CSE8803-EPI Project Proposal

Rachit Bhargava
rachitb@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

Junyan Mao
jmao@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

Pratik Nallamotu
pratiknallamotu@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

In this paper we research the correlation between COVID-19 related tweets and spikes in COVID-19 cases. We propose a machine learning model using natural language processing techniques to analyze keywords and sentiments of COVID-19 related tweets to predict future COVID-19 spikes. We utilized a Kaggle dataset of COVID-19 related tweets and a data repository of COVID-19 time series data across the globe.

KEYWORDS

machine learning, neural networks, epidemiology, COVID-19, pandemic spread

ACM Reference Format:
Rachit Bhargava, Junyan Mao, and Pratik Nallamotu. 2021. CSE8803-EPI Project Proposal. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease has spread worldwide since its first discovery in Wuhan, China, in December 2019, leading to an ongoing pandemic [13]. The first confirmed case in the United States was discovered on January 20, 2020 [8]. Recently, the death toll of COVID-19 has passed 700,000, with over 43,000,000 total cases in the United States [6].

In such a pandemic, it is important for health professionals to be able to predict upcoming spikes in cases so that they can prepare and distribute necessary medical resources accordingly and make policy adjustment to reflect more strict public health rules. Social media chatters serve as a good data source because there is a strong correlation between COVID-related talks and actual COVID case spikes, and there is usually a lag between people experiencing the COVID-19 symptoms and being officially diagnosed with COVID-19. Therefore, we can use the trend of COVID-related keywords discovered in social media chatters to effectively analyze prior COVID spikes and use learned knowledge to predict future COVID spikes.

This project aims to analyze posts from Twitter (Tweets) with COVID-19 content to capture the connection between the trend of discussion in Tweets and the actual trend of COVID-19 cases, which will make it able to predict future COVID-19 trends.

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

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

2 DATA

For this project, we will use the "COVID-19 Complete Twitter Dataset (daily updates)" from Kaggle [9]. This dataset consists of tweets retrieved from the Twitter Stream related to COVID-19 chatter on a daily basis, from March 22, 2020 to October 12, 2020. This is a subset of a larger dataset hosted on GitHub [2, 10], which is what we are going to use in this project because it has a wider range of daily data and is able to capture multiple known COVID spikes. The data collected from the Twitter Stream captures all languages with English as the most prevalent, Spanish as the second most prevalent, and French as the third most prevalent. The dataset is organized in a tab-separated values (TSV) file, with the three columns being "tweet_id", "date", and "time". The authors of the dataset also provided the top 1000 frequent terms, the top 1000 bigrams, and the top 1000 trigrams in the format of comma-separated values (CSV) file to facilitate natural language processing (NLP) tasks. The whole dataset has a size of 32.43 GB (and growing). Due to the lack of sufficient processing power, we will pick 3-5 days from a week to reduce the data load.

In order to get location data, we will hydrate the data retrieved from Kaggle [9] with the geolocation metadata fetched from Twitter API [1] using the "tweet_id" attribute.

For COVID-19 cases data across the globe, we will be using the COVID-19 data repository provided by Johns Hopkins University [4]. This dataset includes a time series data set report of COVID-19 active and recovered cases, as well as deaths. It includes two directories we can focus on: one for states in the United State, and another for COVID-19 data on a global level.

3 PROBLEM FORMULATION

Given $\mathcal{D} = \{X_i\}_{i=1}^{N}$ micro-blogs (tweets) from real-world users related to the target pandemic (where $X_i = \{a_i, b_i\}$ is the i-th tweet and N is the total number of tweets, with a_i being the vector representation of the i-th tweet and b_i being the date on which the tweet was posted), $C = \{y_j\}_{j=1}^{M}$ data about the number of cases per day (where y_j is the number of cases on the j-th day from when the data started getting recorded and M is the index of the last day on which the data got recorded), and μ threshold that defines a spike in cases if the gradient of the five-days moving average is above it, we are interested in finding a set $\hat{\mathcal{Z}}$ of days where the pandemic observed a spike in number of cases. In other words, given the tweets from a given location about the pandemic of interest (POI), the data about the number of cases for the POI and a definition for spike in numbers, we are interested in predicting the set of days on which the POI might have a spike in the number of cases.

4 APPROACH

The problem can be divided into (i) getting vector representation of each tweet, (ii) getting a set of ground truth pandemic spike set of dates, (iii) performing sentiment analysis on the tweet vectors to add to the tweet vector (a possible correlation between the general sentiment of the tweets and pandemic spikes will be addressed by this) [5], (iv) training a neural network to give probability of a spike coming up (this builds on what a similar paper on predicting flu pandemic suggests [11]), and (v) using the two trained neural network with the input weight α to get a set of dates with a spike coming up.

4.1 Tweet Vectorization

Since tweets are just small sentences of up to 280 characters, we can preprocess them by removing the stop words and then obtaining the stem of the words on the remaining ones by using the Porter's algorithm [12]. Once done, we can perform a one-hot encoding for all the word stems in the tweet, which we can use later in the algorithm.

4.2 Detecting Pandemic Spikes

For each set of five consecutive dates, we check for the change in gradient, and if this change in gradient crosses the given threshold μ, we consider those dates to be in the set of dates with pandemic spikes. Mathematically, this can be written as,

$$\mathcal{Z} = \{z_i; \frac{y_{i+5} - y_i}{5} > \mu, 0 \leq i \leq M - 5\} \tag{1}$$

4.3 Adding Sentiment Analysis Information

We can rate the emotions presented in randomly-taken 5000 tweets as positive or negative. NRC is a list of English words that are commonly used in phrases and each of the words in the word bank relate to one of the eight emotions from psychology (anger,fear, anticipation, trust, surprise, positive, negative, sadness, disgust). We can use the python package NRCLex to take in these randomly-taken tweets to output the emotion classification of each of these tweets, and the classify the sentiment of each tweet as "positive" or "negative" based on the classified emotion. This package uses the NRC word bank to rate each tweet in order of emotion frequencies to assign it as one of the eight emotions, and as "positive" or "negative". We can finally add the sentiment information to the tweet vector. On a side-note, we can train our model to see if sentiment analysis is valid for predicting COVID-19 spikes by passing in a tweet vector with its labeled sentiment and also without and comparing the results.

4.4 Training Spike Predictor NN

We can train a neural network with an Embedding Layer and a Linear Layer, Softmax as the activation function, and AdamGrad (with learning rate of 0.02) as the parameter optimizer. The input would be the tweet vector and the output would be the probability of a day being part of the spike set. Cross Entropy Loss can be used as the loss function (a day is said to have a predicted spike if the model generates a probability of greater than 0.5).

4.5 Predicting Pandemic Spikes

Finally, we can use the probability of a pandemic spike from Section 4.4 to get the set of days with a pandemic spike. Finally, we get,

$$\hat{\mathcal{Z}} = \{z_i; p_i > 0.5\} \tag{2}$$

5 EVALUATION

Our approach can be evaluated with the pandemic spikes that it was able to predict successfully on a holdout data set. We propose to have an $80 - 20$ data split, where the first 80% of the data is used for training the models and the remaining 20% of the data is used for testing the model. At the end, the precision, recall and F1 score of the model will be reported.

6 EXPECTATION

By the end of the semester, we plan to create a model that will be able to analyze the sentiments of tweets and the number of COVID-19 related symptom tweets to predict future COVID spikes. Based on the research papers that we have read, we found that there are multiple existing approaches that have limitations to the data they had access to, so we plan to develop a more holistic and robust prediction model[14]. Our datasets include both Twitter[9] and COVID-19[4] data from the beginning of COVID-19 to present day. This will help us in building a stronger and higher performing predicting model by using existing natural language processing algorithms and time series analysis to predict COVID-19 spikes. By the end of the semester, we will formulate our own mathematical formula to determine what a COVID spike is across time series data. We will clean and process both our twitter COVID-19 dataset, as well as our dataset on COVID-19 cases and deaths across a global level. This step will also include feature engineering, as we will create features such as "sentiment analysis rating (e.g. positive or negative)" and number of times an important COVID-19 keyword appeared in a week . Once we clean and process our data, we plan to analyze our features in the datasets to determine what features of the twitter data are correlated and important to predicting a rapid increase in the number of COVID-19 cases. Our primary objective for this project during this semester is in building our algorithm to determine the correlation between number of tweets, number of keywords relating to COVID-19 symptoms, and geographical location to predict COVID-19 spikes. While building this algorithm and visualizing our data, we want to show our audience and readers the key features we found that were important in helping us predicting COVID-19 spikes using time series data. We plan to use our split training and testing dataset to evaluate the performance of our model and make any adjustments as necessary.

7 WORK DISTRIBUTION

In this project, we will all be splitting the workload evenly between the three group members. All three of us will work on the data processing and cleaning steps to ensure that even when we are developing our model any adjustments that need to be made to the data will be prompt and quick. All three of us will focus on building the machine learning model using natural language processing and time series to analyze and determine predictions of COVID-19 spikes. We will all jointly work together on the final report and presentation due at the end of the semester. We will hold weekly meetings to monitor project progress and clarify any challenges or assumptions that arise during the development phase.

8 EXPECTED TIMELINE

We have roughly 10 weeks until the semester ends to complete our project. The first two weeks will entail data cleaning, processing, and feature engineering. We will also be analyzing and visualizing our data to determine which features are truly of importance to our model in predicting COVID-19 spikes. The new two weeks we will spend on training and testing our model, and fine tuning hyper parameters passed into our model's approaches. The fifth and sixth week we will spend most of our time trying to improve our model's performance and expanding the size of the training dataset to better predict COVID-19 spikes. During this time, we will also experiment and evaluate our model. We include time allotted with the remaining 4 weeks to account for any obstacles or challenges we may face throughout the development process of our model. Once we finish and believe our model is of good performance and accuracy in its prediction ability, we will put together our final report and presentation to discuss with the class the need of our model, assumptions made throughout the process, design and implementation of our approach, the results and performance of our model, and any further improvements we would like to make with more time. Throughout the semester, we will also reach out to the professor or TAs to clarify any questions we have on our approach.

9 SURVEY OF RELATED WORK

There have been multiple researches that are focused on alternative approaches to evaluate an emerging health crisis like COVID-19, given that nationwide mass testing is not feasible in the majority of countries. This phenomenon of social media's ability to provide insights into pandemic activity has been referred to as "wisdom of the crowds", where the collective knowledge and experience of individual users are used to predict trend of a pandemic in the absence of a large-scale virus tracking system [3].

Some of the papers we read had different approaches, but a common objective of using COVID-19 symptom related tweets or searches in predicting COVID-19 spikes. Researchers from the University of Guleph [14] and researchers from the University College London (UCL) [3] both had similar approaches of analyzing features within Twitter datasets and identifying anomalies within the number of tweets related to COVID-19 symptoms to predict when the wave or spike of COVID would happen. We can use insights and results made from both of these papers in helping us detect anomalies and correlations between keywords in the twitter dataset and COVID-19 cases to predict a future spike in COVID-19. A weakness from the paper by the researchers from the University of Guleph [14] was that they used twitter data when testing was not as prevalent because it was the beginning of the pandemic, and also regions had travel-based policies that were not identified, which could explain some of the delayed lag between their prediction of a COVID-19 spike and the actual COVID-19 spikes. Another limitation we found between both of these papers in the twitter data they used was that there may be a bias in the tweet data of the age of the people who are tweeting, as well as a bias of the socioeconomic class of people tweeting as both papers do not take into consideration of people who did not have access to tweet or search these highly correlated symptoms (e.g. cough and fever) of COVID-19[3, 14].

Researchers from the Jaipuria Institute of Management researched Twitter Tweet sentiment analysis of COVID-19 related Tweets [5]. Essentially, the authors used a list of English words associated to the eight basic emotions of anger, feat, anticipation, trust, surprise, sadness, joy, and disgust) to classify each tweet as a positive or negative sentiment. Using this approach, they analyzed the ratio of positive to negative COVID-19 related tweets in the month of March of 2020 for 12 countries. They found that most of the 12 countries in this time span had more positive than negative sentiment towards COVID-19 because they had not seen the peak of their wave yet and were more optimistic about fighting the spread of COVID-19. China, however, had more negative sentiment since at this point, they were seeing the peak of the curve of COVID-19 cases, so this was probably due to fact that many people were tweeting their symptoms and frustration with the pandemic leading to negative sentiment. This paper did a good job and can help us in our approach of determining tweet analysis and seeing if there is a correlation between the ratio of positive to negative sentiment and COVID-19 cases in a week[5]. However, the approach this paper took to classify the sentiments of tweets does not take into consideration sarcasm and irony, so when we are incorporating sentiment analysis in our model we will need to take into account the context and number of cases in a week, and the emotion that each tweet was classified from one of the eight emotions to consider a bias weight of tweets that were classified as "positive" but were meant to be sarcastic or ironic.

Some limitations we found that were common between papers was that during period in which they conducted their research or built their prediction models, they did not have much data due to being in the early stages of COVID-19.

Besides social media, recent researches also focused on utilizing human mobility data to detect COVID-19 outbreaks in the early stage [7]. Contact mixing is critical in the spread of COVID-19, therefore mobility restrictions of various degrees have been implemented in over 200 countries in the attempt to slow down the spread of the disease. In the paper, the researches found that, using macro-level human mobility data (cell phone mobility data from Israel) along with health improves predictions of when and where COVID-19 outbreaks are likely to occur. The main concern over this type of work is about its privacy, which is addressed by using anonymized data.

REFERENCES

[1] Twitter API. 2021. *Twitter API Documentation.* Retrieved October 5, 2021 from https://developer.twitter.com/en/docs/twitter-api

[2] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2, 3 (2021), 315–324. https://doi.org/10.3390/epidemiologia2030024

[3] I Cheng, Johannes Heyl, Nisha Lad, Gabriel Facini, and Zara Grout. 2021. Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK. *Scientific Reports* 11, 1 (2021), 1–13.

[4] CSSEGISandData. [n. d.]. CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE. https://github.com/CSSEGISandData/

COVID-19
[5] Akash Dutt Dubey. 2020. Twitter Sentiment Analysis during COVID-19 Outbreak. *Available at SSRN 3572023* (2020).
[6] Centers for Disease Control and Prevention. 2021. *COVID Data Tracker.* Retrieved October 5, 2021 from https://covid.cdc.gov/covid-data-tracker/#datatracker-home
[7] Grace Guan, Yotam Dery, Matan Yechezkel, Irad Ben-Gal, Dan Yamin, and Margaret L Brandeau. 2021. Early Detection of COVID-19 Outbreaks Using Human Mobility Data. *medRxiv* (2021).
[8] Michelle L Holshue, Chas DeBolt, Scott Lindquist, Kathy H Lofy, John Wiesman, Hollianne Bruce, Christopher Spitters, Keith Ericson, Sara Wilkerson, Ahmet Tural, et al. 2020. First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine* (2020).
[9] Kaggle. 2021. *COVID-19 Complete Twitter Dataset (daily updates).* Retrieved October 5, 2021 from https://www.kaggle.com/imoore/covid19-complete-twitter-dataset-daily-updates
[10] The Panacea Lab. 2021. *Covid-19 Twitter dataset.* Retrieved October 5, 2021 from https://github.com/thepanacealab/covid19_twitter
[11] Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In *2010 2nd international workshop on cognitive information processing*. Ieee, 411–416.
[12] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* (1980).
[13] Wikipedia. 2021. *COVID-19.* Retrieved October 5, 2021 from https://en.wikipedia.org/wiki/COVID-19
[14] Samira Yousefinaghani, Rozita Dara, Samira Mubareka, and Shayan Sharif. 2021. Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada. *Frontiers in public health* 9 (2021), 359.

TweetCaster: Predicting Pandemic Trend from Tweets

Rachit Bhargava
rachitb@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

Junyan Mao
jmao@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

Pratik Nallamotu
pratiknallamotu@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

In this paper we research the correlation between COVID-19 related tweets and spikes in COVID-19 cases. We propose a machine learning model using natural language processing techniques to analyze keywords and sentiments of COVID-19 related tweets to predict future COVID-19 spikes. We utilized a dataset of COVID-19 related tweets and a data repository of COVID-19 time series data across the globe.

KEYWORDS

machine learning, neural networks, epidemiology, COVID-19, pandemic spread

1 INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease has spread worldwide since its first discovery in Wuhan, China, in December 2019, leading to an ongoing pandemic [17]. The first confirmed case in the United States was discovered on January 20, 2020 [8]. Recently, the death toll of COVID-19 has passed 700,000, with over 43,000,000 total cases in the United States [6].

In such a pandemic, it is important for health professionals to be able to predict upcoming spikes and have rough estimates of the new cases so that they can prepare and distribute necessary medical resources accordingly and make policy adjustment to reflect more strict public health rules. Social media chatters serve as a good data source because there is a strong correlation between COVID-related talks and actual COVID case spikes, and there is usually a lag between people experiencing the COVID-19 symptoms and being officially diagnosed with COVID-19. Therefore, we can use the trend of COVID-related keywords discovered in social media chatters along with other extrapolated information (like, sentiment attached) and metadata (like, location and time of posting the tweet) to effectively analyze prior COVID spikes and use learned knowledge to predict future COVID spikes.

This project aims to analyze posts from Twitter (Tweets) with COVID-19 content to capture the connection between the trend of discussion in Tweets and the actual trend of COVID-19 cases, which will make it able to predict future COVID-19 trends.

2 CHANGES FROM PROPOSAL

Some feedback that we received from the proposal was to focus on using Regression to classify COVID-19 spikes. Instead of just predicting whether a day was part of a spike or not (binary classification), our model is now a regressor that predicts the number of new cases per 100,000 people in a location. This helped us expand the scope our project to actually look into the numbers and science behind twitter and COVID-19, versus just predicting a binary classification of "spike" or "not spike"

Also another point the professor and teaching assistant's pointed out was about how we were going to be classifying the sentiment of the tweets. Previously we were thinking of using the NRC, a list of English words and their sentiment, to classify sentiment. Instead we are using a python package of TextBlob, which is similar to Vader in that is uses polarity and intensity scores to classify the sentiment of a text phrase.

3 PROBLEM FORMULATION

Given $\mathcal{D} = \{X_i\}_{i=1}^{N}$ micro-blogs (tweets) from real-world users related to the target pandemic (where $X_i = \{a_i, b_i\}$ is the i-th tweet and N is the total number of tweets, with a_i being the vector representation of the i-th tweet and b_i being the date on which the tweet was posted), and $C = \{y_{l,j}\}_{i=1,j=1}^{L,M}$ data about the number of cases per day (where $y_{l,j}$ is the number of cases on the j-th day from when the data started getting recorded for the l-th location and M is the index of the last day on which the data got recorded and L is the index of the last location from which the data was recorded), we are interested in finding a set $\hat{\ddagger}$ of expected number of new cases per 100,000 people per location for the next day. In other words, given the tweets from a given location about the pandemic of interest (POI), the data about the number of cases for the POI and a definition for spike in numbers, we are interested in predicting the set of number of new cases that might show up in the upcoming few days.

4 DATA PREPARATION

For this project, we use the "Covid-19 Twitter chatter dataset for scientific use" maintained by the Panacea Lab at GSU[2]. This dataset consists of tweets retrieved from the Twitter Stream related to COVID-19 chatter on a daily basis, from March 22, 2020 to October 12, 2020. The data collected from the Twitter Stream captures all languages with English as the most prevalent, Spanish as the second most prevalent, and French as the third most prevalent.

We first downloaded the whole dataset from [3], which was about 15 Gigabytes in size and largely exceeded our computing power. To facilitate our research, we used two criteria to filter the raw data: 1) only take data from 2021. 2) only take data with geo information attached, and from the U.S. This gave us 361661 total records to work with. We believe this amount of data is sufficient to provide valuable information about the trend because it is able to capture the second spike and valley of COVID-19 cases in the United States.

The raw data only records "tweet_id", which is the unique identifier used by Twitter to identify tweets. Therefore, we need to hydrate the raw data with their content in order to do text analysis.

To accomplish this, we utilized the Tweepy[16] package, which serves as a wrapper around Twitter's API to provide a smoother experience when working with the API. After some experiment, we found that using the batch GET endpoint was much more efficient than using the single GET endpoint because Twitter blocks access when a user hits 900 requests in a 15-minute interval. Fortunately, Tweepy provides an internal sleeper to help us monitor this threshold, so the hydrating process was fully automatic. Twitter returns an empty response for tweet that has been deleted. We have to account for this case by further filtering out these entries. The final hydrated dataset was 69.4 MB in size, which was reasonably easy to process and analyze.

For Sentiment Analysis, in order to analyze the sentiment of these tweets, we first had to clean up the format of the tweets using regular expressions. We cleaned up the tweets by omitting any mentions in a tweet (signified with an @), hashtags, rt or RT from tweets which signify a user retweeted the tweet, and we removed any tweets that include a URL to process better with text blob.

5 APPROACH

5.1 Tweet Vectorization

In order for computers to understand text information, we have to convert text into a machine-understandable form - vector, this step is typically called vectorization, or encoding.

The first step is to divide tweets into separate words, or "tokens". This step is typically called tokenization. It can be as simple as splitting the text by space, tab, or new line character. Or can be specific to the domain of the text data. For example, Python's Natural Language Toolkit (NLTK) provides a special TweetTokenizer[13] that handles tweet text. It has information about tweet-specific syntax such as that "@" means mention, and that "#" means topic. More recently since the state-of-the-art BERT model was introduced, Huggingface added a BERTweet[9] model, which was pre-trained on a large-scale English tweet dataset, and it also includes a BertweetTokenizer. In this project, we plan to experiment with all three tokenizers to see which one can help us achieve the best performance.

The second step would be to convert the tokens into vectors, which is usually called vectorization or featurization. There are a number of established methods in this area, including bag-of-words, TF-IDF, Word2Vec, BERT, etc. Here's a brief introduction on each of the above:

- Bag of Words (BoW) is very popular because of its simplicity. BoW consists of a dictionary of known words, and a measure (usually count) of known words. This method is concerned with whether a certain word is in a document and its frequency.
- Term Frequency-Inverse Document Frequency (TF-IDF) tends to capture more information than BoW. It also consists of two parts: TF accounts for how frequently a word appears in a document, and IDF measures how infrequently the same word appears in the whole dataset (containing the document). This gives more weight to words which appear in fewer documents and less weight to words which appear in many documents because the word that appears in many documents is unlikely to be important, e.g., stop words.
- Word2Vec is based on a neural network and learns word associations from a large corpus of text, while generating a low dimensionality vector representation of the word, which can be compared with each other to determine semantic similarity by applying cosine similarity.
- Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art NLP model developed by Google. BERT was pre-trained on language modeling and next sentence prediction, so it captures contextual embeddings for words. For this project, we will be using the BERTweet [12] model which has the same architecture as the base BERT and was pre-trained on a large tweet dataset.

We plan to experiment with TF-IDF, Word2Vec, and BERT to see which method can generate the most informative features for our models to learn. For each method, we will be generating two datasets with 200 features and 500 features to see if more features can provide us with more valuable information. Currently, we have generated the two datasets with 200 and 500 features by TF-IDF.

5.2 Sentiment Analysis

Sentiment analysis refers to taking a piece of text and outputting the sentiment of the text with labels such as positive, neutral, or negative. In this project, we will be analyzing tweets and label them as a sentiment. Using this sentiment analysis label for each tweet, we can add this feature to the tweet vectors to note if the sentiment feature is of importance to predicting COVID-19 spikes.

5.2.1 Algorithms and Libraries. We are using two libraries in python to analyze the sentiment of the tweets that we are processing. **TextBlob**[15] is a NLP library that is able to analzye text and analyze whether a certain text is positive, netural, or negative. The way that it can label text is through three metrics of polarity, subjectivity, intensity. Polarity means how "negative" or "positive" a word is. Subjectivity means whether the text is more opinion based or factual based. Intensity refers to how many modifiers were used in the phrase which could "intensify" the sentiment, such as "very great" would be considered as high intensity and of greater subjectivity. The way TextBlob works for text phrases is that the library has a set dictionary or "lexicon" of words. Each word in the lexicon has a polarity, subjectivity, and intensity score. However, it should also be noted that each unique word in the lexicon can have multiple scores for the metrics as one word can have different levels of sentiment based on the context of the word being used. Thus, if TextBlob wanted to find the sentiment of a single word it would average all the metrics of the single word values from the lexicon and output a sentiment and subjectivity score. For text phrases, TextBlob applies the same logic for each word in the phrase and combines polarity score. The phrase is labled negative if the polarity score is < 0, neutral if the polarity score is = 0, and positive if the polarity score is > 0. It should also be noted that polarity values are between -1 to 1, and subjectivity values are between 0 to 1.

We are using the **WordCloud**[11] library in our project to analyze the cluster of words and frequency of these words in these COVID-19 related tweets. This can better help us understand how tweets and the words that were common between tweets during January-August correlate to the number of COVID cases.

5.2.2 Initial Findings. We analyzed the sentiment of 361661 tweets the dataset of January 1, 2021 to August, 23, 2021. After cleaning up the text of the tweets for Sentiment Analysis, we applied TextBlobs subjectivity and polarity functions on each row of text and assigned a sentiment score of negative is the polarity score was < 0, neutral if the score was = 0, and positive if the score was > 0. After labeling each record of COVID-19 related tweets from the dataset we found that:

- 36% of the tweets were classified as Positive
- 45% of the tweets were classified as Neutral.
- 19% of the tweets were classified as Negative.

This shows that almost half of the tweets were considered as neutral and have conflicting sentiment of the words in these tweets.

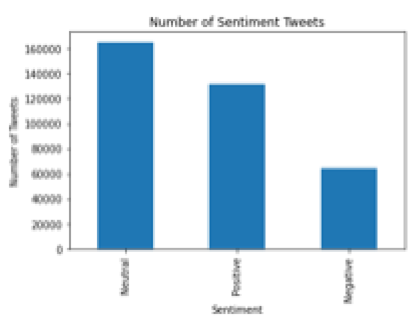

Figure 1: Frequency of Sentiment of the Tweets

Now analyzing the trend of positive tweets for these months:

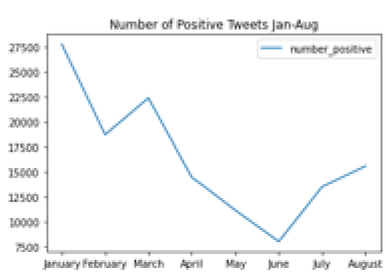

Figure 2: Number of Positive Tweets by month

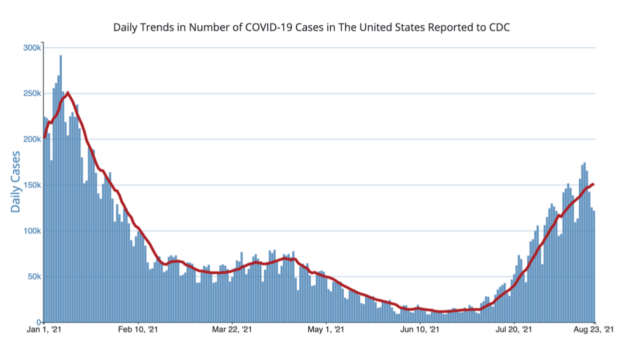

Figure 3: Number of COVID Cases by month [1]

Figure 2 shows the trend of the number of positive tweets from January to August of 2021, and Figure 3 shows the trend of COVID-19 cases from January to August of 2021. From comparing these two graphs, we can see that both graphs follow a similar trend where the number of positive tweets and cases decrease from January to early June, and then start increasing again from June to August 23, 2021.

5.2.3 Related Work on Sentiment Analysis. Researchers from the Jaipuria Institute of Management researched Twitter Tweet sentiment analysis of COVID-19 related Tweets [5]. Using this approach, they analyzed the ratio of positive to negative COVID-19 related tweets in the month of March of 2020 for 12 countries. They found that most of the 12 countries in this time span had more positive than negative sentiment towards COVID-19 because they had not seen the peak of their wave yet and were more optimistic about fighting the spread of COVID-19. China, however, had more negative sentiment since at this point, they were seeing the peak of the curve of COVID-19 cases, so this was probably due to fact that many people were tweeting their symptoms and frustration with the pandemic leading to negative sentiment. However, the approach this paper took to classify the sentiments of tweets does not take into consideration sarcasm and irony, but the library TextBlob does take into account subjectivty which can account for errors due to sarcasm and irony from modifier words.

5.3 Metadata Extraction

To get more information that might be helpful for extrapolating what the number of new cases for some location, we use the available metadata from Twitter's API. This metadata includes,

- Type of tweet expansion, if any (polls, mentions, geolocation), represented by an integer value denoting the type of tweet expansion,
- Type of media included, if any (images, URL's, alternative text), represented by an integer denoting the type of media included,
- Author ID,
- Geolocation information, represented by the location's coordinates (latitude and longitude),
- Tweet creation timestamp, represented by a POSIX timestamp integer value,
- Language of tweet, represented by an integer value denoting the language,
- Referenced tweet, represented by an integer containing the tweet ID,
- Author ID of original tweet poster, if the tweet is a reply,
- Number of likes,
- Number of replies,
- Number of retweets, and
- Number of tweets referenced.

5.4 Vector Creation for Neural Network

We merge all the data discussed thus far in one vector. This vector contains the original tweet vector generated earlier, the sentiment information (represented by an integer that takes the value of -1, 0 or 1, for negative, neutral or positive respectively) and the tweet metadata that we retrieve from Twitter's API. We then merge the tweets from the same day and location to create a 2D vector which can be used by the neural network to predict the target number. Finally, we zero-pad the sequences to have the same dimensions in 2D space. The vectors are then merged along the third dimension to create a series of 3D vectors, each vector representing the tweets posted from any given location on some given day.

5.5 Neural Network Architecture

We use a multi-headed transformer model to predict the numbers for each location on the current and subsequent days. Since activity in different locations or activity on some prior date might influence the number of new cases in a certain region, the self-attention layer takes it into account and calculates the weight for each location and date. Ideally, this layer should be able to assign higher weights to locations closer to the queried location and dates closer to the queried date. The model then returns a 2-D vector of size equal to (number of locations, number of days to predict the numbers for). We use the Adagrad optimizer with a learning rate of 0.01 and no weight decay along with an optimizer scheduler capable of changing the learning rate of the optimizer from 0.01 to 0.0001, as needed. We set the number of heads of the transformer to 8, the number of encoder and decoder layers to 6 each and the dropout rate to 0.1.

6 EVALUATION

To evaluate the effectiveness of our prediction model, we will compare the trend we discover from the tweets to the actual COVID-19 trend (deaths and cases) in a similar fashion described in [4], to see how closely the Twitter data is correlated with the COVID-19 data and see how far ahead tweet information is able to forecast an upcoming outbreak.

We also plan to compare out model against other state-of-the-art works to see if there is any opportunities for improvement, including [14], which used data from Twitter and Google Trend to detect influenza in Greece, and [4, 10], which used Twitter hashtag data to predict upcoming outbreaks of pandemics in the UK.

7 REMAINING WORK

7.1 More Accurate Location for Tweet

Currently, we are filtering all tweet data in the United States which would provide us with a general national trend. However, it could be worthwhile to look into more detailed geo data provided in the response from the Twitter API as it can help us generate more detailed analysis that is specific to a certain region/state, e.g., geo.place_id and geo.coordinates. However, due to Twitter API's regulations and the fact that not all users turned on the "accurate location" feature when they sent a tweet, it can be hard to obtain enough amount of tweet data for us to work with.

7.2 Sentiment Analysis

For the sake of this problem and understanding if sentiment has importance on determining a COVID-19 spike, we can train our model on just the tweet vectors that were classified as only as either positive or negative to see if sentiment is a factor. Then we can compare this accuracy to training the model on tweet vectors that were classified only as neutral.

7.3 Data for Final Vector

As of now, we have just the vectorized tweet, sentiment analysis result and some tweet metadata that is used to predict the normalized number of new cases. We plan to test how much information can our neural network extrapolate from just the available data and add more information related to the tweet by further preprocessing it (for example, frequency of certain popular bigrams and trigrams).

7.4 Neural Network Architecture

We are currently using a transformer model to get information from different time series, but we still need to test the architecture to check if we can extrapolate the required information. We plan to test other RNN models (since RNN's are able to accurately capture information past time steps) and go ahead with what might work the best.

8 SURVEY OF RELATED WORK

There have been multiple researches that are focused on alternative approaches to evaluate an emerging health crisis like COVID-19, given that nationwide mass testing is not feasible in the majority of countries. This phenomenon of social media's ability to provide insights into pandemic activity has been referred to as "wisdom of the crowds", where the collective knowledge and experience of individual users are used to predict trend of a pandemic in the absence of a large-scale virus tracking system [4].

Some of the papers we read had different approaches, but a common objective of using COVID-19 symptom related tweets or searches in predicting COVID-19 spikes. Researchers from the University of Guleph [18] and researchers from the University College London (UCL) [4] both had similar approaches of analyzing features within Twitter datasets and identifying anomalies within the number of tweets related to COVID-19 symptoms to predict when the wave or spike of COVID would happen. We can use insights and results made from both of these papers in helping us detect anomalies and correlations between keywords in the twitter dataset and COVID-19 cases to predict a future spike in COVID-19. A weakness from the paper by the researchers from the University of Guleph [18] was that they used twitter data when testing was not as prevalent because it was the beginning of the pandemic, and also regions had travel-based policies that were not identified, which could explain some of the delayed lag between their prediction of a COVID-19 spike and the actual COVID-19 spikes. Another limitation we found between both of these papers in the twitter data they used was that there may be a bias in the tweet data of the age of the people who are tweeting, as well as a bias of the socioeconomic class of people tweeting as both papers do not take into consideration of people who did not have access to tweet or search these highly correlated symptoms (e.g. cough and fever) of COVID-19[4, 18].

Researchers from the Jaipuria Institute of Management researched Twitter Tweet sentiment analysis of COVID-19 related Tweets [5]. Essentially, the authors used a list of English words associated to the eight basic emotions of anger, feat, anticipation, trust, surprise, sadness, joy, and disgust) to classify each tweet as a positive or negative sentiment. Using this approach, they analyzed the ratio of positive to negative COVID-19 related tweets in the month of March of 2020 for 12 countries. They found that most of the 12 countries in this time span had more positive than negative sentiment towards COVID-19 because they had not seen the peak of their wave yet and were more optimistic about fighting the spread of COVID-19. China, however, had more negative sentiment since at this point, they were seeing the peak of the curve of COVID-19 cases, so this was probably due to fact that many people were tweeting their symptoms and frustration with the pandemic leading to negative sentiment. This paper did a good job and can help us in our approach of determining tweet analysis and seeing if there is a correlation between the ratio of positive to negative sentiment and COVID-19 cases in a week[5]. However, the approach this paper took to classify the sentiments of tweets does not take into consideration sarcasm and irony, so when we are incorporating sentiment analysis in our model we will need to take into account the context and number of cases in a week, and the emotion that each tweet was classified from one of the eight emotions to consider a bias weight of tweets that were classified as "positive" but were meant to be sarcastic or ironic.

Some limitations we found that were common between papers was that during period in which they conducted their research or built their prediction models, they did not have much data due to being in the early stages of COVID-19.

Besides social media, recent researches also focused on utilizing human mobility data to detect COVID-19 outbreaks in the early stage [7]. Contact mixing is critical in the spread of COVID-19, therefore mobility restrictions of various degrees have been implemented in over 200 countries in the attempt to slow down the spread of the disease. In the paper, the researches found that, using macro-level human mobility data (cell phone mobility data from Israel) along with health improves predictions of when and where COVID-19 outbreaks are likely to occur. The main concern over this type of work is about its privacy, which is addressed by using anonymized data.

REFERENCES

[1] [n. d.]. CDC COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker/#trends_dailycases

[2] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2, 3 (2021), 315–324. https://doi.org/10.3390/epidemiologia2030024

[3] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. *A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.* https://doi.org/10.5281/zenodo.5637848 This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates. Release: We have standardized the name of the resource to match our pre-print manuscript and to not have to update it every week..

[4] I Cheng, Johannes Heyl, Nisha Lad, Gabriel Facini, and Zara Grout. 2021. Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK. *Scientific Reports* 11, 1 (2021), 1–13.

[5] Akash Dutt Dubey. 2020. Twitter Sentiment Analysis during COVID-19 Outbreak. *Available at SSRN 3572023* (2020).

[6] Centers for Disease Control and Prevention. 2021. *COVID Data Tracker.* Retrieved October 5, 2021 from https://covid.cdc.gov/covid-data-tracker/#datatracker-home

[7] Grace Guan, Yotam Dery, Matan Yechezkel, Irad Ben-Gal, Dan Yamin, and Margaret L Brandeau. 2021. Early Detection of COVID-19 Outbreaks Using Human Mobility Data. *medRxiv* (2021).

[8] Michelle L Holshue, Chas DeBolt, Scott Lindquist, Kathy H Lofy, John Wiesman, Hollianne Bruce, Christopher Spitters, Keith Ericson, Sara Wilkerson, Ahmet Tural, et al. 2020. First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine* (2020).

[9] Huggingface. 2021. *BERTweet.* Retrieved November 5, 2021 from https://huggingface.co/transformers/model_doc/bertweet.html

[10] Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In *2010 2nd international workshop on cognitive information processing*. Ieee, 411–416.

[11] Zolzaya Luvsandorj. 2021. Simple wordcloud in Python. https://towardsdatascience.com/simple-wordcloud-in-python-2ae54a9f58e5

[12] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200* (2020).

[13] NLTK. 2021. *NLTK.tokenize.TweetTokenizer.* Retrieved November 5, 2021 from https://www.nltk.org/api/nltk.tokenize.casual.html?highlight=tweettokenizer#nltk.tokenize.casual.TweetTokenizer

[14] Loukas Samaras, Elena García-Barriocanal, and Miguel-Angel Sicilia. 2020. Comparing Social media and Google to detect and predict severe epidemics. *Scientific reports* 10, 1 (2020), 1–11.

[15] Parthvi Shah. 2020. My Absolute Go-To for Sentiment Analysis-TextBlob. https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

[16] Tweepy. 2021. *Tweepy: Twitter for Python!* Retrieved November 5, 2021 from https://github.com/tweepy/tweepy

[17] Wikipedia. 2021. *COVID-19.* Retrieved October 5, 2021 from https://en.wikipedia.org/wiki/COVID-19

[18] Samira Yousefinaghani, Rozita Dara, Samira Mubareka, and Shayan Sharif. 2021. Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada. *Frontiers in public health* 9 (2021), 359.

TweetCaster: Predicting Pandemic Trend from Tweets

Rachit Bhargava
rachitb@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

Junyan Mao
jmao@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

Pratik Nallamotu
pratiknallamotu@gatech.com
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

In this paper we research the correlation between COVID-19 related tweets and number of new COVID-19 cases. We propose a machine learning model using natural language processing techniques to analyze keywords and sentiments of COVID-19 related tweets to predict future COVID-19 spikes. We utilized a dataset of COVID-19 related tweets and a data repository of COVID-19 time series data from the United States.

KEYWORDS

machine learning, neural networks, epidemiology, COVID-19, pandemic spread

1 INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease has spread worldwide since its first discovery in Wuhan, China, in December 2019, leading to an ongoing pandemic [18]. The first confirmed case in the United States was discovered on January 20, 2020 [9]. Recently, the death toll of COVID-19 has passed 700,000, with over 43,000,000 total cases in the United States [7].

In such a pandemic, it is important for health professionals to be able to predict upcoming spikes and have rough estimates of the new cases so that they can prepare and distribute necessary medical resources accordingly and make policy adjustment to reflect more strict public health rules. Social media chatters may serve as a good data source for this task because there is a strong correlation between COVID-related talks and actual COVID case spikes, and there is usually a lag between people experiencing the COVID-19 symptoms and being officially diagnosed with COVID-19. Therefore, we can use the trend of COVID-related keywords discovered in social media chatters along with other extrapolated information (for example, sentiment attached) and metadata (for example, time of posting the tweet) to effectively analyze prior COVID case trends and use learned knowledge to predict future new COVID cases.

This project aims to analyze posts from Twitter (tweets) with COVID-19-related content to capture the connection between the trend of discussion in Tweets and the actual trend of COVID-19 cases, which will make it able to predict future COVID-19 trends.

2 RESPONSE TO MILESTONE COMMENT

From our milestone report, we received three main comments.

The first comment was in regards to one of the references we used. We omitted this reference and any discussion we had using this reference as a source. Instead, we replaced it with a more reputable source about the applications of sentiment analysis in fighting COVID-19.

The second comment we received was redundancy in one of our paragraphs in our approach section and related work section. We fixed this by removing this paragraph since it related to the source from the first comment. We also gave the entire report another pass to remove any redundant parts.

The third comment sought more results. We had only visual results on data processing and sentiment analysis of the dataset. We addressed this comment by running a series of experiments aimed at finding some correlation between data we could extract from the tweets and the new COVID-19 cases. These experiments include, running our dataset through a linear regression model with Sentiment Analysis, some time series models, an ensemble model combining our linear regression and time series models, and a TF-IDF and BERT encoding of tweets passed into a neural network to predict COVID-19 cases.

3 PROBLEM FORMULATION

Given $\mathcal{D} = \{X_i\}_{i=1}^{N}$ micro-blogs (tweets) from real-world users related to the target pandemic (where $X_i = \{a_i, b_i\}$ is the i-th tweet and N is the total number of tweets, with a_i being the vector representation of the i-th tweet and b_i being the timestamp at which the tweet was posted), and $C = \{z_j\}_{i=1,j=1}^{M}$ data about the number of cases per day (where z_j is the number of cases on the j-th day from when the data started getting recorded for the United States and M is the index of the last day on which the data got recorded), we are interested in finding $\hat{y}$, the expected number of new cases in the United States for the seventh day from the present day. In other words, given the tweets from the United States about the pandemic of interest (POI) and the data about the number of cases for the POI, we are interested in predicting the number of new cases that might show up in seven days.

4 DATA PREPARATION

For this project, we are using the "Covid-19 Twitter chatter dataset for scientific use" maintained by the Panacea Lab at Georgia State University [4]. This dataset consists of tweets retrieved from the Twitter Stream related to COVID-19 chatter on a daily basis, from January 04, 2020 to October 12, 2021 (ongoing). The data collected from the Twitter stream captures all languages with English as the most prevalent, Spanish as the second most prevalent, and French as the third most prevalent.

We first downloaded the whole dataset that Panacea Lab had provided [5], which was roughly 15 Gigabytes in size and largely exceeded our computing power. To facilitate our research, we filtered the raw dataset using two criteria: 1) only take data from the year of 2021 (specifically from March 01, 2021 to August 22, 2021 because this period captures the second spike and valley of COVID-19 cases in the U.S.), and 2) only take data with United States being the country information attached. This gave us 361661 total records to work with. We believe this amount of data is sufficient to provide valuable information about the trend because it is able to capture the second spike and valley of COVID-19 cases in the United States.

The raw data only records “tweet_id”, which is the unique identifier used by Twitter to identify tweets. Therefore, we attached the content to the “tweet_id” (hydration) in order to do any text-related analysis. To accomplish this, we utilized the Tweepy[17] library, which serves as a wrapper around Twitter’s API to provide a smoother experience when working with the API. After some experiment, we found that using the batch GET endpoint was much more efficient than using the single GET endpoint because Twitter temporarily blocks access when a user hits 900 requests in a 15-minute interval. Fortunately, Tweepy provides an internal sleeper to help us monitor this threshold, so the hydrating process was fully automatic. Twitter returns an empty response for tweet that has been deleted. We account for this case by further filtering out such entries. The final hydrated dataset was 69.4 MB in size, which was reasonably easy to process and analyze. We originally planned to also retrieve the geo information about a tweet to do some further analyses. However, this turned out to be infeasible because there was rarely specific geolocation information (coordinates) attached to a tweet and we could only get the “place_id” of a tweet. This is problematic because Twitter hosts their geolocation data on the Standard V1.1 API that allows up to 75 requests every 15 minutes.

For sentiment analysis, in order to analyze the sentiment of these tweets, we first had to clean up the format of the tweets using regular expressions. We cleaned up the tweets by omitting any mentions in a tweet (signified with an @), hashtags, rt or RT which signify a user retweeted the tweet, and we removed any tweets that include a URL to process better with text blob.

We were able to get COVID-19 cases data for the USA through the New York Times COVID-19 data repository. This dataset included information on cases and deaths during COVID-19, and by using this data we created a new column of new cases which refers to the difference in number of cases for a current day compared to its previous day.

5 PROPOSED METHODS

5.1 Intuition

We propose a five-step process for forecasting the number of new COVID-19 cases from tweets.

First, we perform sentiment analysis on all tweets available to us in the given date range. Prior work has shown some correlation between social media posts from users and new pandemic or epidemic case numbers [3]. Next, we aggregate these numbers by day and retrieve the number of tweets linked with positive, negative and neutral sentiments.

Second, we retrieve some metadata that might be useful in predicting number of new COVID-19 cases. The first input point we use is the day of the week a tweet is posted. We had a hypothesis that the day of the week when a tweet is posted might directly impact the number of cases reported in exactly seven days from when the tweet is published. As we mention in the experiments section later, this metric is pretty handy for keeping track of the weekends. Since certain states do not report new case numbers over the weekend, our model learned to use this input point to mimic the actual data. The second input point we use is the hour of the day when a tweet is posted. We had a hypothesis that more COVID-19 related tweets might be posted at odd hours in the day if people are not feeling well or if they want to express concerns about people getting sick in areas surrounding them. As we discuss later, our model uses this metric to give better predictions. Also, since we aggregate tweets by day, we get the number of tweets posted at every hour of the day (for example, 12:00 AM - 01:00 AM, 01:00 AM - 02:00 AM, and so on) and present it as a data point.

Third, we vectorize the tweets using a suitable embedding (we use BERTweet for our work) for making the predictions. Due to the limitation of computational resources, finding the BERTweet encoding of each tweet and storing it in memory is not possible. Because of this, we aggregate all the tweets from the same day in one line (multiple sentence paragraph, each tweet separated by a comma) and then encode it. This would allow us to take into account what people might be talking about and not rely on just the sentiment or tweet’s metadata, and it would thus allow our model to be able to make even better predictions.

Fourth, we merge the above-mentioned data together to get vectors for each day.

Finally, we use this vector to pass to our neural network that then predicts the number of new COVID-19 cases to expect in seven days from the current day. Here, we use a rolling window of training data, meaning that only a certain window of training data is available to the model to train itself on and then make a prediction based on the tweets from the day immediately following the window. The prediction is compared with the actual number of new cases and we then move the window forward by one day. The general intuition behind this is that the trend of people talking about the disease continuously changes and what was trending two months back might not be trending anymore. The rolling window helps the model make sure that it does not learn from outdated information.

On a sidenote, we also attempted to vectorize the tweets using a range of embeddings to use for making predictions, but we could not get the best performance from any of them. We hypothesize that BERTweet [13] might be able to provide us better results, but we could not use it due to the limitation of computation resources available to us. Instead, we aggregated all the tweets from the same day into one big paragraph, but this is unable to achieve the best results. We discuss this further in the Experiments section.

5.2 Sentiment Analysis

Sentiment analysis refers to taking a piece of text and outputting the sentiment of the text with labels such as positive, neutral, or negative. In this project, we analyze tweets and label each tweet with its respective sentiment. as a sentiment. Using this sentiment analysis label for each tweet,we wanted to see if analysis would have an impact on predicting To a rise in the number of COVID-19 cases day by day. Thus, to see if sentiment of tweets for a day is correlated to the number of new cases we add three new features to our dataset representing the number of positive,negative, neutral tweets for each day.

Sentiment analysis is important to study in regards to predictions of new COVID-19 cases, because various research have indicated that outbreaks and pandemics could have been predicted and controlled if experts studied social media data [2]. These authors hypothesize that by studying the role of social media we can understand symptoms being exhibited, visits to hospitals, and sentiments to predict trends in COVID-19 cases.

5.2.1 Algorithms and Libraries. We use two libraries in python to analyze the sentiment of the tweets that we pass into our regression and neural network models. **TextBlob**[16] is a NLP library that is able to analyze text and analyze whether a certain text is positive, neutral, or negative. The reason of why we chose to use TextBlob for sentiment analysis is because of it has a very similarly accuracy score compared to VADER, and higher accuracy score than Flair [15]. The way that it can label text is through three metrics of polarity, subjectivity, intensity. Polarity means how "negative" or "positive" a word is. Subjectivity means whether the text is more opinion based or factual based. Intensity refers to how many modifiers were used in the phrase which could "intensify" the sentiment, such as "very great" would be considered as high intensity and of greater subjectivity. The way TextBlob works for text phrases is that the library has a set dictionary or "lexicon" of words. Each word in the lexicon has a polarity, subjectivity, and intensity score. However, it should also be noted that each unique word in the lexicon can have multiple scores for the metrics as one word can have different levels of sentiment based on the context of the word being used. Thus, if TextBlob wanted to find the sentiment of a single word it would average all the metrics of the single word values from the lexicon and output a sentiment and subjectivity score. For text phrases, TextBlob applies the same logic for each word in the phrase and combines polarity score. The phrase is labeled negative if the polarity score is < 0, neutral if the polarity score is = 0, and positive if the polarity score is > 0. It should also be noted that polarity values are between -1 to 1, and subjectivity values are between 0 to 1. The subjectivity score for each tweet helps us understand whether it is more of an opinion or fact. This plays a role into its sentiment classification as subjectivity and intensity are used to help make the sentiment classification for each tweet.

We use the **WordCloud**[12] library in our project to analyze the cluster of words and frequency of these words in these COVID-19 related tweets. This helps us understand how sentiments were being assigned to their respective tweets and the most common words in tweets during this time period.

5.3 Tweet Vectorization

In order for computers to understand text information, we have to convert text into a machine-understandable form - vector, this step is typically called vectorization, or encoding.

The first step is to divide tweets into separate words, or "tokens". This step is typically called tokenization. It can be as simple as splitting the text by space, tab, or new line character. Or can be specific to the domain of the text data. For example, Python's Natural Language Toolkit (NLTK) provides a special TweetTokenizer[14] that handles tweet text. It has information about tweet-specific syntax such as that "@" means mention, and that "#" means topic. More recently since the state-of-the-art BERT model was introduced, Huggingface added a BERTweet[10] model, which was pre-trained on a large-scale English tweet dataset, and it also includes a BertweetTokenizer. In this project, we plan to experiment with all three tokenizers to see which one can help us achieve the best performance.

The second step would be to convert the tokens into vectors, which is usually called vectorization or featurization. There are a number of established methods in this area, including bag-of-words, TF-IDF, Word2Vec, BERT, etc. Here's a brief introduction on each of the above:

- Bag of Words (BoW) is very popular because of its simplicity. BoW consists of a dictionary of known words, and a measure (usually count) of known words. This method is concerned with whether a certain word is in a document and its frequency. We are not using BoW in this project.
- Term Frequency-Inverse Document Frequency (TF-IDF) tends to capture more information than BoW. It also consists of two parts: TF accounts for how frequently a word appears in a document, and IDF measures how infrequently the same word appears in the whole dataset (containing the document). This gives more weight to words which appear in fewer documents and less weight to words which appear in many documents because the word that appears in many documents is unlikely to be important, e.g., stop words.
- Word2Vec is based on a neural network and learns word associations from a large corpus of text, while generating a low dimensionality vector representation of the word, which can be compared with each other to determine semantic similarity by applying cosine similarity. We are not using Word2Vec in this project because it only provides word embeddings, while we need sentence embeddings. It's worth mentioning that there is a method called Doc2Vec, which can be thought of as a sentence version of Word2Vec.
- Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art NLP model developed by Google. BERT was pre-trained on language modeling and next sentence prediction, so it captures contextual embeddings for words. For this project, we will be using the BERTweet [13] model which has the same architecture as the base BERT and was pre-trained on a large dataset with 850M English tweets.

In this project, we utilize TF-IDF and BERT to generate embeddings that we then use to predict baseline predictions. We use TF-IDF to generate two variations of tweet embeddings using 200 features and 500 features. We also use BERTweet to generate embeddings by running a mean pooling procedure on the "last_hidden_state" returned by the model, which gives us a vector of length 768 for every set of aggregated tweets - this is defined by BERT's internal architecture. The features generated by BERTweet on the original dataset were too large to fit in the RAM of the Google Colab notebook that we used. As a result, we aggregate the tweets for each day into a single line of text, which is simply achieved by grouping the original data by date. This leaves us with a final vector of shape (253, 768), which we are able to fit in the RAM.

5.4 Metadata Extraction

To get more information that might be helpful for predicting the number of new COVID-19 cases, we use the available metadata from Twitter's API. This metadata includes,

- Day of the week the tweet was created, and
- Hour of the day the tweet was created

We think this information may be helpful because there are more reported cases on weekdays than weekends, and more cases during day hours than night hours based on our observation. We share further intuition behind this decision in the Intuition section. While more metadata might have been useful, it was not feasible to access it due to Twitter's API restrictions of not serving more than 75 requests in a window of 15 minutes. To aggregate the data by day, we calculate the number of tweets posted at every hour of the day and present it as 24 different data points. We make no modifications to day of the week because it remains the same for each tweet.

5.5 Vector Creation

After classifying each tweet's sentiment and extracting the relevant metadata, we merge all the data into one vector where each row represents a day from March 1, 2021 until August 22,202, and contains the following features: number of positive tweets, number of negative tweets,number of neutral tweets, and the number of tweets for each hour of the day. We maintain a separate vector that contains the number of new cases per day.

5.6 Neural Network Architecture

The neural network attempts to find a mapping from the data points that we extract from the tweets to the number of new COVID-19 cases. For this, we use a two-layer network, both of the layers being linear layers. We avoid usage of any activation function because it led to poor performance. For our work, we found 20 to be the ideal number for the number of hidden dimensions.

$$H = A \times X + b$$
$$\hat{y} = C \times H + d$$

Here, A and C are the weight matrices and b and d are the bias vectors. H represents the hidden layer.

As discussed earlier, we use a rolling window to predict the number of new COVID-19 cases for the seventh day from the present day. Let us refer to the rolling window size by r. We also determine the number of days in future for which we want the prediction. Let us refer to this number by f.

To give an example of how this method works, let's say we start at day t. Here, we can assume that we know everything about days before the day t. Since we are at day t, we are looking for the number of cases that might be reported on day $t+f$. While we train our model, we cannot have it go past day t because we never have future data available. As a result, the last prediction that the model can make must be of day $t-1$. Since the model predicts f days in the future, we limit one end of the rolling window to be at $t-1-f$. As a result, the other end of the rolling window would need to be at $t-1-f-r$ because the size of the rolling window is set to be r. Hence, we would have the rolling window (or otherwise referred to as the x matrix) would have data about tweets from day $t-1-f-r$ until day $t-1-f$. The ideal values that we attempt to match for such a rolling window would be the number of COVID-19 cases for day $t-1$ after seeing the data about tweets from day $t-1-f$. This is further illustrated in figure 1. Once the model has been trained, we use the data about tweets from day t to predict the number of the new COVID-19 cases on day $t+f$.

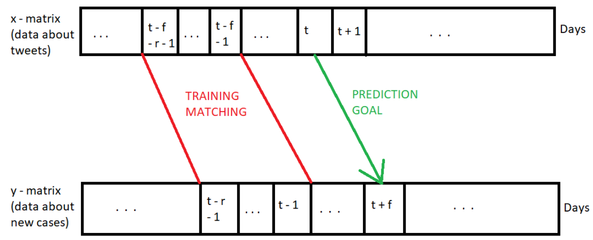

Figure 1: Figure illustrating how rolling window works

For training our model, we use the Adam optimizer [11] with a learning rate of 0.1 and number of epochs for each training session (for each rolling window) to be 10. We set the rolling window size to be 7 days (we discuss why we chose the size to be 7 days in the experiments section) and the number of days for which to make the prediction to also be 7 days.

6 EXPERIMENTS, RESULTS & EVALUATION

6.1 Evaluation

For our work, we use root mean squared error (RMSE) to get a measure of how far our model is from the goal. This is often also referred to as our loss function.

6.2 Analysis of Tweet Sentiments

We analyze the sentiment of 361661 tweets the dataset of March 1, 2021 to August, 22, 2021. After cleaning up the text of the tweets for sentiment analysis, we apply TextBlob's subjectivity and polarity functions on each row of text and assign a sentiment score of negative if the polarity score is less than 0, neutral if the score is equal to 0, and positive if the score is greater than 0.

As seen in figure 3, after labeling each of the COVID-19 related tweets from the dataset we found that,

- 36% of the tweets were classified as Positive,
- 45% of the tweets were classified as Neutral, and
- 19% of the tweets were classified as Negative.

This shows that almost half of the tweets are considered to be neutral or have conflicting sentiment of words.

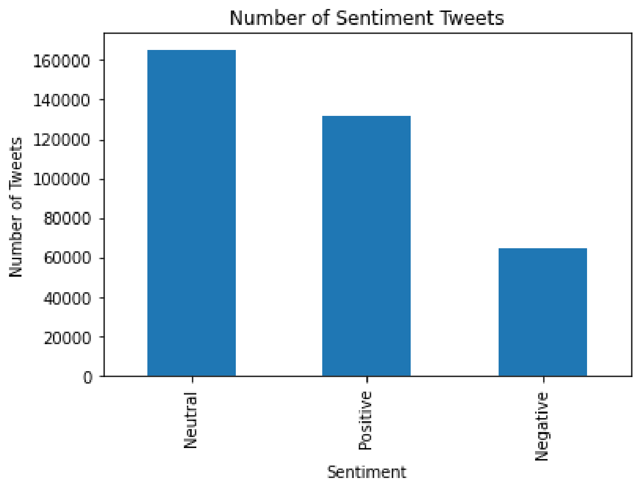

Figure 2: Frequency of Sentiment of the Tweets

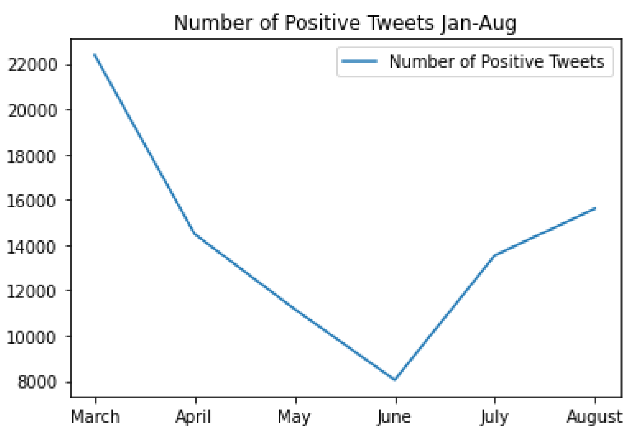

Figure 3: Number of Positive Tweets by month

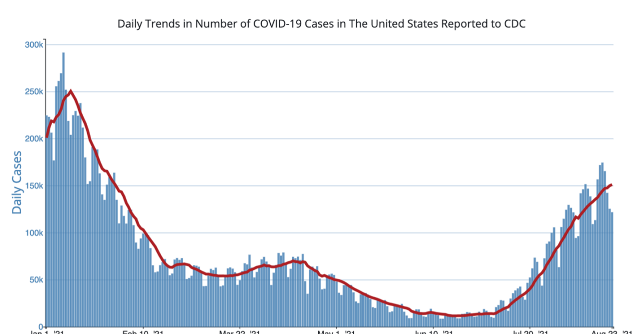

Figure 4: Number of COVID Cases by month [1]

We analyze the trend of positive, negative and neutral tweets over months and compare them with number of new COVID-19

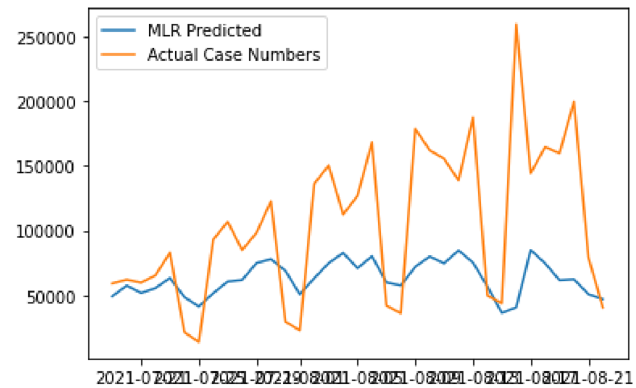

Figure 5: Predicted Number of Cases with Multi-variable Linear Regression with Sentiment Analysis Data

cases and found some pattern between the new COVID-19 cases and positive tweets. Figure 3 shows the trend of the number of positive, negative and neutral tweets from March to August of 2021, and Figure 4 shows the trend of COVID-19 cases from March to August of 2021. From comparing these two graphs, we can see that both graphs follow a similar trend where the number of positive tweets and cases decrease from March to early June, and then start increasing again from June to August 23, 2021.

The above-mentioned observation encouraged us to try to regress the number of new COVID-19 cases from just the sentiment information. The overall goal of this experiment was to see if we can find a direct correlation between the tweets' sentiments and the number of new COVID-19 cases.

We do this by performing an 80:20 train-test split and graphing the predictions along with nothing the RMSE value for the method. We use Adam optimizer with a learning rate of 0.1 and 1000 epochs. Figure 5 shows our observation. The RMSE recorded for the test set was 68441.

Interpretation: Such high RMSE value is probably a result of lack of information. While the model is able to get a sense of rise and fall in number of cases over the week and weekend (probably through change in number of tweets over the weekend), it is still unable to predict what might happen on the next day with just the aggregated sentiment analysis data.

6.3 Time Series Models

Since the number of new COVID-19 cases is a time series dataset, we also attempted to fit time series models, like ARIMA, OLS and their ensembles, to predict the new cases. The purpose of this experiment was to check if we can directly use time series models to predict the number of new COVID-19 cases. However, we found that neither of the three approaches could get us close enough to the actual number of new cases. ARIMA got us close to the true number, but it always had a delay in its predictions.

As mentioned earlier, we again use an 80:20 train-test split for the data. We use ARIMA model with hyperparameters being $p = 2, d = 0, q = 2$ because we found them to work the best with COVID-19 dataset. For the ensemble of the two methods, we use the inverse

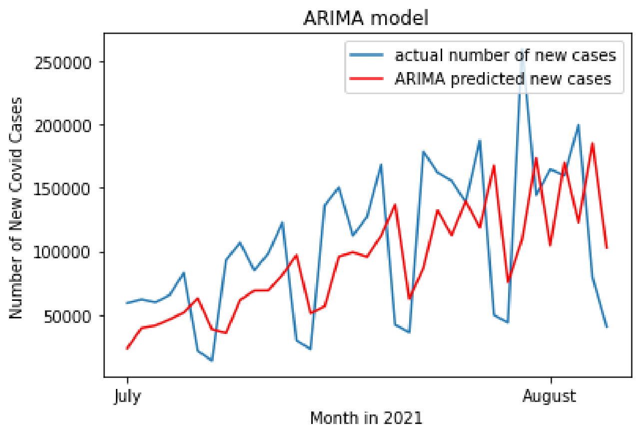

Figure 6: ARIMA Model

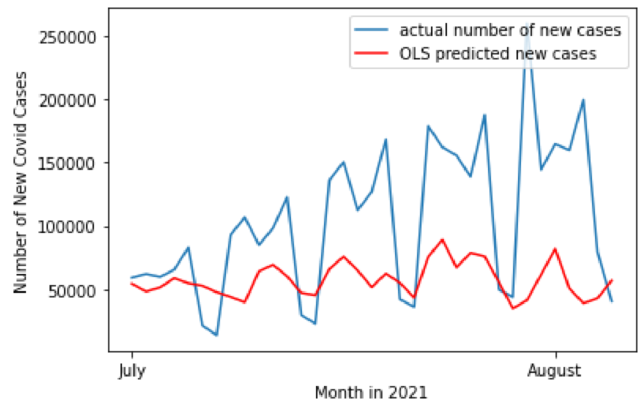

Figure 7: OLS Model

of the RMSE produced by both methods as their weights and then normalize the sum of product of the weight and the prediction of each method.

Figures 6, 7 and 8 display the performance we see with the test dataset. Table 1 shows the RMSE values we found for each method.

Interpretation: While ARIMA and OLS are generally good for fitting time series datasets, they do not account for unknown variables that can cause deviation from the usual trend for such datasets. For example, for COVID-19, such deviations can be a result of introduction of a new COVID-19 variant, some mass gathering or even state or country wide lockdowns. As a result, although ARIMA and OLS predict the drops of cases decently well, they are not able to predict the spikes in cases with high accuracy. ARIMA seems to perform better than OLS probably because ARIMA takes into account the previous time steps and how the target variable is changing over time steps. OLS, on the other hand, is another way to do linear regression which relies on the fact that XX^{-1} is invertible (where X is the input matrix). This might not always be the case which may result in poor performance of OLS. Since both the models seem to underestimate the new COVID-19 case numbers, their ensemble seems to do the same.

To check if adding some sentiment value might help at all, we also try ensembling ARIMA and OLS with the linear regression described in the previous section. Running this experiment while keeping rest of the parameters the same, we get a series of predictions that look like they are getting closer to the actual number of new cases, but they still seem to lack the ability to recognize upcoming spikes. Results of this experiment have been shown in Figure 9 and Table 1.

Interpretation: This might probably be a result of the data available to us because the number of new COVID-19 cases every day start to increase only towards the end of July, while the models are trained on data from March 2021 until the beginning of July 2021, where no such trend is observed.

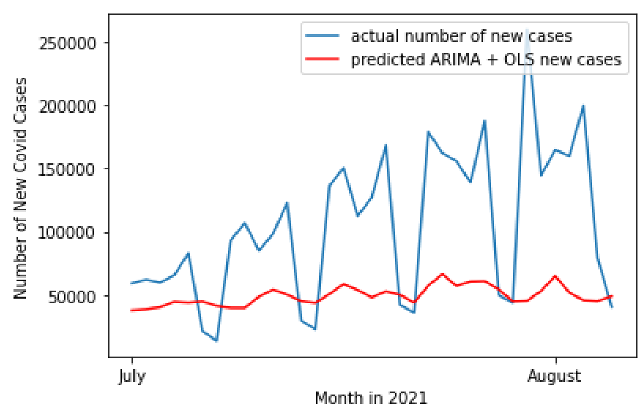

Figure 8: ARIMA + OLS Model

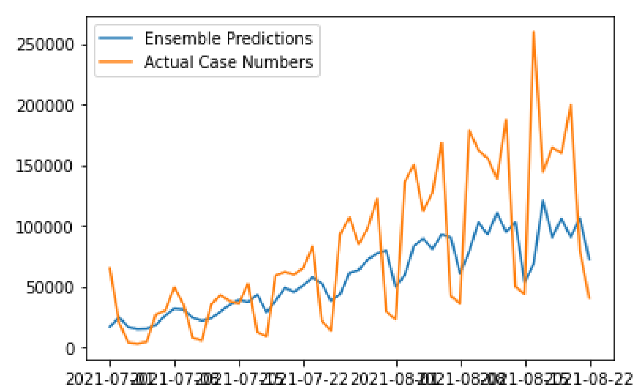

Figure 9: Linear Regression + ARIMA + OLS Model

Table 1: RMSE Values

Model Type	RMSE
ARIMA	58886
OLS	72963
ARIMA + OLS	78597
Linear Regression + ARIMA + OLS	71349

6.4 Neural Network with Tweet Encodings and Metadata

We have been ignoring the contents of the tweets and their metadata thus far. We shall now run an experiment to check if we can get better results if we include the encoded tweets and the metadata discussed earlier. We are also not sure about what encoding method might be the best one to use. As a result, we shall now run experiments with TF-IDF with 200 and 500 features (to see how does increasing the number of features affect the results) and BERTweet to test one of the state-of-the-art ways to encode tweets. For this section, we shall be using the neural network architecture proposed in our work. However, we are not sure about the size of the rolling window yet. Thus, we shall also test it as well with sizes of 7, 14, 21 and 28 days to see what gives the best results. Please note that we take multiples of 7 because they might help our model notice any trends between weeks.

Table 2: RMSE Values - Neural Network with Tweet Encodings

Encoding Type	Rolling Window Size	RMSE
TF-IDF - 200 features	7	25974
TF-IDF - 200 features	14	28504
TF-IDF - 200 features	21	30352
TF-IDF - 200 features	28	30370
TF-IDF - 500 features	7	26025
TF-IDF - 500 features	14	28340
TF-IDF - 500 features	21	29919
TF-IDF - 500 features	28	29893
BERTweet	7	25729
BERTweet	14	27266
BERTweet	21	29243
BERTweet	28	31007

Table 2 shows the results of running the proposed models with the given rolling window sizes.

Interpretation: TF-IDF and BERTweet are able to produce almost equally good results. TF-IDF is probably able to produce similar results with different number of features because the most important 200 features might be able to cover everything that we need to predict the number of new COVID-19 cases because of which adding in more less-important features does not make too much of a difference. As for BERTweet, it is probably not able to do better because we mix the tweets together due to computation limitations. It might be able to do better if we have more computational resources. As far as the size of the rolling window goes, it probably makes sense to limit to a smaller window so that we do not capture outdated trends. This is something that we see from our experiment

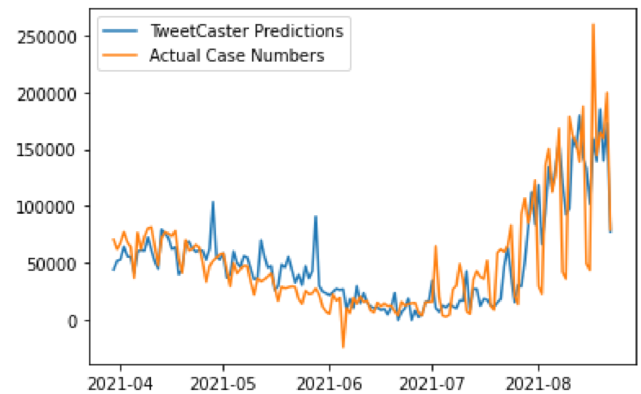

Figure 10: Performance of BERTweet with rolling window of 7 days

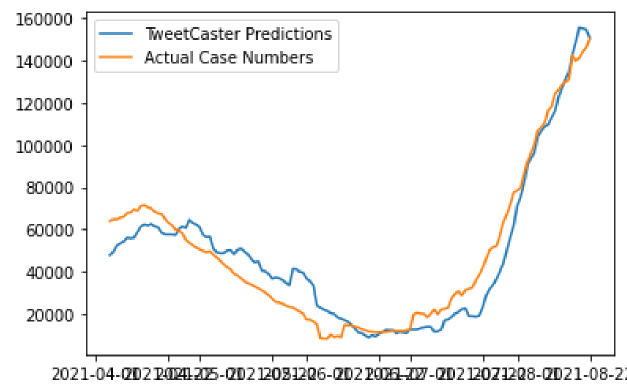

Figure 11: Smoothed graph of performance of BERTweet with rolling window of 7 days

as well where the smaller size rolling window is able to produce better results.

Figures 10 and 11 show how the BERTweet encoding performs with the rolling window of 7 days.

7 SURVEY OF RELATED WORK

There have been multiple researches that are focused on alternative approaches to evaluate an emerging health crisis like COVID-19, given that nationwide mass testing is not feasible in the majority of countries. This phenomenon of social media's ability to provide insights into pandemic activity has been referred to as "wisdom of the crowds", where the collective knowledge and experience of individual users are used to predict trend of a pandemic in the absence of a large-scale virus tracking system [6].

Some of the papers we read had different approaches, but a common objective of using COVID-19 symptom related tweets or searches in predicting COVID-19 spikes. Researchers from the University of Guleph [19] and researchers from the University College London (UCL) [6] both had similar approaches of analyzing features within Twitter datasets and identifying anomalies within the number of tweets related to COVID-19 symptoms to predict when the wave or spike of COVID would happen. We can use insights and results made from both of these papers in helping us detect anomalies and correlations between keywords in the twitter dataset and COVID-19 cases to predict a future spike in COVID-19. A weakness from the paper by the researchers from the University of Guleph [19] was that they used twitter data when testing was not as prevalent because it was the beginning of the pandemic, and also regions had travel-based policies that were not identified, which could explain some of the delayed lag between their prediction of a COVID-19 spike and the actual COVID-19 spikes. Another limitation we found between both of these papers in the twitter data they used was that there may be a bias in the tweet data of the age of the people who are tweeting, as well as a bias of the socioeconomic class of people tweeting as both papers do not take into consideration of people who did not have access to tweet or search these highly correlated symptoms (e.g. cough and fever) of COVID-19[6, 19].

Some limitations we found that were common between papers was that during period in which they conducted their research or built their prediction models, they did not have much data due to being in the early stages of COVID-19.

Besides social media, recent researches also focused on utilizing human mobility data to detect COVID-19 outbreaks in the early stage [8]. Contact mixing is critical in the spread of COVID-19, therefore mobility restrictions of various degrees have been implemented in over 200 countries in the attempt to slow down the spread of the disease. In the paper, the researches found that, using macro-level human mobility data (cell phone mobility data from Israel) along with health improves predictions of when and where COVID-19 outbreaks are likely to occur. The main concern over this type of work is about its privacy, which is addressed by using anonymized data.

8 CONCLUSION

In today's world, we have realized the impact of data science in regards to detecting epidemic outbreaks and case spikes to better help us take actions to minimize them. With the development of new variants of COVID-19, our project can help regional and national officials know when a rise in the number of new COVID-19 cases will happen, and help them promptly develop an action plan and steps to control the rise in cases. In our project, we were able to use Regression and a Neural Network with sentiment analysis and tweet encoding to predict new COVID-19 cases.

9 FUTURE WORK

We plan to make a few improvements to our work in the future:

(1) We were able to run experiments on a date range that would allow us to test with rise and decrease in the number of new COVID-19 cases. We would like to add more training data (a wider date range, more tweets, etc.) to our models and see if we can better observe the trend in COVID-19.
(2) We would also want to extend our model to predict new cases for more countries. As of now, we are predicting new cases for the United States, but the project can be easily extended to another country by switching the language model.
(3) Within the same country, regional prediction is possible given specific location data. This is an important application to research into because it provides more insightful and accurate predictions than the current country-level prediction.
(4) We are also interested in checking how the model would perform if it had a recurrent neural network supporting it because it can somewhat keep track of the past data.

REFERENCES

[1] [n. d.]. CDC COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker/#trends_dailycases
[2] A.H. Alamoodi, B.B. Zaidan, A.A. Zaidan, O.S. Albahri, K.I. Mohammed, R.Q. Malik, E.M. Almahdi, M.A. Chyad, Z. Tareq, A.S. Albahri, Hamsa Hameed, and Musaab Alaa. 2021. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems with Applications* 167 (2021), 114155. https://doi.org/10.1016/j.eswa.2020.114155
[3] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the 2011 Conference on empirical methods in natural language processing*. 1568–1576.
[4] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2, 3 (2021), 315–324. https://doi.org/10.3390/epidemiologia2030024
[5] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. *A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.* https://doi.org/10.5281/zenodo.5637848 This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates. Release: We have standardized the name of the resource to match our pre-print manuscript and to not have to update it every week..
[6] I Cheng, Johannes Heyl, Nisha Lad, Gabriel Facini, and Zara Grout. 2021. Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK. *Scientific Reports* 11, 1 (2021), 1–13.
[7] Centers for Disease Control and Prevention. 2021. *COVID Data Tracker.* Retrieved October 5, 2021 from https://covid.cdc.gov/covid-data-tracker/#datatracker-home
[8] Grace Guan, Yotam Dery, Matan Yechezkel, Irad Ben-Gal, Dan Yamin, and Margaret L Brandeau. 2021. Early Detection of COVID-19 Outbreaks Using Human Mobility Data. *medRxiv* (2021).
[9] Michelle L Holshue, Chas DeBolt, Scott Lindquist, Kathy H Lofy, John Wiesman, Hollianne Bruce, Christopher Spitters, Keith Ericson, Sara Wilkerson, Ahmet Tural, et al. 2020. First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine* (2020).
[10] Huggingface. 2021. *BERTweet.* Retrieved November 5, 2021 from https://huggingface.co/transformers/model_doc/bertweet.html
[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[12] Zolzaya Luvsandorj. 2021. Simple wordcloud in Python. https://towardsdatascience.com/simple-wordcloud-in-python-2ae54a9f58e5
[13] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200* (2020).
[14] NLTK. 2021. *NLTK.tokenize.TweetTokenizer.* Retrieved November 5, 2021 from https://www.nltk.org/api/nltk.tokenize.casual.html?highlight=tweettokenizer#nltk.tokenize.casual.TweetTokenizer
[15] Shahul ESFreelance Data Scientist | Kaggle Master Data science professional with a strong end to end data science/machine learning, deep learning (NLP) skills. Experienced working in a Data Science/ML Engineer role in multiple startups. Kaggle Kernels, Shahul ES, Freelance Data Scientist | Kaggle Master Data science professional with a strong end to end data science/machine learning, deep learning (NLP) skills. Experienced working in a Data Science/ML Engineer role in multiple startups. Kaggle Kernels Master ra, and Follow me on. 2021. https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair
[16] Parthvi Shah. 2020. My Absolute Go-To for Sentiment Analysis-TextBlob. https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524
[17] Tweepy. 2021. *Tweepy: Twitter for Python!* Retrieved November 5, 2021 from https://github.com/tweepy/tweepy
[18] Wikipedia. 2021. *COVID-19.* Retrieved October 5, 2021 from https://en.wikipedia.org/wiki/COVID-19
[19] Samira Yousefinaghani, Rozita Dara, Samira Mubareka, and Shayan Sharif. 2021. Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada. *Frontiers in public health* 9 (2021), 359.