Text Augmentation Techniques in Drug Adverse Effect Detection Task

Pavel Blinov
Sber Artificial Intelligence Laboratory / Moscow, Russia
Blinov.P.D@sberbank.ru

Abstract

The paper researches the problem of drug adverse effect detection in texts of social media. We describe the development of such a classification system for Russian tweets. To increase the training dataset we apply a couple of augmentation techniques and analyze their effect in comparison with similar systems presented at 2021 years' SMM4H Workshop.

1 Introduction

Attention-based neural network models significantly move forward performance frontier for a range of Natural Language Processing (NLP) tasks. Pre-trained models with transformer architectures (Devlin et al., 2018; Liu et al., 2019) essentially changed the way itself of approaching an NLP problem. Fine-tuning such models for a specific task typically yields a solid result. But there are still challenging problems even among the simplest binary text classification tasks. For example, for current state-of-the-art NLP methods, it is not an easy task to differentiate drug Adverse Effects (AE) mentions among real indications for use. Especially if the target text comes from informal data sources (see example in Section 2). For several recent years, this problem stays in research focus and is offered as a shared task during the annual Social Media Mining for Health Applications (SMM4H) workshop (Magge et al., 2021). And the second time it was proposed for the Russian language.

The training data size can be crucial for deep learning algorithms generalization hence the performance metrics (Chen and Lin, 2014). This study explores the ways of gaining additional train data. We describe a couple of such techniques (translation and generation) and apply them to increase the training dataset more than 9 times.

2 Data

The SMM4H workshop organizers released *Train* and *Dev* data (user messages from Twitter) along with target labels. The pair of examples (translated from Russian for readability) are listed below:

> *I finally finished drinking this Tavanik. From which insomnia.* ⇒ ***1***

> *The main symptoms of a lack of thyroxine are just obesity, decreased intelligence, chilliness and insomnia.* ⇒ ***0***

Part	**Count**	**Positive ratio, (%)**
Train	8,184	9.45
Dev	3,425	8.73
Test	9,095	n/a
Augm_Transl	25,678	9.26
Augm_Gen	51,152	9.89
Total	97,534	n/a

Table 1: Dataset statistics.

Statistics about data parts are shown in Table 1. The *Augm_** rows are additional labeled data[1] (see Section 3 for details).

3 System Description

Data augmentation techniques are well presented in the computer vision field (Shorten and Khoshgoftaar, 2019). Distortion of an input image allows getting an additional data sample. Unfortunately for NLP tasks, there are no simple and effective operations to mine new data samples. Mere word order change or replacement of words often leads to loss or change of text meaning. That is because natural language obeys numerous rules and restrictions. To account for most of these rules and 'correctly' transform a text one needs to rely on a language model.

[1] Available for download at https://disk.yandex.ru/d/BQ-YM8MIsni7VQ

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 95–97
June 10, 2021. ©2021 Association for Computational Linguistics

		CV	Test		
System	**Train samples**	**$F_1 \pm F_{std}$**	**Precision**	**Recall**	**F_1**
Median			54.9	55.7	51
Real+Augm_*	86,111	57.2±2.5	39.3	59	47
Real+Augm_Transl	34,959	56.6±2.5			
Real	9,282	55.4±2.2			

Table 2: Systems performance metrics, (%).

3.1 Translate Augmentations

Having a long history of research current neural machine translation methods achieve great success in conveying the meaning and keeping text fluency. This allows the implementation of the idea of back translation for text data augmentation (Edunov et al., 2018). Target text translated from a source to destination language then back to the source language, e.g. ru ⇒ en ⇒ ru. Thus the final translation will contain a slightly different sample.

We apply a shortened version of such pipeline (en ⇒ ru) as we had an English dataset from the previous iteration of SMM4H workshop. In such a way we obtain an additional train part (*Augm_Transl*) of 25,678 samples.

3.2 Generation Augmentations

Besides specialized language models for translation, there is the class of Generative Pre-Training models (e.g. GPT-2) (Radford et al., 2019). Such models, trained for a phrase continuation task, could produce surprisingly plausible and coherent text fragments.

Similar to (Blinov, 2020) we adopt and fine-tune the GPT-2 model for the task of Russian tweet generation. Given a couple of random start tokens, the trained model can complete a tweet message. From this model, we retrieved 100k synthetic unlabeled messages and applied our model (Blinov and Avetisian, 2020) for labeling. Finally, only 51,152 samples with high confidence labels were selected, which comprise the *Augm_Gen* part.

3.3 Modeling

To build the final classifier we used the RuBERT (Kuratov and Arkhipov, 2019) model as a base. It was fine-tuned on the mix of augmented and real labeled data with the mean pooling strategy over contextualized set of token embeddings and binary cross-entropy loss function.

More precisely we prepared 5 of such models according to Cross-Validation (CV) split on the concatenation of *Train* and *Dev* parts. Each fold's train data was joined with *Augm_** parts and a model trained for 5 epochs with a batch size of 128 samples and 3×10^{-5} learning rate.

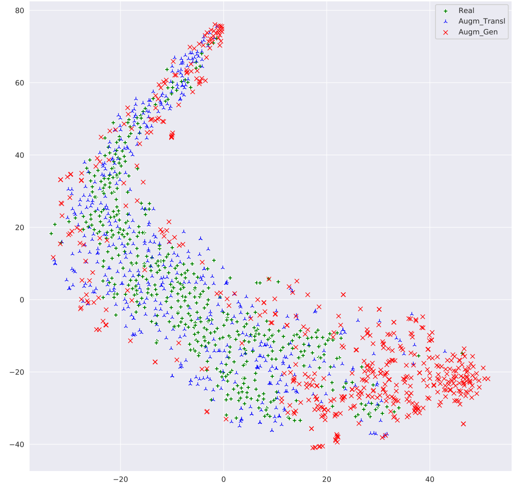

Figure 1: Samples of tweet embeddings from 3 data parts.

As we required to output binary prediction value each epoch training followed by the threshold optimization procedure. In the end, we selected the best model checkpoint and threshold for each of 5 folds. At the test time, input data processed by 5 models, and their output are binarized. The final label for a sample selected as the most common one.

4 Results and Conclusions

F_1-score toward the positive class (Manning et al., 2008) is the main evaluation metric for this task. Table 2 reports cross-validation and test metrics for a number of our systems. As we keep validation sets intact and increase with additional data only train parts we can compare the metric across systems. The *Real** prefix in a system name corresponds to this year's data (*Train* and *Dev* parts from Table 1).

Although we can see clear CV metric improvement it turned out that it did not convert into test performance. Our best system is inferior to even the median metric across participants' systems, overcoming it only in terms of Recall (by the 3% margin).

We hypothesize that this is because of a significant shift in data distribution. Partially it is confirmed by t-SNE (Maaten and Hinton, 2008) plot of randomly sample tweet embeddings from three data parts (see Figure 1), where synthetically generated messages concentrate on the border of the point cloud.

Thus our experiments reveal that procedures of text data augmentation potentially are an interesting tool for obtaining more data. But the successful practical application of these techniques for the AE detection task requires further research.

References

Pavel Blinov. 2020. Semantic triples verbalization with generative pre-training model. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 154–158, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Pavel Blinov and Manvel Avetisian. 2020. Transformer models for drug adverse effects detection from tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 110–112, Barcelona, Spain (Online). Association for Computational Linguistics.

Xue-Wen Chen and Xiaotong Lin. 2014. Big data deep learning: challenges and perspectives. *IEEE access*, 2:514–525.

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*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

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.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.