Automatic classification of tweets mentioning a medication using pre-trained sentence encoders

Laiba Mehnaz
MIDAS Lab, IIIT-Delhi
laibamehnazbt2k16@dtu.ac.in

Abstract

This paper describes our submission to the 5th edition of the Social Media Mining for Health Applications (SMM4H) shared task 1. Task 1 aims at the automatic classification of tweets that mention a medication or a dietary supplement. This task is specifically challenging due to its highly imbalanced dataset, with only 0.2% of the tweets mentioning a drug. For our submission, we particularly focused on several pretrained encoders for text classification. We achieve an F1 score of 0.75 for the positive class on the test set.

1 Introduction

Automatic drug name recognition has mostly been studied in terms of extracting drug names from medical documents and biomedical articles (Liu et al., 2015). However, expanding the same task to extracting drug names from tweets poses a lot more challenges. Tweets are shorter and do not provide enough context compared to academic biomedical articles; they also contain ambiguity, noise, and misspellings, especially in the form of colloquially used terms for the same drugs (Weissenbacher et al., 2019). The shared task of the 5th Social Media Mining for Health Applications specifically aims at tasks that use natural language processing for health applications. We participated in task 1, which is defined as the automatic classification of tweets that mention medications. We use several pre-trained encoders such as BERT (Devlin et al., 2019), BioBERT (Lee et al., 2019), Clinical BioBERT (Alsentzer et al., 2019), SciBERT (Beltagy et al., 2019), RoBERTa (Liu et al., 2019), BioMed-RoBERTa (Gururangan et al., 2020), ELECTRA (Clark et al., 2020) and ERNIE 2.0 (Sun et al., 2019) for the classification of tweets.

2 Dataset

The training and the validation dataset were provided to us by the organizers of SMM4H2020. The training dataset consisted of 55419 tweets, with only 146 positive tweets and 55273 negative tweets. The validation dataset consisted of 13853 tweets, with only 35 positive tweets and 13818 negative tweets. The dataset is highly imbalanced, and the positive tweets account for only 0.2% of the whole dataset. The test set for submitting our system predictions consisted of 29687 tweets.

3 Experiments and System Descriptions

Along with BERT, we use several other pre-trained sentence encoders that are a result of various improvisations over BERT such as BioBERT (Lee et al., 2019), Clinical BioBERT (Alsentzer et al., 2019), SciBERT (Beltagy et al., 2019), RoBERTa (Liu et al., 2019), and BioMed-RoBERTa (Gururangan et al., 2020). For all of the above pre-trained sentence encoders, we use the PyTorch implementation through the transformers library[1] fine-tuning them for 3 epochs with a learning rate of 2e-5, maximum sequence length as 128, and a batch size of 8. Unlike BERT, ELECTRA (Clark et al., 2020) uses an alternative pre-training task, called replaced token detection. It aims to be more sample-efficient than masked language

[1] https://github.com/huggingface/transformers

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 150–152
Barcelona, Spain (Online), December 12, 2020.

modeling used in BERT. Using the implementation of ELECTRA[2] provided by the authors we fine-tuned ELECTRA for 3 epochs with a learning rate of 1e-4, maximum sequence length as 128, and a batch size of 32. ERNIE 2.0 (Sun et al., 2019) provides a continual pre-training framework to incrementally build several pre-training tasks that focus on extracting lexical, syntactic, and semantic information from the training corpora. We use the implementation in PaddlePaddle[3] provided by the authors and fine-tune ERNIE 2.0 for 3 epochs with a learning rate of 3e-5, maximum sequence length as 128, and a batch size of 64.

Model name	F1 score	Precision	Recall
BERT base	0.78	0.83	0.74
BioBERT base	0.80	0.84	0.77
Clinical BioBERT base	0.81	0.82	0.80
SciBERT base	0.83	0.90	0.77
RoBERTa base	0.81	0.82	0.80
BioMed-RoBERTa base	0.85	0.90	0.80
ELECTRA base	0.79	0.81	0.77
ERNIE 2.0	0.83	0.92	0.74

Table 1: F1 score, Precision, and Recall for the positive class on the validation dataset.

Model name	F1 score	Precision	Recall
BioMed-RoBERTa base	0.755	0.770	0.740

Table 2: F1 score, Precision, and Recall for the positive class on the test dataset.

4 Results and Discussion

Due to the imbalance in the dataset, the metric used for evaluating the systems is the F1 score for the positive class, where positive class refers to the set of tweets that mention a drug or a dietary supplement. Table 1 contains the scores of the pre-trained sentence encoders on the validation dataset. As can be seen in Table 1, BERT has the worst performance of all the models. And, BioMed-RoBERTa has the best performance. ELECTRA base performs slightly better than BERT. Compared to BERT, there is a consistent increase in performance of all the models that used domain-specific data for pre-training. Within the group of models using biomedical related data for pre-training, SciBERT performs better than both BioBERT and Clinical BioBERT. SciBERT is trained on a multi-domain corpus, where papers from the computer science domain account for 18% of all the papers, and papers from the biomedical domain account for 82%. Unlike BioBERT and Clinical BioBERT, SciBERT is trained from scratch and has its own vocabulary called the scivocab. These factors could be the reason for SciBERT's better performance compared to both BioBERT and Clinical BioBERT. BioMed-RoBERTa's performance is visibly better than SciBERT. This performance increase could be due to RoBERTa's superior performance over BERT, as well as the additional pre-training data consisting of 2.68M full-text papers from S2ORC (Lo et al., 2020). It is interesting to note that RoBERTa's performance is comparable to BioBERT and Clinical-BERT without any domain-specific pre-training. It is also worth noting that ERNIE 2.0 has the same F1 score as SciBERT without any domain-specific pre-training. It also performs better than RoBERTa. ERNIE 2.0 seems to have learned better representations without any domain-specific training, which could be due to its variety of pre-training tasks aiming to capture lexical, syntactic, and semantic information from the dataset. ERNIE 2.0's performance also leads to an interesting question of the possibility of universal pre-trained models. Table 2 shows the results of our system prediction on the test set. Due

[2] https://github.com/google-research/electra
[3] https://github.com/PaddlePaddle/ERNIE

to time constraints, we could submit only one system prediction for the test dataset. Looking at the performance on the validation dataset, we chose to submit the predictions of BioMed-RoBERTa, as it gave the best performance on the validation dataset. Our system predictions on the test set are competitive and achieve above-average scores among the participants' systems.

References

Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connor, Arjun Magge and Graciela Gonzalez-Hernandez. Deep neural networks ensemble for detecting medication mentions in tweets. *Journal of the American Medical Informatics Association*,Volume 26, Issue 12, December 2019, Pages 1618–1626, https://doi.org/10.1093/jamia/ocz156

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng,Di Jin, Tristan Naumann and Matthew B. A. McDermott. (2019). Publicly Available Clinical BERT Embeddings. *ArXiv, abs/1904.03323.*

Iz Beltagy, Arman Cohan and Kyle Lo. (2019). SciBERT: Pretrained Contextualized Embeddings for Scientific Text. *ArXiv, abs/1903.10676.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv, abs/1810.04805.*

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So and Jaewoo Kang. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.*

Kevin Clark, Minh-Thang Luong, Quoc V. Le and Christopher D. Manning. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ArXiv, abs/2003.10555.*

Kyle Lo, Lucy Lu Wang, Mark E Neumann, Rodney Michael Kinney and Daniel S. Weld. (2020). S2ORC: The Semantic Scholar Open Research Corpus. *ACL.*

Shengyu Liu, Buzhou Tang, Qingcai Chen and Xiaolong Wang. Drug Name Recognition: Approaches and Resources. *Information*. 2015; 6(4):790-810.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta,Kyle Lo, Iz Beltagy, Doug Downey and Noah A. Smith. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *ACL.*

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. *ArXiv, abs/1907.11692.*

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu and Haifeng Wang. (2020). ERNIE 2.0: A Continual Pre-training Framework for Language Understanding. *AAAI.*