SemEval-2021 Task 5: Toxic Spans Detection Project Final Report

Yousef Nademi, Ravika Nagpal
CCID: nademi, ravika

1 Abstract

Detection of toxic spans will explain why classifiers may flag a text as toxic. They can also be used on social media platforms to erase only the offensive part of a statement rather than the entire comment or text. As part of this project, we attempted to detect toxic spans in text using the dataset from SemEval-2021 Task 5: Toxic Spans Detection. Since toxic words are tokens of words in a sentence, this task can be solved using the token classification approach. The text is tokenized first, and then each token is classified as Toxic (T) or Not Toxic (NT). The span of T tokens can be obtained based on their index in the text. The RoBERTa based model and the DistilBERT base model were employed as pre-trained token classification BERT models. Both the models performed well as compared to the random baseline model shared in the competition, which gives the F1 score of 12.22 [Pavlopoulos et al., 2020]. The RoBERTa-based model had the best F1 score of 0.561.

2 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 the detection of toxic language is to provide a safe environment for 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 is no explanation associated with these generated scores [Webpage, 2021]. Toxic span detection will fill this gap and explain 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.

3 Related Work/State-of-the-art

Due to the generation of millions of daily comments/posts on 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 a 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 definition 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.

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.

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.

4 Methodology

4.1 Task and Dataset(Available Data)

The dataset was downloaded from the link provided on the competition website [Webpage, 2021]. The dataset includes the toxic comments/posts in the Civil Comments Dataset. The total number of the training 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. Table 1 displays three randomly selected examples from the training dataset, along with their annotations. We have checked this specific Github page[1] to help up develop our idea.

4.2 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. Using SpaCy NER, the sentences were tokenized and stored in a tuple along with their onset. 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. The tokens were then fed into the tokenizer class of the used algorithms to refine them so that they were compatible with the algorithms. Doing so will lead to an increase in the number of tokens. To address this issue, the extra tokens generated for each token were labeled the same as the original label of the token.

The model added the beginning and ending tokens to each sentence. In the last step, tokens were padded to have the same length before training. Afterward, pre-processed tokens were fed to the algorithms for token classification. The reason for converting the problem from span classification to token classification was that many pre-trained models were developed for such tasks, and it can perform better than training a simpler approach or re-training the pre-trained models. Also, the dataset is not large enough to provide a good result for re-training. Therefore, we utilized token classification as an intermediate stage for toxic span detection.

4.3 Algorithm

The pre-trained models of RoBERTa base [Liu et al., 2019], and DistilBERT base model (uncased) [Sanh et al., 2020] were selected for token classification. These models were obtained from the Hugging Face website [Wolf et al., 2020].

RoBERTa base model is a case-sensitive algorithm that was trained on a large English corpus. As stated in the Hugging Face website[2], humans did not play a role in token labeling. Masked language modeling (MLM) was utilized to train this model such that 15 percent of a sentence was masked, and the remaining section was used for training. Then, mask sections were fed into the model, and the classifier needed to label the masked section (https://huggingface.co/roberta-base).

Unlike the RoBERTa base model, the DistilBERT base model (uncased) is case insensitive. Similar to the RoBERTa base model, it was trained on the English corpus, and during the training, no human labeling was involved. It also used MLM

[1] https://github.com/gchhablani/toxic-spans-detection

[2] https://huggingface.co/roberta-base

for training[3].

4.4 Evaluation Metrics

For model evaluation, we used the same metrics as proposed by the competition and described in detail by 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 the system A_i concerning 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 is evaluated using this criterion. We further analyzed the models' performance using statistical significance tests. The student's t-test is used to check the null hypothesis. The null and alternate hypotheses are as follows:

- Null hypothesis: 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. Passing the test confirms a statistically significant difference between models' performance. The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis. A p-value less than 0.05 (typically 0.05) is statistically significant.

4.5 Computational Requirements

The Google Colab's Pro GPU (NVIDIA K80/P100) was employed for fine-tuning and testing the algorithms. Training of each model takes about 6 hours to finish. No hyperparameter tuning was used during training. Following parameters were set during the model's training:

Epochs=3, learning rate=2e-5, batch size=10, optimizer = AdamW(weight decay of 0.01)

Default values were used for other parameters.

4.6 Results

After training, the F1 performance of the models were measured on the test dataset. The result are summarized in Table 2.

Model	F1 score
RoBERTa base model	0.561
DistilBERT base model (uncased)	0.534

Table 2: The model's performance on the test dataset

To evaluate the possibility of observing the same result from either of these models, the statistical significance test was performed on the prediction accuracy of both classifiers on the test set data. The calculated p-value was 0.821, indicating that the difference between these models in terms of accuracy is not statistically significant. However, it should be noted that for P-value calculations, F1 performance was not compared, and the test was performed on the accuracy of the prediction. The accuracy of both classifiers was high, and both classifiers can correctly predict many of the labels.

4.7 Discussion

In this project, the F1 performance of two Bert-based models of RoBERTa based model and DistilBERT base model were investigated on the Toxic Span Detection SemEval-2021 Task 5 dataset.

Both models had comparable performance and were able to correctly label the tokens as T or NT. Using these labels, one can compute the spans after classifying the original tokens of the sentence.

For the specific task, two studies by Devlin et al. [2018] and Chhablani et al. [2021] have achieved the F1-Score of 0.68 for their test dataset. Our performance was lower than the one reported by these authors, which can be attributed to two main reasons:

1. They had more extensive preprocessing on tokens before feeding them to the Bert Based models. For example, they have added special tokens of [SEP] to their sentences when the evaluated text has been made of 1 sentence. Here, special tokens of the start and end of the sentence were only added to the beginning and end of data points, even if it was made of more than one sentence. Furthermore, no lowercase conversion of tokens was performed. The reason was that we wanted to check the performance of case sensitives and case insensitive Bert models.

[3] https://huggingface.co/distilbert-base-uncased

2. Extra layers of Long Short Term Memory(LSTM) were not added to the model architectures. The additional layers will aid the algorithm in learning existing patterns in the toxic span dataset, ultimately improving model performance. Trying these approaches will boost the effectiveness of pre-trained Bert models but will be computationally more expensive to train.

Another important note is that the current dataset is relatively small. If the annotated dataset was larger, the algorithms could learn better patterns and eventually better performance. However, Generating annotated toxic span dataset is a tedious task, and it is time-consuming to develop.

4.8 Conclusion

In this project, two pre-trained models of RoBERTa based model and DistilBERT base model were used to detect the toxic span of a text. Both models performed relatively well. The F1 scores were 0.561 and 0.534, respectively, for RoBERTa based model and DistilBERT base model. As explained in the discussion, further preprocessing of the text, as well as the addition of extra layers of training on top of pre-trained models, will likely improve the performance of toxic span detection.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.