COVID-19 virus state-wise tracking in the United States

Ananth Adisesh
Department of Information Studies
Syracuse University
Syracuse, NY 13210 USA
aadisesh@syr.edu

Abstract

Data is critical to understanding the COVID-19 outbreak. We track three related kinds of data that can help us understand the US outbreak and the disease itself: testing data, hospitalization, outcomes data, and mortality rate. This data can further be enhanced by performing data cleaning, visualization, machine learning and statistical modelling techniques as well as model performance evaluation techniques, which can help further understand the data and the spread of the disease across the states in the United States.

1 Introduction

COVID-19 outbreak was first reported in Wuhan, China and has spread to more than 50 countries. WHO declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on 30 January 2020. Naturally, a rising infectious disease involves fast spreading, endangering the health of large numbers of people, and thus requires immediate actions to prevent the disease at the community level. This paper is done for research in the community and aims to predict and forecast COVID19 cases, deaths, and recoveries through predictive modelling. The model helps to interpret patterns of the spread of the virus on disseminating related health information, and assess political and economic influence of the spread of the virus.

2 Related Works

SEIR refers to Susceptible, Exposed, Infectious, and Removed or Recovered, respectively. It is based on the SIR model but adds the Exposed compartment as a variable. Susceptible refers to individuals who can catch the infection and may become hosts if exposed, Exposed are individuals who are already infected but are asymptomatic, Infectious are individuals who are showing signs of infection and can transmit the virus, Removed or Recovered are individuals who are previously infected but are no longer infectious and already immune to the virus . Once the compartments of SIR or SEIR models are determined, modelling can be done using a variety of methods. A Conditional Autoregressive (CAR) was used to account for epidemics with a spatial or transportation-related vector and modelled with MCMC. Demographic effects such as birth and death rates were added to the SEIR to model equilibria with vital dynamics. Sentiment analysis is a supervised machine-learning problem. There are different types of sentiment analysis including fine-grained sentiment analysis, emotion detection, aspect-based sentiment analysis and multilingual sentiment analysis. In binary sentiment classification, the possible categories are positive and negative. In fine-grained sentiment classification, there are five groups (very negative, negative, neutral, positive, and very positive). Sentiment analysis is one of the most popular tasks in natural language processing, and there has been a lot of research and progress in solving this task accurately. Deep neural networks are widely used in sentiment polarity classification; however, it often requires huge numbers of training data, and the size of training data varies

quite significantly among domains. It was found that a dual-module approach is the best method that encourages the learning of models with promising generalization abilities. Bidirectional Encoder Representations from Transformers (BERT) is an embedding layer designed to train deep bidirectional representations from unlabelled texts by jointly conditioning on both left and right context in all layers. It is pre-trained from a large unsupervised text corpus such as Wikipedia or Book Corpus. There are 15% of the words in the input sequence are masked out which is one of the objectives of BERT. Then, a deep bidirectional Transformer encoder is fed by the entire sequences so that the model learns to predict the masked words. Moreover, this small model has been trained on SST-2 dataset which is a common dataset for sentiment-analysis. However, there are few disadvantages in this method as it is based on SST-2 dataset which is for movie reviews and our dataset is about coronavirus news. It is a similar task which is for sentiment analysis but it does not perform that well because sentiment for movies and news might be different. However, it is the fastest way to get results and act as a benchmark or starter for further research. It can also be easily improved by adding more dataset for our domain (coronavirus news). Last but not least, it can do prediction instantly compared to previous methods that need bigger computer resources.

3 Data pre-processing and preparation

For the preparation, we will consider data cleaning to start with. The data cleaning is done by checking the data types for all attributes in the dataset. This helps us understand the data and prepare it for further processing. It is important to check the data of all the columns individually and understand how the data can be used with other columns for processing and modelling. The first five rows of the data are printed to show how the data is related to each other.

The data cleaning is done by replacing the null values with zeroes since the data missing is usually zero. The next step is to check for any duplicate rows in the dataset and eliminate them to reduce the repetition and bias in the data. This is done using the NumPy library imported at the beginning of the processing. Then, the count of rows and columns is taken after each step to check how much data has been expunged.

```
Out[1]: date                        int64
        state                      object
        positive                  float64
        negative                  float64
        pending                   float64
        hospitalizedCurrently     float64
        hospitalizedCumulative    float64
        inIcuCurrently            float64
        inIcuCumulative           float64
        onVentilatorCurrently     float64
        onVentilatorCumulative    float64
        recovered                 float64
        dataQualityGrade           object
        lastUpdateEt               object
        dateModified               object
        checkTimeEt                object
        death                     float64
        hospitalized              float64
        dateChecked                object
        totalTestsViral           float64
        positiveTestsViral        float64
        negativeTestsViral        float64
        positiveCasesViral        float64
        positiveIncrease            int64
        negativeIncrease            int64
        totalTestResults            int64
        totalTestResultsIncrease    int64
        posNeg                      int64
        deathIncrease               int64
        hospitalizedIncrease        int64
        dtype: object
```

Figure 1

```
In [9]: df.count()                    #This shows that there are no duplicate values

Out[9]: date                        6681
        state                       6681
        positive                    6666
        negative                    6480
        pending                      954
        hospitalizedCurrently       3936
        hospitalizedCumulative      3378
        inIcuCurrently              2066
        inIcuCumulative              884
        onVentilatorCurrently       1837
        onVentilatorCumulative       310
        recovered                   3698
        dataQualityGrade            5580
        lastUpdateEt                6326
        dateModified                6326
        checkTimeEt                 6326
        death                       5983
        hospitalized                3378
        dateChecked                 6326
        totalTestsViral             1834
        positiveTestsViral           608
        negativeTestsViral           612
        positiveCasesViral          3558
        positiveIncrease            6681
        negativeIncrease            6681
        totalTestResults            6681
        totalTestResultsIncrease    6681
        posNeg                      6681
        deathIncrease               6681
        hospitalizedIncrease        6681
        dtype: int64
```

Figure 2

4 Visualization and Exploratory Data Analysis

Visualization is done to understand the data and express the relation between the different attributes in the dataset. Visualization requires three main attributes to consider: clarity, accuracy and efficiency. These attributes help define the depth of understanding of data through visualization. Data visualization has many advantages:

- It possesses built-in themes for better visualizations and data understanding.
- It has tools built in statistical functions, which reveal hidden patterns in the data set.
- It has functions to visualize matrices of data, which become very important when visualizing large data sets.

Visualization is done to further understand the hidden patterns in the dataset, which helps in future processing and modelling techniques. Let us see the visualizations used.

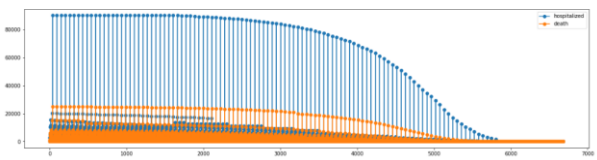

Figure 3

The line plot in Figure 3 shows the number of hospitalizations to the number of deaths due to the virus. This helps understand the mortality rate of the population due to COVID-19 virus.

Figure 4

The heat map in Figure 4 helps us understand the correlation between the different attributes in the dataset. This can be used later for machine learning models and statistical modelling.

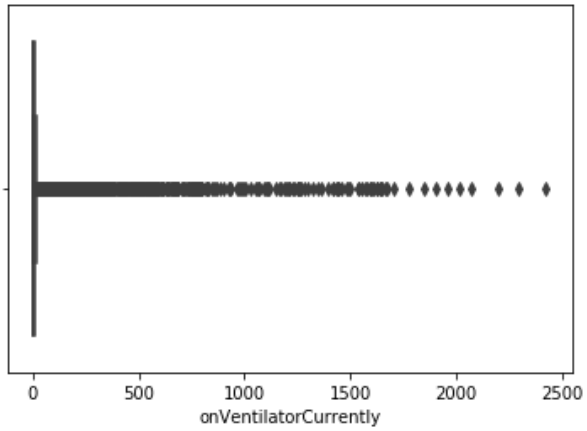

Figure 5

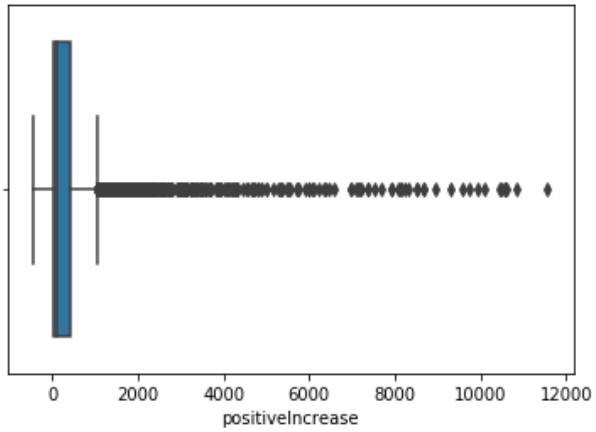

Figure 6

The box plots from Figure 5 and Figure 6 show that there are outliers in the data, which can be eliminated for accurate modelling of the data. These outliers can be removed from the data using the interquartile function. This interquartile function first requires the data to be segregated into quartiles and then removing the outliers using the given function. This helps reduce the errors while processing the data.

5 Machine Learning Methods

Machine learning involves a computer to be trained using a given data set, and use this training to predict the properties of a given new data. For example, we can train a computer by feeding it 1000 images of cats and 1000 more images which are not of a cat, and tell each time to the computer whether a picture is cat or not. Then if we show the computer a new image, then from the above training, the computer should be able to tell whether this new image is a cat or not.

The process of training and prediction involves the use of specialized algorithms. We feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data.

In this project, linear regression was used to understand the linearity between the variables and their relationship between each other. The linear regression between 'totalTestResults' and 'positiveIncrease' shows the relationship between the two variables and how they affect each other. The positive increase in the number of cases due to the virus is linked to the total number of test results in a way. This linear regression shows the amount of dependency of one variable to another. The advantage of linear regression is that the modelling speed is fast, does not require very complicated calculations, and runs fast when the amount of data is large.

The plot in figure 7 shows the dependency of the variables along a linear axis.

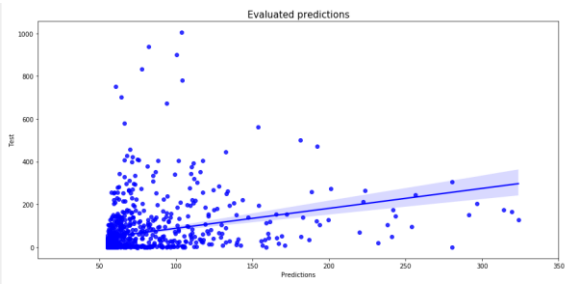

Figure 7

Another machine-learning model used is an unsupervised machine learning model called k-means clustering. A cluster refers to a collection of data points aggregated together because of certain similarities. You'll define a target number *k*, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The *'means'* in the K-means refers to averaging of the data; that is, finding the centroid.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

Let us look at the k-means clustering performed in this project.

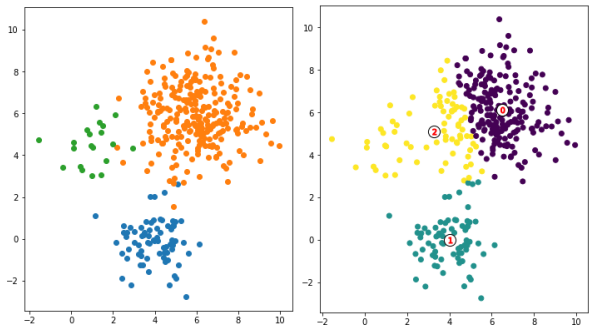

Figure 8

The clustering was performed for five clusters. The variables used were ‘positive’, ‘negative’ and ‘pending’. These variables were used to see how the spread of the types of cases were through the dataset. By observation of the graph in Figure 8, we can see that the ‘pending’ cases were not concentrated around the centre, the ‘positive’ cases were heavily concentrated around the centre of the dataset and the ‘negative’ cases were lightly concentrated around the centre.

6 Results and Discussion

The results produced via machine learning techniques are quite good in comparison to the visual observations made and the standard statistical analysis applied.

The linear regression showed that the relationship between the variables was slightly linear and that one variable is dependent on the other variable in comparison. This helps us understand that the positive increase in a particular state can be predicted with the total number of test results using this linear regression model.

The k-means clustering unsupervised machine learning model shows that the positive cases are more concentrated around the centre and the negative cases are less concentrated around the centre. This is because most cases diagnosed are positive and fewer cases turn out to be negative. Thus, k-means clustering helps in understanding the data attributes which occur more and concentrate around the different parts of the dataset. This shows that the positive cases have been more than the negative and pending cases and that the positive cases have been seen more often throughout the dataset. Thus, this model can be used to predict the future data acquired and reduce the number of positive cases by taking required precautions in the states that have high number of positive cases.

Acknowledgments

I thank Professor Ying Lin for giving me the opportunity to work on this project and for teaching me all the techniques that I have inculcated in this project. I also thank ‘The Atlantic’ for initiating ‘The COVID tracking project’ which collects the data from verified sources to provide updated dataset everyday.

The link to the presentation:
https://drive.google.com/file/d/1wQEt9FTHEiiHuFZfprsh0csUxq_KMkMm/view?usp=sharing

References

The COVID Tracking Project, The Atlantic. Retrieved on July 3, 2020 from https://covidtracking.com/data/download

John Hopkins University, "Coronavirus Map," John Hopkins University, 17 March 2020. [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed 17 March 2020].

A. Rachah and D. F. M. Torres, "Analysis, simulation and optimal control of a SEIR model for Ebola virus with demographic effects," Commun. Fac. Sac. Univ. Ank. Series A1, vol. 67, no. 1, pp. 179-197, 2018.

T. Porter, "A path-specific approach to SEIR modeling," Ph. D Thesis University of Iowa, 2012.

M. Munikar, S. Shakya and A. Shrestha, "Fine-grained Sentiment Classification using BERT," ArXiv, 2019.

The Institute for Disease Modelling, "SEIR and SEIRs models," Institute for Disease Modelling, 2019. [Online]. Available: https://institutefordiseasemodeling.github.io/Documentation/general/model-seir.html. [Accessed 03 March 2020].

J. Passy, "Here's why the U. S. government's effort to contain the coronavirus outbreak was never going to be successful.," Market Watch, 29 February 2020. [Online]. Available: https://www.marketwatch.com/story/heres-why-the-coronavirus-may-spread-in-the-unitedstates-despite-government-efforts-to-contain-the-outbreak-2020-02-27. [Accessed 27 February 2020].