Modeling COVID-19 new cases using twitter and Google mobility trends

Artificial Neural Networks Project

Methods used:

Natural Language Processing: Word Embedding
Binary Classification Model using Logistic Regression
Binary Classification Model using Artificial Neural Networks
Multiple Linear Regression model
Recurrent Neural Network: Long Short-Term Memory Networks

Khwansiri NINPAN: A19

https://github.com/Khwansiri/Deep_Learning/tree/master/COVID-19%20new%20cases%20prediction

Modeling COVID-19 new cases using twitter and google mobility trends

Introduction:

Traditionally, public health is monitored by aggregating statistic obtained from healthcare providers. Such models are costly, slow, and maybe biased. Recently, several studies have been successfully used digital media and social network services like Google Flu Trends, Google search data, and Twitter data to reduce the latency and improve the overall effectiveness of public health monitoring for Influenza surveillance.

Here, I applied machine learning techniques to study the relationships between global twitter trends, Google Maps mobility reports and number of daily COVID-19 new cases.

Data description:

1. Twitter data
   Dataset of tweets is acquired from the Twitter Stream related to COVID-19 chatter from http://www.panacealab.org/covid19/. This dataset contains the top 1000 frequent terms, the top 1000 bigrams, and the top 1000 trigrams in all languages. Daily top 100 bigrams tweets in English were selected for this study.

2. Google Maps mobility reports
   Community Mobility Reports from Google Maps are created with aggregated, anonymized sets of data from users who have turned on the Location History setting. The dataset displays the percentage changes from baseline in visits to retail and recreation places, grocery and pharmacy shops, parks, transit stations, workplaces and residence location. Besides residence mobility, the remaining locations show the same trend of mobility change. In order to decrease the dimension of dataset, I selected transit station mobility as their representative together with the residential data for this study. This dataset is obtained from https://www.google.com/covid19/mobility/

3. COVID-19 new cases
   The website "Our World in Data" collects the statistics on the coronavirus pandemic for every country in the world. For this study, I focused on global daily new cases of COVID-19. All information is updated daily at https://ourworldindata.org/coronavirus.

Duration of the study: March 22nd to May 4th, 2020

Project design:

Part 1: Create binary classification model to label the type of daily tweets

Part 2: Create multiple linear regression model to observe the relationship between twitter trends, global mobility reports and number of COVID-19 new cases

Part 3: Attempt to use long short-term memory networks to predict COVID-19 new cases in advance

Part 1: Create binary classification model to label the type of daily tweets

Twitter dataset contains messages in unigram, bigrams, and trigrams word token. For example, a tweet "COVID 19 pandemic" will be presented in unigram as "COVID", "19", "pandemic". Correspondingly, "COVID 19", "19 pandemic" for bigrams and "COVID 19 pandemic" for trigrams.

I first selected the top 50 English tweets words in bigrams for 1 month period during March 22nd to April 20th, 2020 and manually labeled them into 2 classes, "Others tweets" and "COVID-19 tweets" to see the trend of the pandemic awareness around the world. I chose to work with bigrams word tokens since unigram words are not clear enough to classify whether they are about COVID-19 or not. Trigrams and bigrams word showed approximately similar trends, I decided to do further study with bigrams to get the smallest unit of word as possible.

Before doing classification model, I cleaned the tweets by removing unwanted words like symbols and numbers, transforming words to lowercase since uppercase and lowercase words have different ASCII code that represent text in computer and they will be counted separately even they are the same word. I also did stemming which is the process of reducing derived words to their word base or root, and finally, I removed the stop words in English, which are generally the most common words like "the", "a", "an", etc. The PorterStemmer class for stemming process and list of stop words used are from NLTK (Natural Language Toolkit) in python. The number of labeled tweets after preprocessing step is shown in Figure 1.

Before doing any models, the texts have to be converted into numeric form first. This step is called "Word Embedding". I used Tokenizer function from keras and build-in GloVe word embeddings from Stanford NLP to create embedding matrix. Then I created a dictionary that contain words as key and the corresponding numerical vectors as values, in the form of an array.

I split 20% of the data as test set and first created classification model by using Logistic Regression from scikit learn package. Based on information of confusion matrix, this model has approximately 79% accuracy. The colormap classification is shown in Figure 2. The result shows some mixing between red and green dot that cannot be separated by linear classifier suggests that the variables are not linearly distributed.

Therefore, I did another classification model by using deep learning in keras with artificial neural networks (ANN) principle. I added 3 hidden layers with rectifier activation function and 50 neurons/ hidden layers with 20% dropout rate to avoid overfitting problem. Since I needed the output as probability between 0 and 1, the activation function of the output layer was set to sigmoid function. The selected optimizer of algorithm to find the best weight was adam, the loss function used was binary cross entropy and the criteria that I used to verify and improve the quality of the model was model accuracy. I did the training in 100 epochs with batch size 32. With this setting, the classification model showed approximately 93% accuracy. The colormap classification is shown in Figure 3.

The ANN classification model was selected to label the additional 2,400 bigrams tweets word to get the final tweets dataset (4,400 tweets word) that contains labeled top 100 bigrams word from total duration 1 month and half between March 22nd to May 4th, 2020.

Part 2: Create multiple linear regression model to observe the relationship between twitter trends, global mobility reports and number of COVID-19 new cases

World Cloud of top 50 tweets is shown in Figure 4, trends of COVID-19 new cases, transit station versus residential Google mobility reports and tweets of others topic versus COVID-19 topic are shown in Figure 5. and confusion matrix of the pair wise correlation between variables is shown in Figure 6.

After split 20% of the data as test set with random state 10, multiple linear regression model between all features and COVID-19 new cases shows approximately 77% accuracy. The comparison between the prediction of COVID-19 new cases and the real data is shown in Figure 7. See that the prediction and real COVID-19 new cases have quit similar trend and the distance between the prediction and the real data suggests that our model is not overfitting.

Part 3: Attempt to use long short-term memory networks to predict COVID-19 new cases

After seeing the relationship between COVID-19 new cases, Google mobility reports and twitter data. I tried to forecasting COVID-19 new cases by using the information of all features at 1, 3, 5 and 7 days ahead.

Since each feature have different scale (Google mobility reports is shown in percent change from base line while others is shown as total count number). I did the feature scaling by normalization before creating the model.

Keras is used to build the recurrent neural networks (RNN) with 4 long short-term memory layers, 50 neuron units and 20% dropout in each layer. I used adam function as optimizer and detected mean squared error as loss function. The regression was trained in 100 epochs with batch size 32. The prediction is shown in Figure 8. Even the model didn't show the high accuracy and has the inverse trends for the prediction in day 6, we can see some correlate trend and the possibility to forecast COVID-19 new cases by using twitter data and Google mobility reports.

Summary and Discussion:

Concordance with several previous study, the information of digital media and social network services can be used to create model for health care surveillance. This project can be considered as one of possible models for COVID-19 prediction. The information in longer time period, higher amount of data, and streaming twitter in sentence before divided in bigrams should improve the classification accuracy and model prediction.

Figures

Figure 1. Number of tweets in each class after manually labeled.

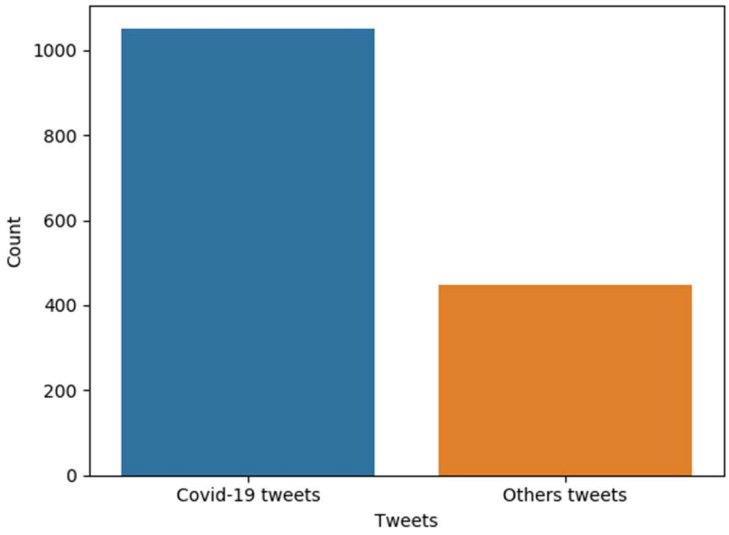

Figure 2. Colormaps classification from Logistic Regression model of scikit learn Other tweets were labeled as 0 and COVID-19 tweets were labeled as 1.

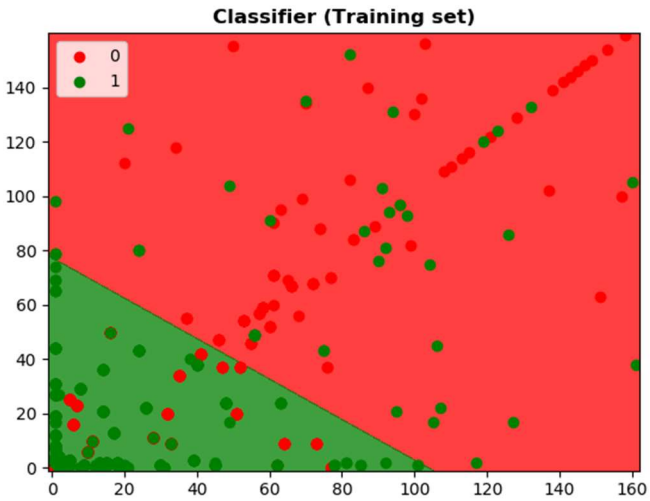

Figure 3. Colormaps classification from deep learning with Keras.
Other tweets were labeled as 0 and COVID-19 tweets were labeled as 1.

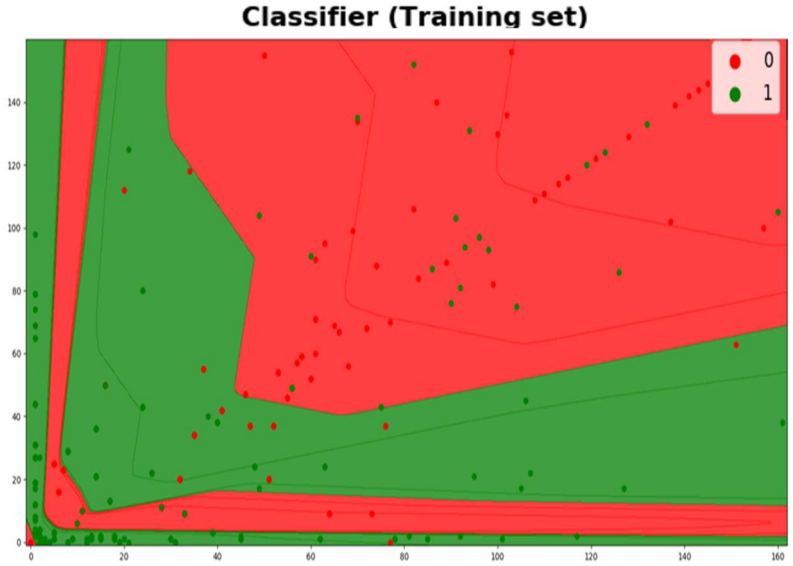

Figure 4. World Cloud of global top 50 tweets during March 22nd to May 4th, 2020

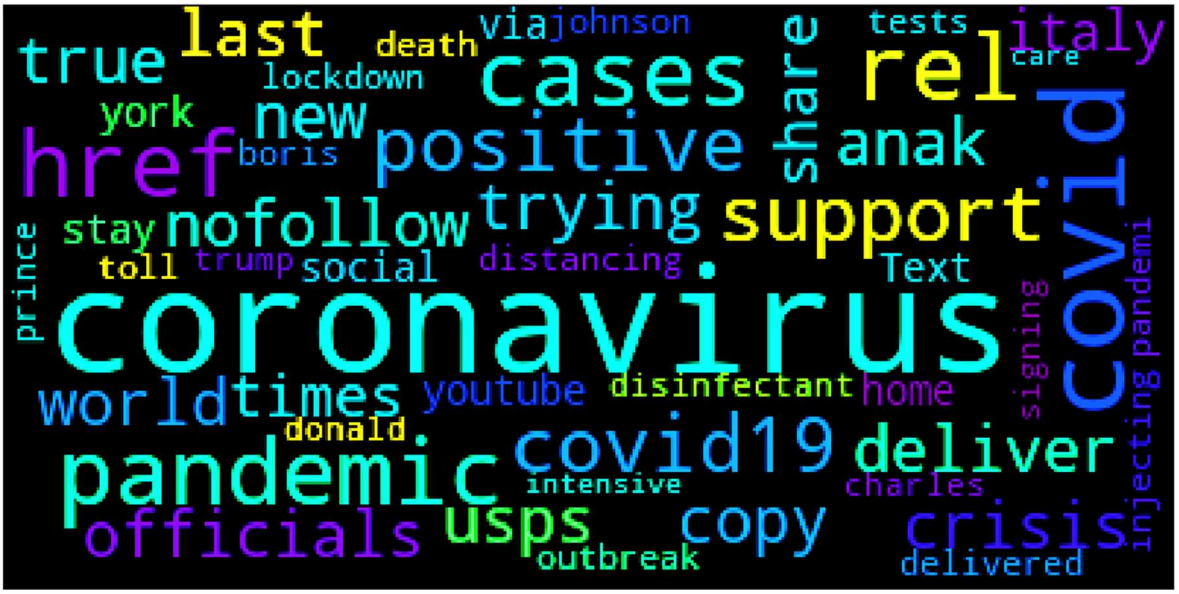

Figure 5: Trends of COVID-19 new cases, transit station versus residential Google mobility reports and tweets of others topic versus COVID-19 topic during March 22nd to 4th May, 2020.

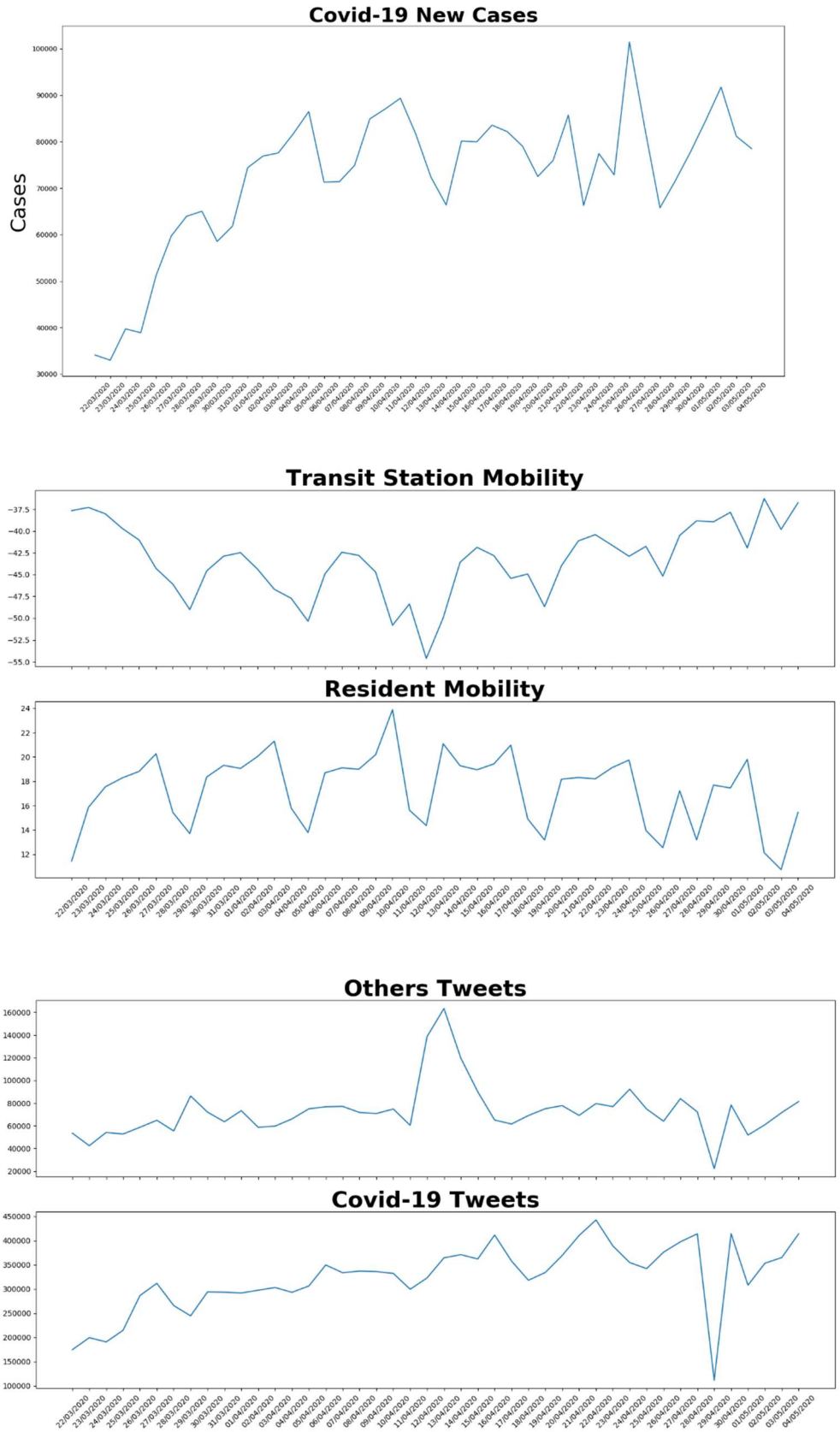

Figure 6: Confusion matrix shows the pairwise correlation between variables

Figure 7. Comparison between the predicted COVID-19 new cases and the real data using multiple linear Regression model of Google mobility reports, twitter data and COVID-19 new cases

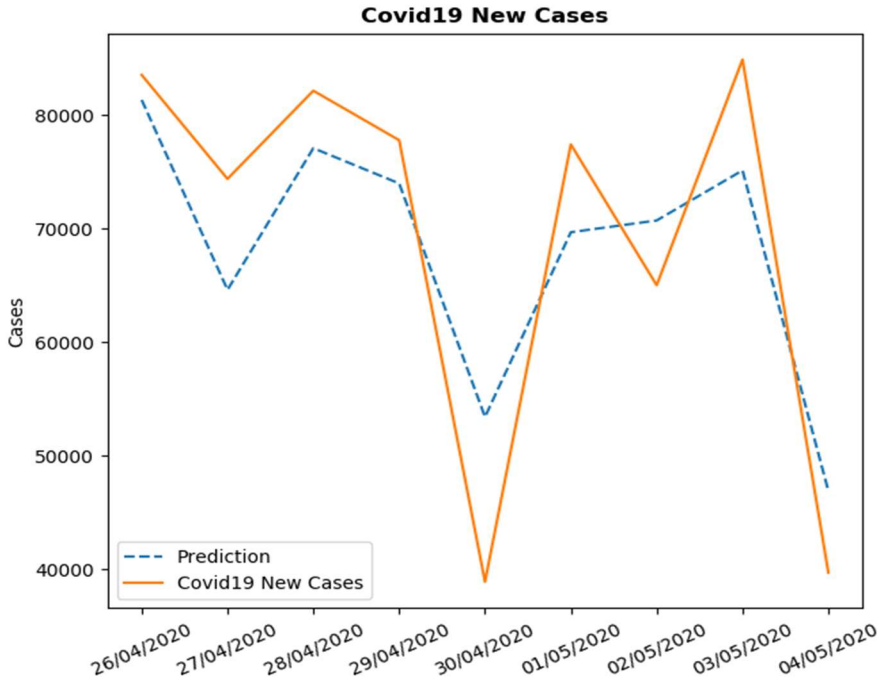

Figure 8. Recurrent neural networks for the prediction of COVID-19 new cases by using information of Google mobility reports and twitter data of 3 days ahead.