COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter

Martin Müller
Digital Epidemiology Lab
EPFL
Geneva, Switzerland
martin.muller@epfl.ch

Marcel Salathé
Digital Epidemiology Lab
EPFL
Geneva, Switzerland
marcel.salathe@epfl.ch

Per E Kummervold
FISABIO-Public Health
Vaccine Research Department
Valencia, Spain
per@capia.no

May 15, 2020

Abstract

In this work, we release COVID-Twitter-BERT (CT-BERT), a transformer-based model, pretrained on a large corpus of Twitter messages on the topic of COVID-19. Our model shows a 10–30% marginal improvement compared to its base model, BERT-LARGE, on five different classification datasets. The largest improvements are on the target domain. Pretrained transformer models, such as CT-BERT, are trained on a specific target domain and can be used for a wide variety of natural language processing tasks, including classification, question-answering and chatbots. CT-BERT is optimised to be used on COVID-19 content, in particular social media posts from Twitter.

***K*eywords** Natural Language Processing · COVID-19 · Language Model · BERT

1 Introduction

Twitter has been a valuable source of news and a public medium for expression during the COVID-19 pandemic. However, manually classifying, filtering and summarising the large amount of information available on COVID-19 on Twitter is impossible and has also been a challenging task to solve with tools from the field of machine learning and natural language processing (NLP). To improve our understanding of Twitter messages related to COVID-19 content as well as the analysis of this content, we have therefore developed a model called COVID-Twitter-BERT (CT-BERT)[1].

Transformer-based models have changed the landscape of NLP. Models such as BERT, RoBERTa and ALBERT are all based on the same principletraining bi-directional transformer models on huge unlabelled text corpuses [1, 2, 3, 4]. This process is done using methods such as mask language modelling (MLM), next sentence prediction (NSP) and sentence order prediction (SOP). Different models vary slightly in how these methods are applied, but in general, all training is done in a fully unsupervised manner. This process generates a general language model that is then used as input for a supervised finetuning for specific language processing tasks, such as classification, question-answering models, and chatbots.

Our model is based on the BERT-LARGE (English, uncased, whole word masking) model. BERT-LARGE is trained mainly on raw text data from Wikipedia (3.5B words) and a free book corpus (0.8B words) [2]. Whilst this is an impressive amount of text, it still contains little information about any specific subdomain. To improve

[1] https://github.com/digitalepidemiologylab/covid-twitter-bert

performance in subdomains, we have seen numerous transformer-based models trained on specialised corpuses. Some of the most popular ones are BioBERT [5] and SciBERT [6]. These models are trained using the exact same unsupervised training techniques as the main models (MLM/NSP/SOP). They can be trained from scratch, but this requires a very large corpus, so a more common approach is to start with the trained weights from a general model. In this study, this process is called domain-specific pretraining. When trained, such models can be used as replacements for general language models and be trained for downstream tasks.

2 Method

The CT-BERT model is trained on a corpus of 160M tweets about the coronavirus collected through the Crowdbreaks platform [7] during the period from January 12 to April 16, 2020. Crowdbreaks uses the Twitter filter stream API to listen to a set of COVID-19-related keywords[2] in the English language. Prior to training, the original corpus was cleaned for retweet tags. Each tweet was pseudonymised by replacing all Twitter usernames with a common text token. A similar procedure was performed on all URLs to web pages. We also replaced all unicode emoticons with textual ASCII representations (e.g. :thumbs_up: for) using the Python emoji library[3]. In the end, all retweets, duplicates and close duplicates were removed from the dataset, resulting in a final corpus of 22.5M tweets that comprise a total of 0.6B words. The domain-specific pretraining dataset therefore consists of 1/7th the size of what is used for training the main base model. Tweets were treated as individual documents and segmented into sentences using the spaCy library [8].

All input sequences to the BERT models are converted to a set of tokens from a 30 000-word vocabulary. As all Twitter messages are limited to 280 characters, this allows us to reduce the sequence length to 96 tokens, thereby increasing the training batch sizes to 1024 examples. We use a dupe factor of 10 on the dataset, resulting in 285M training examples and 2.5M validation examples. A constant learning rate of 2e-5, as recommended on the official BERT GitHub[4] when doing domain-specific pretraining.

Loss and accuracy was calculated through the pretraining procedure. For every 100 000 training steps, we therefore save a checkpoint and finetune this towards a variety of downstream classification tasks. Distributed training was performed using Tensorflow 2.2 on a TPU v3-8 (128GB of RAM) for 120 h.

2.1 Evaluation

To assess the performance of our model on downstream classification tasks, we selected five independent training sets. Three of them are publicly available datasets, and two are from internal projects not yet published. All datasets consist of Twitter-related data.

2.1.1 COVID-19 Category (CC)

This dataset is a subsample of the data used for training CT-BERT, specifically for the period between January 12 and February 24, 2020. Annotators on Amazon Turk (MTurk) were asked to categorise a given tweet text into either being a personal narrative (33.3%) or news (66.7%). The annotation was performed using the Crowdbreaks platform [7].

2.1.2 Vaccine Sentiment (VS)

This dataset contains a collection of measles- and vaccination-related US-geolocated tweets collected between March 2, 2011 and October 9, 2016. The dataset was first used by Pananos et al. [9], but a modified version from Müller et al. [7] was used here. The dataset contains three classes: positive (towards vaccinations) (51.9%), negative (7.1%) and neutral/others (41.0%). The neutral category was used for tweets which are either irrelevant or ambiguous. Annotation was performed on MTurk.

2.1.3 Maternal Vaccine Stance (MVS)

The dataset is from a so far unpublished project related to the stance towards the use of maternal vaccines. Experts in the field annotated the data into four categories: neutral (41.0%), discouraging (25.3%), promotional (43.9%) and ambiguous (14.3%). Each tweet was annotated threefold, and disagreement amongst the experts was resolved in each case by using a common scoring criterion.

2.1.4 Twitter Sentiment SemEval (SE)

This is an open dataset from SemEval-2016 Task 4: Sentiment Analysis in Twitter [10]. In particular, we used the dataset for subtask A, a dataset annotated fivefold into three categories: negative (15.7%), neutral (45.9%) and positive (38.4%). We make a small adjustment to this dataset by fully anonymising links and usernames.

2.1.5 Stanford Sentiment Treebank 2 (SST-2)

SST-2 is a public dataset consisting of binary sentiment labels, negative (44.3%) and positive (55.7%), within sentences [11]. Sentences were extracted from a dataset of movie reviews [12] and did not originate from Twitter, making SST-2 our only non-Twitter dataset.

The dataset split size is predefined for the SST-2 and SE datasets. For the SST-2 dataset, the test dataset is not released. For the other datasets, we aimed at a split of around 50%-30% between the training and development

[2] wuhan, ncov, coronavirus, covid, sars-cov-2

[3] https://pypi.org/project/emoji/

[4] https://github.com/google-research/bert

Dataset	Classes	Train	Dev	Labels			
COVID-19 Category (CC)	2	3094	1031	Personal	News		
Vaccine Sentiment (VC)	3	5000	3000	N	Neutral	Positive	
Maternal Vaccine Stance (MVS)	4	1361	817	Disc	A	N	Promotional
Stanford Sentiment Treebank 2 (SST-2)	2	67 349	872	Negative	Positive		
Twitter Sentiment SemEval (SE)	3	6000	817	Neg	Neutral	Positive	

Table 1: Overview of the evaluation datasets. All five evaluation datasets are multi-class datasets with sometimes strong label imbalance, visualised by the proportional bar width in the label column. N and Neg stand for negative; Disc and A stand for discouraging and ambiguous, respectively.

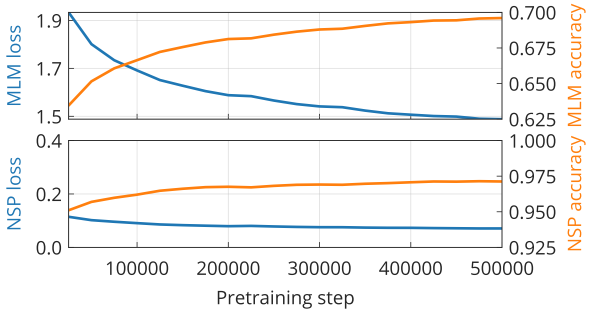

Figure 1: Evaluation metrics for the domain-specific pretraining of CT-BERT. Shown are the loss and accuracy of masked language modelling (MLM) and next sentence prediction (NSP) tasks.

sets, leaving a test set of 20% which was not used in this work. Our intention was not to optimise the finetuned models but to thoroughly evaluate the performance of the domain-specific CT-BERT-model. We experimented with different numbers of epochs for each training dataset for BERT-Large (i.e. checkpoint 0 of CT-BERT) and selected the optimal one. We then used this number in subsequent experiments on the respective dataset. We ended with three epochs for SST-2, CC and SE, five epochs for VC and 10 epochs for MVC, all with a learning rate of 2e-05. The number of epochs was dependent on both the size and balance of the categories. Larger and unbalanced sets require more epochs.

3 Results

3.1 Domain-sepcific pretraining

Figure 1 shows the progress of pretraining CT-BERT at intervals of 25k training steps and the evaluation of 1k steps on a held-out validation dataset. All metrics considered improve throughout the training process. The improvement on the MLM loss task is most notable and yields a final value of 1.48. The NSP task improves only marginally, as it already performs very well initially. Training was stopped at 500 000, an equivalent of 512M training examples, which we consider as our final model. This corresponds to roughly 1.8 training epochs. All metrics for the MLM and NLM tasks improve steadily throughout training. However, using loss/metrics for these tasks to evaluate the correct time to stop training is difficult.

3.2 Evaluation on classification datasets

To assess the performance of our model properly, we compared the mean F1 score of CT-BERT with that of BERT-Large on five different classification datasets. We adapted the number of training epochs for each dataset according to its size in order to have a similar number of training steps for each dataset. Our final model shows higher performance on all datasets (a mean F1 score of 0.833) compared with BERT-Large (a mean F1

Dataset	BERT-LARGE	CT-BERT	ΔMP
COVID-19 Category (CC)	0.931	0.949	25.88%
Vaccine Sentiment (VC)	0.824	0.869	25.27%
Maternal Vaccine Stance (MVS)	0.696	0.748	17.07%
Stanford Sentiment Treebank 2 (SST-2)	0.937	0.944	10.67%
Twitter Sentiment SemEval (SE)	0.620	0.654	8.97%
Average	0.802	0.833	17.57%

Table 2: Comparison of the final model performance with BERT-LARGE. CT-BERT shows improvements on all datasets. The marginal improvement is the highest on the COVID-19-related dataset (CC) and lowest on the SST-2 and SemEval datasets.

score of 0.802). As the initial performance varies widely across datasets, we compute the relative improvement in marginal performance (ΔMP) for each dataset. ΔMP is calculated as follows:

$$\Delta\text{MP} = \frac{F_{1,\,\text{BERT-LARGE}} - F_{1,\,\text{CT-BERT}}}{1 - F_{1,\,\text{BERT-LARGE}}}$$

From this metric, we can observe the largest improvement of our model on the COVID-19-specific dataset (CC), with a ΔMP value of 25.88%. The marginal improvement is also high on the Twitter datasets related to vaccine sentiment (MVS). Our model likewise shows some improvements on the SST-2 and SemEval datasets, but to a smaller extent.

3.3 Evaluation on intermediary pretraining checkpoints

So far, we have seen improvements in the final CT-BERT model on all evaluated datasets. To understand whether the observed decrease in loss during pretraining linearly translates into performance on downstream classification tasks, we evaluated CT-BERT on five intermediary versions (checkpoints) of the model and on the zero checkpoint, which corresponds to the original BERT-LARGE model. At each intermediary checkpoint, 10 repeated training runs (finetunings) for each of the five datasets were performed, and the mean F1 score was recorded. Figure 2 shows the marginal performance increase (ΔMP) at specific pretraining steps. Our experiments show that downstream performance increases fast up to step 200k in the pretraining and only demonstrates marginal improvement afterwards. The loss curve, on the other hand, shows a gradual increase even after step 200k. We also note that for the COVID-19-related dataset, most of the marginal improvement occurred after 100k pretraining steps. SST-2, the only non-Twitter dataset, improves much more slowly and reaches its final performance only after 200k pretraining steps.

Amongst runs on the same model and dataset, some degree of variance in performance was observed. This variance is mostly driven by runs with a particularly low performance. We observe that the variance is dataset dependent, but it does not increase throughout different pretraining checkpoints and is comparable to the variance observed on BERT-LARGE (pretraining step zero). The most stable training seems to be on the SemEval training set, and the least stable one is on SST-2, but most of this difference is within the error margins.

4 Discussion

The most accurate way to evaluate the performance of a domain-specific model is to apply it on specific downstream tasks. CT-BERT is evaluated on five different Twitter-based datasets. Compared to BERT-LARGE, it improves significantly on all datasets. However, the improvement is largest in datasets related to health, particularly in datasets related to COVID-19. We therefore expect CT-BERT to perform similarly well on other classification problems on COVID-19-related data sources, but particularly on text derived from social media platforms.

Whilst it is expected that the benefit of using CT-BERT instead of BERT-LARGE is greatest when working with Twitter COVID-19 text, it is reasonable to expect some performance gains even when working with general Twitter messages (SemEval dataset) or with a non-Twitter dataset (SST-2).

Our results show that the MLM and NSP metrics during the pretraining align to some degree with downstream performance on classification tasks. However, compared with COVID-19 or health-related content, out-of-domain text might require longer pretraining to achieve a similar performance boost.

Whilst we have observed an improvement in performance on classification tasks, we did not test our model on other natural language understanding tasks. Furthermore, at the time of this papers writing, we only had access to one COVID-19-related dataset. The general performance

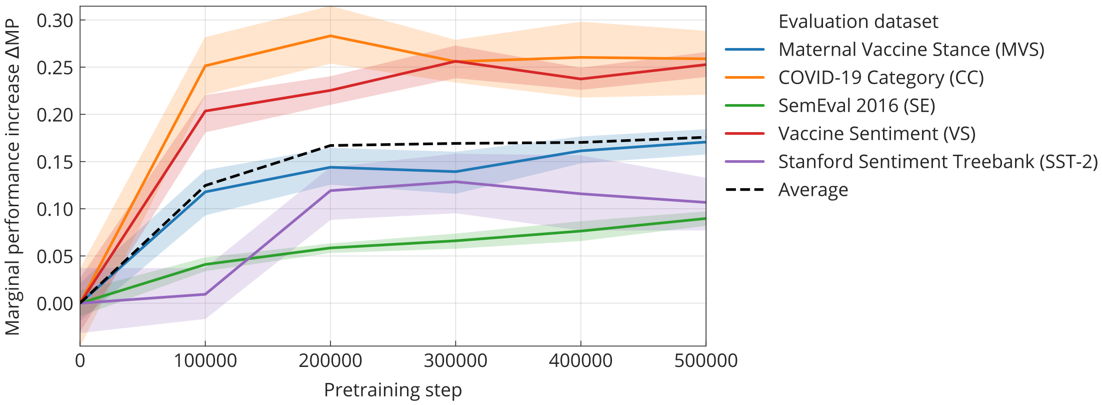

Figure 2: Marginal performance increase in the F1 score (ΔMP) on finetuning on various classification tasks at increasing steps of pretraining. Zero on the x-axis corresponds to the base model, which is BERT-LARGE in this case. Our model improves on all evaluated datasets, with the biggest relative improvement being in the COVID-19 category dataset. The bands show the standard error of the mean (SEM) out of 10 repeats.

of our model might be improved further by considering pretraining under different hyperparameters, particularly modifications to the learning rate schedules, training batch sizes and optimisers. Future work might include evaluation on other datasets and the inclusion of more recent training data.

The best way to evaluate pretrained transformer models is to finetune them on downstream tasks. Finetuning a classifier on a pre-trained model is considered computationally cheap. The training time is usually done in an hour or two on a GPU. Using this method for evaluation is more expensive, as it requires evaluating multiple checkpoints to monitor improvement and on several varied datasets to show robustness. As finetuning results vary between each run, each experiment must be performed multiple times when the goal is to study the pretrained model. In this case, we repeated the training for six checkpoints, 10 runs for each checkpoint on all the five datasets. A total of 300 evaluation runs were performed. The computational cost for evaluation is therefore on par with the pretraining. Large and reliable training and validation sets make this task easier, as the number of repetitions can be reduced.

All the tests are done on categorisation tasks, as this task is easier in terms of both data access and evaluation. However, transformer-based models can be used for a wide range of tasks, such as named entity recognition and question answering. It is expected that CT-BERT can also be used for these kinds of tasks within our target domain.

Our primary goal in this work was to obtain stable results on the finetuning in order to evaluate the pre-trained model, not to necessarily optimise the finetuning. The number of finetuning epochs and the learning rate are, for instance, have been optimised for BERT-LARGE, not for CT-BERT. This means that there is still great room for optimisation on the downstream task.

5 Data Availability

The model, code and public datasets are available in our GitHub repository: `https://github.com/digitalepidemiologylab/covid-twitter-bert`.

6 Funding

PK received funding from the European Commission for the call H2020-MSCA-IF-2017 and the funding scheme MSCA-IF-EF-ST for the VACMA project (grant agreement ID: 797876).

MM and MS received funding through the Versatile Emerging infectious disease Observatory grant as a part of the European Commissions Horizon 2020 framework programme (grant agreement ID: 874735).

The research was supported with Cloud TPUs from Google's TensorFlow Research Cloud and Google Cloud credits in the context of COVID-19-related research.

7 Conflicts of Interest

The authors have no conflicts of interest to declare.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages

5998–6008, 2017.

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

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[4] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[5] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[6] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[7] Martin M Müller and Marcel Salathé. Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *Frontiers in public health*, 7, 2019.

[8] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1), 2017.

[9] A Demetri Pananos, Thomas M Bury, Clara Wang, Justin Schonfeld, Sharada P Mohanty, Brendan Nyhan, Marcel Salathé, and Chris T Bauch. Critical dynamics in population vaccinating behavior. *Proceedings of the National Academy of Sciences*, 114(52):13762–13767, 2017.

[10] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.01973*, 2019.

[11] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[12] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.