Kevin Roa
ACL Paper Summary
CS4395.001

Information

- Title: CONDA: a CONtextual Dual-Annotated dataset for in-game toxicity understanding and detection
- Link: https://aclanthology.org/2021.findings-acl.213
- Authors:
    - Henry Weld
    - Guanghao Huang
    - Jean Lee
    - Tongshu Zhang
    - Kunze Wang
    - Xinghong Guo
    - Siqu Long
    - Josiah Poon
    - Soyeon Caren Han
- Affiliations: School of Computer Science, The University of Sydney, NSW, Australia

Problem Summary

Toxicity within modern games is becoming an increasingly bigger problem and current solutions are inadequate to handle them. Traditional toxicity detection models only focus on a surface-level understanding of the text without taking into account the deeper context or semantic clues of the conversations. If the conversations are annotated at both an utterance and token level while also utilizing the prior chat history, a more robust dataset can be built. By applying a natural language understanding (NLU) approach, a more applicable dataset can be created to combat the increase in player harassment from toxic behavior in games.

A major point of interest for this paper is the focus on in-game language rather than generalized toxicity detection. General purpose datasets are not able to cover the unique aspects of in-game language such as game terminology, slang, and context dependant conversations, etc. The framework they propose is more robust for this type of use case and can be applied elsewhere. While prior approaches only analyzed toxicity at an utterance level, this new approach analyzes toxicity in the context of entire conversations and accounts for both utterances and tokens.

Prior Works

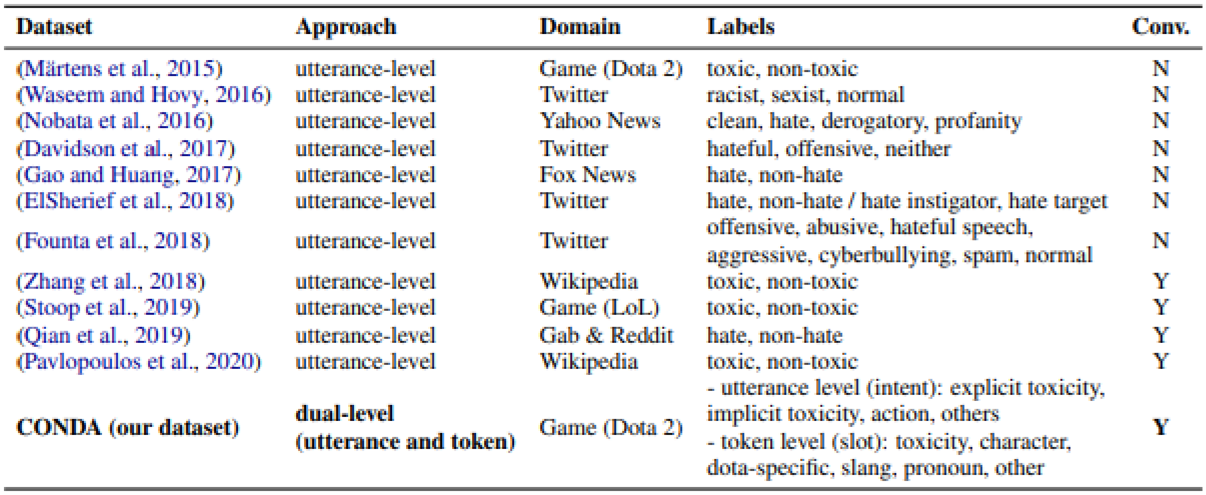
Toxicity Datasets in Online Games

- Analyzed anti-social and disruptive behavior at a single utterance level.
- Typically focused on concepts such as cyberbullying and griefing.
- Rudimentary data annotation using previously established lexicon categories.
    - Not robust enough to handle the unique characteristics of online communication in games.

Toxicity Datasets in Online Community

- Analyzed hate speech and abusive language at a single utterance level.
- Focused on creating solutions that detect warning signals or generate intervention responses

Table:



Dataset	Approach	Domain	Labels	Conv.
(Märtens et al., 2015)	utterance-level	Game (Dota 2)	toxic, non-toxic	N
(Waseem and Hovy, 2016)	utterance-level	Twitter	racist, sexist, normal	N
(Nobata et al., 2016)	utterance-level	Yahoo News	clean, hate, derogatory, profanity	N
(Davidson et al., 2017)	utterance-level	Twitter	hateful, offensive, neither	N
(Gao and Huang, 2017)	utterance-level	Fox News	hate, non-hate	N
(ElSherief et al., 2018)	utterance-level	Twitter	hate, non-hate / hate instigator, hate target	N
(Founta et al., 2018)	utterance-level	Twitter	offensive, abusive, hateful speech, aggressive, cyberbullying, spam, normal	N
(Zhang et al., 2018)	utterance-level	Wikipedia	toxic, non-toxic	Y
(Stoop et al., 2019)	utterance-level	Game (LoL)	toxic, non-toxic	Y
(Qian et al., 2019)	utterance-level	Gab & Reddit	hate, non-hate	Y
(Pavlopoulos et al., 2020)	utterance-level	Wikipedia	toxic, non-toxic	Y
CONDA (our dataset)	**dual-level (utterance and token)**	Game (Dota 2)	- utterance level (intent): explicit toxicity, implicit toxicity, action, others - token level (slot): toxicity, character, dota-specific, slang, pronoun, other	**Y**

Table 1: Comparison of CONDA with other toxicity datasets (Conv.: Conversation).

Unique Contributions

- Attempts to build a toxicity detection dataset using both intent classification and slot filling. It takes a Natural Language Understanding (NLU) approach toward toxicity detection rather than Natural Language Processing (NLP).
- Uses a robust dual semantic level toxicity framework that can handle utterance and token level patterns. It takes into account the rich chat history rather than single utterances.
- Created formalized metrics for toxicity detection which can be used to evaluate other models and datasets.

Evaluation Criteria

- UCA
    - Utterance classification accuracy measure
    - Sentence-level classification is based on the ratio of the number of correctly predicted utterances to the total number of utterances
    - Mainly used in abusive language detection
- U-F1
    - Utterance F1 score
    - Combines precision and recall classifiers into a single metric
        - Utterance score based on relevance
        - Attempt to seek a certain balance between precision/recall
- T-F1
    - Token F1 score
    - Combines precision and recall classifiers into a single metric
        - Token score based on relevance
        - Attempt to seek a certain balance between precision/recall
    - Get prediction performance for slot tokens and averages it for each class
- JSA
    - Joint semantic accuracy
    - Overall prediction performance over the semantic hierarchy
    - Combination of utterance level and token level predictions
        - Correct analysis is when both prediction levels are correct

Citations

This paper has only been cited a total of 4 times, though, the individual authors of the paper have been cited multiple times elsewhere.

- Henry Weld (lead author)
    - 54 Citations
    - Worked on:
        - joint intent detection
        - slot filling models
        - Intent classification
- Josiah Poon (Advisor)
    - 2691 citations
    - Has played a role in the creation of various AI-related papers
    - Covers areas such as NLP/text classification/sentiment analysis/recommendation systems
        - Using techniques such as convolutional networks/deep learning/deep reinforcement learning agents
- (The other authors of the paper have also made notable contributions in related areas, however, they won't be discussed here)

References

[1] Henry Weld, Guanghao Huang, Jean Lee, Tongshu Zhang, Kunze Wang, Xinghong Guo, Siqu Long, Josiah Poon, and Caren Han. 2021. 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. Association for Computational Linguistics.
[2] https://scholar.google.com/citations?user=I-u_06gAAAAJ&hl=en&oi=ao
[3] https://scholar.google.com/citations?hl=en&user=Q7U0O0gAAAAJ&view_op=list_works&sortby=pubdate