Text Mining for COVID-19

Divya Dongala  
Computer Science *Montclair state University, New Jersey, USA dongalad1@montcalir.edu*

Pragati Gupta

Information Technology

*Montclair State University , New Jersey, USA  
gupta5p@montclair.edu*

*Abstract*—The purpose of this paper is to develop text classification analysis, method on covid-19 database collected from Kaggle.

Keywords— covid-19, data mining, text classification, wolrdometer, analysis, prediction.

# Introduction

The outbreak of the COVID-19 pandemic has mobilized scientists from various fields around the world to embark on research to find ways to eliminate the coronavirus and its effects. Results from this research are been published in a vast and growing number of scientific articles, making it difficult for stakeholders to keep abreast of the latest results. This brings about a growing urgency to apply recent advances in AI, such as text mining and natural language processing, to support researchers in the extraction of new insights in the ongoing fight against this infectious disease Ease of Use.COVID-19 is caused by a coronavirus called SARS-CoV-2.

Data of covid-19 is huge in includes number of deaths, recovered people, new cases. In this paper shows the text classification of covid-19 dataset over the world as well as in United States.

This data is collected from Kaggle and used for text mining analysis. The data set contain 628319 metadata rows.

This paper analyzes the following questions.

1. Show the number of confirmed deaths and recovered cases in USA.
2. Show the number of confirmed Deaths and Recovered cases in world
3. Remove all the records where the confirmed case is less than 10,000
4. The place where the maximum number of cases are recorded
5. The place where the minimum number of death cases were recorded.
6. Sort the entire data with respect to the confirmed cases in ascending order
7. Sort the entire data with respect to number of recovered cases in descending order

This paper concludes all the answers for the above-mentioned questions including the preprocessing, visualization, and data analysis.

.

# Related work

[1] Data mining plays an important role in various fields such as artificial intelligence, machine learning, and database systems. Data mining also used in the medical field for mining of healthcare systems that help to discover hidden patterns and is very useful for disease prediction. Data mining is the technique in which useful information is extracted from the raw data. The data mining is applied to accomplish various tasks like clustering, prediction analysis and association rule generation with the help of various Data Mining Tools and Techniques. In the approaches of data mining, clustering is the most efficient technique which can be applied to extract useful information from the raw data. The clustering is the technique in which similar and dissimilar types of data can be clustered to analyze useful information from the dataset. The clustering is of many types like density-based clustering, hierarchical clustering, and partitioning based clustering. Data-mining capabilities in Analysis Services open the door to a new world of analysis and trend prediction.

These research works mainly focus on the design and implementation of a COVID-19 prediction the confirmed, recovered and death cases on the whole world. The predictive data-mining model predicts the future outcomes based on past records. Data mining will help figure out the future credit risk of the applicant and predict future credit history of the applicant by using past data. Classification is known as the procedure used to locate a model that best suits identified data sets or ideas. The model helps predict the class of objects when class labels are not available . The most widely used for data mining prediction is time series forecasting methods . It is also one of the most popular models in traditional time series forecasting and is often used as a benchmark model for comparison with any other forecasting method. It is often difficult to identify a forecasting model because the underlying laws may not be clearly understood. In addition, hydrological time series may display signs of seasonality and nonlinearity which traditional linear forecasting techniques are ill equipped to handle, often producing unsatisfactory results .

[1] Paper shown the prediction of covid19 by using a time series data mining technique based on the current dataset on the proposed combination of three major pillars to analyze the outbreak of the COVID-19 virus. The coronavirus disease has terrifically affected the lives of people around the globe. Many people have lost their loved ones with the number of deaths worldwide currently goes beyond approximately 100,000 in 24 hours increasing. While Data mining techniques and related technologies have penetrated into our daily lives with many successes, they have also contributed to helping humans in the extremely tough fight against COVID-19. This paper has presented a predicted and analysis of confirmed, recovered, and death cases of COVID-19 and forecasting based on the number of cases time series based on the current data. Although various studies have been published, we observe that there are still relatively limited applications and contributions of Data mining in this battle. This is partly due to the limited availability of data about COVID-19 whilst Data mining methods normally require large amounts of data for computational models to learn and acquire knowledge. However, we expect that the number of Data mining studies and research areas related to COVID-19 and other things increase significantly and play great roll for people especially these kinds of worst time. This paper mainly predicted the COID-19 outbreak for the last 10 days and analysis graphically by using the data mining time series technique for both confirmed, recovered and death cases. When we see the perdition it achieved 99%, for the three cases means confirmed, recovered and death cases. The Constraints/limitation of this research is it was difficult to get related works because as we know COVID-19 virus is new virus. So, I could not get organized articles and related sources.

[2] Predictions were generated using a gradient-boosting machine model built with decision-tree base-learners. Gradient boosting is widely considered state of the art in predicting tabular data and is used by many successful algorithms in the field of machine learning. As suggested by previous studies, missing values were inherently handled by the gradient-boosting predictor. We used the gradient-boosting predictor trained with the Light Python package. The validation set was used for early stopping, with ROC as the performance measure.

To identify the principal features driving model prediction, SHAP values were calculated. These values are suited for complex models such as artificial neural networks and gradient-boosting machines. Originating in game theory, SHAP values partition the prediction result of every sample into the contribution of each constituent feature value. This is done by estimating differences between models with subsets of the feature space. By averaging across samples, SHAP values estimate the contribution of each feature to overall model predictions.

The model was scored on the test set using the ROC. In addition, plots of the PPV against the sensitivity (precision–recall curve) were drawn across different thresholds. Metrics were calculated for all the thresholds from all the ROC curves, including sensitivity, specificity, PPV and negative predictive value, false-positive rate, false-negative rate, false discovery rate and overall accuracy. Confidence intervals (CI) for the various performance measures were derived through resampling, using the bootstrap percentile method with 1000 repetitions.

[3] In this research paper, naive Bayes, logistic regression, and decision tree supervised learning algorithms would be used to develop the prediction model of COVID-19 infection using an epidemiology labeled dataset for positive and negative COVID-19 cases in Mexico. The ML Naïve Bayes algorithm is used for classification learning tasks in which instances of the dataset are discriminated based on the specified feature.

Logistic Regression ML algorithm is used for classification learning tasks in which the association versus categorical dependent features against independent features are determined. The learning algorithm is used when the dependent features has binary values such as 0 and 1, true or false, negative or positive, and no or yes.

Decision tree ML algorithm is used to divide learning activities where the tree is constructed by dividing the dataset into smaller sets until each partition is clean and pure and the data classification depends on the type of data. The partition of the dataset attribute of numerical data type (B)≤z, where z is the value of B domain for the entire categorical attribute of the data type partition C, form the values of (C), D∈E when E is a small set (B).

Support vector machine (SVM) is a learning algorithm that is being used for regression and classification learning tasks. The dataset points are represented in space in SVM and are divided into points and groups with similar structures that fall into the same groups [49]. The data are considered p-dimensional for linear SVM that can be partitioned by the size of p-1 planes known as hyper planes . Therefore, the planes divide the set of boundaries and data space among the data groups for regression or classification learning task.

ANN imitates the functions and activities of the brain of human being which is identified as the nodes, which is technically known as or called artificial neurons. The neurons communicate and transmit data and information among themselves in form of 0 s and 1 s or combination and each neuron has a specific weight given to it, which indicates its functions and roles to play in the system.

Correlation coefficient analysis is used to determine a strong relationship between two sets of dataset features which can be either dependent and independent features or variables.

The supervised ML models for COVID-19 infection were developed with a decision tree, logistic regression, naïve Bayes. SVM and ANN machine learning algorithms with an epidemiology dataset for positive and negative COVID-19 cases in Mexico. Before the development of the model, the correlation coefficient analysis between the various dependent and independent features was carried out to determine a strong relationship between each dependent feature and independent feature of the dataset.

# Models & Methods

*Dataset*

This data is collected from Kaggle and used for text mining analysis. The data set contain 628319 metadata rows.

This COVID-19 dataset contains confirmed cases, recovered, cases, number of deaths occurred due to pandemic. It has list of regions including countries, states and WHO regions. The dataset is segregated based on date.

*Pre-Processing the data set*

Chart

Description automatically generated

*Fig.1.Heat Map before preprocessing*

Chart, bar chart

Description automatically generated

*Fig.2.Heat Map after preprocessing*

To convert the raw data into the clean dataset in this experiment we removed all the missing values and null values. Pprofile information of the dataset which includes a minimum value, maximum value, mean value and standard deviation of each feature of the dataset. We removed all duplicate rows and the data and finally we start our analysis on the clean data.

Diagram

Description automatically generated

*Fig.3.The process of evaluating the data*

This research paper will answer all the questions mentioned below for the respective dataset. The number of confirmed deaths and recovered cases in USA. The number of confirmed Deaths and Recovered cases in world. Remove all the records where the confirmed case is less than 10,000. The place where the maximum number of cases are recorded. The place where the minimum number of death cases were recorded. Sort the entire data with respect to the confirmed cases in ascending order. Sort the entire data with respect to number of recovered cases in descending order.

Chart, bar chart, treemap chart

Description automatically generated

*Fig.4. a. covid-19 cases visualization in percentage*

We analyzed the worldwide covid-19 data with the number of recovered cases is 69.49pct, the number of active cases 27.01pct and the number of death cases 3.5pct.

Chart, bar chart

Description automatically generated

*Fig.4. b. covid-19 cases visualization in percentage*

We analyzed the worldwide covid-19 data with the number of new cases is 79.00pct, the number of new deaths 7.00pct and the number of new recovered cases 13.00pct.

Chart, bar chart

Description automatically generated

*Fig.5.covid-19 top 10 countries with active cases*

In our world data report, we observed to 10 countries with the highest number of covid-19 cases in US, Brazil, India, United Kingdom, Russia, South Africa, Colombia, France, Canada, and Peru.

Chart, bar chart

Description automatically generated

*Fig.6.covid-19 top 5 states in US with active cases*

In our US covid data report we observed to 10 states with the highest number of covid-19 cases New York, California, Arizona, Florida, Illinois, Texas, Nevada, Pennsylvania, Michigan, and Massachusetts.

Chart, histogram

Description automatically generated

*Fig.7. a. worldwide covid cases graphical representation*

Chart

Description automatically generated

*Fig.7. b. worldwide covid cases graphical representation*

Chart, histogram

Description automatically generated

*Fig.7. c. worldwide covid cases graphical representation*

In fig.7. c. is showing the total number of cases, the total number of recovered cases over the population in among the countries.

In this paper prediction analysis where the data divided into “attributes” and “labels”. Attributes are the independent variables while labels are dependent variables whose values are to be predicted. We split 80% of the data to the training set while 20% of the data to test set. The test size variable is where we specify the proportion of the test set.

To train our algorithm we used the Linear Regression technique which is represented by an equation 𝑌 = 𝑎 + 𝑏𝑋 + 𝑒, where a is the intercept, b is the slope of the line and e is the error term.

We will use our test data and see how accurately our algorithm predicts.

Chart, scatter chart

Description automatically generated

*Fig.8. 2-D graph of covid-19 active and death cases*

Chart, bar chart

Description automatically generated

*Fig.8. bar graph of covid-19 active and death cases*

# Results & Discussions

In our text mining analysis, we found the most affected country from the covid-19 is United states. In United States the most affected state is New York as per our analysis. Our prediction is satisfying with the actual data for active and death cases as per shown in Fig.8.

# CONCLUSION

Using a text mining approach this study was able to predict the Covid results. There are various data mining techniques used to predict an outbreak. It is hard to deal with this pandemic it requires lots of preparation to manage the unpredicted situation in that way our research helps to predict the future active cases and the number of deaths so that we will the prepared to face the outbreak.

# Acknowledgment

We thank the anonymous reviewers for their valuable feedback, and professor Jiayin Wang, who provide their very helpful comments on this survey.

##### References

1. Tajebe Tsega Mengistie ,“ COVID-19 Outbreak Data Analysis and Prediction Modeling Using Data Mining Technique,” NITM, Bijni Comple,Laitumkhrah, Shillong, Meghalaya 793003,India. International Journal of Computer (IJC) ISSN 2307-4523
2. Zoabi, Y., Deri-Rozov, S. & Shomron, N. “Machine learning-based prediction of COVID-19 diagnosis based on symptoms”. npj Digit. Med. 4, 3 (2021)
3. L. J. Muhammad1, Ebrahem A, Algehyne2, Sani Sharif Usman3, Abdulkadir Ahmad4, Chinmay Chakraborty5, I. A. Mohammed6, “Supervised Machine Learning Models for Prediction of COVID‑19 Infection using Epidemiology Dataset” SN Computer Science (2021) 2:11, Springer Nature Singapore Pte Ltd 2020