uestcc@SMM4H'22: RoBERTa based Adverse Drug Events Classification on Tweets

Chunchen Wei[1], Ran Bi[1], and Yanru Zhang[1, 2]

[1]University of Electronic Science and Technology of China, China
[2]Shenzhen Institute for Advanced Study, UESTC, China

Abstract

This is a description of our participation in the *ADE Mining in English Tweets* shared task, organized by the Social Media Mining for Health *SMM4H* 2022 workshop. We participate in the subtask a of shared Task 1, and the paper introduces the system we developed for solving the task. The task requires classifying the given tweets by whether they mention the Adverse Drug Effects. We utilize RoBERTa model and apply several methods during training and finetuning period. We also try to improve the performance of our system by preprocessing the dataset but improve the precision only. The results of our system on test set are 0.601 in F1-score, 0.705 in precision, and 0.524 in recall.

1 Introduction

Adverse Drug Events (ADEs) refer to negative side effects related to the drug. In the area of Social Media Pharmacovigilance, mining ADEs from social media is one of the most studied topics. In the Social Media Mining for Health 2022 (SMM4H) shared Task 1a (Weissenbacher et al., 2022), we focus on classifying tweets reporting ADEs and labeling them with ADE or noADE. In training process, we input 18000 labeled tweets provided by organizers to RoBERTa model and finetune the model by several practical methods such as Stochastic Weight Averaging (SWA), then output the prediction results. The system is shown in Figure 1.

We also attempt to preprocess the tweets by deleting some stop words and emojis, the performance of all approaches is shown in the paper.

2 Data

The dataset (Magge et al., 2021) provided by the organizer includes a training set of 17,385 tweets, a validation set of 915 tweets, and a test set of 10,984 tweets. It is worth noting that the sample proportion is highly unbalanced, tweets mentioning ADEs are only 7% of the training data.

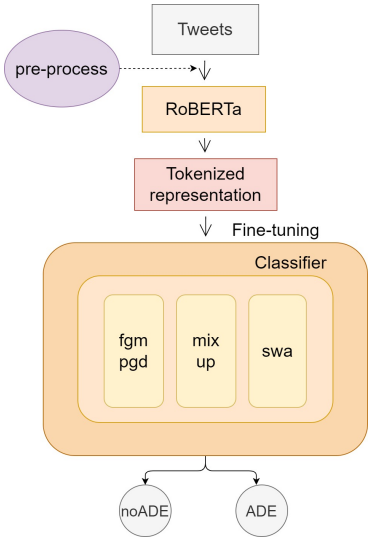

Figure 1: The architecture of system.

For data preprocessing, we save the original training data at first containing the redundant symbols and emojis in the tweets. Then we perform following preprocessing on the original dataset to construct another dataset as a comparison:

a) lowercase all words and remove whitespace.

b) remove instances of '@USER' followed by '_'.

c) remove extranous characters and emojis.

d) remove stopwords[1].

3 Method

In this section, we introduce the methodology we propose to accomplish the task of the competition.

We directly use the original dataset as input of the system at first. Then we pass the input into the BERT-based model RoBERTa (Lee and Toutanova, 2018) and get the token representations. By inputting the representations through the fully connected layers and finetuning the model, we obtain the final model and use it to predict the test set.

[1]https://www.nltk.org/nltk_data/

Proceedings of the 29th International Conference on Computational Linguistic, pages 35–37
October 12–17, 2022.

During the period of finetuning, we add several methods to improve the performance of our model:

- Fast Gradient Method (FGM) was first proposed by Miyato et al. (2016), we use this method to add a disturbance to the original input training samples and conduct adversarial training, so as to improve the robustness and generalization ability of the model.
- Project Gradient Descent (PGD) proposed by Madry et al. (2017) is an iterative attack and each iteration will project the disturbance into the specified range, we apply this method to further improve the generalization ability of the model.
- Mix-up is a simple and effective data augmentation method (Zhang et al., 2017), we apply this method on the provided dataset to prevent the model from overfitting.
- Stochastic Weight Averaging (SWA) is a method to improve the generalization ability of deep learning-based model through gradient descent and does not require additional computation (Izmailov et al., 2018), we use it in finetuning phase.

4 Experiments

In the shared Task 1a, the RoBERTa model is trained for 10 epochs with a learning rate 5×10^{-5} using Adam optimizer (Kingma and Ba, 2014). We set the batch size in training duration to 16 and 64 in validation and test duration, respectively. In addition, embedding size, output size, and sequence length in the experiments are set up as 768, 768, and 256.

We also conduct the ablation experiments to test the performance of several methods mentioned in last section. We first set the parameter alpha to 0.2 during the process of mix-up. Then we add FGM and PGD methods into the training process and use SWA with a learning rate 5×10^{-5}. As the results of ablation experiment shown in Table 1, the performance of the model on validation set improves gradually as methods add.

In order to verify the effectiveness of data preprocessing, we use the datasets before and after data preprocessing to train the model respectively. The results on validation set of two datasets are in Table 2, it shows that the application of data preprocessing increases precision but reduces recall.

Method	P	R	F1
RoBERTa	0.844	0.721	0.756
RoBERTa^{+}	**0.891**	0.640	0.745
RoBERTa$^{\#}$	0.851	0.718	0.779
RoBERTa*	0.859	**0.765**	**0.809**

$^{+}$RoBERTa with mixup.
$^{\#}$RoBERTa with mixup and swa.
*RoBERTa with mixup, swa, fgm and pgd.

Table 1: Ablation experiments on the validation set.

Dataset	P	R	F1
Dataset$_{raw}$	0.859	**0.765**	**0.809**
Dataset$_{cleaned}$	**0.873**	0.750	0.761

Table 2: Task1a results on validation set.

As for the prediction results of test set, points of three evaluation metrics significantly drop compared with results on validation set. In our two submitted results of prediction, one uses data preprocessing and another not, both are predicted by the RoBERTa model added with three methods. In addition, full training set (validation set + training set) is used to train the model. Results of our submissions are demonstrated in Table 3, mean results of all submissions by all participants are in the last row of table.

5 Conclusion

In this work, to complete the task of binary classification on tweets, we propose a system based on RoBERTa model applying with several methods to improve the generalization ability of our model, and get the results of 0.731 in precision, 0.524 in recall, and 0.601 in F1 score. We also validate the effectiveness of finetuning methods and data preprocessing in the experiments. The performance of our proposed system on test dataset compared with the mean results indicates the robustness and generalization ability of our system, and we will continue to improve it in the future.

	P	R	F1
Dataset$_{raw}$	0.705	**0.524**	**0.601**
Dataset$_{cleaned}$	**0.731**	0.339	0.463
mean	0.646	0.497	0.562

Table 3: Task1a results on test set.

References

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

J Devlin M Chang K Lee and K Toutanova. 2018. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Deepademiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Rual Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications (# smm4h) shared tasks at coling 2022. In *In Proceedings of the Seventh Social Media Mining for Health (# SMM4H) Workshop and Shared Task*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.