Transformer Models for Drug Adverse Effects Detection from Tweets

Pavel Blinov, Manvel Avetisian
Sberbank Artificial Intelligence Laboratory, Moscow, Russia
{Blinov.P.D, Avetisyan.M.S}@sberbank.ru

Abstract

In this paper we present the drug adverse effects detection system developed during our participation in the Social Media Mining for Health Applications Shared Task 2020. We experimented with transfer learning approach for English and Russian, BERT and RoBERTa architectures and several strategies for regression head composition. Our final submissions in both languages overcome average F_1 by several percents margin.

1 Introduction

In the world of pandemic threats it is important to pay a special attention to the process of drug discovery. For the pharmaceutical industry it was always important not only to develop a new drug, but also to detect its possible Adverse Effects (AEs) even to smallest group populations as soon as possible. Although early stages of a clinical trial reveal most of a drug AEs, constant monitoring and feedback collection at the last pharmacovigilance stage is mandatory as it allows to identify rare or slowly evolving AEs. With current development of social media networks and artificial intelligence methods it is tempting to mine this information from the publicly available user data such as Twitter messages. If successful one can get an additional source of valuable information.

The above stated problem got into the research focus of 2020 Social Media Mining for Health Applications (SMM4H) Workshop (Klein et al., 2020). The organizers divided this problem into three consequent sub-tasks: 1) medical tweets detection; 2) classification of tweets that report AEs and 3) extraction and normalization of AEs from text. Specifically the second task was to build a binary classification model able to distinguish tweets that report an AE of a medication from those that do not. We concentrate our efforts only on this task as it is the only one that offers multilingual version (English, French and Russian).

In the recent years, methods for Natural Language Processing (NLP) developed rapidly. Nowadays transformer-based neural architectures (Vaswani et al., 2017) dominate the field. Our goal for this study was to benchmark BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) architectures and several strategies for regression head composition on this practical problem from medical domain with real-world data.

2 Data

The organizers provided labeled data alone with the Train / Validation split. Table 1 summarize statistics about the data for our languages of interest. As one can see the subset of the data for Russian language is very limited which affects the results (see Section 4).

In our experiments we don't use the provided split, but join train and validation parts for each language and perform 5-fold Cross-Validation (CV) procedure (Bishop, 2006). This allows us to estimate standard deviations for the models.

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 110–112
Barcelona, Spain (Online), December 12, 2020.

Language	Train	Validation	Test
English	20,544 (9.4)	5,134 (9.84)	4,759 (-)
Russian	6,090 (8.75)	1,522 (8.74)	1,903 (-)

Table 1: Data statistics (number of examples (% of positive)).

The overall tendency of recent deep learning models is to avoid intermediate tasks and preprocessing, so the only preprocessing steps we perform with the labeled data is to remove user names and drop duplicate tweet messages (about 3.5% for the English).

3 Our approach

As the train data were very limited we resort to a transfer learning approach. That is, we take an NLP model pre-trained on a large corpus of texts (a backbone model) and fine-tune it for a specific task at hand. In this study as such backbone models for English we used BERT-base and RoBERTa-base from Transformers library (Wolf et al., 2019). For the Russian BERT model we fine-tuned RuBERT (Kuratov and Arkhipov, 2019). There was no Russian RoBERTa model, so we trained one from scratch on the public data (Shavrina and Shapovalova, 2017) and made it available.[1] This model was trained with Adam optimizer (starting learning rate 3×10^{-6}) on 16 GPU (Tesla V100) and the batch size of 1024 (with gradient accumulation) for about 60K steps.

Originally BERT and RoBERTa architectures designed a special classification token (cls) to summarize the state of a whole input text sequence (into a single embedding) and allow to configure downstream tasks (thought classification or regression head). It is known that in NLP-related tasks max-pooling operation over the embeddings associated with sentence keyword selection and the mean-pooling is a way of encoding the general meaning of the sentence. So we experimented with this pooling strategies over the last encoder states. Additionally the option with concatenation of embeddings (cls & max & mean) were explored. The resulting hidden state followed by additional fully connected layer of the same size and ends with regression output. Based on CV metric maximization strategy the threshold were selected to make final binarization of predictions.

4 Results and Conclusions

The primary evaluation metric for the task was F_1-measure toward the positive class as a trade-off between Precision (P) and Recall (R) (Manning et al., 2008). The 5-fold CV results of our experiments are shown in Table 2 (best regression heads for architecture and language are in **bold**).

		Regression head			
Language	**Architecture**	**cls**	**max**	**mean**	**cls & max & mean**
En	BERT	63.89 ± 1.3	$\mathbf{64.35 \pm 1.25}$	63.72 ± 1.33	64.02 ± 1.26
	RoBERTa	66.57 ± 0.19	66.63 ± 0.61	66.32 ± 1.2	$\mathbf{67.45 \pm 0.79}$
Ru	BERT	47.12 ± 3.94	47.87 ± 3.94	$\mathbf{48.64 \pm 3.82}$	48.1 ± 3.57
	RoBERTa	42.31 ± 3.98	43.01 ± 4.41	42.5 ± 4.74	$\mathbf{44.80 \pm 4.18}$

Table 2: 5-fold CV results for language, architecture and specific regression head ($F_1 \pm F_{\text{std}}$, %).

First of all, Table 2 shows that the English model performs much better with respect to the Russian one, probably because of the training data size (large standard deviation is in favor of this hypothesis). Second, the right choice of a regression head do help to get better performance.

For the English final submission we selected RoBERTa model with (cls & max & mean) regression head. Based on the test data it showed $F_1 = 57\%$ ($P = 50\%, R = 66\%$). That is better (by 11%) than average performing system $F_1 = 46\%$ ($P = 42\%, R = 59\%$).

[1] https://huggingface.co/blinoff/roberta-base-russian-v0

For the Russian submission we selected BERT model with mean regression head. In this case our system metrics is perfectly matched between CV and test, $F_1 = 48\%$ ($P = 36.1\%, R = 70.5\%$). Which is also 5.3% improvement compared to average scores for this task $F_1 = 42.7\%$ ($P = 36.2\%, R = 58.3\%$).

The results allow us to conclude that the drug adverse effects detection in raw text still remains challenging problem and requires more elaborate methods to reach the acceptable quality.

References

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805. version 2.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (smm4h) shared tasks at coling 2020. *Proceedings of the Fifth Social Media Mining for Health Applications (SMM4H) Workshop Shared Task*.

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

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. *Computing Research Repository*, arXiv:1907.11692.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: "taiga" syntax tree corpus and parser. In *Proceedings of the International Conference CORPORA2017*, pages 78–84, Saint-Petersbourg, Russia.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *Computing Research Repository*, arXiv:1910.03771.