Are BERT-based models really king in domain specific scenarios? A Twitter COVID-19 Sentiment Analysis Case Study

Abstract

The field of natural language processing (NLP) is evolving quickly, and it is seeming dominated by ever larger models. Particularly in the subfield of sentiment analysis, state-of-the-art results are almost entirely obtained by BERT-based models, that require many hours to train and are beyond the computing capacity of most organizations. However, research on Twitter sentiment analysis indicates that domain adaptation may key for classifying sentimenton tweets. To further investigate this issue, the performance of three models was examined on a Twitter dataset which contains tweets about COVID-19. The models investigated were a roBERTa-based model, an RNN-based model and a traditional random forest classifier trained with a dictionary of terms. The results show that the RNN-based model performed best, with an accuracy of 86%, followed by the traditional random forest classifier and the roBERTa-based model with 72% and 46% accuracy, respectively. These results suggest that indeed, domain adaption is very important in specific scenarios, such as Twitter COVID-19 sentiment analysis. While the RNN-based model had a performance that was 14% superior to second best model, it took 690 times longer to train, which shows that, when computation capacity is a limitation or in cases in which training must be sped up, traditional NLP feature extraction techniques may still be useful.

1 Introduction

Sentiment analysis is a Natural Language Processing (NLP) task that has been around for over 20 years and has widespread commercial applications in various domains [6]. Yet, a search for benchmark datasets to perform sentiment analysis on the Papers with Code website returns many datasets, but only one is related to Twitter data [8]. Tweets have special characteristics and, as pointed out by researchers [10, 6], results obtained for other kind of datasets may not transfer well to Twitter data. Tweets are different to other text in that they are rife with internal slangs, abbreviations, and emoticons [4]. Moreover, tweets are compact, often employ novel language with Twitter-specific communication elements, have a strong sentiment class imbalance and are stream-generated [6] – all aspects that add complexity of mining their opinions. So, while very large models, such as BERT-based models, have been the state-of-art for many datasets used in the sentiment analysis task, they may not be the best for Twitter data. In reality, earlier work by Zimbra et al. [11] indicates that domain adaptation is particularly important to properly classify tweets. To investigate this issue further, this study assesses the performance of three different models on a recently released Twitter dataset that contains tweets about COVID-19. The studied models are a roBERTa-based model, a recurrent neural network (RNN)-based model and a traditional random forest classifier trained with a dictionary of terms). Results show that models trained with specific domain data outperformed very large models trained only on tweets (but not on tweets related to COVID-19).

Preprint. Under review.

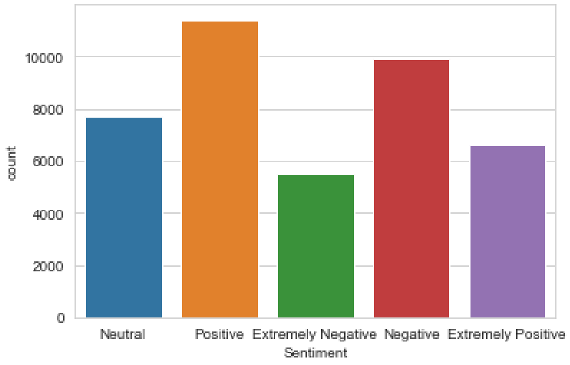

Figure 1: Data imbalance in the training data.

The remainder paper is organized as follows: section 2 presents related work, section 3 introduces the experimental setup used in the experiment, section 4 describes the results, and section 5 concludes with suggestions for future work.

2 Related Work

Sentiment analysis is often perceived as a mature field, though much is still to be explored [11]. As it happened in the field of computer vision, sentiment analysis moved from feature extraction to deep learning approaches, with convolutional neural networks (CNNs) and recurrent neural networks (RNNs) playing an important role in this transition [10]. Bidirectional Encoder Representations from Transformers (BERT) were created in 2018 and soon topped the performance on the sentiment analysis for many datasets [6]. Regarding dataset, sentiment analysis on Twitter data was historically studied using the ACL 14 Twitter and SemEval 2014 Task 4 2 datasets; however, these datasets pre-date the introduction of BERT. In 2020, a new Twitter related dataset was released [2] and the current state-of-the-art for it is a BERT-based model. Still, researchers argue that for Twitter data, aspects other than the model may be more relevant, such as the pre-processing of the data [4] or training the model for a specific dataset of interest [11]. The next section presents the experimental setup used to investigate this latter question.

3 Experimental Setup

The aim of this research is to investigate how large BERT-based models compare to simpler domain adapted models for a new Twitter COVID-19 dataset. The following subsections describe the dataset, pre-processing and chosen models for the study.

3.1 Dataset

The dataset chosen for the experiment was made available by Aman Miglani at Kaggle [5] and consists of tweets collected between Dec 31st, 2019 and Sep 8th, 2020 about the COVID-19 pandemic. The dataset metadata does not contain details on the criteria for collecting the dataset or how it was labelled into one of the five categories: extreme negative, negative, neutral, positive, and extreme positive. Figure 1 shows the number of tweets in each category and, as it can be seen, it is an imbalanced dataset. The dataset is available for download at Kaggle in the form of train and test data with 41,157 and 3,798 tweets, respectively. This original train data was subsequently split into training and validation sets with a 75/25 ratio. The new training set was used for training the RNN-based model and the random forest model. The next subsection presents the models used in the study.

3.2 Models and Hyper-parameters Selection

The models assessed in this study are presented below. All models were trained using an Intel Core i7-8665U CPU equipped with memory of 16 GB.

RNN-based model was developed by Cho et al. [3] and consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other RNN decodes the representation into another sequence of symbols. The two RNNs are jointly trained to maximize the conditional probability of a target sequence given a source sequence [3]. This model was introduced during the Empirical Methods in Natural Language Processing (EMNLP) of 2014 and has gained attention of the NLP community [10], especially because it performs well on small datasets. It a slight variation of a Long Short Term Memory network (LSTM) and it is ofte refered to as a Gated Recurrent Unit (GRU) [10]. It combines the "forget" and "input" gates into a single update gate and merges the cell state and hidden state [10]. The model was implemented using Keras 2.4.3. The embedding dimension used was 16 and number of units 256, following from Tripathi [9] work on the COVID-19 dataset on Kaggle. The return of the last output in the output sequence was set to true, following from [9]. Likewise, global average pooling, a dense layer of dimension 64, relu activation and a drop-out layer of 0.4 were selected for the same reason. The model was trained using the Adam optimization algorithm.

Random Forest model was develped using traditional NLP techniques for text classification which involve building a sparse matrix with the words present in the sentences that one wishes to classify. In this study, these steps were performed using the scikit-learn 0.24.1. The random forest classifier was set with the number of estimators (trees) equal to 40, criterion 'gini', maximum depth of 25,000 and a minimum sample split of three.

roBERTa-based model was developed by Qudar and Mago [7] in a paper published at Empirical Methods in Natural Language Processing (EMNLP) of 2020. The model is currently the best performing model (state-of-the-art) in the recently launched TWEETEVAL benchmark, that is comprised of seven tasks related to Twitter data. The model was run using the transformer library 4.5.1 in Pytorch through the Hugging Face website [1].

3.3 Pre-processing

As it is usual in NLP sentiment analysis tasks, the extreme categories were merged with the less extreme ones, that is, for example, extreme negative was merged with the negative category. Moreover, the tweets were cleaned to remove stop words, website pages and symbols. The same pre-processing was applied to the data used to evaluate all three models.

3.4 Metrics

As it is common in sentiment analysis [6], the results were evaluated using accuracy as the main metric and the F1-score as a secondary metric. The chosen metrics are calculated as follows:

- Accuracy: true positive + true negative / total number of rows/tweets in the dataset
- F1-score: true positive / (true positive + ½ (false positive + false negative))

4 Results

The comparative analysis of the roBERTa-based model, random forests and the RNN-based model shows that the latter performed better than the other ones. The RNN-based model had a total accuracy across all labels of 86%, while the random forest model and the roBERTa-based model had an accuracy of 72% and 46%, respectively. Table 4 and 4 summarizes the results for each label (negative, neutral, and positive) and for the whole dataset.

These results show that domain-specific training plays an important role in the accuracy of the model, as the models trained using the Twitter COVID-19 dataset had a better performance. The results were obtained for a fraction of the validation data due to limitations in the usage of Hugging Face, which hosts the roBERTa-based model. The RNN-based model was also evaluated in the test set and yield an accuracy similar to the one found on the validation set, 83%.

Table 1: Results for the validation data

Model	Negative Acc.	Neutral Acc.	Positive Acc.	Total Accuracy
roBERTa-based	57%	76%	28%	46%
Random forests	64%	87%	73%	72%
RNN-based	90%	83%	84%	86%

Table 2: Further results for the validation data

Model	Negative f1-score	Neutral f1-score	Positive f1-score
roBERTa-based	0.61	0.36	0.42
Random forests	0.72	0.61	0.79
RNN-based	0.84	0.84	0.88

Figure 2 shows the loss and accuracy for the training and validation sets during training of the RNN-based model. As it can be seen, after the second epoch the model starts to overfit the training data and the performance on the validation stagnate. Because the difference in performance in the validation set was less than 1.53%, the final model (epoch 5) was used for evaluation on the test data.

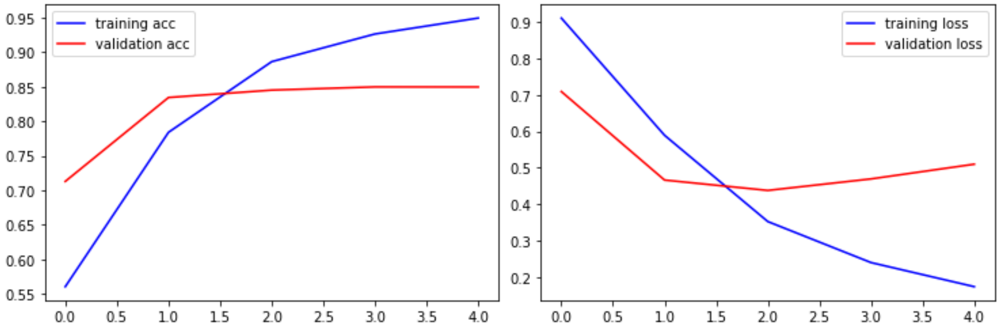

Figure 2: Accuracy and loss during training of the RNN-based model.

Regarding computation times, the RNN-based model took 3 hours and 19 min to train, while the random forest model took 17.3 seconds. Indeed, the fact that RNNs/LSTM/GRU take a long time to train has been a source of criticism and research for lighter models. For instance, the current state-of-the-art model in the only Twitter related dataset for sentiment analysis on Papers with Code, specifically proposes a light-weight model as a counterpart to heavy models [8].

The accuracies found are in line with the ones reported by Zimbra et al. [11] in their review about the state-of-the-art in Twitter Sentiment Analysis. Moreover, Zimbra et al. [11] also found domain-specific models to perform better for Twitter data, which was confirmed for the dataset used in this study as well.

As mentioned in section 3, all models were evaluated on the same data, that had been pre-processed the same way. However, the original paper in which the roBERTa-based model was proposed [7] employed a much simpler pre-processing (stop words were not removed, for example). It is possible that the applied pre-processing may have impacted the performance of the model. As a matter of fact, research shows that the pre-processing of text plays a very important role in Twitter sentiment analysis, but that it is a step that is often ignored by researchers [4].

Lastly, a comparison of the instances wrongly classified by two best performing models indicate that at least 35% of the tweets mistakenly classified by the RNN-based model had been correctly classified by the random forest model. This suggests that an ensemble approach would be desirable for this dataset. Indeed, the use of ensembles had been already recommended by researchers who investigated the performance of different models for Twitter sentiment analysis classification [11].

Related to the wrongly classified tweets, work published by Song et al. [8] in 2019 highlights the importance of manually inspecting misclassified tweets for faulty labelling. As the authors explain, sentiment labelling can be very subjective and in the absence of clear guidelines for labeling, it may be the case that some of the training and test data were labelled in an incompatible way. As mentioned in section 3, the dataset does not contain metadata on the criteria used in the labeling process, so Song et al. [8]'s observation may be relevant for the dataset used in this study.

5 Conclusion & Future Work

In this work the performance of three different models was assessed to investigate whether BERT-based models without domain adaptation can outperform simpler models trained on a specific dataset. A dataset of tweets related to COVID-19 was used as a case study. Results show that domain adaptation is key for good performance in domains like Twitter COVID-19 sentiment analysis classification.

In the future, it would be interesting to further investigate how different hyper-parameters impact the performance of the RNN-based model (the best performing model), the impact of different pre-processing steps, the performance of ensembles of models in this dataset and the quality of the dataset labels.

References

[1] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke. cardiffnlp/twitter-roberta-base-sentiment · Hugging Face, 2020. URL `https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment`.

[2] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *arXiv:2010.12421 [cs]*, Oct. 2020. URL `http://arxiv.org/abs/2010.12421`. arXiv: 2010.12421.

[3] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs, stat]*, Sept. 2014. URL `http://arxiv.org/abs/1406.1078`. arXiv: 1406.1078.

[4] Z. Jianqiang and G. Xiaolin. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access*, 5:2870–2879, 2017. ISSN 2169-3536. doi: 10.1109/ACCESS.2017.2672677. Conference Name: IEEE Access.

[5] A. Miglani. Coronavirus tweets NLP - Text Classification, 2020. URL `https://kaggle.com/datatattle/covid-19-nlp-text-classification`.

[6] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. ISSN 1949-3045. doi: 10.1109/TAFFC.2020.3038167. Conference Name: IEEE Transactions on Affective Computing.

[7] M. M. A. Qudar and V. Mago. TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis. *arXiv:2010.11091 [cs]*, Oct. 2020. URL `http://arxiv.org/abs/2010.11091`. arXiv: 2010.11091.

[8] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao. Attentional Encoder Network for Targeted Sentiment Classification. *arXiv:1902.09314 [cs]*, 11730:93–103, 2019. doi: 10.1007/978-3-030-30490-4_9. URL `http://arxiv.org/abs/1902.09314`. arXiv: 1902.09314 version: 2.

[9] H. Tripathi. COVID 19 Tweets Analysis, 2021. URL `https://kaggle.com/himanshutripathi/covid-19-tweets-analysis-97-accuracy`.

[10] L. Zhang, S. Wang, and B. Liu. Deep Learning for Sentiment Analysis : A Survey. *arXiv:1801.07883 [cs, stat]*, Jan. 2018. URL `http://arxiv.org/abs/1801.07883`. arXiv: 1801.07883.

[11] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Transactions on Management Information Systems*, 9(2):5:1–5:29, Aug. 2018. ISSN 2158-656X. doi: 10.1145/3185045. URL `https://doi.org/10.1145/3185045`.