Adverse drug reaction detection in Twitter using RoBERTa and Rules

Sedigheh Khademi, Pari Delir Haghighi, Frada Burstein
Monash University
{Sedigh.khademi, Pari.Delir.Haghighi, Frada.Burstein}@monash.edu

Abstract

This paper describes the method we developed for the Task 2 English variation of the Social Media Mining for Health Applications (SMM4H) 2020 shared task. The task was to classify tweets containing adverse effects (AE) after medication intake. Our approach combined transfer learning using a RoBERTa Large Transformer model with a rule-based post-prediction correction to improve model precision. The model's F1-Score of 0.56 on the test dataset was 10% better than the mean of the F1-Score of the best submissions in the task.

1 Introduction

The enormous quantity of self-reporting about medication effects in social media provides an opportunity for additional surveillance of medication effects, supplementing traditional health-related reporting systems (Edo-Osagie et al., 2020). Task 2 of the SMM4H Shared Task challenge (Klein et al., 2020) was defined as a binary classification task to identify tweets reporting an adverse effect (AE) from taking a medication. The language differences between AEs and indications (the reasons for using a medication) in the data could be quite subtle and increased the difficulty of the task. Datasets were available in English, French and Russian, but we only worked with the English dataset.

Fine tuning of state-of-the-art language models for classification has proven to be highly effective when applied to small but varied text collections (Weissenbacher et al., 2019). We chose a Hugging Face RoBERTa Large Transformer model (Liu et al., 2019) based on our previous experience in identifying vaccine adverse reaction mentions in tweets. Informed by previous studies that have used lexicons to improve classification results (Asghar et al., 2017), we developed a post-prediction rule-based correction based on a lexicon of phrases that mostly appeared in what we judged were non-AE tweets in the test data. This improved the F1-Score by 2%.

2 Data description and preparation

The Shared Task English data was supplied as two tab-delimited text files, one with 20,544 tweets as training data, and another of 5,134 tweets as a validation dataset. Each tweet was accompanied with a tweet id, a user id, and a "positive" class label of 1 for an AE, and a "negative" class label of 0 for a non-AE. Eventually an unlabeled test dataset of 4,759 tweets was provided by the challenge organizers for submission to the task. Around 9.3% of the data was designated as the positive AE class in the labeled datasets.

A comparison of tweets indicates how difficult this challenge is, there sometimes seemed to be little difference between an AE and something like it. For instance, these consecutive entries with their classes, both describing a side effect of pain following rivaroxaban intake: *negative class*: "04.40 just taken flecainide and 2 paracetamol to dull pain side effect of rivaroxaban."; *positive class*: "20.05 day 16 rivaroxaban diary. intense back and knee pain most of day; taken more paracetamol. had to shop, walking painful".

There were several aspects of the data which had a bearing on data preparation and classifier training, and on our conclusions about possible future directions: (i) We identified 423 texts that were in both the training and validation datasets, and (ii) within these datasets there were also duplicates; (iii) the data was highly imbalanced.

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 113–117
Barcelona, Spain (Online), December 12, 2020.

To fix records that existed in both datasets, we removed them from the training data. We then also removed remaining duplicates from within each dataset. After cleaning, there were 1,896 positive labels in the training data and 473 in the validation data. The training dataset then consisted of 20,110 tweets – the 1,896 positive labels with the rest being all the negative labels. The de-duplicated validation dataset consisted of 2,167 negative and 473 positive labels, a total of 4,807 tweets.

We addressed the class imbalance in the training dataset by randomly sorting the data then splitting the negative labels of the training data into 7 datasets and combining each with all the positive labels. Therefore, each training dataset comprised of 2,602 negative labels and 1,896 positive labels, a total of 4,498 tweets.

The tweets text was not altered, and we added no features. Data was exported as a tab-delimited format suitable for use by the Transformer model. The Hugging Face libraries we used provided routines to tokenize and vectorize the text to Transformer model requirements.

3 Model

We used Hugging Face's RoBERTa Large Pytorch Transformer model (Liu et al., 2019) for classification, specifically the class *transformers.RobertaForSequenceClassification*, which is described in the Hugging Face documentation as a "RoBERTa Model transformer with a sequence classification/regression head on top (a linear layer on top of the pooled output)". RoBERTa is an acronym for "Robustly Optimized BERT Pretraining Approach" and was developed by Facebook to improve on Google's original BERT - "Bidirectional Encoder Representations from Transformers" (Devlin et al. 2019). A very summary explanation is that the models are based on Transformer language models and utilize multi-headed encoder/decoder attention mechanisms, which dispense with recurrence and convolutions entirely (Vaswani et al., 2017). They use intentionally masked sections of text to learn to predict the most probable words in sentences. RoBERTa improves on BERT by removing its next-sentence pretraining objective; by using larger mini-batches and learning rates; and using an order of magnitude more data and for a longer time than BERT was trained on. RoBERTa Large was the largest model of the available RoBERTa models on the Hugging Face site, we chose it because our tests on similar Twitter data have shown improvement in F1-Scores compared to BERT.

4 Method

4.1 Best model vs an ensemble

Models were learned on the larger imbalanced dataset and on the seven smaller balanced datasets, with F1-Scores calculated on the validation dataset. Experimentation showed that no more than 10 epochs were required to find the best model, and often only 2 epochs were needed. The best scoring model was retained from each of the eight datasets. An additional two models were kept because of their superior score, and the poorest scoring model was discarded, resulting in nine models. One of the retained models was learned on the large imbalanced dataset, the rest were learned on the smaller datasets, and the best individual performing model was one of these – it became our "Top 1" model. We then collated the predictions over the validation dataset for all odd-numbered (i.e. 3, 5, 7 and 9) groupings, to find out which maximum voting combinations most consistently matched the class. An ensemble of predictions from seven models had the best F1-Score - this was due to a reduction of false positives, leading to a greater precision. However, a consequence of the ensemble approach was a reduction in recall, and the single best model trained on one of the balanced datasets had the highest recall. Therefore, we considered how we might retain a high recall while reducing false positives to improve the F1-Score, and to assist with evaluation of recall significance we included an F1-beta score using a beta of 1.3.

4.2 Rule-based correction

Our data analysis included an examination of frequently appearing words and phrases in the classes. In the *training* data we noted that some phrases either inevitably or almost always were used by the negative class. Although this pattern did not appear consistently in the validation data, it was found in the test data. Analysis of our model predictions on the test data showed these phrases were present in what looked like false positive predictions. The phrases included mentions of lawyers and lawsuits, news and research, political figures, addictions, natural therapies, and encouragements to consult medical experts.

Therefore, we created post-prediction rules which we applied to the test data to enforce the negative class for tweets that contained these words or phrases.

For example, what we judged as the positive class and that our models had predicted as positive did *not* tend to contain phrases enjoining consultation with health experts or procuring medications. That is, phrases such as "ask your", "consult", "with your", "your doctor", "get your", and "try our" were almost always in discussions or in advertisements but *not* in reporting of adverse effects. To evaluate the effect of applying these rules in the test data we first labelled the test data with our model's predictions, then adjusted the labels to what the rules considered correct. This indicated that a 2% improvement in F1-Score was possible. We submitted a model with the correction, and in fact it did benefit by 2% over the same model without the correction.

5 Generated texts

When analyzing the data, we found two users in the training data who between them had 251 tweets that consisted of what looked like generated text. The tweets contained non-sensical phrases intermingled with key medication-related words, e.g.: "scuba dive ventolin hfa two brett butlers. the former"; only two of these texts were labelled as AE. We assessed the functionality of the 251 generated texts by learning classifiers on seven balanced datasets *without* the generated tweets, retaining the best model from each dataset. Compared to the models trained on all the data we found a higher recall to precision ratio, but that F1-scores were uniformly worse by at least 0.01. In this case we found an ensemble of the top 5 models gave the best F1-Score, but at 0.661 on the validation data it was 0.024 worse than the 0.685 of the best ensemble trained on all the data.

Although it was not performant, we wanted to know how this worked with the test data and so made a submission using the ensemble with a manual correction. Its test F1-Score was 0.46, which although identical with the mean best test dataset F1-Score for the challenge, was 10% below our best score of 0.56 obtained when the generated text was retained.

It was apparent that the generated text had a correcting effect by removing false positives. It seemed possible that this was due to their containing text with potentially AE indicative words surrounded with words and language structure that was not indicative of AE. A 10% difference was significant and so it seems likely that these texts have been introduced by the challenge organizers as a corrective measure.

6 Results

Scoring and Ensembles: Table 1 illustrates the scores obtained on validation data. The first column is the best model learned on the imbalanced dataset. Under the heading "Ensembles" the "Average" column is the average of the scores obtained from the models used in the various ensembles, and the remaining columns are scores obtained from ensembles, starting with "Top 1" which is the single best model trained on a balanced dataset.

	Imbalanced dataset	Ensembles						Without generated
		Average	Top 1	Top3	Top 5	Top 7	Top 9	
TP	330	328	**366**	339	337	337	334	**380**
TN	4,146	4,139	4,083	4,147	4,149	4,160	**4,165**	4,037
FP	188	195	251	187	185	174	**169**	297
FN	143	145	**107**	134	136	136	139	**93**
Precision	0.637	0.627	0.593	0.644	0.646	0.659	**0.664**	0.561
Recall	0.698	0.693	**0.774**	0.717	0.712	0.712	0.706	**0.803**
F1-Score	0.666	0.659	0.672	0.679	0.677	**0.685**	0.684	0.661
F1-Beta (1.3)	0.674	0.667	**0.695**	0.688	0.686	0.692	0.690	0.692

Table 1: Model scores on validation data

The table shows that the best model for recall (F1-Beta at beta 1.3) is the "Top 1" model, but that the highest F1-Score belongs to an ensemble of 7 models due to its balance of recall and precision. The final

"Without generated" column contains the scores of the best, 5-model ensemble when models were trained without the generated texts – it has a higher recall than any other model, but also the poorest precision.

Submitted Models: Table 2 shows the test scores obtained from the models we submitted. The first column contains the means of the scores of the best models submitted to the task. The second column has all the scores from our best submitted model, which was the Top 1 model described above, but with the rule-based post-prediction corrective step applied. Its F1-Score it 10% better than the mean best F1-Score of the task. The third column is from the same model but without the corrections – as we had estimated there was a 2% difference between the two. The third column is the best model trained without the generated text in the training data – its F1-Score matches that of the mean best score. We were only given precision and recall scores for our best submission.

	Mean of Task best	Top 1 model corrected	Top 1 model uncorrected	Without generated
Precision	0.42	0.56		
Recall	0.59	0.55		
F1-Score	0.46	**0.56**	0.54	0.46

Table 2: Test scores of submitted models

7 Analysis

The best F1-Score of 0.685 we achieved on validation data outperformed the best F1-Score of 0.665 on validation data in the equivalent 2019 task. The model we submitted was our "Top 1" model with a manual correction applied, its (uncorrected) validation F1-Score of 0.672 was also above last year's best validation score. Our best F1-Score of 0.56 on the test data exceeded the average 2020 year's best test F1-Score of 0.46 by 10%.

An increased score over last year was possible as the RoBERTa Large model is superior to the BERT model used by some of last year's entrants. We did not do any extra training or tuning specific to the task (e.g. tuning with texts containing drug names from the training set). We did not evaluate any difference that might have been made by the removal of the duplicated tweets, but it is possible that some of our result is due to that data cleaning.

Under-sampling the negative class by splitting the data into 7 reasonably balanced datasets improved model performance, but it came with the price of decreasing the training examples. Ameliorating the data splitting by creating an ensemble of the best models' predictions increased precision, but, as noted above, this was at the expense of recall. In the end, the model used for our submission was the one that was trained on the best randomly created individual split dataset, as it had the highest recall and we wanted to evaluate manual corrections with it. We submitted predictions from this model with and without the rule-based post-prediction corrective step applied. The corrected model gained our top F1-Score of 0.56, the uncorrected score was 0.54. The extra 2% gained by the corrective step was useful and suggests additional development of such corrective processes.

The misclassified tweets on the validation data are numerous. Sometimes the difference in the language between the positive and negative classes was so subtle that neither we as non-experts, nor indeed the models, could pick the class correctly. For instance, all the models predicted the positive label for this negative class: "i am on only 25mg lamictal but it makes me anxious and insomniac so i skip sometimes and get depressed". We can understand that the model finds anxiety, insomnia and depression as side effects, and although we as human judges can see that depression is in the context of not taking the medication the other two symptoms still pertain. However, we think that this was labelled as a negative class because it is not reporting a presently experienced side effect - the tense indicates an ongoing general experience with the medication.

8 Future work

Having established a benchmark, in the future we should try to understand the data further and see what can be done to improve the data as well as the models' ability to learn the data idiosyncrasies, as there

is a lot of potential improvement in these areas. We are inspired by our observation of the effect of inserting generated texts into the data, and would like to research the applicability of that approach for improving training data - for instance, by inserting examples of the use of tense around side effects to indicate whether a current AE is being experienced. Additionally, we intend to explore the improvements possible with steps such as retraining the underlying language model with AE and applicable drug mentions, including using word masking on these texts and additional examples like them.

References

Asghar, M. Z., Khan, A., Ahmad, S., Qasim, M., & Khan, I. A. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PloS One*, *12*(2), e0171649.

Edo-Osagie, O., De La Iglesia, B., Lake, I., & Edeghere, O. (2020). A scoping review of the use of Twitter for public health research. *Computers in Biology and Medicine*, 103770.

Klein, A. Z., Alimova, I., Flores, I., Magge, A., Miftahutdinov, Z., Minard, A.-L., O'Connor, K., Sarker, A., Tutubalina, E., Weissenbacher, D., & Gonzalez-Hernandez, G. (2020). Overview of the Fifth Social Media Mining for Health (SMM4H) Shared Tasks at COLING 2020. *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*. https://doi.org/10.18653/v1/w19-3203

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. *1*. http://arxiv.org/abs/1907.11692

Weissenbacher, D., Sarker, A., Magge, A., Daughton, A., O'Connor, K., Paul, M., & Gonzalez, G. (2019). Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, 21–30.