# Covid Cases Forecast Using Twitter Data

Stephone Christian[1]

[1]christian.stephone@gmail.com

December 3, 2021

**Abstract**

This paper details an approach to leveraging Covid related Tweets to forecast State-level Covid case numbers. The forecast is constrained to the United States. Inputs to the model include the daily averaged Glove Vectors of Covid related tweets per state, the state (encoded into a numerical class), and the state's population. The final model consists of a two layer LSTM model with a fully connected output layer. The model was able to attain an $r^2$ value of 0.99, as well as a MAPE (mean absolute persentage error) score of 0.11. The strong performance of the model suggests that similar data and model architectures can be leveraged to forecast Covid cases at higher geographic granularity, and potentially provide farther in future forecasts.

## 1 Introduction

Covid-19 in the United States has had a drastic effect on our current day. It has taken lives, separated families, and changed our every day interactions. Thanks to herculean efforts by researchers and first responders, interventions such as social distancing and vaccine distribution provide hope during these bleak times. In the United States, unfortunately, taking such actions as to minimize the impact of Covid-19 are highly political, and are thus influenced by public opinion and sentiment. The purpose of this project is to understand whether social media, specifically twitter, can be used to understand Covid-19 sentiment, and whether Covid-19 related tweets can be used to predict number of cases at the state level in the United States.

## 2 Dataset

The dataset for this investigation was a combination of 4 datasets. (1) the covid twitter dataset, (2) Population data from the United States Census, (3) Covid case numbers by state from the NY Times, and (4) pre-trained GloVe vectors provided by Stanford NLP researchers[1]. The twitter dataset consists of Covid related tweet Ids compiled by researchers at University of Melbourne[2]. In the data-processing step of this investigation, these tweet

Ids were hydrated using the Twitter API, and filtered based on whether 1. latitude and Longitude values were available and 2. the tweet was made by a User in the United States. 25 dimensional GloVe vectors pre-trained on the Twitter Corpus were used in this investigation. The Twitter corpus consists of 27B tokens and a vocabulary size of 1.2M. Lastly, state-level population for the 2019 Census[3] and Covid case rates from the NY Times[4] were collected. These datasets were joined together and spanned the months of October and November, 2020. Only two months were collected given the slow processing time due to the Twitter API Rate Limiter.

## 2.1 Dataset Introduction and Exploration

After data filtering and stitching as described above, 20k data points over the months of October and November from 2020 were collected. The distribution of datapoints where roughly proportional to the state population, with the exception of the state of Louisiana. This particular state was unfortunately excluded, give the poor quality of data that the prepossessing step yielded. Below is a word cloud from a uniformly sampled subset of the hydrated twitter data. What's interesting are the regional names that appear. While the distribution of the data is fairly uniform (once normalized by population), it's possible that the news coverage of the pandemic was unequally distributed, leading to individuals in other states tweeting about the states of national focus (California, New York, Florida, etc)

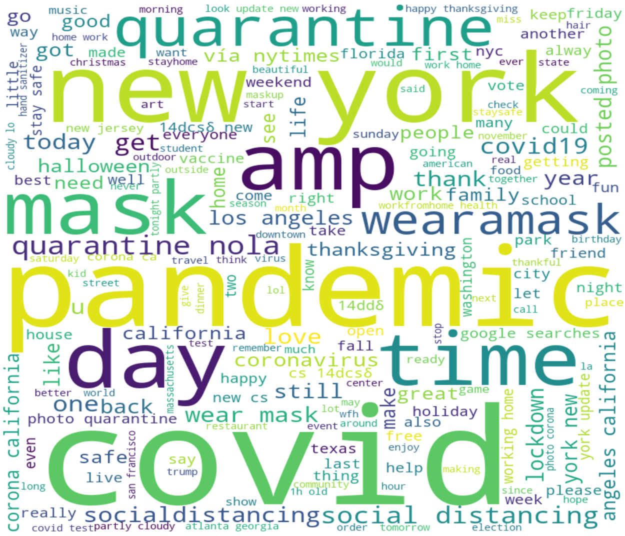

Figure 1: Covid Tweet Analysis

Tweet sentiment is also a promising angle of investigation. Stop words from the tweet text were removed, and the polarity of the text was scored using an open source lexicon and rule-based sentiment analysis tool, VADER[5]. The following plot was created using the complete dataset

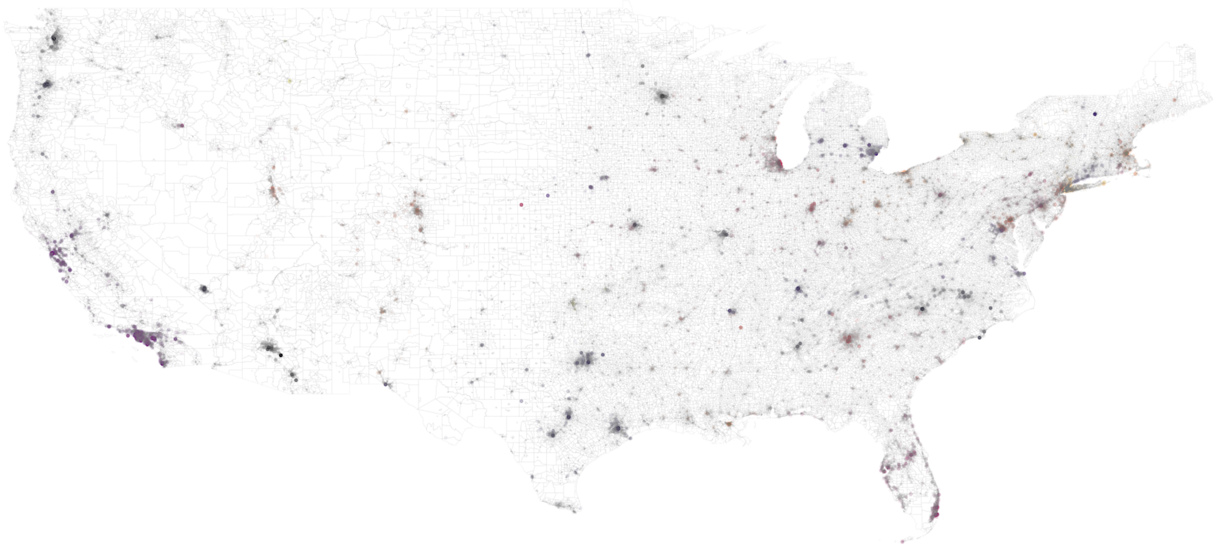

Figure 2: Sentiment Analysis and Mapping

# 3 Approach

### 3.0.1 Data Preprocessing

Since the investigation was performed on a local machine, GloVe vector representation of the text was computed during the data preprocessing step, instead of employing an embedding layer in the network during the training process. The GloVe vector was memory-mapped and stored on disk using numpy, and a vocabulary dictionary was created to map a word to the index of the GloVe vector representation. For each state and each date, all of the tokens in the tweets where mapped to GloVe vectors and averaged. The dimension of a single training example was 28 x 1 (the date, in epoch format, the state, encoded as a number, the population of the state, and the 25 dimensional GloVe vector). State abbreviations were sorted in alphabetical order, and mapped to their index in the sorted list. Preprocessing step yielded torch variables X and y that were then fed into the network using the built in Torch DataLoader.

## 3.1 Model Trianing and Evaluation

The problem was framed as a regression analysis one. given data from day $t_0$ for a given state, the y values of interest where the number of Covid cases for a given states for days $t_1$,$t_2$,$t_3$,$t_4$,$t_5$. Three models were evaluated for the task. (1) a simple Linear regression model as a base model, (2) a single layer LSTM and (3) a 2 layer LSTM with a fully connected layer. The LSTM models were implemented in pytorch, and the linear regression was implemented using open source sklearn modules. Each model was evaluated using rolling cross validation, a cross validation technique compatible with timeseries data. 4 folds where evaluated, and the metrics from each fold averaged to yield the final metric scores. Given the relatively small amount of data, a validation set was forgone for soley a train / test splot with 4 fold cross validation. The following table describes the fold partition scheme.

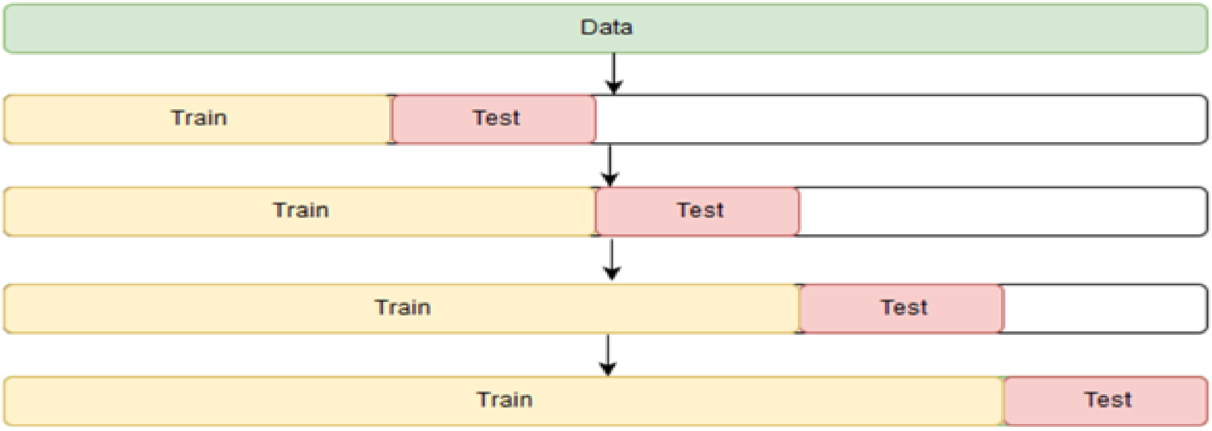

Figure 3: Rolling Cross Validation for Time-Series data

| | Training Data Percentage Split | Test Training Percentage Split |
|---|---|---|
| 1st Fold | [0, 0.25) | [0.25, 0.40) |
| 2nd Fold | [0, 0.40) | [0.40, 0.55) |
| 3rd Fold | [0, 0.60) | [0.60, 0.75) |
| 4th Fold | [0, 0.80) | [0.80, 1.00] |

Table 1: Rolling Cross Validation, 4 folds

### 3.1.1 Linear Model

This base line model was deployed to get a general sense of how a simple model could perform. A linear model for data in each state was created through Ordinary Least Squares. For each state, the fold test metrics were evaluated and averaged. The performance of 50 linear models, provides a coarse lower bound for the performance of a single machine learning network tasked with creating regression outputs for 50 states.

| | MAPE score | $r^2$ score |
|---|---|---|
| Linear Model | 0.60 | 0.618 |

Table 2: Linear Model evaluation metrics

### 3.1.2 Single Layer LSTM

The second model that was developed in pytorch is a single layer LSTM. For this model, a parameter search for the number of hidden units, learning rate, batch size, and dropout layer keep rate was performed. The best results were yielded from 256 hidden units, a learning rate of .005, batch size of 64, and a drop out layer with a keep rate of 0.20. Adam optimization was used for batch gradient descent. Again, rolling cross validation with 4 folds was used on the test set to evaluate the model.

| | MAPE score | $r^2$ score |
|---|---|---|
| Single Layer LSTM | 0.16 | 0.88 |

Table 3: Single Layer LSTM evaluation metrics

### 3.1.3 Stacked LSTM with FC layer

The final model that was explored was a 2 layer stacked LSTM with a fully connected (FC) layer. While the previous single layered LSTM had promising results, a more complex model was considered to potentially capture more information from the data. After a grid search for hyper parameter tuning, a batch size of 32, hidden size of 128, learning rate of 0.009, and a drop out rate of 0.25.Adam optimization was used for batch gradient descent. Below are the final metrics on the test set, again using rolling cross validation.

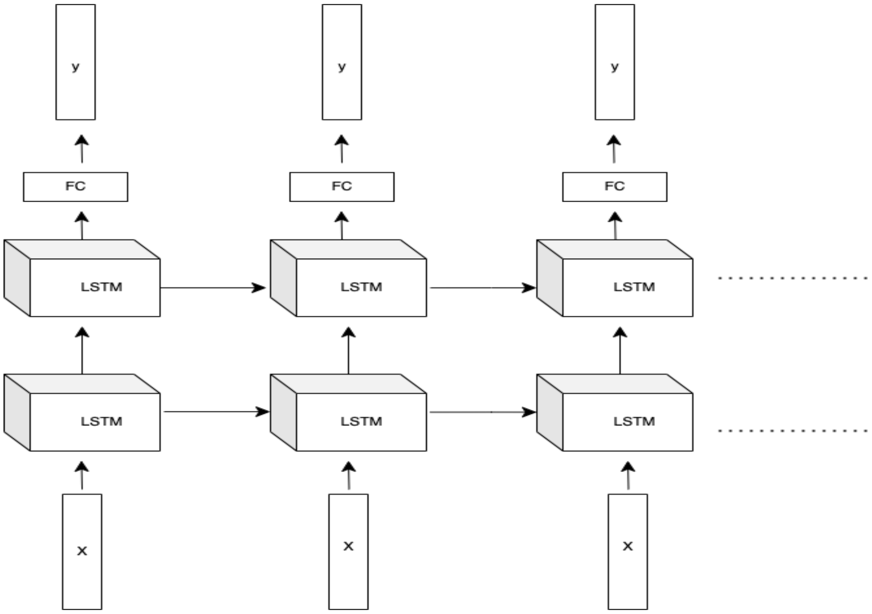

Figure 4: Final Model Architecture

| | MAPE score | $r^2$ score |
|---|---|---|
| 2 Stacked LSTM w/ FC layer | 0.09 | .99 |

Table 4: 2 Stacked LSTM w/ FC layer model evaluation metrics

# 4 Conclusions and Future Work

This study demonstrates how social media can be leveraged for regression prediction of Covid Cases at the state level. Many future directions for this work exists. With more data, Covid predictions at a higher geographic granularity can be generated, which could help inform community organizers on target areas to help increase vaccination rates, information dissemination, and medical center preparation. While one should always default to more experienced and more technical predictions provided by the CDC for such topics, social media can be used to forecast on different topics that may not be in the immediate focus of government and researchers. Such topics include domestic violence, racial violence, and economic distress. Applications of such a network can extend to even the commercial. A benevolent investigator for a company many be able to forecast sales of a certain product in certain areas, and better their product if dips in consumption or dropping sentiment is detected. Regardless of application, one would hope that a deployment of such a model would be used in a generous, benevolent way to help communities with problems that are not being addressed by larger powers. From a technical lens, there are many ways future work can better the model. Apart from increasing the amount of data available for training, a more robust hyperparameter search given more compute power may yield even better results. Furthermore, if more time permits, examining the model to see for what state(s) the predictions are strongest for, as well as which states the predictions are weakest for, may shed light on ways to improve model performance.

# 5 References

[1] Pennington, Jeffrey, Socher, Richard and Manning, Christopher D. ”Glove: Global Vectors for Word Representation..” Paper presented at the meeting of the EMNLP, 2014.

[2] Rabindra Lamsal, March 13, 2020, ”Coronavirus (COVID-19) Tweets Dataset”, IEEE Dataport, doi: https://dx.doi.org/10.21227/781w-ef42.

[3] Data.census.gov. 2021. Explore Census Data. [online] Available at: https://data.census.gov/cedsci/ [Accessed 2 December 2021].

[4] New York Times, https://github.com/nytimes/covid-19-data

[5] Hutto, C.J. Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

# 6 Appendix

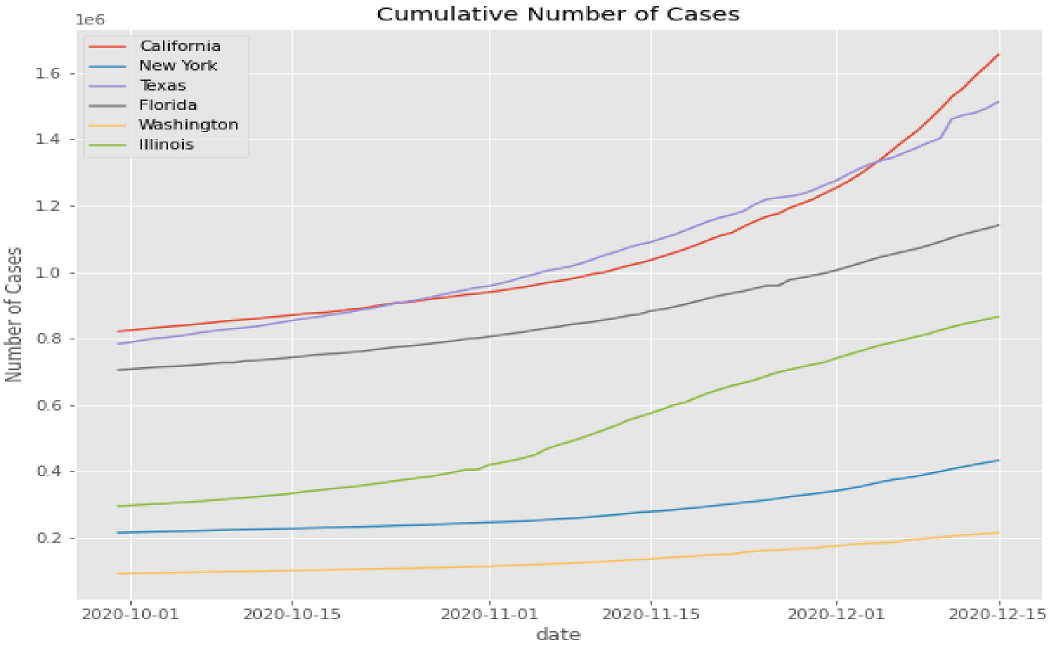

Figure 5: Covid Cases, selected States