Effectiveness of BERTweet on Modern Tweet Classification Tasks: A Study in the Importance of the Pre-Training Dataset

Sofija Kotarac
Department of Computer Science
The University of Toronto
Toronto, ON M5S 1A1
sofija.kotarac@mail.utoronto.ca

Alex Pejovic
Department of Computer Science
The University of Toronto
Toronto, ON M5S 1A1
alex.pejovic@mail.utoronto.ca

Abstract

The BERTweet model [1] presents an opportunity for a great case-study on the importance of the specificity of the pre-training dataset for pre-trained NLP models. Trained on Twitter-data, having a sister model trained on regular English language data, RoBERTa [2], we use these two models to test whether domain-specific pre-training can provide better results for an NLP model. Through experimental evidence, we arrive at conclusions which seemingly contradict previous papers on the topic, thus we advise more studies be done on the significance of domain-specific pre-training.

1 Introduction

BERTweet [1] is the first public large-scale pre-trained language model for English Tweets. The model has the same architecture as BERT$_{base}$ [3], and it is trained using the RoBERTa pre-training procedure [2]. Since RoBERTa is a publicly available pre-trained language model, being a near state-of-the-art model – partially evidenced by its position as 19th on the GLUE [4] leader-board[1], as well as variations of the model being in the top 10 – with one of the few differences between it and BERTweet being the dataset it was pre-trained on, we are given an opportunity for a great case-study on the importance of the relationship between the pre-training dataset and the desired tasks we want our model to handle.

In this paper, we compare performances between the RoBERTa$_{base}$ model – which is pre-trained on 160GB of text from books, Wikipedia, news articles, etc. [2] – and the BERTweet$_{base}$ model – which is pre-trained on 80GB of Tweets [1]. We compare the performances of these models against many Twitter related text-classification tasks, and a few text-classification tasks not related to Twitter. We hypothesize that the BERTweet model will perform better on the Twitter related tasks due to it most likely having a better semantic understanding of Twitter-language due to Twitter-English having many irregularities in comparison to English written in more formal settings [5]. We additionally suspect our RoBERTa model to perform better on the non-Twitter tasks, as it is pre-trained on more formal English.

[1] https://gluebenchmark.com/leaderboard

Preprint. Under review.

2 Architectures and Tasks

In this section, we briefly go over the two models we use for conducting our experiments, explaining how they are similar, how they are different, and which parts of the models are relevant to our experiment. Additionally, we go over the test tasks used to compare the models.

2.1 Pre-training Architecture

For a detailed overview on how these two models are pre-trained, it is highly recommended that you read the original papers where the models were published [1] [2]. All that is relevant to our experiment, as discussed in the introduction, is that BERTweet and RoBERTa follow the same architecture. However, with an exception of differences in datasets used for pre-training, and the method of tokenizing the data – BERTweet tokenizing with the "TweetTokenizer" from the NLTK tool-kit [6], and RoBERTa tokenizing with a slight variation of BPE (byte-pair-encoding) [7].

2.2 Fine-tuning Architecture

All of the tasks done in our experiments are done by fine-tuning the model for each task. Since all of our tasks are single-sentence text-classifications, we can use the same fine-tuning architecture for each task. We append a linear prediction layer on top of the pooled output of our models in order to make a classification. We learn the parameters of this output linear layer using the training split of a given task dataset, and test the accuracy of our predictions using the test split of the data. All models learned during pre-training are frozen in order to ensure the integrity of the pre-training model, thus the only parameters we learn in our tests are those belonging to our final linear classification layer, whose weights are randomly initialized. The final layer has the same number of outputs as there are possible labels for our given task. Argmax is used to make a prediction for the label.

We use the transformers library [8] to aid us in fine-tuning both models, and we do an independent fine-tuning for each task. Each fine-tuning is done for 7 epochs over the training set. We use AdamW [9], with a linearly increasing learning rate for the first 10% of training steps, with a maximum learning rate of 0.005, and a weight decay of 0.01. Cross-Entropy loss is used.

2.3 Tasks

We compare these two models based on their testing accuracies on 13 different single-sentence test classification tasks. We use two tasks from the GLUE benchmark [4], specifically the Corpus of Linguistic Acceptability task (CoLA) [10] and the Stanford Sentiment Treebank task (SST2) [11]. These two tasks are both non-Twitter related, used as a sort of control group for non-Twitter task performance. Then, we use all 11 tasks from the TweetEval benchmark [12].

For most of these tasks, BERTweet's performance has never been tested in a published paper – excluding the irony detection task [13] and the sentiment analysis task [14] from TweetEval, both tested in the original BERTweet paper – thus we are given the additional opportunity to test the BERTweet model on new Twitter NLP tasks. All of these tasks have fixed training, validation, and testing splits, thus we use those to conduct our experiment. We compare these models based on their test accuracies after fine-tuning.

For every task, we tokenize the data with each model's respective tokenizer, and batch the data with a batch size of 32. For every task with both tokenizers, we ensure each row of the tokenized data has a maximum length of 30, as very few rows of data for any task's dataset are longer than 30 words long.

3 Related Works

It has been previously argued that Domain-Adaptive Pre-training (pre-training a model with a different database based on the task-domain) yields better results when applied to the RoBERTa model [15], however a test on the efficacy of Domain-Adaptive Pre-training has never been done with the BERTweet model. Additionally, the paper that presented these results [15] pre-trained with both the general-purpose datasets and the domain-specific datasets, arguably leading to unfair results considering the Domain-Specific models would receive more pre-training time compared to the non-domain-adapted models.

Table 1: Non-Twitter Task Results (GLUE Subset)

Task	RoBERTa$_{base}$	BERTweet$_{base}$
CoLA	**0.765**	0.691
SST2	**0.835**	0.825
Average	**0.800**	0.758

Table 2: Twitter Task Results (TweetEval)

Task	RoBERTa$_{base}$	BERTweet$_{base}$
Emoji	**0.409**	0.370
Emotion	**0.696**	0.627
Hate	**0.567**	0.541
Irony	**0.647**	0.640
Offensive	**0.790**	0.774
Sentiment	**0.656**	0.617
Abortion	**0.678**	0.665
Atheism	**0.782**	0.716
Climate	**0.838**	0.759
Feminist	**0.614**	0.520
Hillary	**0.682**	0.628
Average	**0.669**	0.623

The public BERTweet$_{base}$ models performance has only been represented in two studies: the original paper that presented the model [1], as well a paper published a year later which used the model to classify sentiments of Tweets [16]. The original BERTweet paper did show that BERTweet preformed much better in Twitter-specific tasks compared to RoBERTa, however it did not undergo many tasks that were tested in this paper since it did not use the TweetEval benchmark [12], and, more substantially, it did not test any tasks that were non-Twitter-specific, which is necessary in order to show that it is the pre-training dataset that is causing the improved performance. The lack of training non-Twitter-specific tasks can be thought of as the lack of a control group. The later paper which used BERTweet was even farther away from showing the significance of the pre-training dataset, as instead of comparing BERTweet with its twin model RoBERTa, it compared BERTweet with models which are known to perform worse than RoBERTa, thus making the comparisons obviously unfair.

Other BERT-based domain-specific models have been introduced, including BioBERT [17] and SciBERT [18], however both models are further from the state-of-the-art compared to BERTweet, as they are based on the original BERT model, therefore they don't benefit from the significant benefits that RoBERTa has over the original BERT model [2]. Lastly, the original BERTweet paper drew inspiration for similar Twitter-based BERT models to be released, thus a French language BERTweet was published which seemed to perform better compared to BERT models that weren't pre-trained on Twitter-like datasets [19].

4 Experimental Results

In this section, we briefly observe and interpret the results of our experiment presented in Tables 1 and 2. These tables exhibit the final test accuracies for each task after fine-tuning for each model, with the higher test accuracy between the two models in bold. We review how these results compare to our expectations before conducting the experiments, and how the method could be improved in the future to obtain possibly more meaningful data.

4.1 Non-Twitter Related Tasks

The results from Table 1 are unsurprising and align with our expectations for these tests. Our RoBERTa model was pre-trained on 160GB of English data, while the BERTweet model was pre-trained on 80GB of data from English language tweets. Thus, RoBERTa had not only twice as much data to use for pre-training, but RoBERTa's English language data had much more proper English in it compared to the "bad language" in the Twitter data [5]. Since these non-Twitter tasks dealt with more formal English [4], these would all be advantages the RoBERTa model would have in its favour for these tasks. RoBERTa outperforming BERTweet in these tasks have simple and logical explanations, results like these were completely expected.

4.2 Twitter Related Tasks

The results from Table 2 are very surprising, as they do not align with our expectations, and they present us with evidence which appears to contradict established knowledge about pre-trained language models. As discussed in the previous subsection, the RoBERTa model was pre-trained with twice as much data, however we expected this fact to be counteracted by the fact that the BERTweet model was pre-trained under Twitter data, which is thought to be a great advantage for Twitter related tasks as Domain-Adaptive Pre-training is considered to provide substantial positive results [15].

The original BERTweet paper performed a similar test that we did with the tasks corresponding to the rows Irony and Sentiment in Table 2 [1]. However, it reached opposing results, as in the paper, the BERTweet model performed better than the RoBERTa model. This could possibly be because they had a better performing and a computationally more demanding fine-tuning procedure, however this does not necessarily explain the difference as we used the same fine-tuning procedure for both models in our experiments. Although, improving our fine-tuning procedure would not hurt our results, and it would at least make us more confident in making more meaningful conclusions about our experiments.

5 Discussion

If we are to accept the results from Table 2 as significant, it would imply interesting contradictions to already accepted knowledge in the sphere of pre-trained NLP models. It is widely-accepted that Domain-Adaptive pre-training is useful and helps models perform better on tasks specific to the domain they were trained in [15] [1] [16] [17] [18] [19] [5]. These results call this into question, as our BERTweet model did not perform significantly better in domain-specific tasks compared to other generic tasks.

Additionally, if we are to accept that domain-specific pre-training holds no significant importance, then this calls into question whether companies like Twitter want to use domain-specific models or not. Sentiment analysis is one of the most studied tasks related to Twitter [20], and our results raise doubts on how one would want to conduct experiments predicting sentiments on Twitter.

6 Conclusion

We have furthered the study of pre-trained NLP models, a highly influential sphere of the modern day study of neural networks. Our experiments lead to some unexpected results, whose implications could contradict many previously published papers on this topic. We thus recommend more studies be done on the effectiveness of domain-specific pre-training datasets, so that we can better understand the importance and effect pre-training datasets have on pre-trained NLP models.

References

[1] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020* (Q. Liu and D. Schlangen, eds.), pp. 9–14, Association for Computational Linguistics, 2020.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers) (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[4] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *ICLR (Poster)*, OpenReview.net, 2019.

[5] J. Eisenstein, "What to do about bad language on the internet," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA* (L. Vanderwende, H. D. III, and K. Kirchhoff, eds.), pp. 359–369, The Association for Computational Linguistics, 2013.

[6] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly, 2009.

[7] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, The Association for Computer Linguistics, 2016.

[8] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020* (Q. Liu and D. Schlangen, eds.), pp. 38–45, Association for Computational Linguistics, 2020.

[9] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

[10] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 625–641, 2019.

[11] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1631–1642, ACL, 2013.

[12] F. Barbieri, J. Camacho-Collados, L. E. Anke, and L. Neves, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020* (T. Cohn, Y. He, and Y. Liu, eds.), vol. EMNLP 2020 of *Findings of ACL*, pp. 1644–1650, Association for Computational Linguistics, 2020.

[13] C. V. Hee, E. Lefever, and V. Hoste, "Semeval-2018 task 3: Irony detection in english tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018* (M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, eds.), pp. 39–50, Association for Computational Linguistics, 2018.

[14] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," *CoRR*, vol. abs/1912.00741, 2019.

[15] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, eds.), pp. 8342–8360, Association for Computational Linguistics, 2020.

[16] T. Macrì, F. Murphy, Y. Zou, and Y. Zumbach, "Classifying tweet sentiment using the hidden state and attention matrix of a fine-tuned bertweet model," *CoRR*, vol. abs/2109.14692, 2021.

[17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *CoRR*, vol. abs/1901.08746, 2019.

[18] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), pp. 3613–3618, Association for Computational Linguistics, 2019.

[19] Y. Guo, V. Rennard, C. Xypolopoulos, and M. Vazirgiannis, "Bertweetfr : Domain adaptation of pre-trained language models for french tweets," in *Proceedings of the Seventh Workshop on Noisy User-generated Text, W-NUT 2021, Online, November 11, 2021* (W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, eds.), pp. 445–450, Association for Computational Linguistics, 2021.

[20] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of twitter research: Data model, graph structure, sentiment analysis and attacks," *Expert Syst. Appl.*, vol. 164, p. 114006, 2021.

7 Contributions

As a unified effort, both author's listed contributed equally to all parts of this paper, including everything from research, knowledge, writing, theorizing, coding, and experimenting.