SemEval-2021 Task 5: Toxic Spans Detection Project Proposal

Yousef Nademi, Ravika Nagpal
CCID: nademi, ravika

1 Task/Introduction

The definition of toxic language is broad and includes *offensive, abusive, hateful, racist*, and any other negative language form. Detection of toxic language is a challenging task for any online and interactive platform. The goal of toxic language detection is to promote free and healthy conversations while also protecting marginalized voices [Pavlopoulos et al., 2020]. Toxic spans detection is one of these tasks which, focuses on evaluating systems that recognize the spans in a text that make it toxic. Identifying these spans can assist human moderators (for example, news portal moderators) who frequently deal with lengthy comments to not rely solely on a system-generated unexplained toxicity score per post.

Much literature and annotated datasets have been developed to identify any form of toxic language, and tremendous progress has been made in this area [Pitenis et al., 2020]. Detection of the particular text spans that make a comment offensive has been mainly neglected. The current state-of-the-art offensive language identification models flag the entire comment rather than highlighting the offensive parts [Ranasinghe et al., 2021]. In this project, we will address the detection of whole toxic words from the text provided.

2 Related Work/State-of-the-art

Detection of toxic language attracted a lot of attention. However, the focus was primarily on labeling the entire comment. Pavlopoulos et al. [Pavlopoulos et al., 2020] investigated the role of text context, assuming that comments cannot be judged independently. To that end, they investigated two specific research questions: (1) does context affect human judgment, and (2) does taking the context of the comment into account improve the performance of toxicity detection systems? They have used the Wikipedia conversations dataset by limiting the context to the previous post in the thread and the discussion title. Their result showed that the context of comments could both amplify or mitigate the perceived text's toxicity (around 5% change in the label). However, the context's consideration did not improve the performance of toxicity classifiers. Others have looked into the detection of toxic language, but there haven't been many studies that focus on toxic spans detection. Ranasinghe et al. [Ranasinghe et al., 2021] worked on the SemEval-2021 Task 5: Toxic Spans Detection competition. Lexicon-based Word Match was used as their baseline algorithm that had a poor performance. They also utilized various Neural Transformers, including Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018]. They achieved an **F1-Score of 0:68** for their best performing model.

Chhablani et al. [Chhablani et al., 2021] explored a simple version of the above two approaches, i.e., token classification and span prediction. BERT-based models - BERT, RoBERTa, and SpanBERT - were used in this paper for both tasks. When these models were combined, the prediction of ToxicSpans improved by 3 percent. Results were investigated on four hybrid approaches - Multi-Span, Span+Token, long short-term memory(LSTM)-conditional random field (CRF), and a combination of predicted offsets using union/intersection. Another work by Luu and Nguyen [Luu and Nguyen, 2021] used the combination of Bidirectional LSTM(BiLSTM)-CRF model [Lample et al., 2016] and the ToxicBERT [Hanu and Unitary team, 2020] for classifying toxic words in the post.

Text	Toxic Spans
So, who did that stupid Photoshop work? It's awful!	[17, 18, 19, 20, 21, 22]
And the damn apartments are down the street by Waikele - talk about interesting coincidence!!!!!	[8, 9, 10, 11]
I loved the paintings that he bought!	[]
Rotten to the core. Put them in prison until they perish. They should never have a day of freedom again.	[0, 1, 2, 3, 4, 5]

Table 1: Four comments from the dataset along with their annotations. The offensive words are displayed in the red color and the spans are indicated by the character position in the instance.

3 Methodology

3.1 Task and Dataset(Available Data)

The dataset was downloaded from the competition website[1]. The dataset consists of posts in the Civil Comments Dataset that were considered to be toxic. It includes 690 instances, out of which 43 instances do not contain any toxic spans. The training dataset has a total of 7,939 instances, and 485 of these are without any toxic spans. Each data point includes a list of abusive spans and the post. All text are only in English and no other language is considered.

Four randomly selected examples from the training dataset, along with their annotations are shown in Table 1.

3.2 Available Code

Upon searching the internet, there are some repositories on the same task. However, we neither ran nor carefully read them. If we decide to use some of their code in our project, we will cite them properly in both the check-in and final reports.

3.3 Algorithm

First, a simple baseline algorithm will be chosen, such as the SpaCy EntityRecognizer(NER) Tagging model. The SpaCy NER Tagging model is a NER classifier that is based on SpaCy Language Models. Using the spans provided, it predicts the entities labeled as TOXIC in the text. Then, two or three cutting-edge pre-trained algorithms, such as BERT, Bi-LSTM, ToxicBERT, and others, will be chosen and fine-tuned for our task. The exact algorithms selected will be discussed in the check-in report.

3.4 Evaluation Metrics

For model evaluation, we will utilise the same metrics as proposed by the competition. A brief summary of these metrics is provided below.

Let system A_i return a set $S^n_{A_i}$ of character offsets for parts of a text that was found to be toxic. Let Gr_n be the character offsets of the ground truth annotations of n. Using these information, the $F1$ score of system A_i with respect to the ground truth Gr for post n will be calculated (See Equation 1).

$$F^t_1(A_i, Gr) = \frac{2.P^n(A_i, Gr).R^n(A_i, Gr)}{P^n(A_i, Gr) + R^n(A_i, Gr)} \quad (1)$$

Where P^n and R^n are the precision and the recall, respectively.

The performance of selected models will be evaluated using this criterion. We will further analyze the models' performance using statistical significance tests. These tests can show the likelihood of observing the same performance among different algorithms, thereby confirming that the difference in models' performance is statistically significant.

4 Repository URL

We have created the repository for our project and the link for that is:
https://github.com/UOFA-INTRO-NLP-F21/f2021-proj-ravikanagpal

References

Gunjan Chhablani, Abheesht Sharma, Harshit Pandey, Yash Bhartia, and Shan Suthaharan. NLRG at SemEval-2021 task 5: Toxic spans detection leveraging BERT-based token classification and span prediction techniques. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 233–242, Online, August

[1] https://competitions.codalab.org/competitions/25623#learn_the_details-data

2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.27. URL https://aclanthology.org/2021.semeval-1.27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1030. URL https://aclanthology.org/N16-1030.

Son T. Luu and Ngan Nguyen. UIT-ISE-NLP at SemEval-2021 task 5: Toxic spans detection with BiLSTM-CRF and ToxicBERT comment classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 846–851, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.113. URL https://aclanthology.org/2021.semeval-1.113.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*, 2020.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*, 2020.

Tharindu Ranasinghe, Diptanu Sarkar, Marcos Zampieri, and Alex Ororbia. Wlv-rit at semeval-2021 task 5: A neural transformer framework for detecting toxic spans. *arXiv preprint arXiv:2104.04630*, 2021.