Covid-19 Prediction Model

Kumar Saurabh
MSc Big Data Analytics and Artificial Intelligence
L00157097

***Abstract*— Coronavirus disease (Covid-19) is a highly infectious disease which was caused by a new type of virus called as corona virus in the Wuhan city of China. There were many theory which states that the virus was leaked from a lab in Wuhan but WHO recently denied that theory due to lack of evidence found in the lab. The most common symptom of this virus is mild to moderate illness. Though there is no cure for coronavirus as of now, people recover without getting any type of special treatment. It started spreading all around the world after February 2020 when the WHO actually release a press conference to warn about this deadly virus. After March 2020, strict lockdown were imposed all around the world. Both domestic and international travel were banned and people were asked to remain in their houses and come out only when it is very important with face masks on and use of hand sanitizers were made mandatory. After this in May, most companies started to test this virus samples to bring cure as by then, millions of people were affected and thousands had already lost their lives. This paper offers clear details of Covid-19 functionality as well as key Big Data analysis of the infected people, recovered ones, the amount of deaths till date and further pandemic forecast evaluation. This model discusses challenges and implentation to fight with the Covid-19 using Big data.**

***Index Terms*— Coronavirus, Covid-19, Virus, China, Big Data analysis, WHO, Face Masks, Sanitizers, Lockdown.**

I. Introduction

IN recent past, there has been a disease in every century which is considered very harmful for human beings. If I talk about Spanish Flu, Asian Flu, Black death, AIDS, SARS originated through bats, Ebola and now Corona virus (Covid-19) which is also a form of SARS. The first traces of this virus was found in Wuhan city of China in November 2019 however, China didn't disclosed it and kept on spreading throughout the city in a quick time. WHO earlier, in January said that people all around the world don't have to worry about it but later changed their statement in February stating that everyone needs to take care of themselves and face masks need to worn while going outside, sanitizers should be used to prevent ourselves from any contaminated surface. People were also asked to take maintain a social distancing when are in public places. Soon, in March, they announced that countries should close their borders as it started to spread all around the world from China [1]. Lockdown was imposed throughout the world. International travel as well as Domestic travel were stopped to stop the spread of this virus. The most contaminated countries in the initial days were countries of Europe such as Spain, France, etc. Later on the cases in USA and India kept on rising along with Brazil. In about 6-7 months, the infected cases rises to more than 1 billion in these countries.

The business throughout the world was impacted a lot because major working industries which requires on-site jobs were forced to shut down to avoid more spreading of disease. To avoid impact to businesses, people were asked to work from home and work for more hours because they have nowhere to go. Most of the IT companies have still asked their employees to work from home in order to make sure this pandemic doesn't impact the economy of the country.It was forecast by various agencies all around the world that this pandemic would cost the global GDP by 1 percent which is a huge amount. Though the recovery started at a much faster pace after October in India. The race to get cure/ vaccine for Covid-19 started and every country started to do the testing phase of trail. In a small amount of time, there was a large amount of data related to the testing phase, vaccination phase and also related to the Covid-19 infections. Various countries started a race to make the vaccine with more efficiency and as soon as possible. The WorldoMeter data in February reads as below:

Fig 1: WorldoMeter Latest Data

II. Characteristics and Challenges of Covid-19

The WHO termed this pandemic as "major and unexpected incident of international concern" [2]. It is a public health concern for all the countries worldwide and requires major international coordination. They warned everyone and divided the pandemic into 6 levels of which level 6 is considered

as the highest risk. The outbreak of major public health incidents usually has many characteristics.

Some of the characteristics explained by WHO are listed below:

Characteristics of COVID-19 incidents. WHO: World Health Organization.

Feature	Explanation	COVID-19
Sudden	The incident can suddenly erupt without warning.	Sudden outbreak.
Uncertainty	Knowledge of viruses may be limited.	Unknown, new coronavirus.
Unpredictability	The impact and sustainability of the event cannot be predicted quickly and accurately.	Political, economic, social, cultural, and other influences.
Highly hazardous	Damage to people's health and property.	More than 21 million cases have been diagnosed worldwide, and more than 7,700,000 deaths [6].
High social attention	Arouse widespread and in-depth public attention.	Baidu index daily average of 35,083 [7]. The Google trends index reached 100 (represents the hottest search) [8].
Chain reaction	The incident occurred beyond its administrative area, expanding the scope of its impact.	Spread quickly to 188 countries in the world [9].
Timely disposability	Governments respond quickly with strict controls.	A WHO Global Emergency [6].
Preventive actions	Minimize the pandemic loss, so that the pandemic is gradually controlled; resume work and production on the basis of establishing strict avoidance and prevention experience.	Strictly control the source of infection; cut off the route of infection; protect susceptible people; introduce quarantine and curfew.

Fig 2: WHO explained table

All above are classified by the WHO in various aspects. When it was spreading, the government in various countries tried to avoid the community transmission of this disease as after that it spreads with a much faster rate. The mortality rate of this disease was much less than SARS but it was spreading very fast. For e.g. a data released in China during the first half of March shows that out of 80,735 cases, 58,600 were cured and around 3200 lost their lives [3]. At that time, it wasn't spread around the world on a large scale. Therefore, WHO released a warning and it forced the government to stop transportation, close shopping centres and restrict public movement to stop the spread. Combating Covid-19 has become a long term issue for the entire world, and it is changing the economical, social and political landscapes. With this emergency situation, most of the government acted quickly and deployed the best resources they have available in their respective country to stop the spread of this disease.

III. Big Data Resource associated with Covid-19

Big data is a technology which can obtain, discover and analyze high volume of data to extract values.

Fig 4: Big Data

There is rapid enhancement of information technologies which are widely used in healthcare, manufacturing, logistics, transportation and various other industries. Big data has become an important technical factor in growth of modern era government. Most of the things are predicted and acted upon after analyzing and making predictions through various big data collected from different sources.

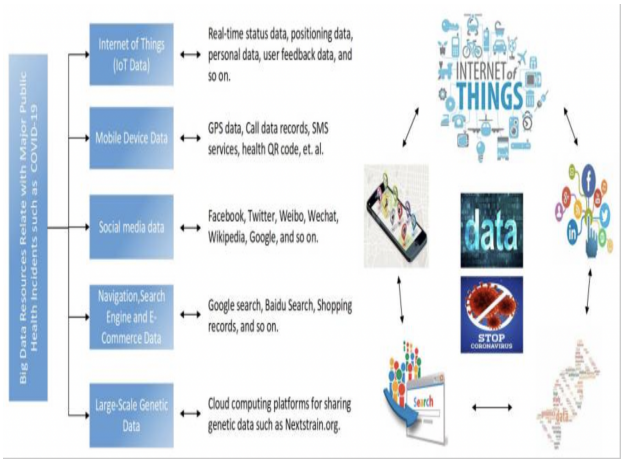

Fig 2: Covid-19 Data sources

The newer technology of storing and analyzing data provides the great aspect to take key decision based on large amount of data. Various data sources are classified as:

A. Internet

In the Internet, there is a keyword named as "IoT" which is powerful source for data extraction and visualisation. It can be used for intelligent identification, tracking and management of various things. For collection of Covid data, IoT (Internet of Things) is really helpful [4]. It contains real time data, data extracted from GPS, it can be personal data as well and also the data which is extracted from the feedback provided by the users. It can also gather data from facial recognition and gives you a lot of information for e.g. if a person is wearing mask or if the mask is not worn properly, it can use the recognition feature to detect all of those. It can also analyse and inform you about a person's behavior during the lockdown.

B. Mobile phones

Nowadays, everyone has a smartphone and is a major source of information. The data generated from the mobile phone of a user gives you a lot of information about his activity or daily routine. If a person has switched on his location data, we can easily track his movements and monitor based on his activity. Various Covid tracking apps were launched by the countries all around the world. People were asked to download that app and that app asks them to switch on the Bluetooth feature of the mobile so that it can tell us that whether we were in close contact of any infected person or if we are in a contaminated area, that app is going to tell us that it's a risk zone. That data from users are monitored by the government agencies to keep track of the infected people. These data not only help

the government to keep track of the infected people but also make them predict how fast or slow the virus is spreading in a particular region and later throughout the country.

C. Social Media

Social media is a big platform for data. Almost everyone has one or more social media account and use them regularly. Social media has provided a new way to Big data to maximise it's chances to track, control and prevent any kind of viruses. Many social media platform has opened up a new feature for Covid. These app not only help someone see the latest figure of the cases in their country but it also has a various other feature to help user prevent themselves from getting infected to the virus. US once used Twitter data to detect a seasonal flu which was spreading in the country to predict when will it reach at its peak by creating the flu prediction algorithms.

D. Large Scale Data

The data which were in form of genetic which were extracted from pathogens usually plays an important role in looking out for the source of the virus along with its clinical trails and the development of the vaccine which can be used as a cure for that virus. Initially, the vaccine trail started late in most of the countries because the researchers, doctors and scientists were not able to get the genetic data of the virus which would provide them sufficient amount of information to start developing the cure. A type of genetic data information source is shown below.

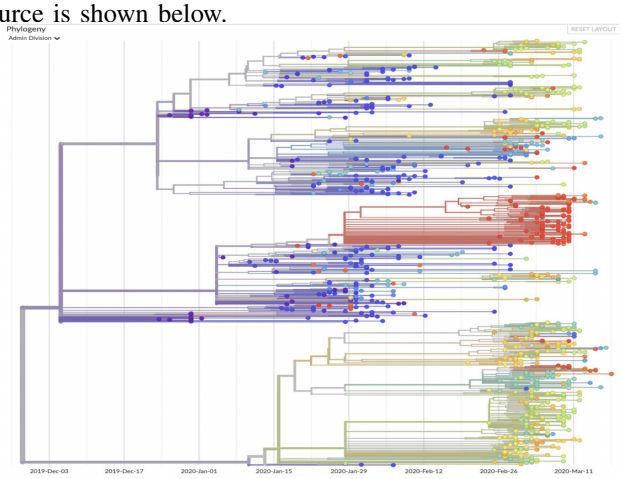

Fig 5: Covid-19 Data sources

In above figure, the sequence represents the vertical line. When a sequence of data will sit on its own line that means that in that sequence, we didn't found any new variation. If the line is longer, that means that there are more variation.

IV. Data Pre-processing and Modeling

Soon after December 2019, the virus started to spread all around the world. The people from China used to travel all around the world and until then no one was knowing about the virus. WHO earlier in January 2020, said in a press conference that no one needs to worry about the virus but later on in February back off from their statement and finally in March 2020, declared the virus as a health emergency in the world. Later on it was declared as global pandemic. At the time, this virus was spreading, an effective Big Data model was required to be implemented to predict the spreading of the virus, the mode of spreading of the virus and also what should be the next stage of the spreading of the virus. A strong policy was required to be implemented at the time to restrict the spread of that virus. This report discusses about the prevention and control from the virus using various mechanism. With this technical report, I will try to create a model which is able to make powerful predictions to help fight with this disease. The model is divided into various phases:

A. Dataset Used

The first and foremost thing which is required for creating a model is to get data. I have used multiple data sources. I have used 'json' data which will be extracted from a link. Apart from this, I have used data from WHO website, Kaggle and few other sources. All the data used apart from 'json' are 'csv' which is comma separated values required to train and test the model. The size of all the datasets used is around 30MB. The WHO global dataset used consists of more than 99000 rows and similarly the other big dataset used consists of more than 75000 rows. Both of these two datasets are the primary dataset for creating the Big data model.

B. Visualization

If we talk about visualization and processing, MapReduce technology and also Hadoop which is used on a large scale for such type of distributed infrastructure [5]. Nowadays, government agencies all around the world are using the Big Data models to keep track of all the information related to the Covid-19. Further open source analytical engines like Spark is used to handle Big Data and cloud platforms such as Amazon AWS or Databricks are used to execute the Spark libraries for creating a model.
I am using Databricks as the cloud platform and Spark as the source engine to visualize and create an effective model for Covid-19 predictions. I have also used Python for more visualization of data. Some of the visualization are:

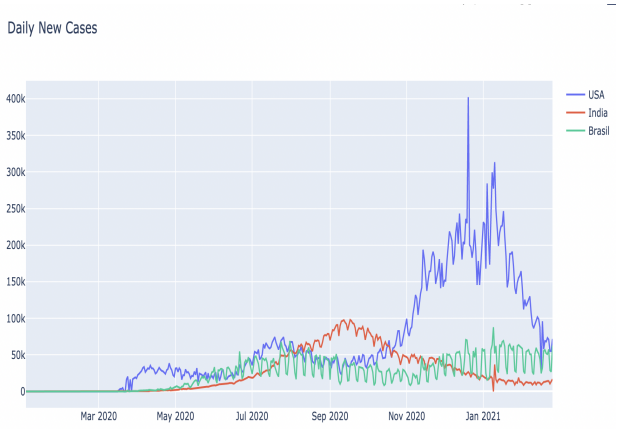

Fig 6: Daily New Cases

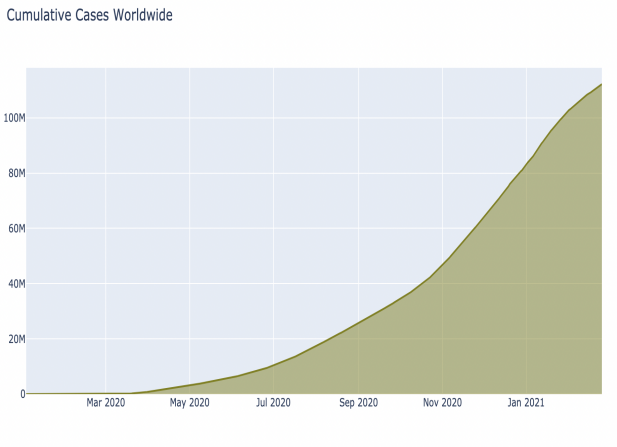

Fig 7: Cumulative New Cases

C. Modeling

I have used Linear Regression and Decision Tree to create the model. I have designed a scatter matrix plot to visualise relationship between combination of variables. With the help of that we would be able to allow many other relationship to explored in that one chart. It is a great place for visualizing a lot of thing at one place. It is shown below:

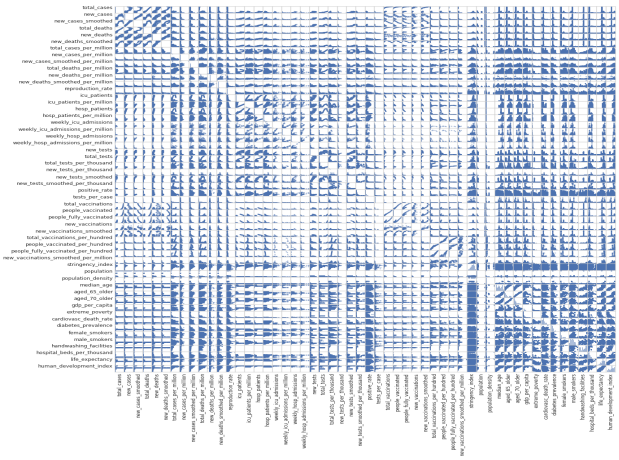

Fig 9: Scatter Matrix

D. Prediction

With the help of Linear Regression and Decision Tree Model, we are able to get below results:

1. Linear Regression Model
a) RMS: 5.858056487295799e-05
b) R2: 1.0
c) Accuracy: 0.884118383533988

2. Decision Tree Model
a) Prediction: 44.731320624321115

3. Random Forest Model
a) Prediction: 50.32201939557669

With above we can say that the Linear Regression Model is working better than the other two models. So, we can use the Linear Regression model to make a better prediction for our Covid-19 model.

V. Challenges with Covid-19 Model

The most important challenges related with Big data is it's trustworthiness. As they are large in size, there are a lot of question on it's accuracy, credibility, consistency and validity. If the data is not good, you will end up getting poor results and you won't be able to implement that model for training purposes [6].
In my case, the data I have used is from various sources to it becomes a very difficult and challenging to implement algorithms on it as the nature of the data would be different from the other. I have implemented my models on a single data frame and used other two for visualization purpose. The data keeps on updating everyday so we need to keep a track on that as well as we will end up getting same results again and again. The data should be smooth as well when we implement algorithms on it [7]. So it is very important for data pre-processing, its cleaning and also removed missing or noisy data.
After that we would be able to get a good model which can be implemented on a large scale. After this pandemic, various organisations all around the world has released their data based on their tracking, so it is very important to be selective in such case. The data extracted from WHO, John Hopkins University, etc. can be more reliable compared to the other datasets.
One of the biggest challenge with Big data is that its misinterpretations can lead in dangerous consequences which can be harmful for human rights as well as public health. It happened in 2014 during Ebola when there was some miscalculation lead unappropriated movement of people. So, to find the real truth behind the Big data, qualitative analysis needs to be done in terms of Covid-19.

VI. Conclusion

This report focuses on prevention, challenges and control related to Covid-19 with the help of Big data. We also discuss about the definition of Covid-19, how it spread out and what measures needs to be taken to prevent humans from it. Along with that we discussed what measures government need to take to tackle this situation and how they can create an effective model with the help of Big data. Government can make use of the big data to make important decision to fight with global pandemic like Covid. With the help of Big data analysis, they can create a repair mechanism and promote it to large scale to remove the fear among people. Big data has a lot of qualities but it has some limitations too if not handled properly. Generating big data takes time so it is nearly impossible to make an early and accurate prediction. Also if the predictions are made with insufficient amount of data, it is pretty much possible that it can cause harm to public health. So, it becomes very important to handle the Big data carefully to make an effective predictions and act according to that. Lastly, in order to ensure operability, the use of big data in epidemic

prevention and control must take into account administrative rights, privacy security, cost, and other factors, as well as the balancing of interests with public epidemic prevention. With that I conclude my report related to Covid-19 predictions.

VII. References

[1] Jia Q, Guo Y, Wang G, Barnes SJ. Big Data Analytics in the Fight against Major Public Health Incidents (Including COVID-19): A Conceptual Framework. Int J Environ Res Public Health. 2020;17(17):6161. Published 2020 Aug 25. doi:10.3390/ijerph17176161

[2] World Health Organization Coronavirus Disease (COVID-19) Outbreak. [(accessed on 18 August 2020)]; Available online: https://covid19.who.int/

[3] National Health Commision of the People's Republic of China The Latest Situation of New Coronavirus Pneumonia. [(accessed on 16 August

[4] Li X.L., Zhang L., Li K., Wang Y.Y. A Data Allocation Strategy for Sensor of Internet of Things. Comput. Res. Dev. 2013;50:297–305.

[5] Maguire D.J. An overview and definition of GIS. Geogr. Inf. Syst. Princ. Appl. 1991;1:9–20.

[6] Toh A. Big Data Could Undermine the COVID-19 Response. Wired. [(accessed on 15 August 2020)]

[7] CSSEGISandData/COVID-19. [(accessed on 15 August 2020)]

Github link: https://github.com/srb7600/COVID$_REPORT$.