Sentiment Analysis During COVID-19

Connor Weeks
Virginia Tech
Arlington, VA, USA
crweeks.edu

Niloofar Shadab
Virginia Tech
Blacksburg, USA
nshadab@vt.edu

Abstract

The long term psychological effects of COVID-19 remain largely unknown. In this project we seek to better understand the pandemic's effects, through sentiment analysis of Twitter data. To accomplish this, we first train 4 different text classifier models using a labeled sentiment dataset. Of these, we found BERT to be the best performing using accuracy and F1-score as evaluation metrics. Then using an aggregated set of COVID-related tweets, we collect a large text dataset using the Twitter API and use our BERT model to make sentiment predictions. Finally we collected more recent data from the Twitter API using keywords and preformed a similar analysis. We then compare the results of the two datasets to evaluate the possible changes that COVID-19 pandemic caused to people's emotions.

Keywords-*Sentiment Analysis, COVID-19, Twitter Data, LSTM, BERT, GRU, Emotional Analysis, Word Embedding, Transformer-based Models*

CCS Concepts: • **Deep Learning** → **Supervised Learning**; *Sentiment Analysis*; *Word Embedding*; • **Deep Learning** → Data Pre-processing.

ACM Reference Format:
Connor Weeks and Niloofar Shadab. 2021. Sentiment Analysis During COVID-19. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Sentiment analysis has been widely used to extract emotional language from text. Recently the effectiveness of sentiment analysis has grown significantly with the usage of deep learning methods in NLP. By applying these methods to social media datasets, especially long-timeframe datasets collected from Twitter, it may be possible to identify long-term trends in sentiment across a population.

In the wake of COVID-19 pandemic and subsequent quarantine, many people faced extreme isolation. While it is widely presumed that levels of stress and depression increased as a result of this long-term social distancing, accurately surveying and estimate these mental effects is difficult. By leveraging sentiment analysis across social media datasets, we may be able to track levels of stress, depression, and other sentiments as a response to the pandemic.

To accomplish this we first created a sentiment classifier using a labeled dataset of text collected from Twitter. Given the importance of the classifier's accuracy we built several models with different deep learning architectures. After comparing these models to other related works, we used the most accurate model to evaluate a larger unlabeled dataset of time-stamped tweets. Finally we measured and visualized the various sentiment levels by the date published.

Our analysis sought to answer several key questions:

1. **Has the quarantine increased the levels of depression?**
   In addition to the measurements from during the pandemic, we measured levels of negativity after the widespread usage of vaccine. For this purpose, we retrieved, collected and analyzed the last 7-day tweets. This gave us a baseline to search for any significant differences.
2. **If so, have these increases been consistent?**
   Has the continued long-term isolation continued to increase levels of depression or has acclimatization to social distancing lowered depression to more normal levels.
3. **If so, have these increases been consistent?**
   We aimed to identify effects coinciding with spikes in the infection rate and stay-at-home orders.

While other works have also sought to use sentiment analysis to understand the pandemic, we believe that the longer time-frame of our dataset could provide more insights. In particular the effects of the vaccine's announcement and deployment have yet to be explored by other studies.

2 Related Works

The main application of machine learning for identifying emotional trends among Twitter users was conducted by [1]. In this work, the researchers analyzed how features obtained by LIWC are related to depression signals on social media and how that can be used for user-level classification on a dataset containing 171 depression users.

Since the COVID-19 hit, various research has been conducted to study the probable impacts of the COVID-19 pandemic on different aspects of social life. Emotion analysis during the pandemic is one aspect of scientific research. One of the sources of this analysis is Twitter data. People exchange their opinions, feelings, and news related to COVID-19 and the pandemic through this social media platform. Therefore, there are several ongoing endeavors to gather, process, and use this data. One of the works in this area was to develop a large-scale COVID-19 Twitter chatter dataset for open scientific research. The preprint paper can be found in [2]. In another work, researchers at Yale University employed Natural Language Processing to investigate what are the aspects/topics of the COVID-19 pandemic that people are depressed about [3]. Anti-lockdown protests, Virus news, death rates, and many other factors were monitored using Twitter data.

In another work, researchers at Rochester University studied the trends of depression in tweets [4]. However, these researchers compare the users' performance in their previous tweets. They chose targeted users to evaluate if they also had shown any depression trends before the pandemic. They also determined other types of psychological disorder patterns such as bipolar, PTSD, and autism during the COVID-19 pandemic.

There have been multiple studies on the economic and social impacts of COVID-19. In [5], the impacts of the COVID-19 outbreak were investigated on the world economy across various industries. This study includes 30 countries and assesses their GDP trends.

3 Datasets Explanation

We had three different sets of datasets. One dataset is for training and it has labels. This dataset was from Twitter and it contains 1.6 million tweets with positive or negative sentences. The second dataset is for our sentiment analysis which was the COVID-19 related tweets. This dataset was retrieved from [1]. This source had all the ids, dates, and locations of the tweets. We used that dataset to do queries and retrieve their corresponding texts through Twitter API. We created the third dataset for the purpose of this project. We extracted the last 7-day of tweets that include keyword "feel" in them. We needed this dataset to compare the baseline for the ordinary proportion of negative and positive tweets. We also didn't want to use old datasets that were available online. Instead, we seek tweets that are recent so that they can be a good representation of the ordinary level of positivity and negativity in tweets.

For the training dataset, we explored the distribution and statistics of the sentences. It can be seen in Figure 1 and Figure 2.

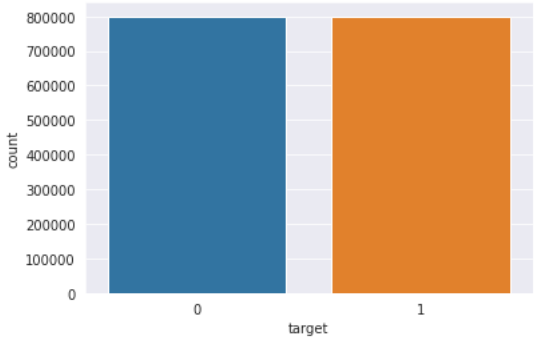

Figure 1. Diagram of the distribution of labels.

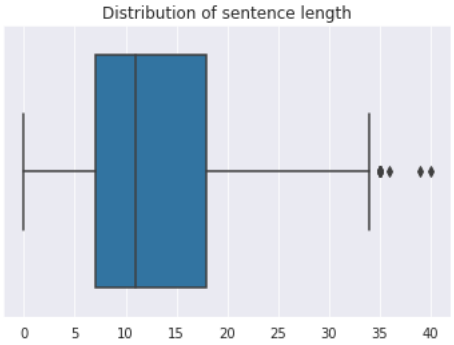

Figure 2. Diagram of the distribution of the length of the sentences.

For the COVID-19 dataset, in Figure 3, we present some statistics of the volume of COVID-19 related tweets per day. Due to limitations of the number of tweets that can be retrieved per hour, we decided to retrieve only 10,000 tweets per day for the entire timeframe of the pandemic.

4 Data Pre-processing

In this section, we talk about the procedures that were taken to prepare data for our analysis. The training dataset was publicly available on internet. However, for the COVID-19 related tweets, we had to query the tweets through Twitter

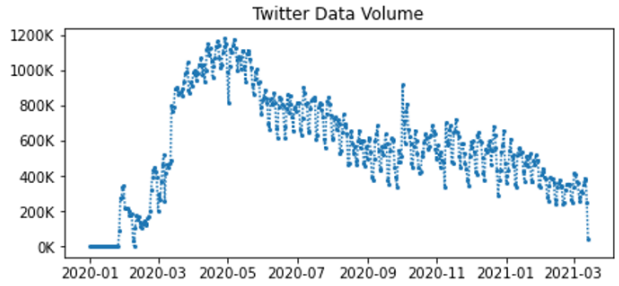

Figure 3. COVID-19 Twitter volume per day

	target	text
0	0	awww thats a bummer you shoulda got david carr...
1	0	is upset that he cant update his facebook by t...
2	0	i dived many times for the ball managed to sav...
3	0	my whole body feels itchy and like its on fire
4	0	no its not behaving at all im mad why am i her...

Figure 4. Formatted Tweets with Labels

API. The process of retrieving tweets from Twitter API will be explained later in the paper.

4.1 Training Data Pre-processing

To pre-process the training data, we conducted the following steps. For this step of training data pre-processing, we used a Kaggle resources [6, 10, 11].

- We first imported data from a csv file. Tweets columns in the csv files were ['target', 'id', 'date', 'flag', 'user', 'text']
- In this project we only used "text" and "target" columns. We therefore, formated texts column to prepare them to be tokenized. We removed all URLs, hashtags, punctuation, all non-character, extra spaces, and mentions in the texts. Then we change all the words to lowercase.
- We tokenized texts using TweetTokenizer. We used this code to tokenize the texts.
  tweet_tokenizer = TweetTokenizer(reduce_len=3)
  formated_tweets.text = formated_tweets.text.apply (lambda x: tweet_tokenizer.tokenize(x))
- We then changed the texts into lists and used padding for the tokenized list of words.
- we used "glove.6B.100d" for the vectorized representation of the words. It has 6 Billion token and 100 Features. It changes word into a meaningful state where the distance between words is related to semantic similarity [7].

The formated tweets after pre-processing is as shown in Figure 18. Target 0 is a representation of a negative tweet and target 1 is a representation of a positive tweet.

4.2 COVID-19 Related Data Pre-processing

For our pandemic analysis we used an extensive dataset of COVID-19 related tweets. This dataset was aggregated by Panacea Labs[1] who gathered the tweets from several other sources. The tweets in the dataset were determined to be relevant to the pandemic by a search for related keywords.

For many academic projects working with social media, data dehydration is common practice. This refers to the removal of most data attributes to reduce storage space required. Because users are forced to re-hydrate the data using an web API, this has the additional benefit of preventing researchers from accessing information from deleted tweets or users. However, this created an enormous problem for project. Since the tweets' text were needed for sentiment analysis, we were forced to collect all our data from the Twitter API instead of a direct download.

While we received free developer licenses within a few days of applying, the rules for academic licenses are much more stringent. This restricted us to 300 API requests per 15 minutes. Since our goal was to collect enough data for a longitudinal analysis, we choose to collect a fixed amount of tweets for each day in our dataset. In total we made 10,000 API calls for each day in our dataset over the course of two weeks. A package called tweepy allowed us to run these calls in python without making the web requests ourselves.

The original COVID-19 Twitter data had 5 columns. There included id, date, time, language, and country. From the Twitter API, we were able to get a full json of each tweet. This included username, updated geo-location, and many other unneeded attributes. Ultimately, the dataset we reduced to 5 attributes, date, time, location, language, and text. Preprocessing was done to the text to remove URLs and non-acsii characters. In Figure 5, we show a sample of the texts for 02/22/2020.

	date	text
0	2020-02-22	g rollout delayed in china due tocoronavirus
1	2020-02-22	is your organization ready to manage the poten...
2	2020-02-22	wisconsin paramedic shares his experience moni...
3	2020-02-22	many have praised china for its strict quarant...
4	2020-02-22	coronavirus isnt stopping china from launching...

Figure 5. Formated Tweets of COVID-19 Dataset

4.3 Last 7-day Twitter Data Pre-processing

To validate our analysis and have better understanding of how COVID-19 related tweets differ from the usual tweets in terms of emotional analysis, we retrieved tweets from the last seven days using **Twitter API**. As we could not access the premium version of Twitter API, we could not have the option for the full archive query search. If we were able to use the premium Twitter API search, we would have used tweets posted right before the pandemic started (approximately tweets from Dec 2020). December 2020 is the best timeframe to evaluate the level of negativity in ordinary Tweets. This way, we had better understanding of how COVID-19 worsen Twitter's users' emotions. Considering this limitation, we decided to choose tweets posted in the last seven days for the purpose of this project. Therefore, we retrieved tweets that had specific key-word "feel". Using this keyword allowed us to better capture negative tweets and reducing neutral contents in the dataset. For this, we used **Tweepy** library to conduct the query. In the query, we filtered all the retweets from the dataset. We were able to retrieve **183,415** tweets containing the keyword "feel". The procedure in using Tweepy and Twitter API was explained in the previous subsection. We pre-processed the tweets using the methods explained in Section 4.1. The query was a simple search command (tweets = tw.Cursor(api.search, q=new_word,lang="en").items()).

The final formated results can be seen in Figure 6.

	texts
0	i went through the same thing and im
1	benji post made you feel
2	hey this account is a safe and quiet place for...
3	oh ur brazilian okay i feel you lol weve alway...
4	at least one of the top women investors im exc...

Figure 6. Formated Tweets with keyword "feel" from the last seven days.

We also gathered another dataset which contains the neutral word "I". This dataset allowed us to compare the COVID-19 datasets to the ordinary level of both emotional contents as well as contents that have emotional and non-emotional contents. This dataset had around 121,000 tweets from the last seven days in Twitter. In Figure 7 you can see the results of the formated texts for this datasets.

	texts
0	dilf damn i love friends
1	i want to know who noel roberts is
2	i like your boots and also your hat your face ...
3	iheart bp for bbmai vote for on
4	im pretty sure i saw the episode with the bibl...

Figure 7. Formated Tweets with keyword "I" from the last seven days.

5 Model Evaluation

Sentiment analysis is a type of sequence classification problem. While it can solved with other approaches like bag-of-words, we focused our efforts on newer deep learning-based methods. More recently a new class of NLP methods using transformer-based architectures has replaced more tradition RNN-based models as the state-of-the-art. However, transformer-based models are much more complex and require extensive computing resources to train from scratch. For this reason, we also focused on other architectures including RNN, LSTM, and GRU.

Our next step after preprocessing the data and selecting a word embedding, is to train our model. RNN, the simplest and oldest model we intend to use, should provide a good baseline for our other methods. Like our other methods, RNN, or recurrent neural networks, use recurrent layers, which store memory values between tokens. These layers can be difficult and time-consuming to implement so we used Keras an open-source deep-learning package for python. Keras gives us access to fast implementations of the different types of layers we intended to use in our experiments.

In addition to our RNN-based classifier, we created LSTM, GRU, and Transformer-based models. We separated a portion of labeled data to use as a test set for our evaluation. Our primary metrics for evaluation were accuracy and F1-score, a combination of precision and recall for a target class. Only our highest scoring classifier was used for our unsupervised evaluation.

5.1 BERT

The BERT architecture is a general language model developed by Google in 2019. We choose to use BERT because it is widely considered state-of-the-art on a wide range of NLP tasks. Unlike our other models, BERT uses a specified word embedding called WordPiece.

As with many modern language models, BERT is first pretrained on a large English corpus before being fine-tuned for more specialized tasks. BERT specifically is pretrained on masked token prediction and next sentence prediction. Masked token prediction, also known as masked LM, works similarly to a de-noising auto-encoder. A small percentage of tokens are replaced with mask tokens, and models makes predictions about the original token.

The second task is next sentence prediction. This is not an encoder/decoder problems but a binary prediction problem. The model was given sentences A and B. 50% of the time B is the actual sentence which follows A, but the other 50% of the time B is a random sentence. The model is trained to determine if sentence B follows A.

After the pre-training, the BERT model can be fine-tuned on a wide number of tasks including, sequence classification, question answering, and token prediction. Since sentiment analysis is a type of sequence classification problem we used, Transformers' BertForSequenceClassification for our model. This adds a dropout layer with 10% dropout rate and a single fully connected linear layer. Since we have two classes, it uses cross-entropy for the loss function during back propagation.

We used the Transformers library by Huggingface for our BERT architecture. We found the default parameters for layer size, dropout, and attention heads sufficient for achieving high accuracy. We used the following hyperparameters for this model.

- We used ADAM optimization algorithm with learning rate of 0.00002.
- We used binary cross-entropy loss function.
- We chose batch size of 5.
- Hidden Layer Dropout 10%
- We used 12 hidden layers in the transformer encoder
- We used 12 attention heads for each attention layer
- GELU activation function for the hidden layers
- We chose 20% of dataset as the validation set.

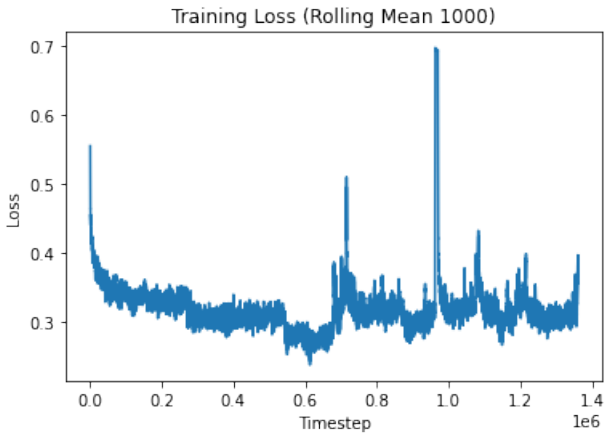

Figure 8. BERT training loss

5.2 RNN

Recurrent Neural Networks (RNNs) is a family of an artificial neural networks that use the output of previous state as an input for the next state and it has a hidden state. It is used for sequential datasets such as texts. We used keras to build a Vanilla RNN model. We used the following hyper-parameters for the model:

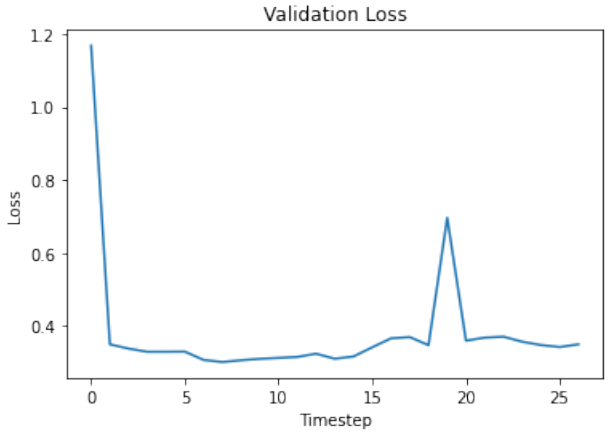

Figure 9. BERT validation loss calculated every 10,000 mini-batches

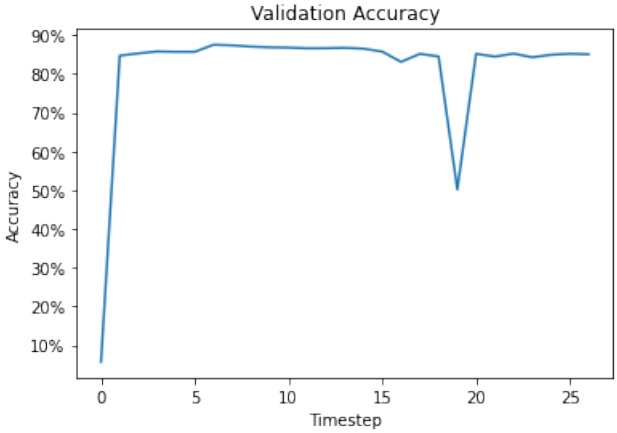

Figure 10. BERT validation accuracy calculated every 10,000 mini-batches

- Dropout 25%
- Embedding layer with output dimension of 100.
- Two bidirectional SimpleRNN layers with 64 parameters
- Dense Layer with ReLU activation and output dimension of 512
- Dense layer with Sigmoid activation and output dimension of 1.
- We used ADAM optimization algorithm
- We used binary cross-entropy loss function.
- We chose batch size of 256.
- We chose 20% of dataset as the validation set.

One important note for the hyper-parameters used for RNN, LSTM, and GRU is that the performance of these models did not improve by decreasing the learning rate. In some cases, we got even worse results by decreasing the learning rate. This problem could be due to overfitting which worsen

the model performance for the validation data. We also tried different dropouts, 50%, 25%, and 10%. 25% dropout had the best performance for the three models.

5.3 LSTM

Long Short-Term Memory (LSTM) is a recurrent neural network that has feedback connections. An LSTM cell consists of input and output gates, forget gates. One of its applications is text classification. LSTM uses the memory cell to keep information from a period of time. This extra information allows the LSTM model to perform classification and prediction tasks more accurately compared to the Vanilla RNN models. We used keras to build an LSTM model. We used the following hyper-parameters for the model:

- Dropout 25%
- Embedding layer with output dimension of 100.
- Two bidirectional LASTM layers with 64 parameters
- Dense Layer with ReLU activation and output dimension of 512
- Dense layer with Sigmoid activation and output dimension of 1.
- We used ADAM optimization algorithm
- We used binary cross-entropy loss function.
- We chose batch size of 256.
- We chose 20% of dataset as the validation set.

5.4 GRU

Gated Recurrent Units (GRU) is similar to LSTM model. However, it has only a reset gate and an update gate. Therfore, it has less parameters due to the absence of output/input gates. We used keras to build a GRU model. We used the following hyper-parameters for the model:

- Dropout 25%
- Embedding layer with output dimension of 100.
- Two bidirectional LASTM layers with 64 parameters
- Dense Layer with ReLU activation and output dimension of 512
- Dense layer with Sigmoid activation and output dimension of 1.
- We used ADAM optimization algorithm
- We used binary cross-entropy loss function.
- We chose batch size of 256.
- We chose 20% of dataset as the validation set.

5.5 Comparing the Results

As it is shown in Table 1, we compared the accuracy, loss, F1-score, precision, recall and the execution time of the above-mentioned models for our COVID-19 sentiment analysis. LSTM had better performance in terms of the average execution time. However, for the rest of the metrics, BERT model's performance is significantly higher than the other three models. Therefore, we chose BERT model for the COVID-19 tweets sentiment analysis. One of the reasons that RNN model has high execution time could be the problem of back propagation for RNN especially for long sequence of data when compared to LSTM and GRU. The reason for the high execution time for BERT is its complex architecture.

Table 1. Comparing Accuracy of Models after Training

Metrics	**BERT**	**LSTM**	**RNN**	**GRU**
Epochs	1	10	10	10
Execution Time	12,500 s/epoch	600 s/epoch	900 s/epoch	700 s/epoch
Training Accuracy	90%	83%	80%	81%
Validation Accuracy	87%	82.6%	80.3%	82.1%
Loss	0.335	0.383	0.426	0.396
F1-Score	0.897	0.824	0.799	0.813
Precision	0.91	0.816	0.803	0.832
Recall	0.88	0.841	0.804	0.804

However, to better compare the performance of the other three models, we created a boxplot (Figure 11) to evaluate their accuracy per epoch results better. This figure helps us evaluate which of these baseline models have more robust performance against our training dataset. For this, we plotted the final validation accuracy after each epoch for the models. It should be mentioned that due to the extended time required for the BERT model to reach 10 epochs, we chose 1 epoch for the model to run instead of 10. As after sometime, we got disconnected from the server and we couldn't use the GPU runtime when running it through GoogleColab. Running only for 1 epoch, we could see that except for the execution time, BERT model outperforms the other three models. In Figure 11, we could see that the median of LSTM model is higher than the other two models. Moreover, the first quartile of RNN had the highest variance. GRU also has the lowest variance between its minimum and maximum value for the accuracy. GRU has the lowest number of outliers. Therefore, we could conclude that GRU is a more robust model and LSTM is a more accurate model among the three baseline models.

In Figure 12, we can see the trend of model loss at each epoch. From Figure 12, we could conclude that RNN model has the highest loss compared to LSTM and GRU. It is also evident that the loss trend for RNN is following a line with very low slope which indicates that with even higher epochs, we could not reach to an acceptable loss for our training dataset. GRU and LSTM models, on the other hand, have similar trend of loss per each epoch. However, overall, LSTM model has better performance than GRU model. We tried 20 epochs as well. The accuracy remained in the same range.

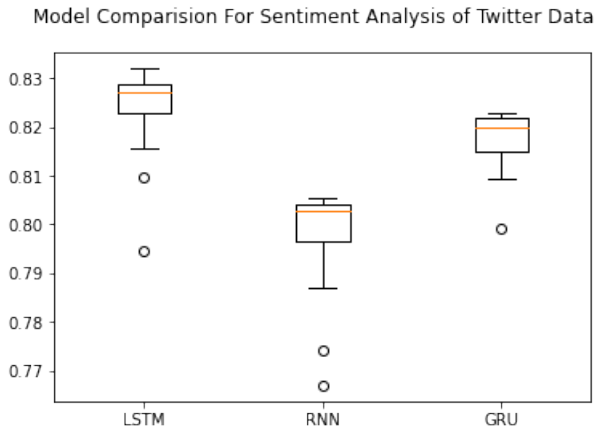

Figure 11. Comparing accuracy of sentiment analysis models for the Twitter data

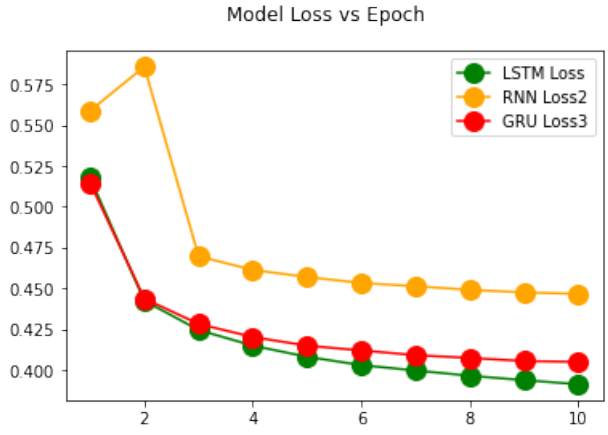

Figure 12. Comparing loss of sentiment analysis baseline models for the Twitter data

6 Analysis

Finally, after selecting BERT as our final classifier, we used it to identify trends in data from the period of the COVID-19 pandemic. In particular, time-stamps was used to partition our data and visualize our data temporally. We intend to use this data for regional and international comparison against corresponding infection spikes, quarantine orders, and the announcements of vaccine deployment.

In this section, we compare the results of COVID-19 tweets and the last 7-day tweets to investigate any insightful difference between the results of these two datasets. Here we are also looking for the impacts different policies such as lockdowns, or vaccinations.

6.1 Classifying COVID-19 Tweets

The first step of evaluation was to create a sentiment prediction for every tweet collected from the API. Given the balanced training dataset, we used a argmax of softmax output rather than a weighted prediction. The softmax confidence was essentially ignored when making these predictions.

Next, the data was aggregated by day and by month.The plots for these results are shown in figures 9 and 10 respectively.

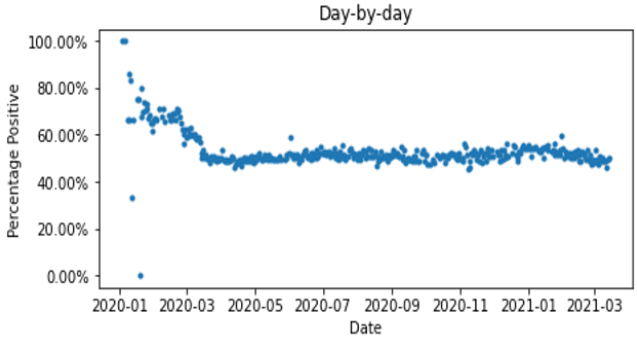

Figure 13. Day-by-day COVID-19 related tweets Over the period of the pandemic.

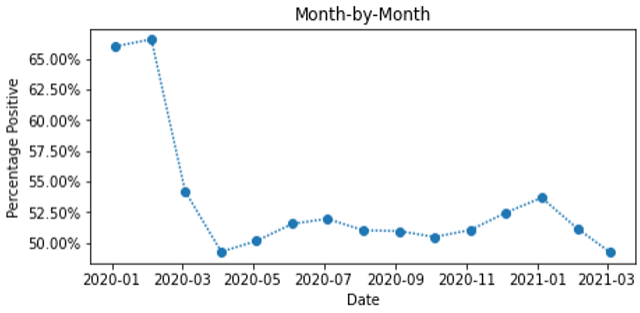

Figure 14. Average of Positiveness in COVID-19 related tweets per month Over the period of the pandemic.

From these chart we were able to make several observations. First it should be noted that extreme outliers during the first few days of our dataset were caused by a low sample size. For the first three weeks of the dataset, the number of tweets available was less than 100. On several days, due to tweet/users deletions or other updates no tweets were available. The first day for which the full 10,000 tweets could be collected was 2020-01-27.

Despite the effect of low sample size during the first month, it's clear that there is a significant positive effect for the first three months of the dataset. This is likely due to the significant media coverage during these months. In general we found the tweets from news sources were far more likely to be given a positive label by our classifier.

The data shows that the lowest point of sentiment was April 2020. This coincides with both the highest level of twitter coverage and the initial spike in cases/deaths in the US. Conversely the highest spike of positive sentiment outside of the initial 3 months happens in January 2021. This also coincides with a spike in deaths, however, we believe the reason for the high sentiment is likely caused by the news and deployment of vaccines.

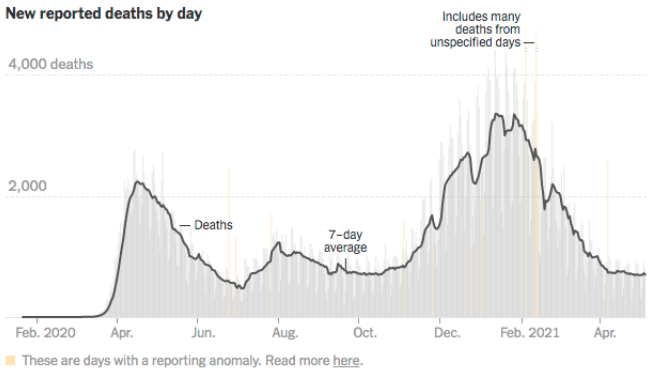

Figure 15. COVID-19 deaths as reported by the New York Times

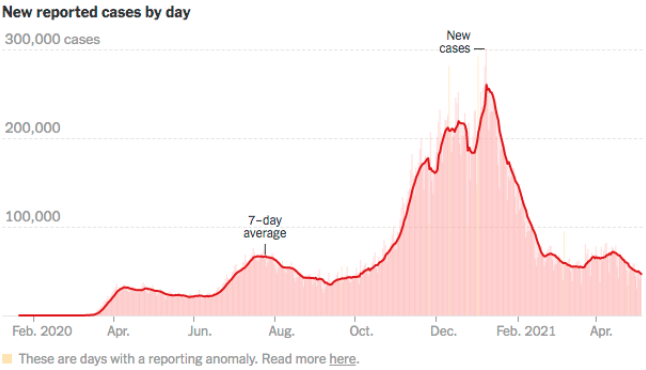

Figure 16. COVID-19 new cases as reported by the New York Times

6.2 Classifying The Last 7-day Tweets

After running the pre-trained BERT model on our 7-day tweets, we got the following results in Table 2. The first dataset contained mostly tweets that express emotions which is because of the keyword "feel". As it can be seen, the proportion of the negative tweets posted in Twitter is higher than the positive tweets. Therefore, we could conclude that most of the tweets that contains emotions have negative contents. This is specifically fascinating when we compare it to the COVID-19 datasets. COVID-19 datasets consist of all the news coverage, and CDC coverage, and WHO coverage of updates. These tweets mostly do not classified as negative as their contexts are neutral and are for public audience.

However, for the second dataset with keyword "I", we can see that the proportion of the negative and positive tweets is the same. 50% of the tweets are positive and 50% are negative. This difference can be explained by the fact that tweets in the second datasets contained neutral texts as well while the first dataset contains mostly emotional contents.

Table 2. BERT model results for the 7-days tweets

Labels	"Feel" Keyword	"I" Keyword
negative:	116255	61043
positive:	67160	60154

In Figure 17, the distribution of the positive and negative tweets based on the keyword "Feel" is demonstrated. The last 7-days is considered from April 20 to April 27.

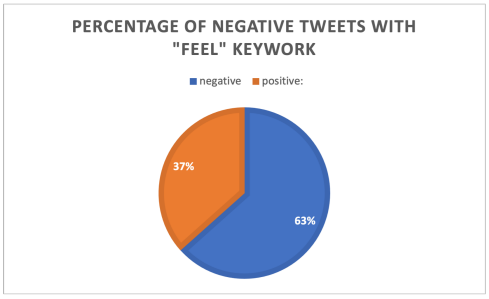

Figure 17. Distribution of the negative and positive tweets over the last 7-days..

6.3 Comparing The Analysis

The key takeaways from our analysis on the COVID-19 dataset are as followed:

- On average, COVID-19 tweets are 5% to 15% more positive than the ordinary tweets with "feel" keyword. This is an interesting fact that shows a large portion of COVID-19 related tweets do not contain emotional contents.
- On average, COVID-19 tweets have similar proportion of the positive and negative tweets compared to the second dataset containing word "I". This is an interesting fact that shows containing keywords related to COVID-19 does not necessarily increase the negativity in the context.
- The positivity in COVID-19 tweets has its lowest in April 2020. This is the timeframe of the peak of the lockdown, curfews, and stay-at-home orders.
- The biggest spike over the course of this pandemic happened in January 2021. This could be linked to both the success of vaccination trial, and the change of political environment in the USA.

- The increase in the positive tweets related to COVID-19, started from Nov 2020. This also could be linked to the political events such as the presidential election during that time and the increase of hope for a better policy to combat the virus.
- The spike of positivity in COVID-19 tweets was despite the increase in the total cases of Coronavirus.
- In the first two months most of the COVID-19 related tweets consisted of news related the virus. Therefore, considering the fact that general news in Twitter have neutral nature in the texts, we can explain why the percentage of the COVID-19 tweets is higher in the first two months of the pandemic.
- Around late June to July 2020, the first round of announcements of successful trials for COVID-19 vaccine was reported. The most important one was from the University of Cambridge successful trial. A spike in positive tweets increased around that timeframe.
- The increase in negativity in tweets increased again in February and March 2021 which could be the result of the increase in the number of coronavirus cases.
- One conclusion from the spikes of positivity from November to January 2021 despite the increase in cases is that the political environment and the news coverage around it had bigger effects on the people's perception of this pandemic and its effects on their life.

7 Data Visualization

7.1 Labeled Data

To visualize the training data, we used **WordCloud** package and then using the plt.imshow() function, we plotted the words created by WordCloud package.

Figure 18. Visualizing tweet words with Labels

7.2 COVID-19 Data

Similar to the training dataset, we were able to use WordCloud for the COVID-19 datasets. We, then, plotted the words from the date 02/22/2020. In Figure 19, you can see the visualization of the common words.

Figure 19. Visualizing tweets words for COVID-19 dataset

8 Lesson Learned

During our data collection for COVID-19 related tweets we discovered that queries were only getting text truncated to 144 characters. This discovery was made several days into our data collection so we needed to re-collect thousands of samples. We realized how important it was double check accuracy for long running programs and data collection.

9 Future Work

As a future work, we could analyze the COVID-19 data by language, or region to get better insights on the tweeter trends over the course of the COVID-19 pandemic. Geolocations were of particular interest for our project. But after discovering that only about 1% of tweets are geotagged, we decided not to do any further analysis given the small sample sizes. However, with enough time to fully gather the dataset it might be possible to do a geographically-based analysis.

For the training data, we could prepare more training labels to increase the accuracy of the BERT model. Moreover, if there is enough time and resources, BERT model can be run for more than 1 epoch to increase the accuracy of the BERT model. Finding a machine was a particular problem for this part. While we planned to train the model on Colab, the time required for just 1 epoch was longer than Colab's timeout duration. Because of this we were forced to train the model on a machine for one of our research groups for less time than we planned.

For the last-7-day tweets, we could try more datasets with different keywords to have better understanding of how the COVID-19 dataset can fit in and compared with other contexts in Twitter platform.

References

[1] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. In Seventh international AAAI conference on weblogs and social media.
[2] Banda, Juan M., et al. "A large-scale COVID-19 Twitter chatter dataset for open scientific research–an international collaboration." arXiv preprint arXiv:2004.03688 (2020).
[3] Li, Irene, et al. "What Are We Depressed About When We Talk About COVID-19: Mental Health Analysis on Tweets Using Natural Language Processing." International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer, Cham, 2020.
[4] Zhang, Yipeng, et al. "Monitoring Depression Trend on Twitter during the COVID-19 Pandemic." arXiv preprint arXiv:2007.00228 (2020).
[5] Fernandes, Nuno. "Economic effects of coronavirus outbreak (COVID-19) on the world economy." Available at SSRN 3557504 (2020).
[6] https://www.kaggle.com/jackttai/twitter-sentiment-classification-keras-lstm
[7] https://en.wikipedia.org/wiki/GloVe_(machine_learning)
[8] https://www.nytimes.com/interactive/2021/us/covid-cases.html
[9] Jacob Devlin., et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" arXiv preprint arXiv:1810.04805 (2019)
[10] https://www.kaggle.com/jarxrr/sentiment-analysis
[11] https://github.com/viritaromero/Detecting-Depression-in-Tweets

10 Contributions

The work was shared equally between the two members. Dataset pre-processing was done by both Niloofar and Connor in parallel to ensure the process could be validated. Connor trained BERT model and Niloofar trained RNN, LSTM, and GRU models. Connor retrieved the COVID-19 tweets, and Niloofar retrieved the last 7-day tweets. The analysis was performed by brainstorming and reaching to a mutual conclusion.