Cross-lingual toxicity classifier

Valerii Baianov **Svetlana Gabdullina** **Dmitrii Leshchev** **Gleb Mezentsev**

Abstract

Nowadays the Internet is full of toxic comments and moderatio is crucial to promoting healthy online discussions between people or people and chat-bots. There is a variety of methods for toxicity detection. The goal of this project is to answer a question whether English data can improve detection of toxicity in Russian language. We take pretrained XLM-Roberta as a base model, finetune it with English/Russian/English-Russian data and compare the results.

Code: Google Colab notebook with code
Video: YouTube video presentation

1 Introduction and motivation

Detecting toxic posts on social network sites is a crucial task to keep a clean and friendly space for online discussion. To identify and classify toxic online commentary, the modern tools of data science transform raw text into key features from which predictions for monitoring offensive conversations can be made.

In this project we study how can adding a new language improve the results of text classification (toxicity detection). The motivation behind it is the some kind of parallel or similar data between different languages. Transformer-based models (Vaswani et al., 2017) show the state of the art results in the modern natural language processing. As long as we want to solve a cross-lingual transfer tasks, we need a multilingual language model (Conneau et al., 2020).

We look at 3 cases:

- pretrained multilingual model + finetune on Russian data + test on Russian data;
- pretrained multilingual model + finetune on English data + test on Russian data;
- pretrained multilingual model + finetune on Russian and English data + test on Russian data.

2 Related work

Toxic Language Detection There is a well known Kaggle competition task[1] - predicting the toxicity score for a document with the Jigsaw Unintended Bias dataset. In (Morzhov, 2020) models based on Convolutional Neural Networks (CNNs) (Kim, 2014) and Recurrent Neural Networks (Cho et al., 2014) are compared with a Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2019), obtaining the best performance from an ensemble of all used models. In (Gencoglu, 2021) and (Richard and Marc-André, 2020) used the same dataset to improve on the automatic detection of cyberbullying content.

Multilingual Language Models From pretrained word embeddings (Mikolov et al., 2013b; Pennington et al., 2014) to pretrained contextualized representations (Peters et al., 2018; Schuster et al., 2019) and transformer based language models (Radford et al., 2018; Devlin et al., 2019), unsupervised representation learning has significantly improved the state of the art in natural language understanding. Parallel work on cross-lingual understanding (Mikolov et al., 2013a; Schuster et al., 2019; Lample and Conneau, 2019) extends these systems to more languages and to the cross-lingual setting in which a model is learned in one language and applied in other languages.

In (Conneau et al., 2020) it is shown that pretraining multilingual language models at scale leads to significant performance gains for a wide range of crosslingual transfer tasks. They introduce their

[1] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

best model XLM-RoBERTa. It has 3 different types of embeddings:

- language embeddings;
- position embeddings;
- token embeddings.

3 Methodology

The project is based on the XML-RoBERTa. It is a Transformer model trained on 100 languages with the multilingual MLM objective (Devlin et al., 2019; Lample and Conneau, 2019). The authors of XML-RoBERTa sample streams of text from each language and train the model to predict the masked tokens in the input. For the training they build a clean CommonCrawl Corpus following the (Wenzek et al., 2019). All languages are processed with the same shared vocabulary created through Byte Pair Encoding (BPE) (Sennrich et al., 2016).

We take the pretrained XML-RoBERTa model from Hugging Face[2] and add a linear classifying head.

Our hypothesis is checked with the following datasets:

- Jigsaw Unintended Bias dataset;
- Toxic Russian Comments Dataset[3];
- Russian Language Toxic Comments[4] dataset.

Each of the datasets has several labels for toxic sentences. For the simplicity we unite them and solve a binary classification problem. Also, we combine the last two datasets in one – ”Russian”. For tokenizing our data we use a special AutoTokenizer for our model also from Hugging Face. All data was truncated to have not more than `max_len` tokens.

As it was said before, we consider 3 cases of training data, but the last one (Russian and English) is extended by 3 more situations with different ratio of Russian and English samples in the dataset: ($1 : 1$, $2 : 1$, $3 : 1$ respectively). The description of the datasets used in the experiments are provided in the Table 1.

dataset		number of samples	share of toxic labels
train	ru	183891	0.18
	en	312735	0.07
	ru-en	496626	0.11
	ru-en (1:1)	367782	0.12
	ru-en (2:1)	275836	0.14
	ru-en (3:1)	245188	0.15
val	ru	26795	0.18
test	ru	52016	0.18

Table 1: Description of the datasets.

	accuracy	precision	recall	F1
ru	0.9722	0.9258	0.9191	0.9224
en	*0.8259*	0.9359	*0.0359*	*0.0690*
ru-en	0.9718	0.9298	0.9124	0.9210
ru-en (1:1)	0.9729	0.9206	0.9296	0.9250
ru-en (2:1)	0.9722	0.9257	0.9194	0.9226
ru-en (3:1)	0.9726	0.9327	0.9137	0.9231

Table 2: Evaluation metrics.

We do not freeze any parameters of the model and train it for 10 epochs with a mini-batch gradient descent with the following settings:

- batch size: 16
- max length: 64
- loss: Cross Entropy
- optimizer: Adam
- learning rate: 10^{-5}

Among all steps of training process we choose the best model according to the validation loss and the validation accuracy. To evaluate the models we use accuracy, F1-score, precision and recall metrics.

4 Results, discussion, and conclusion

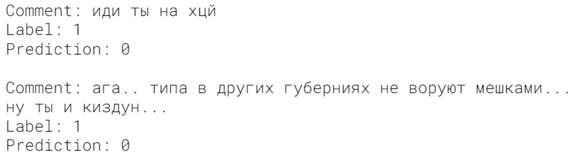
The results of our experiments are presented in the Table 2. Let us talk about them in more detail. The first row shows us that our classifier works good with the data from the same dataset as was used for training.

Actually, the table contains very similar numbers except for the second row. This is a total failure of toxicity detection and we can say for sure that training dataset should contain samples from the language the classifier is supposed to work with.

[2] https://huggingface.co/

[3] https://www.kaggle.com/alexandersemiletov/toxic-russian-comments

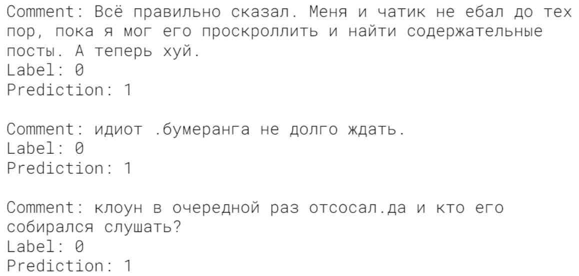
[4] https://www.kaggle.com/blackmoon/russian-language-toxic-comments



```
Comment: иди ты на хцй
Label: 1
Prediction: 0

Comment: ага.. типа в других губерниях не воруют мешками...
ну ты и киздун...
Label: 1
Prediction: 0
```

Figure 1: Misprints in comments.



```
Comment: Всё правильно сказал. Меня и чатик не ебал до тех
пор, пока я мог его проскроллить и найти содержательные
посты. А теперь хуй.
Label: 0
Prediction: 1

Comment: идиот .бумеранга не долго ждать.
Label: 0
Prediction: 1

Comment: клоун в очередной раз отсосал.да и кто его
собирался слушать?
Label: 0
Prediction: 1
```

Figure 2: Incorrect labeling.

On the one hand, from the other rows we can say that all models give close results and the difference between them is negligible to say that adding the English data to the training dataset benefits. On the other hand, we also cannot say that the English data somehow harms our model. They are equal. We suppose that the method of adding data from a different language can be more usefull for closer languages, for instance, Italian and Spanish.

We also analyzed the misclassified samples and detected 3 different groups of mistakes.

- The first group of comments have a misprint in the only offensive word. Model thinks that some of them are not toxic (e.g., 1).

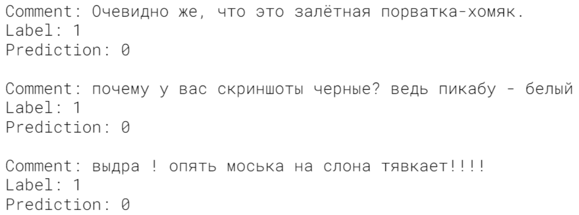
- The second are obviously toxic, but they were labeled incorrectly in the dataset (e.g., 2).

- And the last ones have ambiguous meanings. They are difficult to classify even for a human (e.g., 3).

From our experiments and their results we can make the following conclusions:

- Language of the training data does matter and it impossible to train a classifier for Russian language without the Russian data.

- Adding English samples to the training dataset does not affect on the model in the case of Russian language, but this approach should be checked with another languages.



```
Comment: Очевидно же, что это залётная порватка-хомяк.
Label: 1
Prediction: 0

Comment: почему у вас скриншоты черные? ведь пикабу - белый
Label: 1
Prediction: 0

Comment: выдра ! опять моська на слона тявкает!!!!
Label: 1
Prediction: 0
```

Figure 3: Ambiguous meaning.

References

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Oguzhan Gencoglu. 2021. Cyberbullying detection with fairness constraints. *IEEE Internet Computing*, 25(1):20–29.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Sergey Morzhov. 2020. Avoiding unintended bias in toxicity classification with neural networks. In *2020 26th Conference of Open Innovations Association (FRUCT)*, page 314–320.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Khoury Richard and Larochelle Marc-André. 2020. Generalisation of cyberbullying detection.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data.