SemEval-2021 Task 5: Toxic Spans Detection Project Check-in Report

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 objective of Detection of toxic language is providing a safe environment for a beneficial and constructive conversations [Pavlopoulos et al., 2020]. The task of Toxic spans detection is finding the span of toxic and offensive words, thereby empowering human moderators (for example, sport moderators) to moderate better [Webpage, 2021]. Moderators frequently encounter lengthy comments which may contain toxic words. To find these comments, they utilize a machine generated score for toxicity. However, there are no explanation associated with these generated scores [Webpage, 2021]. Toxic span detection will fill this gap and provide an explanation on why a comment/text is toxic. In this project, we will address the span of the toxic word(s) by using advanced pre-trained models for the data provided in the SemEval-2021 Task 5: Toxic Spans Detection.

2 Related Work/State-of-the-art

Due to generation of millions of daily comments/posts in the social media, this area has attracted the attention of many scholars. However, the focus was primarily on labeling the entire comment. Pavlopoulos et al. [2020] investigated the role of text context, assuming that comments cannot be judged as standalone independent text. 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. They also limit their difinition of context to the previous post in the thread and the discussion title. Their results 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 have not been many studies that focus on toxic spans detection. 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 performed poorly. 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. [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 Toxic Spans 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 [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.

3 Methodology

3.1 Task and Dataset(Available Data)

The dataset was downloaded from the link provided in the competition website [Webpage, 2021]. The dataset includes the toxic comments/posts in the Civil Comments Dataset. Total number of the train-

Text	Toxic Spans
And the damn apartments are here.	[8, 9, 10, 11]
I loved the paintings that he bought!	[]
Rotten to the core.	[0, 1, 2, 3, 4, 5]

Table 1: Three 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.

ing data is 7,939, where 485 of them are not toxic. For each datapoint, toxic spans were stored in a list along with the post that includes toxic words with the given toxic span. It is worth mentioning that the only language present in the dataset is English. Three randomly selected examples from the training dataset, along with their annotations are shown in Table 1. There are some github page on the same task. We have checked this specific Github page (https://github.com/gchhablani/toxic-spans-detection) to help up develop our idea.

3.2 Preprocessing

The dataset was preprocessed using SpaCy EntityRecognizer (NER) Tagging model [Honnibal and Montani, 2017] . The SpaCy NER Tagging model [Honnibal and Montani, 2017] is a NER classifier that is based on SpaCy Language Models. We first tokenize each word of the sentence along with its onset in a tuple. Then, if the onset of the word was in the provided toxic span, the word was labeled as T (Toxic), and NT (Not Toxic) otherwise. Following this step, we have calculated the length of longest sentence in the dataset. Using this length, we padded the shorter sentences with zero so that all sentences have the same input length. The preprocessed text will be feed to the algorithms for token classification.

3.3 Algorithm

Two or three cutting-edge pre-trained algorithms, such as BERT, Bi-LSTM, ToxicBERT, will be applied and fine-tuned for our task. The result will be presented in the final report of our project.

3.4 Evaluation Metrics

For model evaluation, we will utilise the same metrics as proposed by the competition and described in details in the paper of [Chhablani et al., 2021]. A short description of the metric is reproduced below.

Let system A_i return a set $S^n_{A_i}$ that includes the character offsets of toxic words. Let Gr_n be the character offsets of the ground truth annotations of n. Using these data, the $F1$ score of system A_i with respect to the ground truth Gr for post n can be measured using 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. The student's t-test will be used to check the null hypothesis. The null and alternate hypotheses are as follows:

- Null hyphothesis: The mean of the test accuracy of pair of an algorithm is the same.
- Alternate hypothesis: The mean of the test accuracy of pair of an algorithm is different.

The test will show the likelihood of observing the same performance among different algorithms. If the test is passed, it will confirm that there is a statistically significant difference between models' performance.

3.5 Results

As of now, we have completed the pre-processing of the data that was explained in detail in the pre-processing section of the report which includes, tokenisation, padding of the text, and labeling of tokens. We will be applying the Bert-based models to this processed data.

4 Workplan

Task	Date
Implementation of Bert(Yousef)	7 Nov
Implementation of Bi-LSTM(Ravika)	10 Nov
Implementation of ToxicBERT(Ravika)	17 Nov
Comparison of Models(Yousef)	25 Nov
Final Report's preparation(Ravika,Yousef)	30 Nov

5 Repository URL

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 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.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

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.

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.

Competition Webpage. SemEval 2021 Task 5: Toxic Spans Detection: https://competitions.codalab.org/competitions/25623#learn_the_details. 2021.