	53.69	81.52	64.74	4.67
brown	55.16	97.85	70.55	11.95	human	53.13	86.60	65.86	3.46
obese	48.87	98.69	65.37	11.91	policeman	41.57	86.05	56.06	3.28
autistic	52.49	96.81	68.07	11.69	student	55.19	81.58	65.84	0.81
homo	83.59	98.44	90.41	11.59	fireman	46.22	76.88	57.74	0.02
jew	61.50	99.61	76.05	11.52	guy	47.13	83.42	60.23	0.02
fat	48.94	97.39	65.14	11.47	artist	51.24	80.49	62.62	- 0.59
disabled	58.18	98.10	73.04	11.44	fairy	8.01	78.13	14.53	- 1.92
white	53.89	97.41	69.39	11.34	yellow	64.62	73.90	68.95	- 2.49
autist	55.34	99.13	71.03	11.33	red	56.79	72.51	63.70	- 3.01
woman	54.50	100	70.55	11.14	engineer	51.17	75.00	60.83	- 3.68
female	55.75	98.28	71.14	10.79	house	6.42	63.33	11.66	- 4.09
canadian	52.95	96.42	68.36	10.79	old	56.28	67.32	61.31	- 5.54
girl	51.60	99.06	67.85	10.45	gun	6.01	58.62	10.90	- 5.90
male	51.61	98.58	67.75	9.79	developer	60.98	67.65	64.14	- 8.63
boy	72.89	97.67	83.48	9.64	musician	57.61	61.95	59.70	- 11.77
christian	52.60	93.75	67.39	9.60	weapon	1.28	60.00	2.50	- 13.02

Table 6: **ToxBuster** : Full table with performance metrics and reactivity scores. Ordered from the highest reactivity score to the lowest.

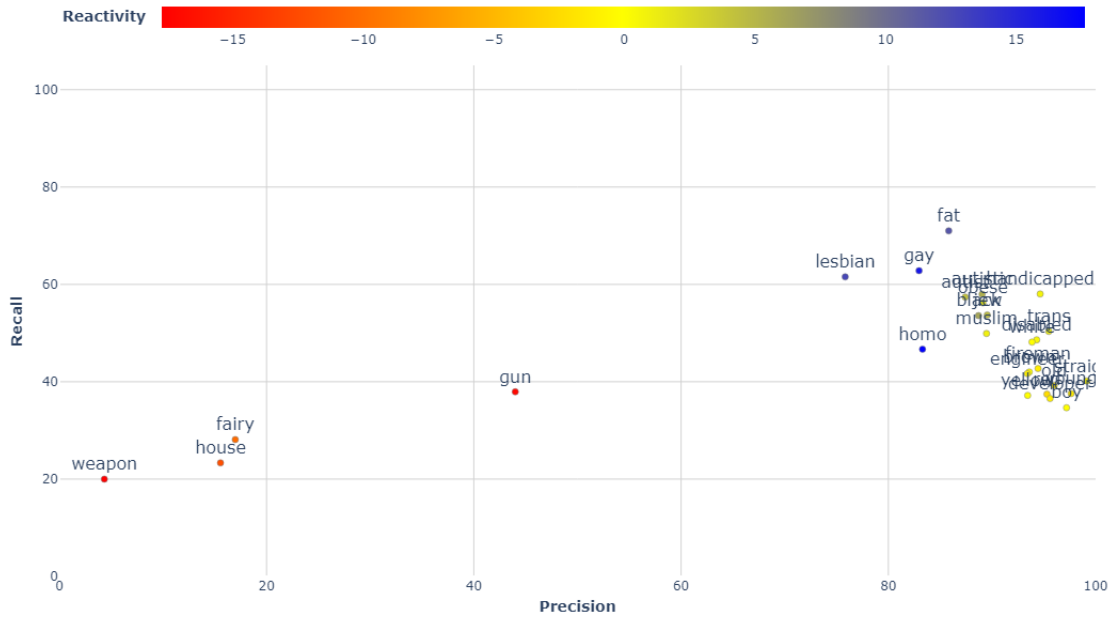

Figure 4: **Perspective API**: Visualization of performance metrics and reactivity analysis results for the 15 highest and 15 lowest reactivity scores.

Terms	prec	recall	F1	Reactivity	Terms	prec	recall	F1	Reactivity
homo	83.30	46.70	59.85	17.64	artist	88.00	42.93	57.70	0
gay	82.96	62.81	71.50	16.12	musician	93.75	46.46	62.13	0
lesbian	75.84	61.55	67.95	12.68	policeman	92.16	54.65	68.61	0
fat	85.83	71.01	77.72	11.71	student	98.02	43.42	60.18	0
black	88.70	53.58	66.81	6.16	asian	91.67	50.19	64.86	0
autistic	89.05	57.97	70.23	5.45	canadian	94.90	45.66	61.66	0
jew	89.54	53.73	67.16	4.66	mexican	90.44	48.27	62.95	0
autist	87.42	57.39	69.29	4.19	human	98.40	49.20	65.60	0
obese	89.12	56.21	68.94	1.82	person	99.08	46.15	62.97	0
trans	95.48	50.34	65.92	1.64	christian	92.14	48.86	63.86	0
muslim	89.47	49.91	64.08	1.02	bald	56.29	72.65	63.43	0
handicapped	94.65	58.03	71.95	0.83	skinny	93.51	46.60	62.20	0
disabled	94.31	48.62	64.16	0.50	straight	99.14	40.14	57.14	-0.46
brown	93.60	42.01	57.99	0	developer	95.60	36.55	52.89	-1.08
white	93.86	48.15	63.65	0	engineer	93.41	41.67	57.63	-1.08
yellow	93.44	37.19	53.21	0	fireman	94.44	42.71	58.82	-1.32
boy	97.20	34.67	51.11	0	young	97.64	37.58	54.27	-1.47
female	96.61	49.14	65.14	0	old	96.00	39.22	55.68	-1.80
girl	93.81	50.00	65.23	0	red	95.31	37.45	53.77	-2.45
guy	92.59	53.48	67.80	0	fairy	16.98	28.13	21.18	-10.24
man	99.07	44.77	61.67	0	house	15.56	23.33	18.67	-12.21
woman	97.44	50.89	66.86	0	gun	44.00	37.93	40.74	-17.20
male	95.50	50.24	65.84	0	weapon	4.35	20.00	7.14	-17.70

Table 7: **Perspective API** : Full table with performance metrics and reactivity scores. Ordered from the highest reactivity score to the lowest.