# Transfer Learning BERT for Analysis of Twitter Sentiment Towards Contact Tracing During COVID-19

**Christina Knight**
Stanford University
cqknight@stanford.edu

**Coen Armstrong**
Stanford University
crrarmstrong@stanford.edu

## Abstract

During the COVID-19 pandemic, governments have often made subjective decisions about basic social trade-offs, such as privacy versus public health. These decisions are made too quickly to be subject to elections, and often are not covered by political polling; sentiment analysis on Twitter could be one way to check public opinion about these decisions. We train a transfer-learning model off BERT to classify public opinion as positive versus negative. We achieve 69% accuracy, 18% higher than a BERT implementation without transfer learning as a baseline.

## 1 Introduction

Nations placed varying emphasis on values such as individual freedom and privacy when determining government COVID-19 responses. We want to see how public opinion on Twitter reacted. We are building a fine-tuned BERT-based model to analyse the sentiment of contact-tracing related tweets.

Inferring public opinion is a vital check on governments, given recent transformations in the basic moral trade-offs underlying their policies. Measuring restrictions' popularity also indicates likely compliance, helping governments pick policies or forecast cases. Hence, recent publications such as O'Callaghan et al., in the Irish Journal of Medical Science, have studied public opinion towards COVID-19 contact tracing apps through the lens of survey evidence [1].

Our sentiment analysis complements other barometers of public opinion, like surveys. Sentiment analysis is much cheaper, so we can ask more specific questions, generate time series data, and iterate. It also targets an audience, with often no house phone, that some polls miss: polls over-sample the old, who may have very different attitudes towards COVID because of their different political positions and risk profiles. Additionally, some countries do not have strong traditions of polling, partly because governments might want to control public opinion, so it can be complemented with evidence on government restrictions.

Singapore and the United States took polar approaches to the use of surveillance during COVID-19. While parts of the United States were famously libertarian, eschewing mandatory contact-tracing apps, the Singapore government implemented harsh restrictions and government oversight. The government mandated contact-tracing apps very early in the pandemic, appointed "social distancing ambassadors" to take pictures of rulebreakers and report them to the government, and had a large bank of volunteers and army soldiers to pre-emptively quarantine possible cases and monitor quarantined people. According to survey data 49.2% of Singaporeans support the government using people's phones to track their movements without their consent [2].

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

By using Twitter to compare public sentiment towards government surveillance in the United States and Singapore, we aim to further the conversation surrounding "ethical"disease surveillance and to what extent citizens' condone government surveillance when used for the public good.

The input to our algorithm is a tweet and a sentiment label (0 for negative and 1 for positive sentiment. So it is {tweet, sentiment}. We then use transfer learning off BERT, Google's bidirectional transformer-based model, to output a predicted label of 0 or 1. We normally have one or two hidden layers that we train on our own on top of the weights from BERT.

## 2 Related work

Research during the COVID-19 pandemic attempts to analyze responses, trends, and behaviour and form epistemological models. Papers [9], [12], [1], and [10] all model public sentiment and discourse using different techniques such as CNN, LSTM and LDA. In [7], Nemes and Kiss use an RNN to analyze Twitter users' sentiment towards COVID-19 based on keyword trends. These approaches are accurate first attempts but often do not capture the complex structure of meaning. Accordingly, Sosa's combined CNN and LSTM approach is especially impressive, combining the advantages of both network architectures, but invovles a very large number of hyperparameters to tune off a fairly limited dataset. Hence in our approach we favour transfer learning off BERT, since it optimises for our fairly limited labelled datasets.

As mentioned in our introduction, sentiment towards contact tracing is also a topic of research in other related fields, like political science and medicine, such as in [8].

Naseem *et al.* achieve very high classification accuracy in their excellent paper but their dataset has high class imbalance: 67385/90000 tweets are neutral, so a classifier that just returned 'neutral' the whole time would achieve 75% accuracy already. Their results are similar to ours: most of their supervised techniques see accuracies in this 75% to 80% range, while their best BERT-based techniques see a 15% improvement from this 75% benchmark of picking neutral the whole time (while ours is 18% above the analogous benchmark). Finally, there are only 6300 positive examples, reflecting the underlying distribution of tweets where most tweets about COVID-19 are negative for obvious reasons. Performance on positive examples is severely de-prioritised as less than 10% of their dataset—which is a good scientific decision for classifying COVID related tweets in general. However, on our specific problem more tweets are likely to be positive—we could even have a situation where most tweets are positive (because they want certain contact tracing measures), so we would rather not rely on the inherent data distributions in twitter data to recur. Hence our model needs to have high accuracy but also good F1 scores on both positive and negative sentiment.

Since we are interested in governmental decision making, we are most interested in comparing negative and positive examples rather than just neutral ones: in a voting system, neutral examples do not point in either direction and thus cannot inform any decision making. Thus we approach this problems in terms of 2-class classification, while the other papers under discussion often have a substantial bias toward neutral samples. It is a general problem that most tweets are easily classed as neutral, but this distracts the focus in training on what we most care about: the ratio of positives to negatives. Hence we pose our problem as 2-class classification.

## 3 Dataset and Features

We have two different datasets for our project: NLP_Corona_Train/Test and COVIDSenti.

We used NLP_Corona_Train/Test for our first run-through of fine-tuning BERT. [3] We pre-processed this data by altering the labels from words to integers (i.e. changing negative to 0 and positive to 1) and by arranging it into the form (text of tweet, sentiment), where text of tweet is a string and sentiment is an integer. For our initial run-through, we simplified this data labelling scheme by transferring it from a 5 class labelling scheme (i.e. extremely negative, negative, neutral, positive, extremely positive) to a two-class numerical scheme (negative and positive). We have 33444 tweets, 54% positive. This was partly for initial ease of access and partly because our problem of inferring public opinion centrally depends on positive and negative examples rather than neutral ones. The other reason to use only 1 negative class and 1 positive class is so that it can be combined with COVIDSenti. We also truncated tweets at 45 characters, as is standard, in order to save memory and

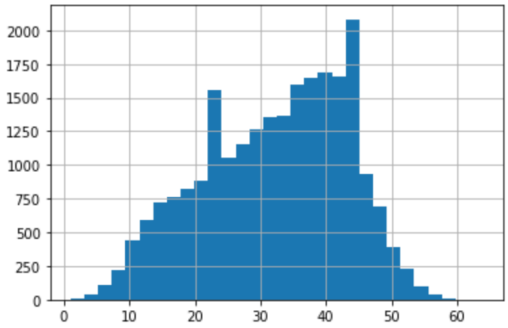


Figure 1: Data Distribution in initial Dataset

run more iterations: since when plotting a histogram of our data distribution, the vast majority of tweets were totally captured. This saves us using too much computational power on blank spaces:

The second dataset, COVIDSenti, we acquired from Dr. Razzak, the professor that conducted the study, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis" [6]. We have combined this dataset with NLP_Corona_Train, for a total dataset of 56059 positive/negative tweets, 31732 of which are positive. The tweet length of this new dataset skews smaller but there are still enough tweets in the 40-45 category on the histogram, purely coming from the first dataset, that we keep max length at 45. To resolve the slight imbalance in the data (it's 56.6% positive) we weight each class proportionately.

## 4 Methods

We use transfer learning off Bidirectional Encoder Representations from Transformers (BERT), which was created and published by Google in 2018. This state-of-the-art Machine Learning model is most commonly used for NLP tasks because it uniquely applies the bi-directional aspect of Transformers to language modelling tasks. Since we had limited data available, we froze the majority of BERT's layers and added the architecture below.

The model architecture we used has (1) a fully connected dense layer of 512 hidden units, (2) a dropout layer, (3) a relu layer, (4) another fully connected dense layer of 512 hidden units, and finally (5) a softmax output layer. Sometimes we experiment with more hidden units (1024), or with adding another hidden layer. The basic logic of this transfer learning paradigm is that the BERT model will learn general features of the NLP problem through the corpus it has already, and those early weights are frozen so the general features are not corrupted. Our model uses an AdamW optimizer and the negative log likelihood loss.

Very briefly, the Adam optimisation algorithm combines both the momentum and RMSprop optimization techniques. We used this sophisticated optimization technique because it is often used in NLP tasks.

The negative log likelihood (NLL) loss is:

$$\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

.

This is useful for classification problems like ours, as the interpretation of NLL in terms of cross-entropy shows—it has a natural information theoretic meaning. This also explains why we need a softmax output layer in our model architecture.

## 5 Experiments/Results/Discussion

As a baseline we compare with a simple BERT implementation with no transfer learning. This has the following confusion matrix performance:

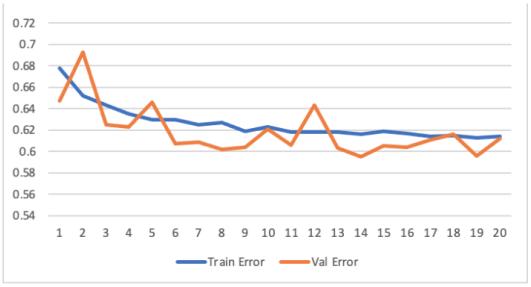


Figure 2: Negative Log Likelihood Loss (x axis) vs. Epochs (y axis); Train (green) and Validation (orange) Error for 20 Epochs, Batch Size 16

| | 0 | 1 |
|---|---|---|
| 1 | 1135 | 1175 |
| 0 | 1261 | 1446 |

Since BERT is already an extremely sophisticated NLP model, the fact that it has F1-score of 0.48 on the 0 class and 0.54 on the 1 class shows that the NLP_Corona_Train/Test dataset is very difficult to perform well on. This was only 51% accuracy. Our primary metric in this will be accuracy.

We took our initial code from Prateek Joshi's github repository, "Fine Tuning BERT for Spam Classification" [4]. Then, we modified this code to be able to use our first dataset, NLP_Corona_Train (with the pre-processing steps mentioned above), to fine tune BERT. We also modified to streamline memory management. We initially set it to run for 10 epochs with a batch size of 16 and learning rate of 1e3.

On our initial training run-throughs (with the simplified NLP_Corona_Train modified dataset), but our transfer learning technique, we reached 64% classification accuracy. Then, with preliminary hyperparameter tuning, we increased our accuracy to 66% by changing sequence length to better reflect the structure of the data (because for memory management reasons it is conventional to cut off all sequences at some certain length). By decreasing batch size to 8, doubling the number of epochs to 20, and further preprocessing of the data we improved to 69% classification accuracy (picture below). Strangely, the validation error (drawn from the same distribution) was consistently lower than the training error. We hypothesised this was because we applied regularisation in training but not in testing: indeed, if we set the dropout probability to 0, validation error is above training error (though some epochs are still exceptions); this decreases accuracy to 68% though. One reason our results are relatively inaccurate could be that our data is slightly imbalanced (ratio of 46-53 (pos-neg)).

We recovered 69% accuracy much faster through more careful preprocessing of data: by manual error analysis we realised the dataset needed to be shuffled. With a batch size of 16 on a reshuffled dataset, after 10 epochs we had the following precision and recall:

| | Precision | Recall |
|---|---|---|
| 1 | 0.65 | 0.71 |
| 0 | 0.73 | 0.67 |

Hence F1 scores of 0.68 for 0 and 0.70 for 1. Hence our model actually overstates negative emotions, since our confusion matrix is:

| | 0 | 1 |
|---|---|---|
| 1 | 1649 | 661 |
| 0 | 896 | 1811 |

So on the training set it overstates the number of 0-class examples by 200. This is reasonable because it is a consistent legal principle to err on the side of caution in infringing upon liberty.

Because the high training set error and low gap between training set error and validation error implies a bias issue, we increased the size of our layers to 1024 hidden units. We know that peak human performance is close to 100% on this task, by manual analysis. However, performance was almost identical with a F1 score of 0.68, though underlying this the model had much higher recall on class 0,

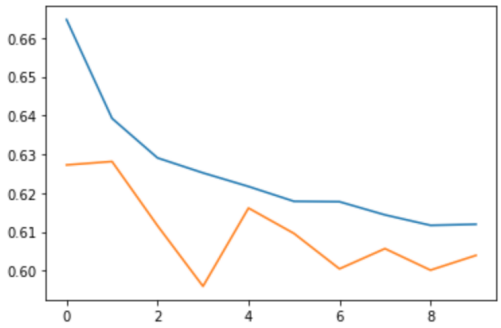


Figure 3: Negative Log Likelihood Loss (x axis) vs. Epochs (y axis); Train (green) and Validation (orange) Error for 10 Epochs, Batch Size 16, on the Amalgamated Dataset

higher precision on class 1, and lower scores on the other two components. With an extra hidden layer, F1 score decreases to 0.65 suggesting an overfitting problem, although we use a validation set and also have a dropout layer to mitigate that.

Tuning learning rate to 3e-3 rather than 1e-3 led to catastrophic accuracy loss, as displayed here:

| | 0 | 1 |
|---|---|---|
| 1 | 2128 | 182 |
| 0 | 2523 | 184 |

This is a recall of 0.92 for the 0 class and 0.07 for the 1 class. To check whether tuning the Adam optimizer learning rate helps, we also used a learning rate 3e-4, which gives accuracy of 68%. Hence having checked the order of magnitude around 1e-3, we are reasonably confident that this is a good learning rate for the problem.

This detailed hyperparameter investigation suggests that the fundamental road-block to getting past 69%, which we have reached through many different hyperparameter combinations, is the data. This is confirmed by manual error analysis which finds a few errors. Hence we use the amalgamated dataset; however, this yet again has only 69% accuracy and F1 scores of 66% and 71%, this time with the following matrix:

| | Precision | Recall |
|---|---|---|
| 1 | 0.75 | 0.68 |
| 0 | 0.63 | 0.70 |

See Figure 3 for a plot from training, which also shows the effect of the dropout on relative validation and training errors.

## 6 Conclusion/Future Work

Transfer learning off BERT performed best, coupled with careful preprocessing of data to achieve better results. It reached 69% accuracy, whereas BERT as a baseline was only 51% accurate. This is because twitter sentiment analysis is a difficult problem, as COVID related tweets are often bitterly sarcastic, or refer to other tweets and internet tropes in fairly inscrutable ways: it is by no means obvious how to analyse them.

There are three major future directions. First, we could build this into a full end-to-end system, that takes tweets from the twitter API under certain hashtags like #contacttracing, and returns a real-time result. Second, we can explore implementations other than BERT that are nevertheless capable of handling these complex representations; there have been good recent results on Twitter data using hybrid CNN biLSTM models, for example. Third, the essential limitation seems to be having enough data, and we could try to label much more data, or work on weakly supervised methods.

## 7 Contributions

Both team members contributed equally to all aspects of the project.

## References


[1] Chakraborty, Koyel et al. “Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media.” (September 2020): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7521435/.

[2] Chen, Nan, and Wang, Peikang. “Advanced Combined LSTM-CNN model for Twitter Sentiment Analysis”. Proceedings of CCIS2018.

[3] Chok, Lazarus. “The Policy Black Box in Singapore’s Digital Contact Tracing Strategy”, LSE Southeast Asia Blog. (September 2020). https://blogs.lse.ac.uk/seac/2020/09/22/the-policy-black-box-in-singapores-digital-contact-tracing-strategy/

[4] Dong et al. “Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification”. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), (June 2014) pp. 49–54.

[5] Joshi, Prateek, “Fine Tuning BERT for Spam Classification.” GitHub, (July, 2020): https://github.com/prateekjoshi565/Fine-Tuning-BERT/blob/master/Fine_Tuning_BERT_for_Spam_Classification.ipynb

[6] Naseem, Usman et al. "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis.” IEEE Transactions on Computational Social Systems, (January 2021): https://ieeexplore.ieee.org/abstract/document/9340540.

[7] Nemes, Laszlo Kiss, Attila. “Social Media Sentiment Analysis Based on COVID-19.” Taylor Francis Online, (July 2020): https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1790793.

[8] O’Callaghan et al. A national survey of attitudes to COVID-19 digital contact tracing in the Republic of Ireland. Irish Journal of Medical Science (October 2020): https://pubmed.ncbi.nlm.nih.gov/33063226/

[9] Sakun Boon-Itt, Skunkan, Yukolpat. “Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study.” JMIR Publications: Advancing Digital Health Open Science, (November 2020): https://publichealth.jmir.org/2020/4/e21978.

[10] Sanders, Abraham et al. “Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse.” medRxiv: The Preprint Server for Health Sciences, (March 2021): https://www.medrxiv.org/content/10.1101/2020.08.28.20183863v3.

[11] Sosa, Pedro. “Twitter Sentiment Analysis using combined CNN/LSTM models”. (June 2017)

[12] Xue, Jia et al. “Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter.” Plos One Journal, (September, 2020): https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0239441.