Twitter Sentiment Analysis: Global Attitudes Towards COVID-19 Policies

Stanford CS224N Custom Project

TA Mentor: Anna Goldie

Christina Knight
Department of Symbolic Systems
Stanford University
cqknight@stanford.edu

Chiara Biondi
Department of Mathematical Computational Science
Stanford University
cbiondi@stanford.edu

Elyse Cornwall
Department of Computer Science
Stanford University
cornwall@stanford.edu

Abstract

This project aims to capture widespread sentiment about COVID-19 and COVID-19 related policies on Twitter—the world's most popular text-based social media platform. To accomplish this task, we fine-tuned a pretrained BERT model: Twitter-RoBERTa-Base-Sentiment [1] on COVID-19-related Tweets labelled for 3 sentiment classes. After training, this model achieves a test accuracy of 65.47%, out-performing the baseline Twitter-RoBERTa-Base-Sentiment by 19.47%. Then, we applied our model to classify Twitter sentiment globally in English speaking nations (United Kingdom, United States, and Canada) related to two of the most prevalent COVID-19-related policies: digital contact tracing (DCT) and vaccine mandates. Our results show that, overall, people viewed digital contact tracing more favorably than vaccine mandates, and that both DCT and vaccines are more supported by Twitter users in the United Kingdom than those in the United States.

1 Introduction

As the COVID-19 pandemic has evolved, countries around the world have instated different national policies to contain the spread of the disease among citizens. Among these strategies are vaccine mandates and digital contact tracing, both of which impose some restriction and/or surveillance on citizens. Socially, these policies reveal a balance between individual freedom and sacrifice with the intent of benefiting the greater good, which we explore in the present paper. While data like case counts and vaccination rates address the quantitative side of the pandemic, there is a qualitative, experiential aspect of COVID-19 that is harder to report through these population-level metrics. The present paper attempts to answer the more nuanced question, "How do people feel about COVID-19 and the policies instantiated to prevent its spread?" which current data fails to comprehensively tackle. Presenting this problem statement in the form of a classic NLP task, we wish to classify the sentiment of Tweets related to COVID-19 policies.

Our driving mission was to draw from the opinions of individuals as expressed in Tweets from throughout the COVID-19 pandemic to fine-tune a model to excel at classifying the sentiment of such Tweets. We decided to use a Bidirectional Encoder Representations from Transformers (BERT) model as our baseline, because such models are pre-trained with massive amounts of data, often specializing in some domain that allows them to outperform more general models on NLP tasks in that domain. Applying this to our problem, we sought a baseline already pre-trained to handle our data: Tweet text.

After experimenting with a few different models, we landed on Twitter-RoBERTa-Base-Sentiment [1], because it showed the highest accuracy in predicting the sentiment of Tweets about COVID-19 vaccines.

We hypothesized that by fine-tuning a Twitter BERT model on a sentiment analysis task with COVID-19 Tweet data, we would produce a model that would have improved performance over the baseline when analyzing the sentiment of COVID-19 policy related Tweets. We then used our model to investigate trends in sentiment about different COVID-19 policies, namely vaccine mandates and contact tracing, by filtering unlabeled COVID-19 Tweet datasets by relevant keywords. Additionally, we evaluated the sentiment of COVID-19 Tweets from datasets collected in different nations—the United States, United Kingdom, and Canada—to see how cultural norms, government structures, and value systems change ideological trends. Our results answered our research question by representing how the public felt about their government's COVID-19 policies through sentiment scores, with some measure of confidence given by the accuracy reported by our model (65.47%). We intend for this model to contribute to the growing sphere of COVID-19 data by providing qualitative metrics about the public's reception of pandemic-era protocols.

2 Related Work

2.1 Generalized BERT Models

Transformer-based models like BERT [2] provide general language models which can be applied to a variety of NLP tasks. Because they are pre-trained on unsupervised tasks with billions of words, they succeed at an array of NLP applications. Importantly, BERT is created using bidirectional pre-training, which is fundamentally more powerful than left-to-right or combined left-to-right and right-to-left models for our task, and BERT's success is attributed to this property [2]. Models like BERT [2] and RoBERTa [3] often serve as the basis for models trained more extensively in a domain of interest, for example biology or science in BioBERT and SciBERT models respectively [2, 3, 4, 5].

2.2 Pre-trained Domain-Specific BERT Models

This paper adds to a growing number of domain-specific pre-trained BERT models, and domain-specific pre-trained transformers in general. Existing pre-trained BERT have offered notable performance improvements in their target domains, for example, BioBERT outperforms the baseline BERT model, achieving a higher F1 score than BERT in biomedical contexts as shown in its introductory paper [4]. COVID-Twitter-BERT contributes a model that outperforms the baseline in a new domain, and promises to be valuable tool to analyze COVID-19 data specifically, as the pandemic progresses [6].

The model showed non-negligible marginal performance improvements in classification tasks on test datasets that were not COVID-19 focused. Namely, the model's improvement over the base $BERT_{LARGE}$ was 25.27% for a vaccine sentiment dataset, 17.07% for a maternal vaccine stance dataset, and 10.67% for a general Tweet dataset, and 8.97% for a generic sentiment bank not originating from Twitter [6]. In comparison to the model's 25.88% marginal performance improvement on a COVID-19 dataset, these values suggest that COVID-Twitter-BERT can be more broadly applicable to health and medical related NLP tasks, as well as tasks performed on Tweet-based datasets.

The present paper aims to make a similar contribution in its domain: sentiment classification of COVID-related Tweet data. COVID-Twitter-BERT provides a promising example of a baseline BERT model improved with additional training data, and this paper adopts a similar methodology to address a new problem.

3 Approach

3.1 Baseline Model

To determine which BERT model to use as a baseline upon which to perform pre-training, we conducted a small experiment on a set of BERT models and compared their performances. We selected three BERT models from Hugging Face, the Twitter-RoBERTa-Base-Sentiment model [1],

the Sentiment-RoBERTa-Large-English-3-Classes model [7], and the BERTweet-Base-Sentiment-Analysis Model [8], to evaluate on a sentiment analysis task for COVID-19 vaccine sentiment data [9]. Based on the resulting accuracy of these models in classifying sentiment, we chose to proceed with Twitter-RoBERTa-Base-Sentiment model as our baseline model for the final project. Our baseline model, Twitter-RoBERTa-Base-Sentiment, has 12 hidden layers and was trained on approximately 58 million Tweets. As shown below, Twitter-RoBERTa-Base-Sentiment outperformed the other models on COVID-19 related Tweets, and therefore we chose to continue the project with Twitter-RoBERTa-Base-Sentiment as our baseline.

Model	Accuracy
Twitter-RoBERTa-Base-Sentiment	46.09%
Sentiment-RoBERTa-large	38.28%
BERTweet-Base-Sentiment	36.72%

Figure 1: Initial Performance of Baseline BERT Models

3.2 Fine-Tuning Approach

This paper improves the baseline Twitter-RoBERTa-Base-Sentiment-Analysis model [3] by fine-tuning on COVID-19 Tweets labeled for sentiment analysis. We experimented with different data collection and distribution methods, batch sizes, number of epochs, and optimizer hyper-parameters. After formatting our COVID-19 Tweet data so that each example contained two fields - TweetContext (the text content of each Tweet) and Sentiment (the sentiment score) - we split the examples into a train, evaluation, and test set using a 70-15-15 ratio respectively. Our best fine-tuned model (with the highest reported accuracy) used data from both the vaccine sentiment dataset [9] and a general COVID-19 dataset [10], had 3 epochs, a batch size of 16, and used an AdamW optimizer with a learning rate of $5e^{-5} = 0.0337$.

3.2.1 Fine-Tune Data Selection

To select our fine-tuning dataset, we experimented with datasets of different size, content, and distribution of sentiment. Ultimately, we selected examples from a COVID-19 vaccine sentiment dataset [9] of 1192 examples and a more general set of 3170 Tweets about COVID-19 [10] (4362 total Tweets).

We compared how training and evaluating more domain-focused sets of data affected our model's accuracy. For instance, we compared our model's accuracy when trained on the vaccine dataset to a combined dataset of these vaccine Tweets with more general COVID-19 data. We selected a subset of 160 Tweets from each of these datasets to perform our experiment. After training, the vaccine model had an dev accuracy of 45.81% accuracy, while the vaccine plus general COVID-19 data model had an accuracy of 58.32% (both on small datasets without hyperparameter fine-tuning). Therefore, we used a combination of both datasets.

Finally, we observed that our datasets were skewed towards neutral sentiment. To test whether this would cause our model to predict neutral sentiment disproportionately based on its frequency in the training data, we hand-selected a subset of our vaccine dataset by removing some neutral examples to create a more even distribution. We compared this model's performance to that of the overall fine-tuned model, and found no performance improvement.

3.2.2 Fine-Tune Parameters

After creating a combined COVID-19 dataset [10] to fine-tune our model, we adjusted training parameters to achieve a higher accuracy.

We experimented with learning rate to fine-tune our model, comparing a learning rate of $2e^{-5}$ and $5e^{-5}$ when training the baseline on our vaccine dataset. We saw improved accuracy with the smaller learning rate, namely 0.7400 (74.00% accuracy) for $2e^{-5}$ learning rate and 0.5832 (58.32% accuracy) for $5e^{-5}$ learning rate, which corresponded with our knowledge that smaller learning rates improve performance when training on smaller datasets. Since we were not sure if this learning

rate trend would extrapolate to a larger training dataset, we trained two full models with the other hyper-parameters fixed and different AdamW optimizer learning rates.

Additionally, we experimented with the number of epochs we trained for, because we noted that our model's loss and accuracy tended to decrease after the third epoch of training. Because our model output the loss and accuracy after each epoch, we used this experimental data to conclude that 3 training epochs was ideal for our use case.

We also experimented with batch size and settled on the standard amount of 16 examples per batch.

3.3 Sentiment Predictions on Real-World Data

After training our baseline model on the full COVID-19 dataset [10] (with no adjustments of sentiment score distribution) with a batch size of 16 for 3 epochs, using a learning rate of $5e^{-5}$ for the AdamW optimizer (our final model with the best accuracy), we applied this fine-tuned model to make predictions on real-word, unlabeled data. We selected unlabeled Tweet data [11] from three dates over the course of the pandemic to capture sentiment throughout this global crisis from May 2020 to January 2022. We then split our data into groups of three countries where English is the primary language (Canada, the United States, and the United Kingdom), and used keywords to identify Tweets about different COVID-19 policies. Our model outputs its predictions on this real-world data as a vector of three scores for each example, representing its confidence in the Tweet's sentiment being negative, neutral, or positive. We took the argmax of this vector to determine the model's prediction for a given Tweet. This allowed us to compute the total number of positive, neutral, and negative Tweets for each segmentation of data, and examine how these totals changed over time, between different countries, and how sentiment differed by policy.

4 Experiments

4.1 Data

4.1.1 Training Data

As stated in the fine-tuning section, we combined two COVID-19 Twitter datasets to create our training repository. More details about training data collection beyond what is included above can be found in the appendix.

4.1.2 Real World Data

The unlabeled data we classified for sentiment analysis came from a dataset of 1.2 billion Tweets collected from May 2020 to January 2022 (at approx. 6 month intervals) related to the COVID-19 pandemic [11]. Then, we filtered by country code to isolate Tweets from the United Kingdom, United States, and Canada. Finally, we did keyword filtering on both of the date datasets to isolate Tweets related to vaccines[1] and digital contact tracing[2]. These datasets allowed us to explore how COVID-19 sentiment changed from May 2020 to January 2022, how citizens of different English-speaking countries compare in their sentiments about COVID-19, and how sentiment on vaccines and digital contact tracing differ.

4.2 Evaluation method

Our primary evaluation metric throughout the paper is accuracy, which we computed as:

$$\texttt{accuracy} = \frac{\texttt{correct predictions}}{\texttt{total predictions}}. \tag{1}$$

We used this metric to determine which baseline BERT model to use and our fine-tuned hyperparameters and baseline. After selecting a combined dataset of vaccine and COVID-19 data to fine-tune our

[1] Vaccine keywords: "vaccine" "inoculation" "covid shot" "pfizer" "moderna" "johnson and johnson"

[2] DCT keywords: "covid app" "tracing app" "COVID-19 app" "contact tracing" "privacy" "security" "surveillance" "app security" "app privacy" "contain virus speak" "movement tracking" "contain virus speak" "movement tracking" "covid application" "DCT" "digital" "bluetooth" "tracing"

model, 3 epochs, and a batch size of 16, we computed accuracy and F-1 scores for our final model (these hyper-parameters with an optimizer learning rate of $5e^{-5}$), another fully-trained model with an optimizer learning rate of $2e^{5}$), and our baseline Twitter-RoBERTa-Sentiment-Classification model.

We also used the F1 score for to determine the best model. The F1 score for each category is calculated as:

$$\texttt{F1 score} = \frac{2 \times \texttt{precision} \times \texttt{recall}}{\texttt{precision} + \texttt{recall}} \tag{2}$$

and we used the `sklearn` function `f1_score` to calculate this metric for the baseline and fine-tuned model. This additional metric was crucial because we found that a model which always outputs 1 (neutral sentiment) managed to score around 50% accuracy on our test sets, since neutral tweets were more frequent than the other classes. Multi-class F1 scores offer a more holistic assessment of our model's performance in comparison to the baseline, since they compute precision and recall for each class (positive, neutral, negative) to indicate how well the model classifies each of these labels individually.

4.3 Experimental details: Training the Final Model

After fine-tuning the initial hyper-parameters on the smaller fine-tuning datasets (delineated above in methods), we trained our combined COVID-19 and vaccine dataset to create two fine-tuned model versions of Twitter-RoBERTa-Sentiment-Classification. We trained for 3 epochs, with an instantaneous batch size of 16, and 417 total AdamW optimization steps. The only difference between these two models was that they had different learning rates: $2e^{-5}$ and $5e^{-5}$.

We determined that our best test set accuracy came from the model with the optimizer learning rate of $5e^{-5}$. The overall training runtime was 12126.4449 seconds, with 0.549 samples per second, 0.034 training steps per second, and a training loss of 0.6256. This model achieved a test loss of 0.8777 and a test accuracy of 0.6547 (65.47%). The test run-time was 103.9735 seconds, with 4.568 samples per second, and 0.289 test steps per second.

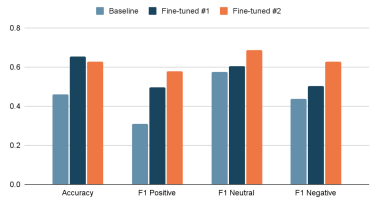
4.4 Results and Analysis

4.4.1 Fine-Tuned Model Results

Overall Results First, as expected, both of our final fine-tuned models achieved a much higher accuracy and F1 score than the baseline model. This is because they were both fine-tuned specifically to analyze COVID-19 Tweets, instead of Tweets in general (like the baseline model). While the baseline achieved an accuracy of 46% and F1 scores of [0.436, 0.575, 0.308] (for negative, neutral, and positive respectively), the model with the AdamW learning rate of $5e^{-5}$ achieved an accuracy of 65.47% and F1 scores of [0.5025, 0.6038, 0.4954], and our model with an AdamW learning rate of $2e^{-5}$ achieved an accuracy of 62.74% and F1 scores of [0.6277, 0.6853, 0.5772]. Interestingly, the model with the AdamW learning rate of $2e^{-5}$ achieved better F1 scores, whereas the model with the AdamW learning rate of $5e^{-5}$ achieved higher accuracy.

	Baseline	Fine-tuned #1	Fine-tuned #2
Accuracy	0.46	0.6547	0.6274
F1 Positive	0.308	0.49536	0.57718
F1 Neutral	0.575	0.60377	0.68534
F1 Negative	0.436	0.50246	0.62766

(a) Table Model Results



(b) Graph Model Results

Figure 2: Accuracy and F1 Scores for Fully Trained Models

AdamW Optimizer Learning Rate While changing the number of epochs, the batch size, and changing the data distribution all provided conclusive results that the hyper-parameters we chose were the best, changing the AdamW optimizer learning rate—controlling how quickly the model optimizes to the problem—was interestingly not as conclusive. Both of the models with these learning rates provided results well above the baseline, but we expected the $2e^{-5}$ rate to work better overall because even though $5e^{-5}$ is the default AdamW optimizer learning rate, we had a relatively small amount of training data. This optimizer rate, since it has adapts more slowly, is a preferred hyperparameter when using a small amount of data. Our smaller fine-tuning experiments (before training our two large models), reflects this trend. Our small train/evaluation set using a $2e^{-5}$ learning rate produced the best evaluation set accuracy out of the entire experiment: 81.25% accuracy. However, once we applied our two fully-trained models to the unseen test dataset, the model with the hyper-parameter of $5e^{-5}$ as the AdamW learning rate performed much better: 65.47% accuracy compared to 62.73% accuracy. Below depicted both models' accuracy over the three epochs and on the test set (which we included as the last stage):

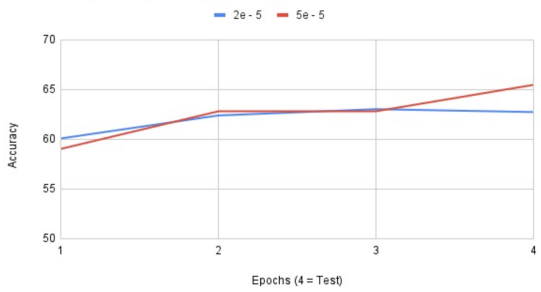


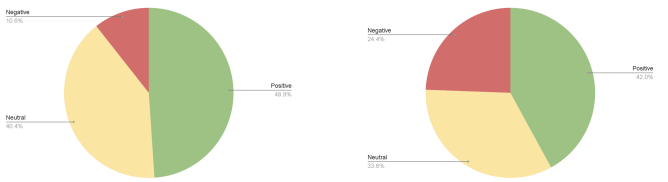
Figure 3: Optimizer Learning Rates by Epochs for Fully-Trained Models

Main Takeaways Overall, these results from the two fully-trained models show us that our approach of performing domain-specific fine-tuning on a BERT model was successful in creating a BERT model for classifying COVID-19 Twitter sentiment. Specifically, the combination of a baseline pre-trained for sentiment analysis of Tweets and fine-tuned on COVID-19 datasets yielded a model that excelled at the intersection of this data. While some fine-tuning decisions delivered conclusive results in terms of accuracy and F1 scores, other parameters, namely learning rate, were not so clear. As we'll explore in the Analysis section, the metrics we use in this paper evaluate different properties of a model which aren't necessarily correlated.

4.4.2 Real World Data Results

Overall Sentiment Overall, our model discovered that sentiment on Twitter towards COVID-19 policies was generally positive. For digital contact tracing, 49.9% of people Tweeted positively on this topic, 40.4% of people posted neutral Tweets, and only 10.6% of people had negative sentiment towards this policy.

Towards vaccines, people had slightly less positive things to say. For instance, 42.0% of people posted Tweets with positive-classified sentiment, 33.6% of people posted neutrally, and 24.4% of people posted negatively.



(a) Sentiment Distribution for DCT Related Tweets

(b) Sentiment Distribution for Vaccine Related Tweets

Figure 4: Sentiment Distributions by Mandate

Country Comparisons We gathered the following data by country, which showed the UK having the highest percentage of positive Tweets overall sentiment in COVID-19 related data.

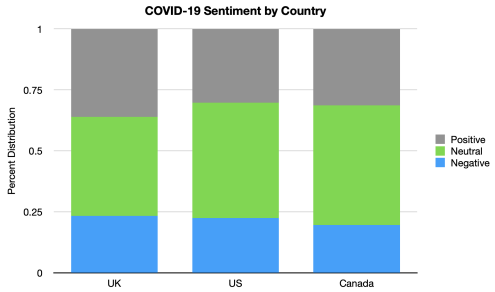


Figure 5: Distributions of Positive, Neutral, and Negative COVID-19 Sentiment by Country

Digital Contact Tracing by Country When comparing digital contact tracing sentiment by country, we discovered that 42.8% of Americans posted positive Tweets about DCT, 32.0% of Americans posted neutrally about DCT, and 25.2% of Americans posted Tweets with negative DCT sentiment. In the United Kingdom, Twitter sentiment towards digital contact tracing was more positive. 56.3% of people posted positive Tweets regarding DCT, 37.5% of people posted neutral Tweets about this technology, and only 6.3% of people posted negative DCT Tweets. These findings, for the most part, corroborate previous research studies that assessed global DCT sentiment through interviews and surveys, which we discuss in Analysis.

5 Analysis

5.1 Fine-Tuned Models

In the process of fine-tuning our models, we found that a combined dataset of COVID-19 data, trained for 3 epochs of optimization, produced the best F1 scores and accuracy. This aligned with our expectation that fine-tuning on a combination of relevant COVID-19 datasets, as opposed to hyper-specific domain data or general Tweet data, would produce a better model for our task. The dataset we used was general enough to make accurate predictions about contact tracing sentiment despite not seeing data overwhelmingly specific to DCT, while still including data specific to our driving question. The number of epochs that yielded the best accuracy also matched our prior beliefs - we expected our accuracy to peak around 3 epochs because after a few more iterations we would begin overfitting to our smaller training dataset. We observed this dropoff in accuracy in our fine-tuning experiments.

When experimenting with the AdamW optimizer learning rate, we found mixed results: the higher $5e^{-5}$ learning rate yielded a better accuracy, while the $2e^{-5}$ achieved better F1 scores. As discussed previously, accuracy provides a measure of overall correct predictions, and this metric can be high even for a broken or uninteresting model when the training dataset is skewed towards one class. In our case, a model that always classified Tweets as neutral managed to achieve around 50% accuracy, while getting every Tweet labeled positive or negative incorrect. We theorize that a similar situation occurred with our higher learning rate, though not so extreme as to cause the model to always output neutral. The higher learning rate may have caused the model to learn to predict neutral more quickly, improving the accuracy because the test dataset was skewed towards neutral, but not the F1 score which identified the underrepresented positive and negative predictions. Since F1 scores are based on precision and recall and are calculated for each class separately, outputting neutral more often will cause lower F1 scores, which penalize over-predicting one class to match a skewed distribution. These results reveal a potential flaw in this model, despite its promising accuracy score. Thus, the higher accuracy for the $5e^{-5}$ learning rate model doesn't contradict our expectation that a smaller $2e^{-5}$ learning rate is better for our smaller training dataset, it just demonstrates the importance of multiple evaluation methods and thoughtful analysis of results.

5.2 Real World Takeaways

Overall COVID-19 Policy Sentiment Our results showed nearly 50% of Tweets pertaining to DCT keywords labeled as positive sentiment, indicating that Twitter users across the countries represented had supported this policy. Notably, only 10.6% of DCT Tweets were labeled negative, indicating that contact tracing policies were largely unopposed in these countries. This result suggests that people felt comfortable sacrificing some of their personal privacy to support their government's effort to slow the spread of COVID-19.

We saw a shift towards neutral and negative sentiment in Tweets about vaccines, with the biggest change from DCT Tweets appearing in the percentage of negatively classified Tweets. We observed 24.4% of Tweets expressing negative sentiment about vaccines, which suggests both apprehension about and opposition to COVID-19 vaccination. To reconcile these results with the more positive sentiment towards contact tracing, we theorize that misinformation surrounding vaccines may account for a significant portion of the negative Tweets. While digital contact tracing isn't a huge leap from the friend maps and location features on social apps, vaccination policies require a physical commitment that, paired with vaccine misinformation, may cause more users to feel negatively about their country's policies.

Digital Contact Tracing in the UK and US Our results showed that Twitter users in the UK supported DCT more than those in the US, and this aligns with ongoing research beyond Twitter data. Multiple academic studies found that more people in the United Kingdom supported digital contact tracing than people in the United States. For instance, in a direct comparison study, Altmann et al. found that a higher percentage of people that would "probably install" a digital contact tracing application than the United States population (approximately 80% compared to 71%) [12]. Other studies suggest that this difference is much higher. For instance, Zhang et al. found that 42% of the US population supported these apps [13], and other researchers found support in the US to be as low as 30%. Conversely, no study conducted in the UK that we are aware of found support below approximately 80%. Therefore, while our findings demonstrated slightly more positive Tweets than expected in the United States and slightly fewer positive Tweets in the United Kingdom, our results follow this same underlying trend: people in the United Kingdom accepted DCT more than people in the United States. A potential explanation for this difference is the deployment of other digital monitoring tools in the UK, such as facial recognition software utilized by the police, that might make citizens more familiar and and comfortable with surveillance technology deployed by the government for the purpose of public safety.

5.3 Mislabeled Sentiment

For the qualitative analysis portion of this report, we looked at Tweets where our model did not accurately predict the sentiment score. One example Tweet reads "RT @LukeMones: ill say it: id love to not get the corona", with a predicted score of 0 (negative) and a true score of 1 (neutral). Our fine-tuned model likely learned that contracting the virus, or "get the corona", is a negative event and labeled it as such. The model fails to notice the preceding "not", and therefore misunderstands the intention and sentiment behind this Tweet.

6 Conclusion

Our fine-tuned RoBERTA-Twitter-Base-Sentiment model achieved significantly increased accuracy and F1 score compared to our baseline. Our main limitation was the size of our training data set, due to a Twitter API error. Although our cited datasets include millions of COVID-19 Tweet examples, we were only able to hydrate a small fraction. We expect that with more training data, the model could achieve a higher accuracy and F1 scores. Further, we applied this model to yield interesting real-world results, specifically in relation to DCT—a protocol that could still be useful in mitigating virus spread today. From our predictions, we found that people regard this system more favorably than vaccine mandates, and that Twitter users in the UK support DCT more than those in the United States. Since epidemiologists have found that 60-80% uptake of DCT would have effectively mitigated COVID-19, the high support we found on Twitter could help guide future government protocol regarding this technology. [14]

References

[1] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. pages 1644–1650, November 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/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. *CoRR*, abs/1907.11692, 2019.

[4] 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. *CoRR*, abs/1901.08746, 2019.

[5] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676, 2019.

[6] Martin Müller, Marcel Salathé, and Per E Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, 2020.

[7] Jochen Hartmann, Mark Heitmann, Christina Schamp, and Oded Netzer. The power of brand selfies. *Journal of Marketing Research*.

[8] Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. 2021.

[9] Izabela Krysińska, Tomi Wójtowicz, Agata Olejniuk, Mikołaj Morzy, and Jan Piasecki. Be careful who you follow: The impact of the initial set of friends on covid-19 vaccine tweets. pages 1–8, 2021.

[10] Rabindra Lamsal. Coronavirus (covid-19) tweets dataset. 2020.

[11] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324, 2021.

[12] Samuel Altmann, Luke Milsom, Hannah Zillessen, Raffaele Blasone, Frederic Gerdon, Ruben Bach, Frauke Kreuter, Daniele Nosenzo, Séverine Toussaert, and Johannes Abeler. Acceptability of app-based contact tracing for covid-19: Cross-country survey study. *JMIR Mhealth Uhealth*, 8(8):e19857, Aug 2020.

[13] Baobao Zhang, Sarah Kreps, Nina McMurry, and R. Miles McCain. Americans' perceptions of privacy and surveillance in the covid-19 pandemic. *PLoS ONE*, 15(12 (Article No.:) e0242652), 2020. Additional information: The publication of this article was funded by the WZB and the Open Access Fund of the Leibniz Association.; doi:10.7910/DVN/5UEFP6.

[14] Dyani Lewis. Contact-tracing apps help reduce covid infections, data suggest. *Nature News*, Feb 2021.

A Appendix

First, we selected 606 of our training data examples from a repository of Tweets collected from October of 2019 to March of 2022 [10]. The Tweets were scraped from Twitter using a set of over 90 COVID-19 keywords and hashtags, and scored for sentiment using TextBlob's Sentiment Analysis module. This Tweet data was originally formatted as Tweet IDs and sentiment scores of $[-1, 0)$ for negative sentiment, 0 for neutral sentiment, and $(0, 1]$ for positive sentiment. To convert this data into the input format needed by our Twitter-RoBERTa model, we hydrated[3] these Tweets using the `twarc`

[3] "Hydrating" Tweets refers to replacing Twitter IDs with the text content of the corresponding Tweet.

Twitter API, and converted the sentiment scores to [0, 1, 2] for negative, neutral, positive as expected by our model.

The other part of our training data came from a vaccine sentiment dataset of 2564 hand-labeled Tweets on Hugging Face [9]. To format this data, we hydrated the Tweets and converted the given sentiment scores in the dataset where 0, 1, 2 represented positive, neutral, negative to the reverse scheme used by our model.