Prediction and Analysis of Covid-19 with various Machine Learning algorithms

Sarthak Maggu, Vijander Singh*

Department of Computer Science and Engineering, Manipal University Jaipur, INDIA

Abstract

The world was stuck by a deadly virus last year, which stopped the world and changed it completely. The virus was given the name "Corona." It has had an everlasting effect on human lives which has caused a lot of people to study about it, following the trend we have chosen "Prediction and analysis of Covid - 19 with different ML algorithms and comparative analysis" as the title of this paper. The title of the paper is self-explanatory to explain what the paper is about and what technology will be used in the paper.

The paper will use data cleaning and plotting techniques to analyse the impact and effect of Covid on human lives and various algorithms to predict how vulnerable the person is to the virus / predict whether a person suffered from Covid or not based on various parameters.

The main ingredient of the paper will be the data on which the model will be built, will be collected through various Google forms or through open source data websites such as Kaggle. The paper would be divided into various parts, which will include data collection, data cleaning, data plotting etc.

After cleaning of data and building of models using various ML algorithms, findings will be reported in a form of report using various plots to favour the findings of the various models.

Keywords: Covid-19; Machine Learning; Pneumonia; ConvNet; CNN; Deep Learning

1. Introduction

The main purpose of the research paper is to establish prediction and analysis of Covid-19 with various machine learning algorithms, since December 2019 the world has been witnessing a great rise in Pneumonia cases, along with Covid. Between December 25, 2019 and February 7, 2020, a single-centre case cohort of the 179 consecutive patients with confirmed and probable COVID-19 pneumonia were hospitalised to Wuhan Pulmonary Hospital, and they were all included in the present study. The information of all patients including demographic data, clinical characteristics, laboratory parameters, and outcomes were collected prospectively. Sentiment analysis for Covid Vaccine, as sentiment analysis have been done regarding Covid it's important to understand the sentiments of people behind the vaccine as the only way to end this pandemic is heard immunity or mass vaccination.

2. Literature Review

Predictors of Mortality for Patients with COVID-19 Pneumonia Caused by SARS-CoV-2: A Prospective Cohort Study [1]

* Corresponding author
E-mail address: vijan2005@gmail.com

ARK: https://n2t.net/ark:/47543/JISCOM2022.v3i1.a29

The main purpose of the research paper is to establish relationship between Pneumonia and Covid, since December 2019 the world has been witnessing a great rise in Pneumonia cases, along with Covid. The paper was based on the data collected in the city of Wuhan.

Design and analysis of a large-scale COVID-19 tweets dataset [2]

The main purpose of this research paper was to design a large-scale dataset to understand the sentiments of people around globe regarding Covid tweets. The paper was based on collection of data using key words from twitter basically scrapping it and then using various Natural Language Libraries available to clean it and use powerful neural networks for analyzing it.

3. Responding to the COVID-19 Pandemic in Developing Countries [3]

The purpose of this research paper was to study the ill effects that Covid has had on developing countries mainly in Asia and South America and how it has impacted the life in these continents. Pneumonia and Covid are directly proportional if talking in simple terms, a person having Covid may or may not have Pneumonia but a person having Pneumonia has 80% chances of having Covid.

People all over the world have different reactions towards Covid and the restrictions that have been implemented by people all the world. Analyzing tweets even helped in understanding the psychological effect Covid has on the world.

Developing countries had to make tough choices between economy and human life. The Covid restrictions hampered the developing and caused a chaos, but if the proper and quick measures were taken the developing countries wouldn't have to make such tough choices.

Figure 1: Sentiment analysis

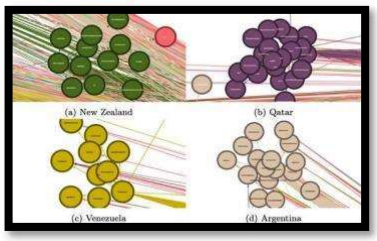

Figure 2: Analysis of Keywords

Figure 3: Keywords taken for analysis

3. Research Objective

Pneumonia detection with neural networks, as the research work has helped us in establishing a relationship between Pneumonia and Covid. Sentiment analysis for Covid Vaccine, as sentiment analysis have been done regarding Covid it's important to understand the sentiments of people behind the vaccine as the only way to end this pandemic is heard immunity or mass vaccination. Studying the impact of Covid where cases have been high and analysing it based on GDP, Urban population, and Hospital beds available per 1000 number of people.

4. Methodology and Framework

Figure 4 depicts the real time machine learning system architecture needed for any real-world prediction. The system architecture also depicts the role of micro services needed in the system architecture.
The architecture would be used for sentiment analysis of Covid vaccine.

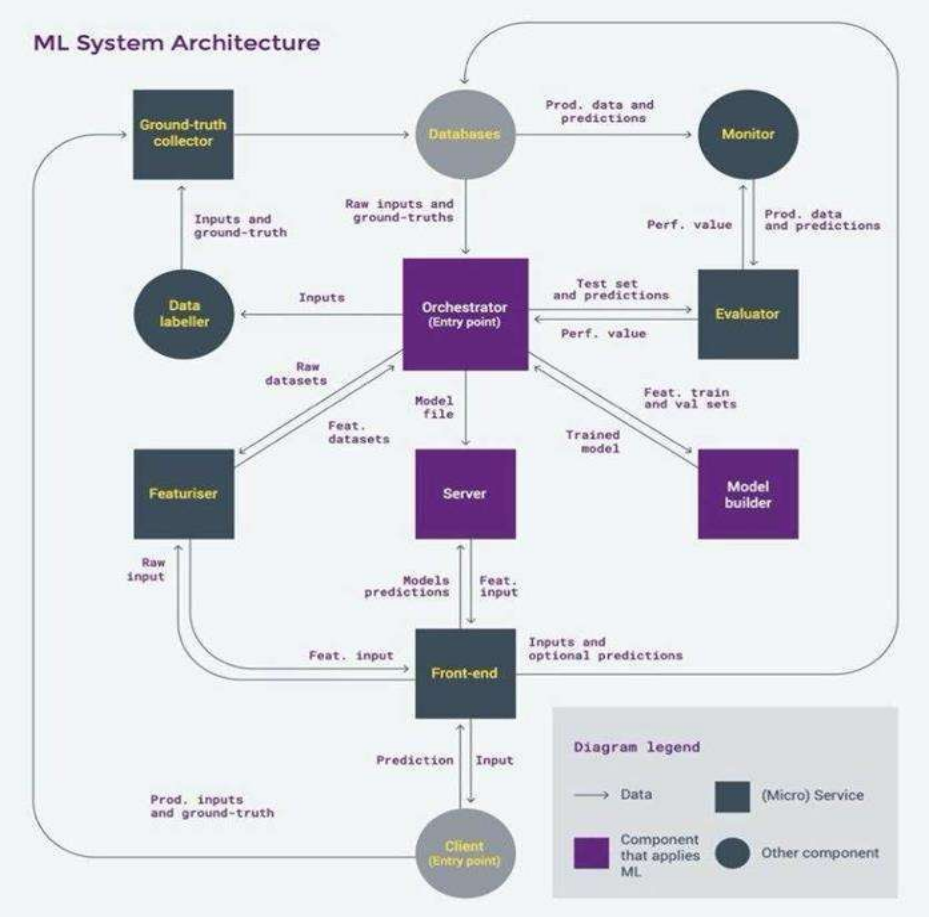

Figure 4 System Architecture [6]

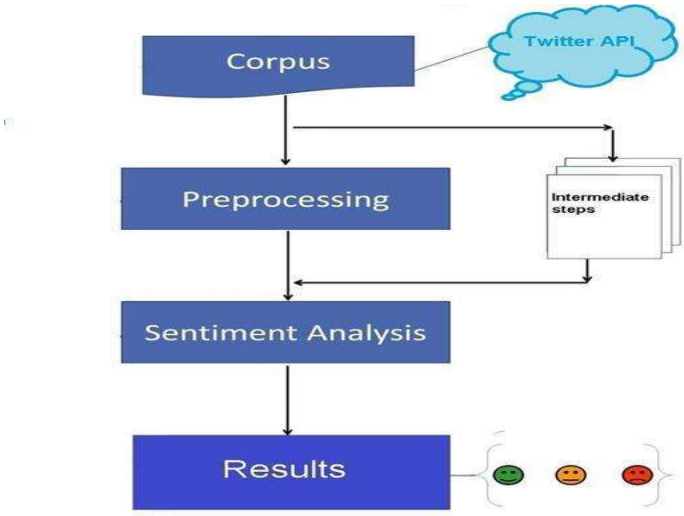

Figure 5 Data flow of the problem solution

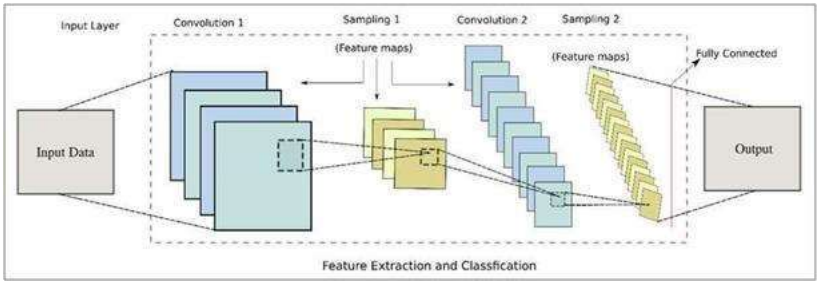

Figure 6 :CNN Architecture

The above architecture would be used for Pneumonia detection.

5. Algorithms and Techniques

Before discussing about the algorithms and techniques, the methodology must be discussed. The paper would be divided into various parts and would follow the traditional flow set by eminent machine learning engineers all over the world. The division is as follows:

Data Collection: The main ingredient of the paper would be data, which will be collected through various Google forms or through open source data sites such as Kaggle etc. The data collected would be stored in .csv or .xlsx format. SQL databases would be used for joining data collected through Google forms and combining them into single .csv xlsx format, or else the model being built would be redundant[5].
Data Pre-processing: Data pre-processing is one of the most important steps, if not done properly it can induce bias making the model useless. Data would be preprocessed with help of various python libraries such as Pandas, Seaborn etc, through this step we will achieve: o Removal of missing values, if not removed our models won't function properly[5].
Removing strongly correlated features, or else they will have maximum influence on our final predictions.
Conversion of categorical features into numerical features, as the existing machine learning algorithms can't deal with categorical features.
Data Plotting: Data plotting will be used to understand how closely the various features are related to each other. Through this step we will be able to evaluate the correlation between various features and select which features to use for model building and which features to avoid. Outliners are like the unwanted ingredients if not removed can spoil the whole model, through this step we will figure out them and remove them to obtain best results.
Model Building: The final step of this cycle is to build model, in which we will be using various machine learning algorithms and analyse them based on different scores such as RMSE,F1 score and etc We will be using various machine learning algorithms ranging from various regression algorithms for calculating the impact to classification algorithms for predicting whether a person will suffer from Covid or not. This step would be done with the help of various sklearn libraries to reduce the human labour and achieve quick results.

The algorithms to be used are as following:

Along with these various traditional machine learning algorithms will be used in the paper, namely regression algorithms and classification algorithms [4, 5].

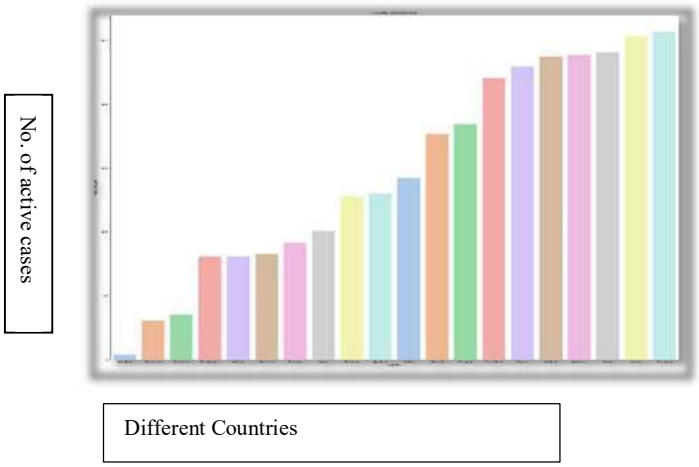

Figure 7 Visualization from scrapped dataset

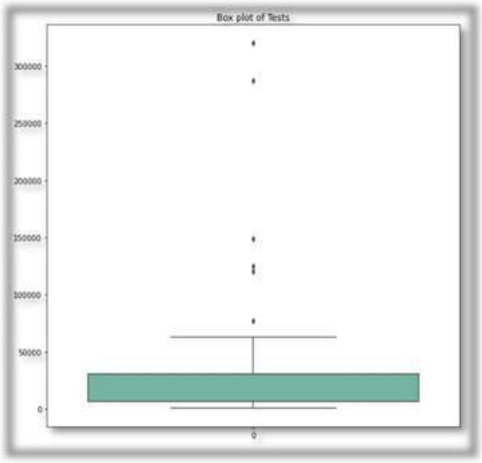

Figure 8 Box Plot visualization

6. Results and Conclusions

Countries with higher GDP, have higher Covid cases as compared to countries with lower GDP. This is due to higher GDP countries have people traveling from all over world, thus the Covid spread much faster in these countries. Higher GDP countries had less beds per 1000 number of people, as most of these countries invested heavily in entertainment and hospitality industry as compared to hospital industry. Countries with higher urban population had lower number of Covid cases as the population was majorly in one area thus isolation was easy as compared to countries with lower urban population. Countries that had higher number of tests, didn't have higher Covid cases as compared to the rest of world. USA has highest number of Covid cases but it's testing isn't even in top 10 in world according to dataset. With transfer learning we were able to achieve 92% accuracy in pneumonia detection, while without it we couldn't achieve even 50% accuracy.We were able to the key words required for vaccine sentiment analysis, both positive and negative keywords included and most used tags. Positive and negative tweets had almost percentage of contribution, thus we must create awareness to speed up the vaccination process. Positive tweets mostly came out from first world countries as they had major number of cases.

This paper is outcome of minor project B.Tech. CSE VI semester.

References

[1] Rong-Hui Du, Li-Rong Liang, Cheng-Qing Yang, Wen Wang, Tan-Ze Cao, Ming Li, GuangYun Guo, Juan Du, Chun-Lan Zheng, Qi Zhu, Ming Hu, Xu-Yan Li, Peng Peng, Huan-Zhong Shi (2020) “Predictors of Mortality for Patients with COVID-19 Pneumonia Caused by SARS-CoV-2: A Prospective Cohort Study” Vol 57 Issue 4

[2] Rabindra Lamsal. "Design and analysis of a large-scale COVID-19 tweets dataset", Applied Intelligence, 2020

[3] Anis Z. Chowdhury & K. S. Jomo(2020) “Responding to the COVID-19 Pandemic in Developing Countries: Lessons from Selected Countries of the Global South.” Development, 2020

[4] Enes Ayan, Halil Murat Inver Diagnosis of Pneumonia from Chest X-Ray Images Using Deep Learning", 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), 2019

[5] towardsdatascience.com

[6] www.cse.unr.edu

[7] erj.erjournals.com