# ToxiSight: Insights Towards Detected Chat Toxicity

**Anonymous ACL submission**

## Abstract

We present a comprehensive explainability dashboard designed for in-game chat toxicity. This dashboard integrates various existing explainable AI (XAI) techniques, including token importance analysis, model output visualization, and attribution to the training dataset. It also provides insights through the closest positive and negative examples, facilitating a deeper understanding and potential correction of the training data. Additionally, the dashboard includes word sense analysis—particularly useful for new moderators—and offers free-text explanations for both positive and negative predictions. This multi-faceted approach enhances the interpretability and transparency of toxicity detection models.

## 1 Introduction

Toxic and harmful speech in online platforms is an escalating concern, impacting the safety and inclusivity of digital spaces. Detecting and mitigating toxic speech is a critical task, where it has evolved significantly over the years. From early approaches that relied on traditional machine learning models with manually engineered features (Watanabe et al., 2018), the field has progressed to the application of deep neural networks (Gambäck and Sikdar, 2017; Zhong et al., 2016; Gao and Huang, 2017; Fehn Unsvåg and Gambäck, 2018) and, more recently, the utilization of pre-trained language models (PLMs) (Yang et al., 2023). These advancements have led to improved performance across a variety of NLP tasks, including toxicity detection.

In parallel with these advances in detection, the importance of explainability in NLP models has grown. As models become more complex, the need to understand and interpret their decisions—especially in sensitive applications like toxicity detection—has led to the development of various explainability techniques. These techniques range from feature importance analysis (Ribeiro et al., 2016) and surrogate modeling (Ribeiro et al., 2016) to example-driven explanations and provenance-based methods (Pezeshkpour et al., 2019). Visualizing these explanations (Bahdanau et al., 2014; Mullenbach et al., 2018) effectively is crucial for both refining models and making them accessible to moderators, particularly new moderators who need to understand the reasoning behind model predictions.

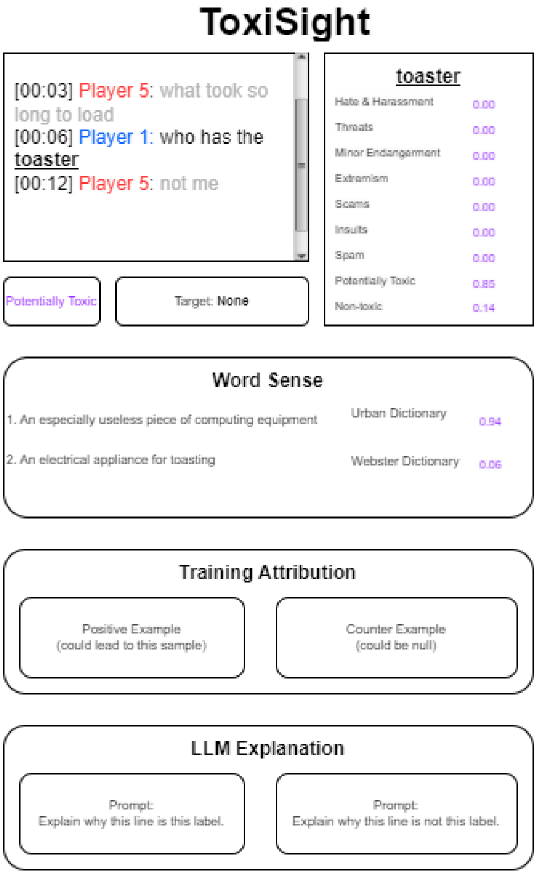
In this context, we introduce ToxiSight, a multi-faceted explainability dashboard designed to enhance the transparency and interpretability of toxicity detection models. ToxiSight integrates existing explainability approaches, including token importance, model output analysis, and attribution to the training dataset with closest positive and negative examples. Additionally, it offers word sense distribution insights and generates free-text explanations for both positive and negative predictions. This comprehensive approach not only aids in the detection and correction of toxic content but also serves as a valuable tool for new moderators, helping them to better understand the nuances of toxicity in online communications.

## 2 Methodology

The development of the ToxiSight dashboard follows a structured approach that integrates various explainability techniques to create a comprehensive tool for understanding and analyzing the inferences made by toxicity detection models. This section outlines the steps involved in implementing the ToxiSight dashboard, detailing each module's function and purpose.

### 2.1 Chat Body

The *Chat Body* module visualizes both the input to the detection model and its corresponding output. The specific chat message under analysis is prominently displayed, with any toxic spans highlighted



ToxiSight

[00:03] Player 5: what took so long to load
[00:06] Player 1: who has the **toaster**
[00:12] Player 5: not me

Potentially Toxic

Target: **None**

| **toaster** | |
|---|---|
| Hate & Harassment | 0.00 |
| Threats | 0.00 |
| Minor Endangerment | 0.00 |
| Extremism | 0.00 |
| Scams | 0.00 |
| Insults | 0.00 |
| Spam | 0.00 |
| Potentially Toxic | 0.85 |
| Non-toxic | 0.14 |

Word Sense

| | | |
|---|---|---|
| 1. An especially useless piece of computing equipment | Urban Dictionary | 0.94 |
| 2. An electrical appliance for toasting | Webster Dictionary | 0.06 |

Training Attribution

Positive Example
(could lead to this sample)

Counter Example
(could be null)

LLM Explanation

Prompt:
Explain why this line is this label.

Prompt:
Explain why this line is not this label.

Figure 1: Dashboard for insights towards detected chat toxicity

through bolding and underlining to draw attention. Beneath the chat message, the dashboard shows the predicted label, indicating whether the content is classified as toxic or non-toxic. If the label is toxic, the module also identifies the target of the toxicity, if applicable. To the right of the chat message, a detailed distribution of the model's output probabilities across different toxicity classes is provided. This offers a granular view of the model's decision-making process.

### 2.2 Word Sense

The *Word Sense* module enhances interpretability by analyzing specific words or phrases identified as potentially toxic. For the highlighted span, ToxiSight determines the most likely meaning of ambiguous or context-dependent words, which is crucial for understanding why certain words were flagged as toxic. The module leverages authoritative sources like Webster, Oxford, and crowd-sourced sources such as Urban Dictionary to provide definitions, displaying multiple senses for each word ranked by relevance to the context, with accompanying confidence scores. This analysis helps to clarify the specific usage and intent behind the language in the chat, contributing to a more nuanced understanding of the model's output.

### 2.3 Training Data Attribution

The *Training Data* Attribution module employs example-driven explainability by tracing the model's prediction back to its training data. This is achieved by identifying a positive example, a similar sample from the training dataset that closely matches the input and supports the same classification. Additionally, the module provides a counter-example, which is a contrasting sample that would have led to a different classification, such as non-toxic. If no suitable counter-example exists, this section may be empty, due to the absence of a sufficiently similar instance. These examples are identified through similarity measures, ensuring that the samples provided are the most relevant and informative for understanding the model's behavior.

### 2.4 Explanation Generation

The *Explanation Generation* module leverages large language models (LLMs) such as LLaMA-3 and GPT-4o to generate free-text explanations that clarify the model's decision. The model is prompted to generate two types of explanations: one that justifies why the input was classified under the given label and another that explores why the input was not classified under an alternative label. These explanations are designed to help researchers and moderators understand the model's reasoning, providing both the justification for its decision and the plausible deniability of alternative classifications. This dual approach offers a balanced perspective, aiding in both the validation and critique of the model's outputs.

## 3 Results

The implementation of the ToxiSight dashboard has significantly enhanced our understanding of toxicity detection models by providing detailed insights into their decision-making processes. Preliminary testing shows that the dashboard effectively highlights ambiguous cases, allowing for more informed moderation decisions. Furthermore, the integration of training data attribution and word sense analysis has improved the model's interpretability, enabling users to trace predictions back to specific examples and understand the contextual nuances that influence toxicity classification.

## Limitations

While the ToxiSight dashboard offers a comprehensive tool for understanding toxicity detection models, there are some limitations to consider. First, the reliance on large language models (LLMs) for generating explanations may introduce biases inherent in the models themselves, potentially skewing the interpretations provided. Additionally, the example-driven approach for training data attribution depends heavily on the quality and diversity of the training dataset. If the dataset lacks representation of certain contexts or language variations, the attributions may be less reliable or informative. The dashboard also assumes that the most relevant word senses and training examples can be accurately identified, which may not always be the case, particularly in complex or highly nuanced conversations. Lastly, while the dashboard aids in interpreting model outputs, it does not guarantee improved performance or fairness in toxicity detection, and there may still be challenges in generalizing its insights across different domains or user groups.

## Ethics Statement

There is a risk that over-reliance on model explanations could lead to unjust outcomes, especially if the explanations are taken as definitive without sufficient human oversight.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint*.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on Twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85, Brussels, Belgium. Association for Computational Linguistics.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3336–3347, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *arXiv preprint*.

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, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023. Towards detecting contextual real-time toxicity for in-game chat. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9894–9906, Singapore. Association for Computational Linguistics.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *International Joint Conference on Artificial Intelligence*.