FBK@SMM4H2020: RoBERTa for detecting medications on Twitter

Silvia Casola
Università di Padova
Fondazione Bruno Kessler
scasola@fbk.eu

Alberto Lavelli
Fondazione Bruno Kessler
lavelli@fbk.eu

Abstract

This paper describes a classifier for tweets that mention medications or supplements, based on a pretrained transformer. We developed such a system for our participation in Subtask 1 of the Social Media Mining for Health Application workshop, which featured an extremely unbalanced dataset. The model showed promising results, with an F_1 of 0.8 (task mean: 0.66).

1 Introduction

Twitter is a valuable source of user-generated data, including health data, and might be used for monitoring drug abuse and adverse effects online. However, tweets that mention medications need to be flagged first. To do so, lists of drugs are not enough: this is due to common spelling mistakes on social networks and language ambiguity (e.g. Xanax might be a drug or a Belgian band). The Substask 1 of the 2020 Social Media Mining for Health workshop (#SMM4H) challenged participants to identify tweets containing drugs or dietary supplements' mentions. The corpus showed a natural, highly imbalanced distribution of labels, with extremely rare positive occurrences. In this paper, we describe our entries, based on a pretrained RoBERTa model (Liu et al., 2019).

2 Datasets

The main dataset (DS2020), for which participants generated predictions, is a corpus of tweets by 112 pregnant women, as described in Weissenbacher et al. (2019). Its label distribution is unbalanced (181+/69091-), with positive instances only constituting 0.2% of the training set. Given the scarcity of positive samples, we also employed another dataset (DS2018), previously used in the 2018 #SMM4H shared tasks (Weissenbacher et al., 2018). It consists of 9611 tweets from 7584 users, annotated for medications' and supplements' mentions; the dataset was artificially balanced (4975+/4647-). Additionally, we used a large corpus of about 5 million general-domain, unlabelled English tweets.

3 Method

We submitted three runs to the leaderboard. All models were based on a RoBERTa pretrained transformer. RoBERTa is an enhancement of BERT (Devlin et al., 2019), with a modified training process (objective and data) and hyperparameters.

The first run (*Base model*) is a simple pretrained RoBERTa base uncased classifier. We first fine-tuned the original model on the DS2018 dataset to classify tweets mentioning medications and supplements. The hyperparameters were selected to maximize F_1 for the DS2020 validation set. Following a transfer learning approach, we used the resulting weights to initialize the final model and trained it on the DS2020 data. The original text was not preprocessed; for both datasets, we only used the text of the tweets, discarding the date, time, and user. No list of medications or other prior knowledge was employed and no explicit training on the medical domain was performed.

In the second run (*MLM model*), we followed a similar pipeline, but pretraining RoBERTa on a large corpus of tweets as a first step. The training was performed with a Masked Language Model (MLM)

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 101–103
Barcelona, Spain (Online), December 12, 2020.

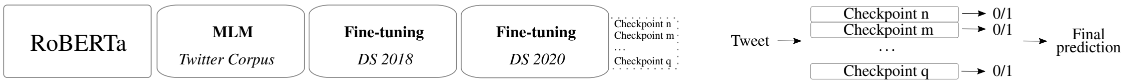

Figure 1: Our training (left) and test (right) pipeline for the *Ensemble MLM* model.

objective, virtually continuing the original transformer's training, to adapt it to the Twitter language specificities. Masked Language Modelling consists of masking out a fraction of the input tokens and training the model to predict the original tokens based on context.

Finally, we observed that, when trained for a varying number of epochs, the *MLM models* exhibited comparable F_1 but made different errors on the validation set. For this reason, we chose an ensemble of five MLM checkpoints (*Ensamble MLM model*) as our third run; we chose all checkpoints to have good results on validation data, but varying precision/recall thresholds. At test time, majority voting was used.

Once the hyperparameters were fixed on the validation set, all final models were trained on both the train and validation data to predict the test set. Figure 1 presents the pipeline employed for the third run.

4 Experiments and Results

In this section, we report our experimental results. Given the skewness of the dataset, system performances were measured in terms of F_1 for the positive class (i.e. tweets containing medications or supplements). Table 1 reports our results on the validation and test sets. The three runs surpassed by 14 points the mean of the submitted systems ($F_1 = 0.66$) and by 2 points the previous state of the art on the dataset (Weissenbacher et al., 2019), which used a complex ensemble model and external knowledge.

For all models, we used a batch size of 32, and experimentally found a constant learning rate of 10^{-6} to perform best. Surprisingly, a long fine-tuning seemed beneficial: training was stopped at epoch 132 (Base) and 73 (MLM) after evaluating the models at each epoch for a maximum of 200.

Figure 2 shows the errors made by the Base model. Note that some tweets might be hard to classify according to the gold standard even for a human annotator.

	Validation			Test		
	F_1	Precision	Recall	F_1	Precision	Recall
Base (DS2020)	0.81	0.82	0.8	-	-	-
Base (DS2018+ DS2020)	0.91	0.91	0.91	0.8	-	-
MLM model	0.86	0.84	0.89	0.8	-	-
MLM ensemble	0.88	0.84	0.91	0.8	0.77	0.83

Table 1: Experimental results

False positives: ”Cayden's already starting to open his eyes! The anesthesia might wear off sooner than we thought…”, ”OBGYN's hate it when you tell them that pulling out is your form of birth control. Like HATE IT.”, ”I have a Man Cold. Well, I'm pregnant and can't take ALL the good drugs so it is like I have a Man Cold.”

False negatives: ”No first thing in the morning and still have this migraine!!! It's been now 4 days and nothing is working! Including a blood patch”, ”So I'm just about to drink castor oil again because at this point I'm tired”, ”Yeah im allergic to codiene idunno if I spelled that right I learned that the hard way”.

Figure 2: Model errors

5 Conclusion

We presented our entries to the #SMM4H shared task. The model obtained promising results albeit trained on tweets' text only. Pretraining on the DS2018 set first highly improves the model performance (+0.1 F_1), while the MLM pretrain seems detrimental. This might be due to the quality of the unannotated dataset, which we used as obtained by the Twitter API, without any filtering or preprocessing.

References

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

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 preprint arXiv:1907.11692*.

Davy Weissenbacher, Abeed Sarker, Michael J. 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 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16, Brussels, Belgium, October. Association for Computational Linguistics.

Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connor, 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.