SemEval-2021 Task 5: Toxic Spans Detection Using BERT-based Model with Data Augmentation

CHONG Cheuk Hei, SINGH Jaspreet Dhaliwal

Abstract

There is an increasing trend of bullying via toxic comments on online discussion such as Reddit and Twitter which may lead to psychological effects on vulnerable victims. To combat this, SemEval-2021 Task 5: Toxic Spans Detection has been introduced. The task is to identify toxic spans from sentences in a document. Analysis has been done on the dataset to extract information. We have solved this task using pre-trained BERT, RoBERTa and DistilBERT models, with adding the feature of data augmentation. Our approach gains the highest test F1-score of 0.698, which remains 0.01 to reach the highest score in this task.

1 Introduction

The democratization of social media has made it easier for people to write their feelings and views on different topics. However, with the rise of social media, comes an increase in bullying and hate speech. People can abuse their 'freedom of speech' and hurt others mentally. To combat this, tasks such as toxicity detection have been done. Categorization of toxicity in terms of level has also been done where the algorithm predicts toxicity and categorizes them into certain levels. All this research has been implemented in several applications, but they have one significant flaw. When a toxic span is detected or identified by the system, moderators or administrators delete the entire tweet. They do not identify the spans/location that make a sentence toxic. We plan to highlight such toxic spans so that they could assist human moderators who often deal with long texts, the moderators do not need to delete the entire message, they can just filter out the toxic spans.

Most of the current research is mostly a binary classification problem. It only identifies if a sentence or a text in a document is toxic or not. The only improvement to this was that toxicity is categorized in levels. The fundamental problem of specific locations that make a text toxic is still not identified. In SemEval-2021 task5 (toxic spans detection), more effort is made in distinguishing what make a word toxic. We have three approaches towards this problem.

The paper contents are listed as follows. Section 2 talks about some works that were done before which we took inspiration from. Section 3 contains key statistics for the dataset. Our approach and results are discussed in task 4 and 5 respectively with Section 6 to summarize our findings.

2 Related Work

2.1 Toxic-related Task

To combat and facilitate the detection of toxic comments, there were different academic tasks which are related to toxic comments classification. Similar tasks like SemEval-2019 Task 5 (Basile et al., 2019) which focuses on Multilingual Detection of Hate Speech Against Immigrants and Women, . There is also task working on Automated Hate Speech Detection (Davidson et al., 2017) with the data provided by Twitter API. However, most of the toxic detection is within sentence-level which might not able to pinpoint which keywords causing to be toxic.

2.2 Detecting Toxicity

There are many approaches on measuring the toxicity of the texts given. It was common to implement convolutional neural networks (CNNs) and Long Short-term Memory (LSTM) networks in the past to benchmark aggression in social media (Kumar et al., 2018). Different pre-trained word embeddings like GloVe (Pennington et al., 2014), FastText(Joulin et al., 2016) are also integrated into DNNs for training (Badjatiya et al., 2017). However, there are long dependencies on toxic

Texts	Spans
Because driving under Ontario laws is **stupid** enough in Ontario.	[38, 39, 40, 41, 42, 43]
Ignorant, selfish, racist, misogynistic snowflakes were THE major part of Trump gaining enough electoral votes. You can't fix **stupid**.	[0, 1, 2, 3, 4, 5, 6, 7, 127, 128, 129, 130, 131, 132]
Says the **hypocrite** who STILL refuses to denounce a disgusting **racist**, or the Dems who reelected him 8 times.	[9, 10, 11, 12, 13, 14, 15, 16, 17, 62, 63, 64, 65, 66, 67]

Table 1: Examples of Dataset

wordings as early part of the comment is the key to detect the toxicity. With the lack of self-attention layer, the above method is challenging for detecting the whole toxic phrases accurately. Currently, with the existence of BERT (Devlin et al., 2018), it further improves the accuracy score on detecting-toxic-related tasks.

3 Dataset

The dataset for SemEval-2021 Task 5 was extracted from Civil Comments Data set (Borkan et al. 2019). The entire data set was off 1.2 million posts and comments. 30 K toxic posts were chosen from 1.2 million posts in the original dataset.

Table 1 shows the example of dataset. If the comment is not toxic, no span is highlighted. It is to be noted that not all toxic spans could be annotated in all toxic posts. Some comments may be inherently toxic, but it is difficult to assign a particular span as the source of toxicity. As a result, nothing is highlighted.

The dataset is split into training and testing data. The training set consist of 7,939 sentences while the test set consists of 2000 sentences. The dataset may consist of one or more toxic spans.

In Figure 1, there is a key finding that majority of the toxic spans in the dataset are single spans. In other words, only one continuous word or phrase contributes to the toxicity. In the training set, single toxic spans make around 67.7% of the training set. The same phenomenon is observed for the validation and test set which contains 65.2% and 70.35% of single toxic spans respectively.

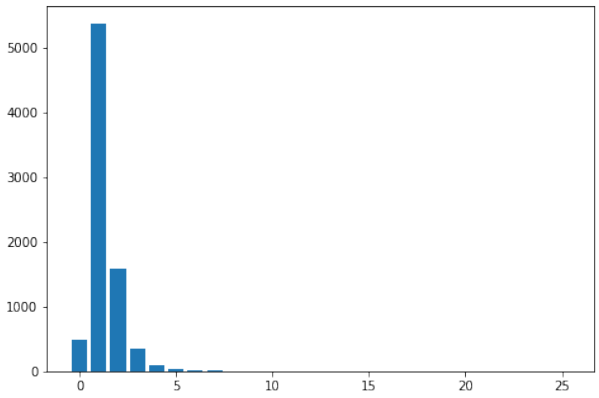

Figure 1: Number of Spans in a sentence

Texts
This is a stupid example.

Table 2: Examples of Toxic Labelling Problems

4 Methods

Our approach is mainly divided into three parts: data prepossessing, model construction and data augmentation. On the first part we will work on cleaning of toxic spans labels for adjustment, then utilizing different types of BERT-based models to compare with the random baselines and some current state-of-the-art methods on toxic token classification. Data Augmentation will also be proposed as a method to increase dataset size which is expected to improve the model performance.

4.1 Data Preprocessing

Fixing Spans Location. In the dataset, it is found that the determination of a character being toxic is not consistent like in Table 2, for example: singletons and whitespaces being highlighted as toxic. Therefore, we have removed the singletons and the whitespaces which are classified as the part of toxic spans so that it can provide accurate ground truth for model to train.

Tokenization. After the spans fixing, the words in the dataset are tokenized and data cleaning is also performed in this process. We have implemented using HuggingFace's tokenizer function to form the word embeddings. Details of tokenizers will be explained in Methods section.

4.2 Models

4.2.1 Baselines

Random. In order to illustrate the performance of our purposed model, we created a random baselines which each character offset will be classified as toxic when the probability p is higher than 0.5. With the random prediction on toxic spans, it is found that the F1 score on **dev** and **test** data are 0.18 and 0.12 respectively.

4.2.2 BERT-based

BERT stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2018), which is one of the state-of-the art models and produces significant improvement on accuracy score. Different with traditional RNN models, self-attention feature is available by transformers which can have a better understanding of a word by looking at other words in the input sequence. Considering the characteristics of toxic spans wording, it is usually a phrase combining with some non-toxic wordings. Therefore, BERT is a good fit on conducting token-level toxic classification for this task.

Apart from using the typical BERT-based model, we select another two of the popular models, which includes RoBERTa and DistilBERT. For RoBERTa (Liu et al., 2019), it is a optimized BERT model which enables dynamic masking during the training epochs. With its much larger datasets and longer time for pre-training, it outperforms with the typical BERT model. For DistilBERT (Sanh et al., 2019), it is used to achieve faster training time by reducing half of the training parameters However, it does not compromise on the accuracy performance, which is valuable on practical toxic detection in real applications. With the comparison of these models, it is expected to have a better performance on RoBERTa and efficient training on DistilBERT.

4.3 Data Augmentation

Data augmentation is commonly used in computer vision to prevent model overfitting which is beneficial to the improvement of testing data results. Easy Data Augmention (EDA) (Wei and Zou, 2019) has been used for this task which enables the following functionality on data augmentation:

- Synonym Replacement (SR): Using synonyms to replace for n words in the sentences randomly
- Random Insertion (RI): Using synonyms to insert n words in the sentences randomly
- Random Swap (RS): Swapping words in different sentences
- Random Deletion (RD): Remove some words in the sentence with probability p.

In this task, augmentation on toxic spans datasets has been implemented to increase the dataset size and enhance the variety of sentence structures. Inspired by EDA, as we found that toxic comments usually are not formally spelled correctly, therefore we used **nlpaug** library for swapping and substituting some characters of words randomly in a sentence for increasing the variety of the sentence patterns. It is expected to be useful for improving the F1 score.

5 Experiment

5.1 Experiment Environment

All the process are operated in Google Colab Pro environment, configured with a NVIDIA T4/P100 GPU, 16GB DDR6/ 12GB HBM2 memory capacity.

5.2 Setup

For the setup of a system, it is first to build the dataset with includes the tensor of the text embeddings and the labels. For the embeddings, it is powered by the Hugging Face pre-trained tokenizers which derives from its own library. For the labels, the value would be either "0" or "1", and "1" represents the token is toxic. When the label is "CLS", "SEP" and "PAD", it will be in negative value. If data augmentation is implemented, there is a probability $p > 0.8$ which the examples will be augmented by replacing and swapping characters of some words. For the model part, token classification will be used in different BERT-based models for the training process.

For the hyper-parameter settings, we have conducted fine-tuning experiment and here are as follows: (highlighted are chosen for the comparison of different models)

- Batch size: 8, **16**, 32
- learning rate: 5e-5
- Epoch: **2**, 3

- Optimizer: Adam (Default)
- weight decay: 0.01
- probability p for data augmentation: 0.2

5.3 Evaluation Metrics

For the evaluation, the Fl score has been followed the task organizers' metrics. (Martino et al., 2019)

Let system A_i return a set $S^t_{A_i}$ of character offsets, for parts of the post found to be toxic. Let G^t be the character offsets of the ground truth annotations of t. $F1$ score of system A_i with respect to the ground truth G for post t are as follows.

$$F_1^t(A_i, G) = \frac{2 \cdot P^t(A_i, G) \cdot R^t(A_i, G)}{P^t(A_i, G) + R^t(A_i, G)} \quad (1)$$

$$P^t(A_i, G) = \frac{S^t_{A_i} \cap S^t_G}{S^t_{A_i}} \quad (2)$$

$$R^t(A_i, G) = \frac{S^t_{A_i} \cup S^t_G}{S^t_{A_i}} \quad (3)$$

If S^t_G is empty for some post t (no gold spans are given for t), we set $F1^t(A_i, G) = 1$ if $S^t_{A_i}$ is also empty, and $F1^t(A_i, G) = 0$ otherwise. We finally average $F1^t(A_i, G)$ over all the posts t of an evaluation dataset T to obtain a single score for system A_i.

5.4 Results

With the fine-tuned hyper-parameters, different types of models have been trained with training data and predicted on the trial and test dataset. From Table 3, it is observed that the all the BERT-based models could achieve the F1 score over 0.675. Due to the larger training parameters in RoBERTa-base model, it achieves the highest F1 score with 0.698, which remains only 0.01 from the highest F1 score ranking of this task.

Although it is expected that the performance on F1 score will improve with data augmentation, it only generally increases the Trial F1 score but slightly decreases the Test F1 score. This might be contributed to the production of noise by the augmented data so that the result is affected. It also proves that quality and diversity of dataset is sometimes more important than the size of the dataset. All in all, BERT-based model is better than other traditional models like LSTM/GRU for detecting toxic spans in token level.

Models	Trial F1	Test F1
Random	0.18	0.12
BERT-Base-Cased	0.668	0.687
+ nlpaug	0.697	0.675
RoBERTa-Base	0.664	**0.698**
+ nlpaug	0.614	0.678
DistilBERT-Base-Cased	0.617	0.676
+ nlpaug	0.661	0.681
HITSZ-HLT		0.708

Table 3: F1 Score in Different Model Configurations, and HITSZ-HLT is the team with achieving the highest F1 score and ranking

5.5 Error Analysis

In Table 4 (see Appendix), we have listed some of the text examples and compare the predicted results with the ground truth. It is observed that the system will classify as toxic wordings when the wordings are more negative. Although "garbage" and "junk" (Example 2) can be toxic wordings, the context is more neutral than other extreme examples. We can also observe the "blockheads" are also toxic in ground truth (Example 1), which means a people regarded as very stupid. However, the system could only identify "ignorant" as toxic. But overall, those are the special examples and the system could generally pinpoint most of the toxic wordings.

6 Conclusion

In this paper, we have introduced using different BERT-based models and data augmentation to perform the toxic spans detection in token level. The RoBERTa model outperforms other models to achieve the highest F1 score. It also proves that data augmentation in NLP does not guarantee to improve the model performance. For the future work, apart from the data augmentation approach, we would like to explore more transformer models, especially those pre-trained by toxic dataset to enhance the performance.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. *arXiv preprint arXiv:1910.02517*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

A Appendix

Texts	Ground Truth	Predicted
Just kind of funny the way liberals who like to carry on about how educated and sophisticated and global they are and how the rest of us are just a bunch of **dumb** parochial hicks always end up unmasking themselves as **ignorant** **blockheads**.	['dumb', 'ignorant blockheads']	['dumb', 'ignorant']
More garbage pro terrorist anti Canadian drivel from the Globe, how's the subscriptions going? anyone still pay for this junk .	[]	['garbage', 'junk']
By the way , Nazis were and are fascists so if you are one, your a **fascist** . Scum like this and Trumpler need to be run out of our country and given to ISIS to use as chew toys	['fascist']	['fascists', 'fascist', 'Scum']

Table 4: Examples - Comparison of Ground-Truth and Prediction Labels, words in red is the ground truth and the words highlighted in yellow is the prediction