Approaching SMM4H 2020 with Ensembles of BERT Flavours

George-Andrei Dima[1,2], Andrei-Marius Avram[1,3], Dumitru-Clementin Cercel[1]
University Politehnica of Bucharest[1]
Military Technical Academy Ferdinand I[2]
Research Institute for Artificial Intelligence, Romanian Academy[3]
andrei.dima@mta.ro, avram.andreimarius@gmail.com, dumitru.cercel@upb.ro

Abstract

This paper describes our solutions submitted to the Social Media Mining for Health Applications (#SMM4H) Shared Task 2020. We participated in the following tasks: Task 1 aimed at classifying if a tweet reports medications or not, Task 2 (only for the English dataset) aimed at discriminating if a tweet mentions adverse effects or not, and Task 5 aimed at recognizing if a tweet mentions birth defects or not. Our work focused on studying different neural network architectures based on various flavors of bidirectional Transformers (i.e., BERT), in the context of the previously mentioned classification tasks. For Task 1, we achieved an F1-score (70.5%) above the mean performance of the best scores made by all teams, whereas for Task 2, we obtained an F1-score of 37%. Also, we achieved a micro-averaged F1-score of 62% for Task 5.

1 Introduction

In recent years, researchers around the world came to realize the usefulness of social media data for extracting health information. The Social Media Mining for Health Applications (#SMM4H) Shared Task 2020 (Klein et al., 2020) brings to the forefront the problem of extracting information from health social media posts. SMM4H was also organized in the previous years and involved several tasks, such as automatic detection of tweets mentioning medication, of tweets describing medication intake or adverse reactions, of tweets mentioning vaccination behaviour, as well as tasks on extraction and normalization of adverse effects utterances (Weissenbacher et al., 2018; Weissenbacher et al., 2019b). This year's competition had five tasks and our team focused on Tasks 1, 2, and 5.

Task 1 was a binary classification one that involved distinguishing between tweets in which some medications or dietary supplements were mentioned (the positive class) and other tweets (the negative class). The challenge of this task, unlike the 2018 similar task, consisted in the highly imbalanced data sets, that is, the tweets had a distribution of the two classes similar with the one encountered in practice. Therefore, the positive class counted for only 0.2% of the examples.

Classification of multilingual tweets that report adverse effects (Task 2) was also a binary classification task that was divided in multiple subtasks, each one concerning a different language. Our team submitted solutions only for the English subtask. For both Tasks 1 and 2, the evaluation metric was the F1-score for the positive class. On the other hand, classification of tweets reporting a birth defect pregnancy outcome (Task 5) was a multi-class classification task with the following classes: *defect*, *possible defect*, and *non-defect*. For this task, the evaluation metric was the micro-averaged F1-score of the first two classes.

We started to approach the competition by studying methods for data preprocessing and for overcoming the class unbalancing issue. Afterwards, we tested multiple language models for classification and we contributed by further pre-training the best language model on social media data. Moreover, we increased the robustness of our approach by constructing an ensemble which managed to obtain 70.5% F1-score, whereas the average of the best scores for the Task 1 was 66.28%.

This paper is further structured as follows. Section 2 presents previous works that helped us in developing our solutions. Section 3 describes the proposed models. In Section 4, we show and interpret the results of the systems. Finally, Section 5 summarizes the conclusions of our work.

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 153–157
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

As related tasks were also organized in previous years of SMM4H, a significant amount of work has been done in developing solutions to the proposed problems. Therefore, previous studies already established some directions for appropriate methods to preprocess social media data, for practices to address imbalanced data sets, or for language models that are more effective for the given tasks. Thus, Ellendorff et al. (2019) showed which data preprocessing steps are more likely to obtain better results on tweets.

The challenge of learning from imbalanced data sets has been reviewed by Chawla et al. (2004). Their paper analyzed some general solutions like over-sampling and under-sampling and also offered useful guidelines in applying these techniques. Moreover, Khosla (2018) approached the problem by assigning different weights to the imbalanced classes and this method proved particularly useful in our experiments.

Mahata et al. (2019) used transfer learning approaches, showing that Bidirectional Encoder Representations from Transforms (BERT) (Devlin et al., 2018) and Universal Language Model Fine-Tuning (ULMFiT) (Howard and Ruder, 2018) are able to handle classification tasks in the medical domain. Also, Gondane (2019) leveraged the focus on biomedical language of BERT for Biomedical Text Mining (BioBERT) (Lee et al., 2020).

The issue of detecting medication mentions in social media data has been previously approached with impressive results. Weissenbacher et al. (2019a) showed that ensemble classifiers can achieve performance close to humans in recognizing mentions of medications in tweets, on balanced data sets. Wu et al. (2018) obtained notable results on social media data mentioning drugs using a neural network based on a multi-head self-attention mechanism.

3 Method

3.1 Text Preprocessing

Data gathered from social media (tweets in our case) implies a specific informal language that is closer to the spoken English, rather than the texts on platforms like Wikipedia. This type of text is rich in grammatical errors, abbreviations (e.g., *"cuz"* instead of *"because"*) and words (e.g., *"lol"* or *"idk"*) that are encapsulating their own meaning and cannot be found in usual dictionaries.

Before feeding this kind of data to a language model, a preliminary step must be done, i.e. preprocessing. It can be noted that this step might strip useful information. Fortunately, previous work (Ellendorff et al., 2019) gave a direction of which preprocessing would provide the best results. For our methods, the best results were obtained by using the following preprocessing steps:

- Replace all URLs with *"url"*;
- Replace all usernames with *"user"*;
- Remove all non-ASCII characters;
- Remove all HTML character references;
- Replace multiple white spaces with one space.

We also experimented with other techniques for spell correction via the Ekphrasis library (Baziotis et al., 2017) but, for our models, the results were not significantly improved.

3.2 Experiments

After preprocessing, data was fed to BERT-based language models that are related to the medical field, namely: BioBERT, Clinical BERT (Alsentzer et al., 2019), BioFLAIR (Sharma and Daniel Jr, 2019), and BioELMO (Jin et al., 2019). Among these four models, the best results were achieved using BioBERT.

BERT was bidirectionally pre-trained using two tasks: Masked Language Modeling and Next Sentence Prediction on a large corpus composed of English Wikipedia and BookCorpus. Starting from the pre-trained BERT, BioBERT was further pre-trained on a biomedical corpus composed of PubMed abstracts and articles.

For Task 1, we performed a series of experiments. First, we fine-tuned the last three layers of BioBERT-Base on the balanced data set of SMM4H 2018 (Weissenbacher et al., 2018), using the *early stopping* technique. Afterwards, we further fine-tuned our model on the actual training set of Task 1. Because this data set was highly unbalanced (55,273 negative examples and only 146 positive examples), the model tended to classify all examples as negative. To overcome this challenge, we experimented with several techniques: over-sampling the positive class, under-sampling the negative class, the *focal loss* (Lin et al., 2017), and adding weights for each class when computing the *binary crossentropy* loss. For the *class weighting* technique, we computed the weights using the formula proposed by Khosla (2018). Our results show that this technique obtains the best performance. We will further refer to this system as **BioBERT-ClassWeights**.

As we mentioned earlier, the language used in tweets is rather different of the language used in the corpora that BioBERT was pre-trained on. Therefore, it seemed intuitive to further pre-train the obtained system on data from social media. Due to the lack of considerable resources, we confined on using the English tweets from the data sets provided within the SMM4H 2020 shared task for the tasks 1, 2, 3, and 5, in order to form a corpus, and we pre-trained BioBERT on it, using the script provided on GitHub[1]. We used this language model in the same system, which we described above, and obtained **BioBERT-PretrainTweets**. Even though this method improved the results, training multiple models resulted in distant scores, thus showing that the model cannot be considered as having sufficient robustness.

In order to improve the robustness of the solution, we constructed two ensembles of multiple classifiers. **Ensemble 1** was formed from three models that performed well on the validation set: BioBERT-Base pre-trained on tweets, Clinical BERT and BioBERT-Large. For **Ensemble 2**, the training and validation sets were combined, shuffled and then splitted in five equally sized folds. Five models of *BioBERT-PretrainTweets* were fine-tuned as for 5-fold Cross-Validation and were afterwards used to form an ensemble. Both ensembles are deciding by averaging the outputs of the composing models.

For Tasks 2 - English and 5, we used *BioBERT-PretrainTweets* fine-tuned for each task. We should mention that for Task 5, concerning multi-class classification, we switched the loss function to *categorical crossentropy*. Yet, because *class weighting* did not improve the results, we decided not to use it.

4 Results

In the practice phase, we submitted predictions on the validation sets for Tasks 1 and 2 - English. For the first task, *BioBERT-ClassWeights* achieved an F1-score of 67.6% with a precision of 69.6% and a recall of 65.7%, while *BioBERT-PretrainTweets* obtained an F1-score of 77.61% with a precision of 81.2% and a recall of 74.2%. For Task 2 - English, *BioBERT-PretrainTweets* achieved 53.87% F1-score on the validation set.

In the evaluation phase, we submitted three solutions for the first task: one prediction from *BioBERT-PretrainTweets* and one prediction from each ensemble described above. The prediction of the *Ensemble 2* scored above the mean scores for Task 1. We also submitted one solution for each Task 2 - English and Task 5. Tables 1, 2, and 3 show the reported scores for each submission, alongside with the averaged score of best submissions of all teams that participated. The precision and recall for the first two submissions on the Task 1 were not reported by the organizers.

Model	F1-score	Precision	Recall
BioBERT-PretrainTweets	55%	-	-
Ensemble 1	66%	-	-
Ensemble 2	$\mathbf{70.5}\%$	79.03%	63.64%
Mean score	66.28%	70.32%	69.48%

Table 1: Test results and the average of best submissions for Task 1.

[1] https://github.com/google-research/bert/blob/master/run_pretraining.py.

Model	F1-score	Precision	Recall
BioBERT-PretrainTweets	37%	26%	60%
Mean score	46%	42%	59%

Table 2: Test result and the average of best submissions for Task 2 - English.

Model	F1-score	Precision	Recall
BioBERT-PretrainTweets	62%	56%	69%
Mean score	65%	62%	68%

Table 3: Test result and the average of best submissions for Task 5.

For Task 1, the scores on the test set indicate that using only *BioBERT-PretrainTweets* is not enough. It performed below average for all tasks, even though the validation scores would have suggested otherwise. On the other hand, the scores of the ensembles improved the prediction significantly. The score of *Ensemble 2* was above the average of the best scores, showing that our best system is an ensemble of BioBERT language models each fine-tunned on both the validation and training sets in a 5-folds manner. Even though the training set is large enough, the positive class is so poorly represented that previously mentioned techniques, which usually address small data sets, significantly improved our system.

5 Conclusion

In this paper, we experimented with ensembles of bidirectional Transformers in the context of social media texts and we studied the value that pre-trained BERT flavours, like BioBERT or ClinicalBERT, bring in solving classification tasks in the medical domain.

We succeeded in obtaining a score above the average of the best scores using *Ensemble 2* on Task 1 and we showed that BERT-based classifiers can give acceptable results even with highly unbalanced data sets. We also showed that a BERT-based language model, pre-trained for a rather colloquial language, improved the results on the given tasks of social media data. Our results on Task 1 show that, in cases where one of the classes is poorly represented, ensembles increase the prediction performance.

Further experiments should consider including an enlargement of the corpus of tweets used for pre-training BioBERT, so that the model will be more capable of representing this type of data. The next step would be to design a new BERT flavour pretrained on social media texts. Another future direction in addressing class imbalance might consist in using data augmentation in order to generate examples of the less represented class (Croce et al., 2020).

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tilia Ellendorff, Lenz Furrer, Nicola Colic, Noëmi Aepli, and Fabio Rinaldi. 2019. Approaching smm4h with merged models and multi-task learning. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 58–61. University of Zurich.

Shubham Gondane. 2019. Neural network to identify personal health experience mention in tweets using biobert embeddings. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 110–113.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89.

Sopan Khosla. 2018. Emotionx-ar: Cnn-dcnn autoencoder based emotion classifier. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 37–44.

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*.

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*, 36(4):1234–1240.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Debanjan Mahata, Sarthak Anand, Haimin Zhang, Simra Shahid, Laiba Mehnaz, Yaman Kumar, and Rajiv Shah. 2019. Midas@ smm4h-2019: Identifying adverse drug reactions and personal health experience mentions from twitter. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 127–132.

Shreyas Sharma and Ron Daniel Jr. 2019. Bioflair: Pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. *arXiv preprint arXiv:1908.05760*.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 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.

Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connor, Arjun Magge, and Graciela Gonzalez-Hernandez. 2019a. 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, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019b. Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 21–30.

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 tweet representation and multi-head self-attention. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 34–37.