Classification of toxic and inappropriate texts on sensitive topics (NLP 2021 Course)

Evgeny Avdotin Evgeny.Avdotin@skoltech.ru
Kundyz Onlabek kundyz.onlabek@edu.hse.ru
Anastasia Sozykina Anastasia.Sozykina@skoltech.ru

Abstract

Toxicity is an important notion in text sentiment analysis, but not all toxic topics can be equally dangerous in terms of toxicity. A more fine-grained notion is inappropriateness, which defines how harmful is the text for speakerś reputation. In this project, we develop a text categorization system which is performing multi-tasking. Namely, we use two systems of categories: inappropriateness and topic. We develop an approach that combines these two labels and aims at improving the existing baselines.

Code: link to code
Presentation: link to video

1 Introduction and motivation

Toxic and inappropriate texts is an essential problem of online society. Internet platforms and chatbots have demand on classification and prevention of toxicity. There are a lot of researches and competitions dedicated to this problem.

However, there is another, to some extent, more complicated problem of inappropriate texts classification. The inappropriate text is not toxic, abuse or offense text but still ambiguous. It's important to define another notion related to inappropriate texts - sensitive topics. We refer to which have a high chance of yielding a discussion which can harm the speaker's reputation as sensitive. The list of inappropriate topics includes gambling, pornography, prostitution, etc.

Inappropriateness includes toxicity and can also be defined as ability to express undesirable views, or prompt listener to dangerous or illegal actions. Hence, inappropriate message is a message on sensitive topic that can frustrate the reader/listener and also can harm the reputation of the speaker. This notion is hard to formalize and in the considered dataset respective labels were obtained by human assessment (Babakov et al., 2021).

In this project, we investigate the problem of inappropriate text classification. Our research is based on the paper of (Babakov et al., 2021).

2 Related work

Although a large amount of English textual datasets labeled for toxicity detection exist, the definition of this term is not agreed among the researchers. One of the largest toxicity corpora operating with multiple labels (*toxic, obscene, threat, insult, identity hate*, etc) was released by Jigsaw based on Wikipedia comments (Jigsaw, 2018). Some works (Waseem and Hovy, 2016) concentrate on specific toxic topics, such as sexism and racism. Another important work is (Banko et al., 2020). Authors suggest a taxonomy of harmful online behavior, mixing toxic topics with other different parameters of toxicity, such as direction or severity.

Different methods for toxic comment identification exist in community. For example, the authors of (Robinson et al., 2018) use classical TF-IDF vectorization combined with different manual feature engineering techniques, and apply classical SVM and neural CNN+GRU approaches for classification. Another important work that covers DL-based approach is (Pavlopoulos et al., 2017). Neural models are compared to word-list baseline and MLP classifier on toxic comment classification task. Authors also propose a novel, deep, classification-specific attention mechanism that improves further the overall results of the GRU-RNN, and can also highlight suspicious words for free, without including highlighted words in the training data.

3 Methodology

3.1 Dataset and task definition

In this work, we use the dataset obtained by the authors of (Babakov et al., 2021). It includes texts from two websites in Russian language: 2ch.hk and otvet.mail.ru. Both these websites are unmoderated and provide a great variety of texts on sensitive topics. Then the obtained data was labeled with the help of Yandex.Toloka crowdsourcing system. Finally, the dataset of sensitive topics and the appropriateness dataset were obtained. Detailed data statistics is provided in (Babakov et al., 2021). Data can be found at GitHub[1].

In appropriateness dataset, each sentence is labeled by appropriateness value between 0 and 1, where 0 denotes ordinary acceptable sentence and 1 totally inappropriate. By selecting a threshold value α, all sentences with appropriateness higher that $1 - \alpha$ can be marked as inappropriate, and with appropriateness lower that α as appropriate. Therefore, appropriateness dataset can be used in binary classification problem. This was the main problem considered by our project.

3.2 Baseline: logistic regression

As a first baseline in our project, we considered the logistic regression model and appropriateness dataset. At preprocessing stage, we used the same technique as in all following approaches. We used pre-trained `BertTokenizer` (fig. 1).

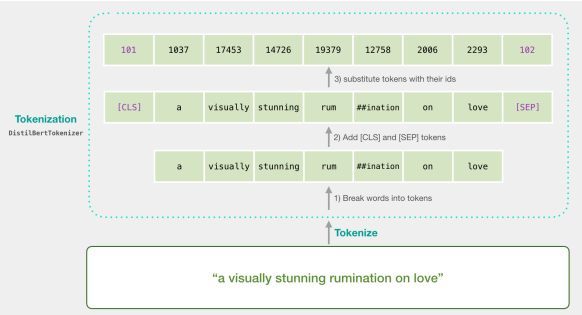

Figure 1: BERT tokenizer

It takes as input collection of raw sentences, splits them into tokens, adds special tokens `[SEP]` and `[CLS]` and encodes each token with its index. More details on BERT tokenization can be found in original paper (Devlin et al., 2018) and in Hugging Face documentation [2]. What is convenient for us here is that `BertTokenizer` pads encoded sentences to same length, so that they can be packed into tensor.

Then tensor with input indexes are fed into embedding layer and then into linear layer. Finally, softmax function is used to obtain class probabilities for each sentence.

3.3 Baseline: BERT

To compare with first baseline and with main method, we decided to include simple BERT-based approach in our work. Introduced in (Devlin et al., 2018), BERT (Bidirectional Encoder Representations from Transformers) is one of the most powerful and well-known language representation models nowadays.

In our work, we used BERT model pre-trained by authors of (Kuratov and Arkhipov, 2019). Preprocessing step here involves the same tokenization as described in 3.2. After token IDs and attention mask is fed into pre-trained BERT, fixed-dimension hidden state vector of a sentence is obtained. It can also be fed into linear layer of corresponding dimension and softmax layer to obtain class probabilities, as described above.

3.4 Main method

The main approach implemented in our project involved four stages.

3.4.1 Multilabel classifier

At this stage, we considered multilabel classification task on sensitive topics dataset. It contains 19 sensitive topics; each sentence can belong to either several of them (one or more than one) or to none of them. Hence, topic labels of sentences form 393 combinations and each of them can be viewed as separate class label. Similarly to binary classification described in 3.3, we train BERT-based multilabel classifier here. Its trained weights will be used further.

3.4.2 Weight averaging trick

From the previous stage we have linear layer matrix of shape 768×393. By applying $\max$ operation to each column and extracting distinct topics, we obtain 786-dimensional embeddings for each of 18 initial topics. We will use these embeddings later.

[1] https://github.com/skoltech-nlp/inappropriate-sensitive-topics/

[2] https://huggingface.co/transformers/main_classes/tokenizer.html

3.4.3 Additional tokens

Then we return to appropriateness dataset. Apart from ordinary tokens constructed by BERT tokenizer we use 18 special tokens denoting dangerous topics. We add to each sentence in the dataset a token corresponding to topic, which this sentence belongs to. Sentence with added topic token may look like `[PL] [RS] [SL] [SEP] hehe, donetskiy zhe...`, where `[PL]` denotes politics, `[RS]` racism etc.

Model	Input Sequence	Label
BERT4TC-S	[CLS] I like this film. [SEP]	{negative, **positive**}
BERT4TC-AQ	[CLS] I like this film. [SEP] What is the result? [SEP]	{negative, **positive**}
BERT4TC-AA	[CLS] I like this film. [SEP] positive [SEP]	{0, **1**}
	[CLS] I like this film. [SEP] negative [SEP]	{**0**, 1}
BERT4TC-AWA	[CLS] I like this film. [SEP] The result is positive. [SEP]	{0, **1**}
	[CLS] I like this film. [SEP] The result is negative. [SEP]	{**0**, 1}

Figure 2: Examples of input sequence constructions

3.4.4 Sentence transformation and final tuning

To improve the method from above, we decided to include approach described in (Yu et al., 2019) in our work. Authors of this work suggest adding auxiliary sentences to original ones to improve BERT classification accuracy. Example is shown in 2.We apply this approach to samples in our dataset.

Then we initialize aforementioned DeepPavlov pretrained BERT. To each weight vector in BERT parameters corresponding to topic token we assign topic embedding value obtained above. Finally, these model is applied for binary classification of transformed dataset sentences.

4 Results and discussion

We trained 3 models in total. Training code can be found at Google Colab [3]. Batch size was equal to 16 for all the models. Logistic Regression baseline model was trained with Adam optimizer with constant learning rate 2×10^{-2}. Its best f1 score on validation dataset was equal to 0.58.

[3] https://colab.research.google.com/drive/13wSJLHKiRYtt0jUla_Mgv4gBeM_GScNn?usp=sharing

BERT baseline model was trained with learning rate 1×10^{-3}, regularization coefficient 1×10^{-2} and 500 warm-up steps. It achieved f1-score 0.81 on validation dataset.

Main approach was trained with the same learning parameters as baseline BERT. It achieved outstanding f1-score level of 0.98 on validation. We tested several approaches shown in 2 and found out that the best quality is provided by BERT4TC-AWA transformation approach.

Model	f1 score
LogReg	0.58
BERT	0.81
Main	0.98

Table 1: Experiments results

5 Conclusion

During this project our team implemented two baseline models and the BERT-based model with various improvements for binary appropriateness classification of sentences. Our implementation showed good results that give an improvement over the baseline.

Although fair results were obtained, further improvements can be made by applying different context-awareness techniques.

References

Nikolay Babakov, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Detecting inappropriate messages on sensitive topics that could harm a company's reputation. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 26–36, Kiyv, Ukraine. Association for Computational Linguistics.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.

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*.

Jigsaw. 2018. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

David Robinson, Ziqi Zhang, and Jonathan Tepper. 2018. Hate speech detection on twitter: feature engineering vs feature selection. In *European Semantic Web Conference*, pages 46–49. Springer.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Shanshan Yu, Jindian Su, and Da Luo. 2019. Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access*, 7:176600–176612.