BioInfo@UAVR@SMM4H'22: Classification and Extraction of Adverse Event mentions in Tweets using Transformer Models

Edgar Morais, José Luís Oliveira, Alina Trifan and **Olga Fajarda**
Department of Electronics, Telecommunications and Informatics (DETI/IEETA),
University of Aveiro, Aveiro, Portugal
{edgarmorais, jlo, alina.trifan, olga.oliveira}@ua.pt

Abstract

This paper describes BioInfo@UAVR team's approach for adressing subtasks 1a and 1b of the Social Media Mining for Health Applications 2022 shared task. These sub-tasks deal with the classification of tweets that contain an Adverse Drug Event mentions and the detection of spans that correspond to those mentions. Our approach relies on transformer-based models, data augmentation, and an external dataset.

1 Introduction

The Social Media Mining for Health Applications shared task (Weissenbacher et al., 2022) addresses challenges relating to the use of social media data for health research. Our team participated in subtasks 1a and 1b. Subtask 1a focuses on the classification of tweets regarding the presence of Adverse Drug Events (ADEs) and subtask 1b focuses on the detection of ADE spans in tweets. In this submission, we performed some experiments that explore the use of transformer-based models to solve the tackled tasks.

2 Datasets

In the experiments executed for these subtasks, we used the datasets provided by the organizers (Magge et al., 2021), which were annotated for each of the subtasks. The training set had 17385 labeled tweets (1239 labeled ADE), the validation set had 915 labeled tweets (65 labeled ADE) and the test set had 10984 unlabeled tweets. Furthermore, an additional dataset was used in the training phase for subtask 1b, along with different balancing techniques.

2.1 Text augmentation

To overcome the class imbalance problem observed in the datasets we used text augmentation to increase the number of positive examples. For each tweet in the minority class, 5 augmented versions of that tweet were created. For text augmentation, we used the TextAttack framework (Morris et al., 2020) to generate new examples, by executing transformations on examples present on the available datasets. The augmenting was done through 4 transformations: replacing characters with random characters, swapping characters with QWERTY adjacent keys, replacing words with synonyms provided by WordNet (Fellbaum, 2010; Miller, 1995), and performing contractions on recognized combinations.

2.2 WEBRADR Benchmark Reference Dataset

The WEBRADR Benchmark Reference Dataset (Dietrich et al., 2020) is a labeled dataset with tweets relating to the presence of ADEs as well as the spans of the identified ADEs in positive examples. Through this dataset, we could extract 31122 tweets (588 labeled ADE). Even though this dataset was not intended for the training of models we explored its use for that purpose.

2.3 Pre-processing

The pre-processing was slightly different for each of the subtasks. For subtask 1a we replaced the special instance "&", with the character "&" and tokenized the tweets with the tokenizer corresponding to the model used. For subtask 1b, besides the steps mentioned for the previous subtask, we also lowercased the tweets.

3 Experiments

3.1 Subtask 1a

In this subtask, we evaluated different transformer-based models, by training them on the training dataset and testing them on the validation dataset. For this task, we evaluated 3 different transformer-based models available on Hugging Face[1]. These

[1] https://huggingface.co/models

Proceedings of the 29th International Conference on Computational Linguistic, pages 65–67
October 12–17, 2022.

Model	Precision	Recall	F1
BERT-large	**0.797**	0.723	0.758
RoBERTa-large	0.778	**0.862**	0.818
BERTweet-large	**0.797**	0.846	**0.821**

Table 1: Comparative results for different transformer modes for subtask 1a.

Training set	Precision	Recall	F1
Base train set	0.797	0.846	0.821
Over-sampling	0.809	0.846	0.827
Under-sampling	0.778	**0.862**	0.818
Augmented set	**0.909**	0.769	**0.833**

Table 2: Results when training a BERTweet-large model classifier with different data for subtask 1a.

Training set	Strict Precision	Strict Recall	Strict F1
Base training set	0.598	0.598	0.598
Base set with pos tweets from WEBRADR	**0.609**	**0.609**	**0.609**

Table 3: Results when training a BERTweet-large model NER pipeline with different data.

	Precision	Recall	F1
Submission	0.839	0.598	0.698
Average	0.646	0.497	0.562

Table 4: Results of submission for subtask 1a.

	Precision	Recall	F1
Overlapping results			
Submission	0.828	0.341	0.484
Average	0.539	0.517	0.527
Strict results			
Submission	0.560	0.235	0.331
Average	0.344	0.339	0.341

Table 5: Results of submission for subtask 1b.

models were Bert-large-uncased (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), and BERTweet-large (Quoc Nguyen et al., 2020). We performed further testing, by training the best-performing model with re-sampled training data or augmented training data. The models were trained with the same hyperparameters and implementation, except for the Bert-large-uncased, which used a batch size of 16 instead of 32 due to memory constraints. All models were trained for 3 epochs with a learning rate of 2e-5.

3.2 Subtask 1b

In this subtask, we used the best performing model from subtask 1a on a Named Entity Recognition (NER) task, and trained it on the training data. In this subtask the experiments focused on evaluating the effect of adding positive examples from the WEBRADR Benchmark Reference Dataset to the training data.

4 Results and discussion

Table 1 presents results obtained from training different models for subtask 1a and testing them with the task validation set. We can observe that the model BERTweet-large presents the best performance. Table 2 presents results obtained from training a BERTweet-large model with different training data obtained through transformations from the original challenge training data. We can observe that the use of text augmentation on the training set yields the best performance. Table 3 presents the results of training a BERTweet-large model for subtask 1b with different data. We can observe that the inclusion of positive examples from the WEBRADR reference dataset in the training set slightly improves the performance of the model.

The submission for subtask 1a was obtained through the training of a BERTweet-large classification model with the augmented training and validation data. For subtask 1b we used a BERTweet-large NER model trained on the training data with positive examples from the WEBRADR dataset. The submission for subtask 1b was obtained by using the NER system on the predictions submitted for subtask 1a. The results from the final submission are in Tables 4 and 5. In subtask 1a, we obtained results above the average in every metric, however in subtask 1b the only metrics above the average values were the precision metrics.

5 Conclusion

We proposed a text classification and NER pipeline which address the challenges posed by subtask 1a and 1b. Our system was able to achieve a F1 score of 0.698 on subtask 1a and a strict F1 score of 0.331 on subtask 1b.

Acknowledgment

This work was supported by FCT – Fundação para a Ciência e Tecnologia within project DSAIPA/AI/0088/2020.

References

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186. Association for Computational Linguistics (ACL).

Juergen Dietrich, Lucie M. Gattepaille, Britta Anne Grum, Letitia Jiri, Magnus Lerch, Daniele Sartori, and Antoni Wisniewski. 2020. Adverse Events in Twitter-Development of a Benchmark Reference Dataset: Results from IMI WEB-RADR. *Drug Safety*, 43(5):467–478.

Christiane Fellbaum. 2010. WordNet: An Electronic Lexical Database. In *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer Netherlands.

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. *CoRR*, abs/1907.1.

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 : JAMIA*, 28(10):2184–2192.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. pages 119–126.

Dat Quoc Nguyen, Thanh Vu, Anh Tuan Nguyen, and VinAI Research. 2020. BERTweet: A pre-trained language model for English Tweets. pages 9–14.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul 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 *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.