Medication Mention Detection in Tweets Using ELECTRA Transformers and Decision Trees

Lung-Hao Lee*, Po-Han Chen, Hao-Chuan Kao
Ting-Chun Hung, Po-Lei Lee and Kuo-Kai Shyu
Department of Electrical Engineering, National Central University, Taiwan
Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan
*lhee@ee.ncu.edu.tw

Abstract

This study describes our proposed model design for the SMM4H 2020 Task 1. We fine-tune ELECTRA transformers using our trained SVM filter for data augmentation, along with decision trees to detect medication mentions in tweets. Our best F1-score of 0.7578 exceeded the mean score 0.6646 of all 15 submitting teams.

1 Introduction

The Social Media Mining for Health Applications (SMM4H) shared task involves natural language processing challenges for using social media data for health research. We participated in the SMM4H 2020 Task 1, focusing on automatic classification of tweets that mention medications (Klein et al., 2020). This binary classification task involves distinguishing tweets that mention a medication or dietary supplement (annotated as '1') from those that do not (annotated as '0'). This task was first organized in 2018 using a data set containing an artificially balanced distribution of the two classes (Weissenbacher et al., 2018). Several approaches have been presented to address this binary classification task (Coltekin and Rama, 2018; Xherija, 2018; Wu et al., 2018). However, this year's task is more challenging. The data set represents a natural, highly imbalanced distribution of the two classes from tweets posted by 112 women during pregnancy, with only approximately 0.2% of the tweets mentioning a medication (Sarker et al., 2017; Weissenbacher et al., 2019).

This paper describes the NCUEE (National Central University, Dept. of Electrical Engineering) system for the SMM4H 2020 Task 1. To deal with highly imbalanced distribution, the support vector machine trained using the training data is used as a filter to crawl and select more tweets for data augmentation. We then fine-tune the pre-trained ELECTRA transformers (Clark et al., 2020), using our augmented data for medication mention detection. In addition, we train the decision tree as a supplementary classifier. Finally, the integrated set of testing instances are detected as a positive class from ELECTRA and decision trees are regraded as label '1' and the remaining cases as label '0' to form our submissions.

2 The NCUEE System

In addition to the training data provided by task organizers, we crawl and select highly related tweets for data augmentation. We manually check small positive tweets to pick up textual terms that may refer to medications, and then use these terms as seeds for query expansions. The pre-trained Word2Vec embedding from Twitter data is used to look up word vectors and compare their cosine similarities. The top 10 similar terms of seeds are collected, where expanded terms are kept if the document frequency (DF) of an expanded term in the positive class exceeds that in the negative class. Each expanded term, along with the query term 'pregnant' is regarded as an individual query to search for possibly related tweets from Twitter. To automatically label highly positive cases, we train the support vector machine (SVM)

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 131–133
Barcelona, Spain (Online), December 12, 2020.

using the provided training data and select crawled tweets predicted to be positive cases by the SVM. Finally, we construct an augmented data set including the original training sets for neural computing.

ELECTRA (Efficiently Learning as Encoder that Classifiers Token Replacements Accurately) is a new pre-training approach that aims to match or exceed the downstream performance of an MLM (Masked Language Modeling) pre-trained model while using less computational loading (Clark et al., 2020). ELECTRA trains two transformer models: the generator, which replaces the tokens in a sequence for training a masked language model; and the discriminator, which tries to identify which tokens in the sequence were replaced by the generator. We use pre-trained ELECTRA transformers and fine-tune them using our augmented data to detect medication mentions in tweets.

According to our empirical results from the validation set, the decision tree (DT) classifiers usually achieved a high degree of precision if the discriminated features had been extracted and learned, but very low recall if the testing cases were significantly different from the trained ones. Hence, we use the same trained SVM as a filter to select the positive cases (predicted as '1') of from the 2018 task training data that may be closely similar to the positive tweets in this task and include these in an augmented set. We then adopt the TF-IDF (Term Frequency-Inverse Document Frequency) weighting method to extract discriminated features of positive tweets from this augmented set and use them with the original training data to train the decision trees.

Finally, based on our error analysis, we found the decision tree classifier was partially complementary to ELECTRA. So, the integrated set of testing instances predicted as the positive class from the ELECTRA transformers and decision trees are labeled '1', otherwise '0 ' in our submissions.

3 Evaluation

We picked up 112 seeds from 181 positive tweets to further expand the dataset by 72 unique terms via cosine similarity through the pre-trained Word2Vec embeddings of the tweets. Without using the SVM as a filter, we have 57,678 positive tweets and 105,273 negative tweets. With SVM, we have 32,619 positive tweets and 65,238 negative tweets. The distribution of tweets after data augmentation (DA) was still remained imbalanced, with a positive to negative ratio close to 1:2. The pre-trained ELECTRA-Large was downloaded from HuggingFace (Wolf et al., 2019). The hyper-parameters used for fine-tuning ELECTRA are as follows: batch size 16; gradient accumulation steps 16; learning rate 1e-5; and number of training epochs 6.

Table 1 shows the results on the validation and test sets. The evaluation metric is the F1-score for the positive class (i.e. tweets that mention medications). For the test set, compared with submission 1 that does not use SVM as a filter for data augmentation and submission 3 that adds the prediction result of the RoBERTa transformer (Liu et al., 2019), our submission 2 achieved the best F1-score of 0.7578. Their relative ranks were identical to those of the validation set.

#	Methods	Validation Set			Test Set		
		P	R	F1	P	R	F1
1	without SVM, ELECTRA + DT	0.9412	0.9143	0.9275	0.7910	0.6883	0.7361
2	with SVM, ELECTRA + DT	0.9444	0.9714	**0.9577**	0.7262	0.7922	**0.7578**
3	with SVM, ELECTRA + DT+ RoBERTa	0.8750	1.0000	0.9333	0.6702	0.8182	0.7368

Table 1. Submission evaluation results.

4 Conclusions

This study describes the NCUEE system participating in the SMM4H 2020 Task 1 for medication mention detection, including system design, implementation and evaluation. Our best F1-score of 0.7578 exceeded the mean score 0.6646 for all 15 teams with at least one submissions.

Acknowledgements

This study is partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3 and MOST 108-2634-F-008-003- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

Reference

Abeed Sarker, Pramod Chandrashekar, Arjun Magge, Haitao Cai, Ari Klein and Graciela Gonzalez. 2017. Discovering cohorts of pregnant women from social media for safety surveillance analysis. *Journal of Medical Internet Research*, 19(10):e361.

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. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task.*

Cagri Coltekin and Taraka Rama. 2018. Drug-use identification from tweets with word and character n-grams. In *Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop and Shared Task*, pages 52-53, Brussels, Belgium.

Chuhan Wu, Fangzhao Wu, Junxin Liu, Sixing Wu, Yongfeng Huang and Xing Xie. 2018. Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tween representation and multi-head self-attention. In *Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop and Shared Task*, pages 34-37, Brussels, Belgium.

Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connorm Arjun Magge and Graciela Gonzalez-Hernandez. 2019. Deep neural networks ensemble for detecting medication mentions in tweets. *Journal of the American Medical Informatics Association*, 26(12):1618-1626.

Davy Weissenbacher, Abeed Sarker, Michael Paul and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In *Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop and Shared Task*, pages 13-16, Brussels, Belgium.

Kevin Clark, Minh-Thang Luong, Quoc V. Le and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR) ,* pages 1-18.

Orest Xherija. 2018. Classification of medication-related tweets using stacked bidirectional LSTMs with context-aware attention. In *Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop and Shared Task*, pages 38-42, Brussels, Belgium.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's transformers: state-of-the-art natural language processing, *arXiv:1910.03771*.

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*. version 2.