MIPT-NSU-UTMN at SemEval-2021 Task 5: Ensembling Learning with Pre-trained Language Models for Toxic Spans Detection

Mikhail Kotyushev
Moscow Institute of
Physics and Technology
Moscow, Russia
mkotyushev@gmail.com

Anna Glazkova
University of
Tyumen
Tyumen, Russia
a.v.glazkova@utmn.ru

Dmitry Morozov
Novosibirsk State
University
Novosibirsk, Russia
morozowdm@gmail.com

Abstract

This paper describes our system for SemEval-2021 Task 5 on Toxic Spans Detection. We developed ensemble models using BERT-based neural architectures and post-processing to combine tokens into spans. We evaluated several pre-trained language models using various ensemble techniques for toxic span identification and achieved sizable improvements over our baseline fine-tuned BERT models. Finally, our system obtained a F1-score of 67.55% on test data.

1 Introduction

Toxic speech has become a rising issue for social media communities. Abusive content is very diverse and therefore offensive language and toxic speech detection is not a trivial issue. Besides, social media moderation of lengthy comments and posts is often a time-consuming process. In this regard, the task of detecting toxic spans in social media texts deserves close attention.

This work is based on the participation of our team, named MIPT-NSU-UTMN, in SemEval 2021 Task 5, "Toxic Spans Detection" (Pavlopoulos et al., 2021). Organizers of the shared task provided participants with the trial, train, and test sets of English social media comments annotated at the span level indicating the presence or absence of text toxicity. We formulated the task as a token classification problem and investigated several BERT-based models using two-step knowledge transfer. We found that preliminary fine-tuning of the model on data that is close to the target domain improves the quality of the token classification. The source code of our models is available at https://github.com/morozowdmitry/semeval21.

The paper is organized as follows. A brief review of related work is given in Section 2. The definition of the task has been summarized in Section 3. The proposed methods and experimental settings have been elaborated in Section 4. Section 5 contains the results and error analysis respectively. Section 6 is a conclusion.

2 Related Work

Computational approaches to tackle text toxicity have recently gained a lot of interest due to the widespread use of social media. Since moderation is crucial to promoting healthy online discussions, research on toxicity detection has been attracting much attention. Our work is also related to hate speech and abusive language detection (Fortuna et al., 2020).

The toxic speech detection task is usually framed as a supervised learning problem. Moreover, fairly generic features, such as bag of words (Harris, 1954) or word embeddings (Mikolov et al., 2013), systematically yield reasonable classification performance (Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017). To better understand the mechanisms of toxic speech detection, some scholars (Waseem et al., 2017; Lee et al., 2018; Karan and Šnajder, 2018; Swamy et al., 2019) compared different techniques for abusive language analysis. Neural architectures and deep learning methods achieved high results in this domain. Thus, Pavlopoulos et al. (2017a,b) explored the possibilities of deep learning and deep attention mechanisms for abusive comment moderation. Park and Fung (2017) proposed an approach to performing classification on abusive language based on convolutional neural networks (CNN). Chakrabarty et al. (2019) used Bidirectional Long-Short Term Memory network. Castelle (2018) experimented with CNN and Gated Recurrent Units. Some recent studies (Mozafari et al., 2019; Risch et al., 2019; Liu et al., 2019a; Nikolov and Radivchev, 2019) utilized pre-trained language models such as Bidirec-

tional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) to detect offensive or abusive language.

In recent years, the task of detecting and analyzing abusive, toxic, or offensive language has attracted the attention of more and more researchers. The shared tasks based on carefully curated resources, such as those organized at the SemEval (Zampieri et al., 2019; Basile et al., 2019), GermEval (Wiegand et al., 2018), EVALITA (Bosco et al., 2018), and OSACT (Mubarak et al., 2020) events, have significantly contributed to the progress of the field and to the enrichment of linguistic resources. In addition to the corpora collected for these shared tasks, Rosenthal et al. (2020) released a large-scale dataset for offensive language identification. Ibrohim and Budi (2018); Leite et al. (2020); Pitenis et al. (2020); Komalova et al. (2021) presented various datasets for abusive speech detection in non-English languages. Most of these datasets classify whole texts or documents, and do not identify the spans that make a text toxic.

3 Shared Task

The task focuses on the evaluation of systems that detect the spans that make a text toxic, when detecting such spans is possible. The goal of the task is to define a sequence of words (character offsets) that attribute to the toxicity of the text, for example:

- **Input.** "This is a stupid example, so thank you for nothing a!@#!@".

- **Output.** [10,11,12,13,14,15,51,52,53,54,55, 56].

The sources of data were various posts (comments) from publicly available datasets. The provided dataset contains 10,629 posts split into training (7939), trial (690), and test (2000) subsets.

Inspired by Da San Martino et al. (2019), the organizers proposed to employ the F1-score for evaluating the responses of a system participating in the shared task. Let system A_i return a set $S_{A_i}^t$ 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. The F1-score of system A_i is calculated with respect to the ground truth G for post t as follows, where $|\cdot|$ denotes set cardinality.

$$F_1^t(A_i, G) = \frac{2 \cdot P^t(A_i,G) \cdot R^t(A_i,G)}{P^t(A_i,G)+R^t(A_i,G)},$$
$$P^t(A_i, G) = \frac{S_{A_i}^t \cap S_G^t}{S_{A_i}^t},$$

$$R^t(A_i, G) = \frac{S_{A_i}^t \cap S_G^t}{S_G^t}.$$

The final F1-score is an average $F_1^t(A_i, G)$ over all the posts t of an evaluation dataset T to obtain a single score for system A_i.

4 Methodology

The stated problem was modified from char-level to token-level binary-classification. The proposed solution utilizes a pre-trained language model with a classification head to classify tokens. Different configurations of BERT pre-trained as masked language models were considered as a backbone.

Due to the lack of available token-level labeled public datasets for toxic comment and the relatively small size and sparsity of dataset provided by the competition, the following training pipeline was proposed to enhance knowledge transfer. First, fine-tune pre-trained BERT on a larger-scale task of *toxic comment classification*, using the Jigsaw dataset[1] from which the competition data were constructed. Second, fine-tune obtained model to solve the actual *toxic tokens classification* problem. The exact training parameters are to be found below.

For the first step:

- remove texts occurred in spans dataset from classification dataset to prevent data leakage (so as spans dataset is sampled from classification dataset);

- 4 epochs, 200 tokens max length, 64 batch size, 10 gradient accumulation, mixed-precision FP16;

- default AdamW (Loshchilov and Hutter, 2017) with lr = 4e-5, Layer-wise Decreasing Layer Rate (Sun et al., 2019) with decay $\eta = 0.95$ and cosine learning rate (LR) schedule with T = 4 epochs and constant LR after epoch 3;

- selected bert-base-uncased as best performance / speed ratio;

- the best model on validation selection each 0.1 epoch by AUC.

For the second step:

- hold-out $\approx 14\%$ of data to train ensemble of models later;

- out-of-5-fold training on the residual ≈ 86% of data;

- 4 epochs, 512 tokens max length, 16 batch size, 10 gradient accumulation, mixed precision FP16;

- default AdamW with lr = 4e-5, Layer-wise Decreasing Layer Rate with decay $\eta = 0.95$ and cosine LR schedule with T = 4 epochs;

- the best model on validation selection each 0.1 epoch by F1-score.

The final solution contains $N \times K$ models, where N is the number of different backbone BERT architectures, K is the number of folds (5 in the current experiments). Obtained models are further to be ensembled using different strategies with validation on the single hold-out dataset:

- hard voting: final spans are selected as at least one model (spans union), as all the models (spans intersection) or as some intermediate methods with at least m models;

- soft voting: final probability is calculated as a weighted sum of models probabilities;

- train meta classifier.

5   Experiments and Results

Three pre-trained backbone BERT architectures were considered: bert-base-uncased, bert-large-uncased (Devlin et al., 2018), and bert-base pretrained for Hate Speech Detection (dehate-bert) (Aluru et al., 2020). First step setup and results:

- select subset of Jigsaw toxic classification data: all the targets with toxicity score ≥ 0.5 ($L = 135168$ objects) as class 1 and randomly sampled $3 * L$ objects with toxicity score < 0.5 as class 0;

- stratified 80% train, 20% validation;

- 0.968 AUC bert-base, 0.968 AUC bert-large, 0.942 AUC dehate-bert.

So as models except bert-base-uncased did not show compatible performance for token classification (and later for tests on the fold 0 did not show good F1-score for the actual task as well), later experiments were continued only for bert-base-uncased pre-trained model fine-tuned on token classification.

For step two results are following:

- train + trial, 8621 comments;

- average F1-score over 5 folds is 0.6714.

The experiments were conducted with Huggingface transformers library (Wolf et al., 2019).

Many patterns in our results are expected, but some stand out. In general, our model is good at detecting obscene language and utterances that demean honor and dignity or denote low moral character. We noticed that our model is not very good at identifying the posts that have no toxic span annotations. According to the corpus description, in some toxic posts, the core message is conveyed may be inherently toxic. Thus, a sarcastic post can indirectly claim that people of a particular origin are inferior. Hence, it was difficult to attribute the toxicity of those posts to particular spans. In such cases, the corresponding posts were labeled as not containing toxic spans. Among our results, there are many examples where the model detected spans in not annotated posts, for example:

- "uhhh Hillary Clinton is a serial killer and thief": [] (true annotation), [26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 44, 45, 46, 47, 48] (our annotation, "uhhh Hillary Clinton is a **serial killer** and **thief**");

- "This goes way beyond just being an asshole skipper, dude must have some serious mental issues": [] (true annotation), [35, 36, 37, 38, 39, 40, 41] (our annotation, "This goes way beyond just being an **asshole** skipper, dude must have some serious mental issues").

In addition, some texts in the dataset raise questions of the annotation credibility, for example:

- "How the hell is this news? Am I supposed to be shocked that the Crown Prince of Bahrain or one of the world's biggest celebrity superstars get's better access to the State Department then I do? During which administration has this ever not been true? The media's desperation to keep this election close is far past ridiculous" (training set, the toxic span annotation is underlined);

- "Yup. NVN used the Press. The Press was USED. Used like their sister on prom night! Idiots. All faux-erudite, not realizing they were being played" (training set, the original annotation is underlined);

915

Rank	Team	F1-score
1	HITSZ-HLT	0.7083
26	UAntwerp	0.67552
27	**MIPT-NSU-UTMN**	**0.67551**
28	NLRG	0.67532
	Avg result	0.57805

Table 1: Results on the test set.

- "And you are a complete **moron** <u>who ob</u>viously doesn't know the meaning of the word **narcissist**. By the way your bias is showing" (test set, the original annotation is underlined, the annotation of our model is highlighted in bold).

The final result of our model is presented in Table 1. As can be seen from the table, the systems of the participants produce close results. Our system achieved 67.55% of F1-score on the test set of this shared task that attracted 91 submitted teams in total. This value exceeded the average result by almost 10%.

6 Conclusion

This paper introduces our BERT-based model for toxic spans detection. As expected, pre-training of the BERT model using an additional domain-specific dataset improves further toxic spans detection performance. Experimenting with different fine-tuning approaches has shown that our BERT-based model benefits from the two-step knowledge transfer technique. An ensemble with spans intersection obtained our best result on the test data.

In our future work, we will evaluate various language models, such as distilled versions of BERT (Sanh et al., 2019; Jiao et al., 2020) and RoBERTa (Liu et al., 2019b).

References

Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

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

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Michael Castelle. 2018. The linguistic ideologies of deep abusive language classification. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 160–170.

Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. 2019. Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 70–79.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5640–5650.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4163–4174.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 132–137.

Liliya Komalova, Tatiana Goloshchapova, Leonid Motovskikh, Rostislav Epifanov, Dmitry Morozov, and Anna Glazkova. 2021. Mca workshop - toxic comments. Mendeley Data, V1.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106.

JA Leite, DF Silva, K Bontcheva, and C Scarton. 2020. Toxic language detection in social media for brazilian portuguese: new dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924. Association for Computational Linguistics (ACL).

Ping Liu, Wen Li, and Liang Zou. 2019a. Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 87–91.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of osact4 arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45.

John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection (to appear). In *Proceedings of the 15th International Workshop on Semantic Evaluation*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017b. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1125–1135.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5113–5119.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model. In *KONVENS*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

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*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.