Data Analysis for COVID-19

I. INTRODUCTION

The outbreak of the new disease in Wuhan, China was caused by novel Coronavirus (2019-nCoV) [1]. This disease is a form of pneumonia. Coronavirus belongs to the *Orthocoronavirinae* subfamily. The first case was observed at the Chinese Center for Disease Control and Prevention (CDC) on 12 December 2019 and was considered as a non-SARS novel coronavirus [2]. The family to which Coronavirus belongs is *Coronaviridae* which consists of a large, single RNA strand of plus sign [3]. Viruses of these family show the symptoms of common cold, diarrhea in human beings. In the year 2003, it was seen the outbreak of coronavirus i.e. severe acute respiratory syndrome coronavirus (SARS-CoV) [4]. In December 2019 at Wuhan, China's symptoms closely resembled the same as pneumonia [5]. Several cases of approximately 1974 were confirmed in China according to the council information office in Beijing, China's capital on 26th January, 2020. Virus started spreading in many other countries like the very first case after China was reported in Thailand, Japan and two cases were also seen in Korea on 16 January 2020. Recent researches have shown some evidence of the origin of the virus from the bat and it was also seen that transmission of the virus is taking place from human to human. The situation started getting worst from 19 January 2020 day by day,

so to take some serious action for the control and prevention from the disease. World Health Organisation (WHO) on 30 January 2020 declared that Coronavirus Disease was an outbreak emergency of international concern after the attack of H1N1 in 2009, the emergence of Ebola virus in 2014, polio in 2014 and Zika virus in 2016 [6] [29]. Finally, on 11 February 2020, World Health Organization (WHO) gave the name of the novel disease which was caused by the corona virus as Corona Virus Disease- 19 (COVID-19) [7] [32]. Record maintenance on 24 February 2020 showed that more than 78, 000 patients were suffering from COVID-19 throughout many countries. The maximum patients were from China according to the World Health Organization (WHO) which were approximately 77,000 and 2500 death [8]. According to the World Health Organization (WHO) the rest of the countries reported 2000 confirmed cases and 300 deaths as on 7 March 2020. In Wuhan, China lockdown orders of all the trains, fights and public transport were passed on 23 January 2020. The exact origin of COVID-19 was not reported but through different researches, it was seen that coronavirus possibly has originated from the bat. According to the Centers for Disease Control and Prevention (CDC), the novel disease COVID-19 was transmitted from person to person through droplets, and the symptoms seen were fever, shortness of breath, and cough which was seen after 14 days [9].

The International Committee on Virus Taxonomy replaced the name of 2019-nCoV as SARS-CoV-2 (severe acute respiratory coronavirus-2 syndrome) [10]. The outbreak of novel SARS-CoV-2 was increasing at an alarming rate in China as global intimidating as pandemic throughout the World. Different methods were used to analyze data regarding epidemiology which were exploratory data analysis (EDA) methods and visualization model. These two methods showed the awareness among the communities and were noticed according to the data analysis that the government, health workers and the public have to cooperate throughout the World to prevent the spreading of the COVID-19 [11].

Data was collected from different sources from different countries regarding COVID-19 [12]. The maximum data related to COVID-19 was available at Google, WHO, CDC, ECDC, NHC of the PRC, JHU CSSE, DXY, QQ websites [13]. With the help of these data from different sources helped to analyze the people getting affected by COVID-19 and the rate of recovery and deaths were also analyzed daily. The dataset was recorded from 22 January 2020. Dataset was analyzed to record the death, survival, recovery, and people who were affected by COVID-19. The very first data suggested that males of age above years were at higher risk of infection with COVID-19. According to two other well-known diseases which are caused by coronavirus i.e SARS and MERS (Middle East respiratory syndrome) the reproduction rate of the virus was 2.5 to 3.0 days in the early stage of the outbreak. After the reproduction rate, it was reported that the doubling time of the virus was 7.5 days [14].

II. METHODS of DATA ANALYSIS

The dataset was retrieved from different sources which were further used for analysis and visualization methods. The methods which were used for analysis were able to track the spreading up of COVID-19 throughout the World. The data included the number of confirmed cases, recovered rate, and death rate in different countries. The dataset can be seen in Fig. 1 as updated on 17th May 2020 [15].

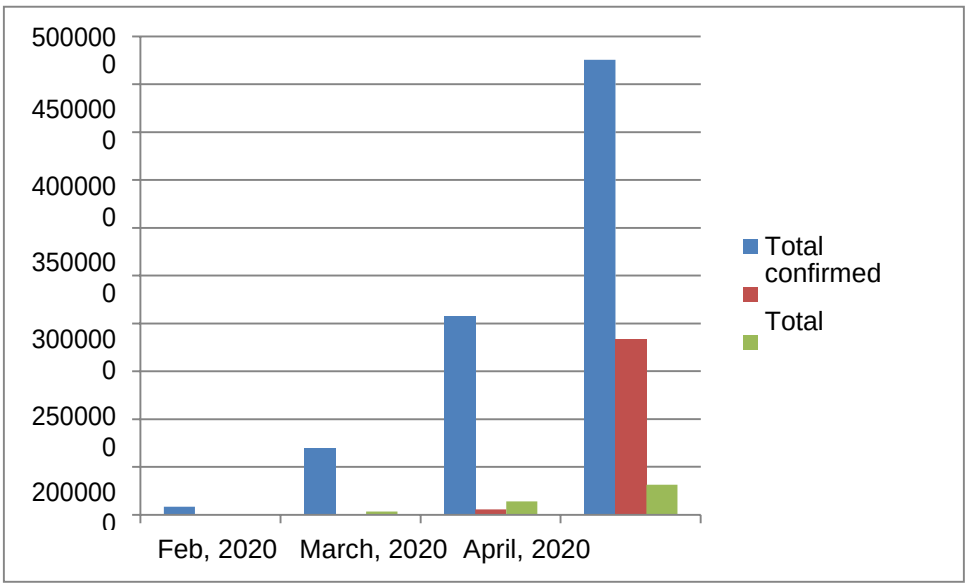

Fig. 1: Outbreak trends over time

A. Data Source

Data related to COVID-19 was retrieved from verified sources like Google, WHO DingXiangYuan, a website that is authorized by the Chinese government [16]. These sites provide information about the confirmed COVID-19 cases, the number of people recovered from the disease and the number of deaths that took place by infection of the virus [17].

B. Data Visualization

The retrieved data from different sites can be used to track the status of the corona [16]. The data collected from a different source can be seen in the table 1 [15]. The updates from the different countries can be seen through the different countries' COVID-19 portal or WHO [18].

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is used to analyze data and visualize the dataset provided by different sources regarding the emergence of the disease. The exploratory data analysis was used to record the dataset of the outbreak of COVID-19 throughout the World [19]. The first dataset was visualized and analyzed between 22 January 2020 to 10 March 2020. It was seen that the rate of people affected by COVID-19 was more from China than the rest of the World, the few affected countries were neighbors of China. After 10 March 2020 more than 30 countries and 32 states in China were affected by COVID-19 [20]. Outside China not many deaths were reported, only ten death reports were noticed until 11 March 2020. It was noticed that the rate of recovery was more than the rate of deaths and by 15 May 2020, more than 200 countries were affected by the corona. Table I shows the number of confirmed cases, the number of deaths and the number of survival till 17th May 2020 [15].

D. Visual Exploratory Data Analysis (EDA)

Visual Exploratory Data Analysis (EDA) is a method used to analyze the rate at which COVID-19 was spreading throughout the globe. In this method, the data was analyzed through a map that helps an individual to understand the epidemiological nature of COVID-19 as shown in Fig. 2. According to the data, it was noticed that China reported the highest rate of cases confirmed with COVID-19 and the highest death rate by the virus (Till 17 March 2020) followed by Italy [21]. EDA provides a piece of good knowledge about the time taken by the virus to spread throughout the globe. The data analysis through EDA is also useful in analyzing the behavior of the disease. EDA helps in understanding the situation of the COVID-19. The data for COVID-19 is available at URL

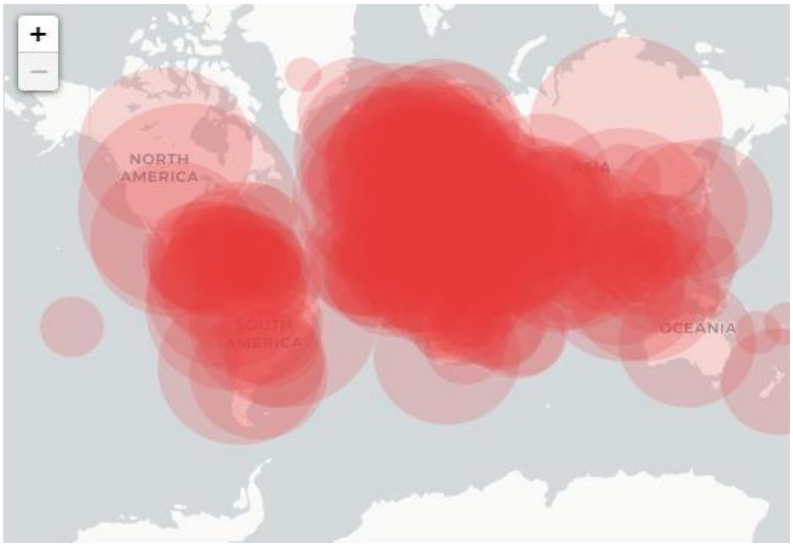

Fig. 2**:** World map of affected region, where the darker red regions in the map predict number of infected cases [15]

E. Predictive Modeling – SEIR Model

Susceptible-Exposed-Infectious-Recovered (SEIR) model is used to predict the time and the rate taken for the spreading up of disease throughout the globe. In this modeling method, real-time data is collected and visualized to forecast the rate of increasing cases for COVID-19 [22]. SEIR model predicts according to the previous data provided to forecast the number of cases that may take place in the future, it also predicts the death rate that may occur in the future because of COVID-19. SEIR model is designed to analyze and classify the news into positive and negative sentiments [31]. The result of news on the behavior of peoples both economically and politically. The properties of Susceptible-Exposed-Infected- Removed (SEIR) system is used to study the outbreak of COVID-19 throughout the World [23]. SEIR is considered to be the model for simulation studies for the disease spreading, where parameters are Susceptible (S), Exposed (E), Infections (I) and Recovered (R). In Susceptible (S), people may or may not have infection were considered, in Exposed (E), people who were incubated after encountering of the virus, in Infections (I) people after incubation showing symptoms were kept and in Recovered (R) parameter it refers to the state where no one is infected with the disease or disease-free people [30].

F. Sentiment Analysis

Sentiment analysis is done to keep a record of data which is neither too long nor too short and it is the result of the SEIR predictive model [25]. Sentiment analysis consists of a summary containing a description of more than eight words of the trained model [26] [27].

G. Statistical challenges of analyzing COVID-19 data

After the outbreak of COVID-19 in Wuhan, China, the statistical model plays a major role in comparing the number of confirmed COVID-19 cases, the number of recoveries and the number of death rate that is taking place throughout the globe as shown in Fig. 4, Fig. 5 and Fig. 6 respectively. The statistical model compares the data from origin i.e China to the data of different countries with respect to time in the form of a bar graph. The data from different countries of the confirmed cases are recorded from the very start of the outbreak of the disease. The separate data is maintained in a statistical model for the cases which are recovered from the infection and the number of deaths caused by COVID-19. The two protocols are maintained under in which closed COVID-19 cases are recorded which are as follows:-

1. International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC) (isaric.tghn.org)
2. Lean European Open Survey on SARS-CoV-2 Infected Patients (LEOSS) (leoss.net).

For COVID-19 patients, the most important clinical endpoints are the record of intensive care, invasive ventilation, and survival. The less relevant endpoint is supportive oxygen. According to these two endpoints data can be analyzed on a statistical model that will be dependent on time. The data is further collected from ISARIC and LEOSS to analyze data in the standard protocol [28].

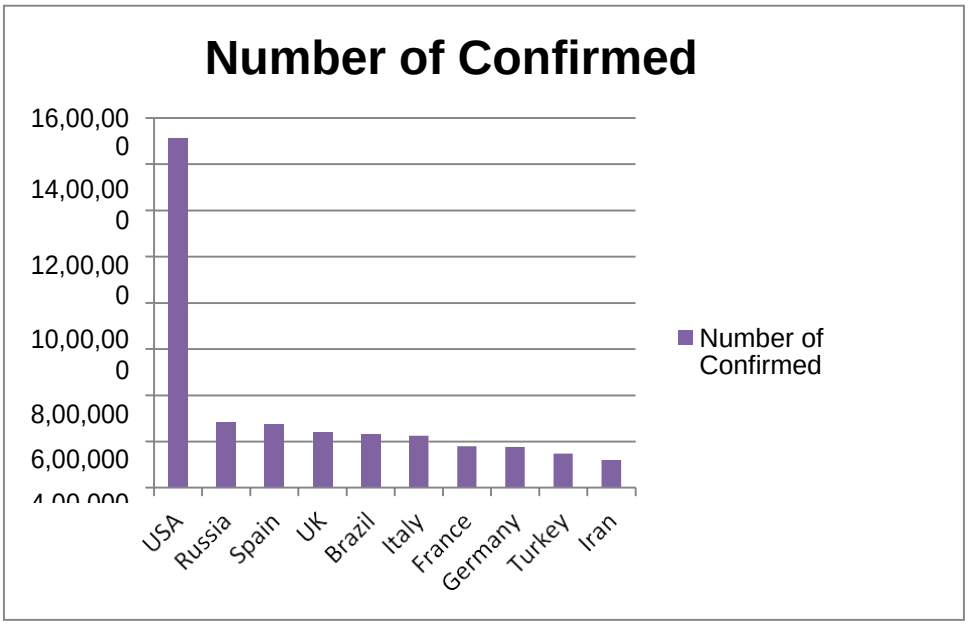

Fig. 4: Most affected countries showing number of confirmed COVID-19 cases.

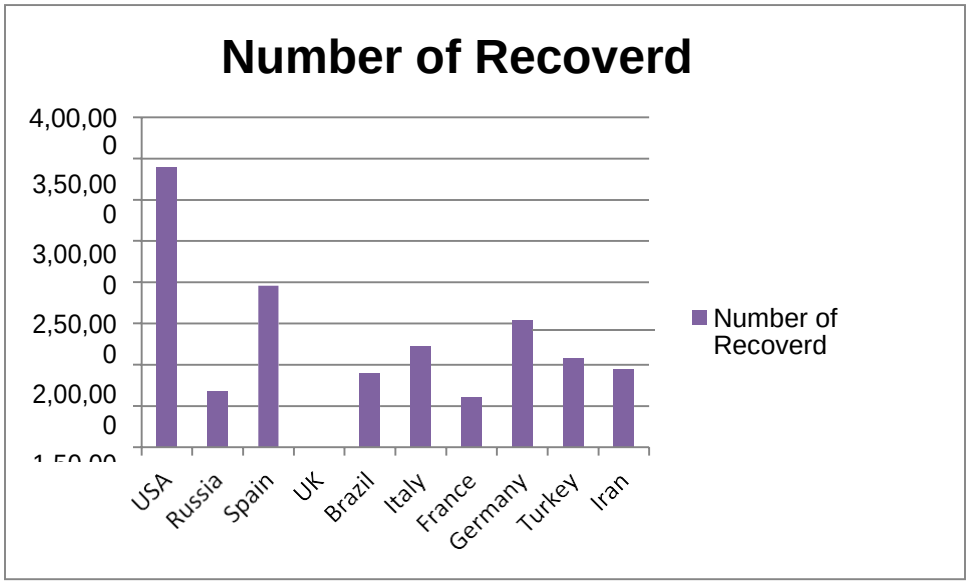

Fig. 5: Most affected countries showing number of recovered COVID-19 cases.

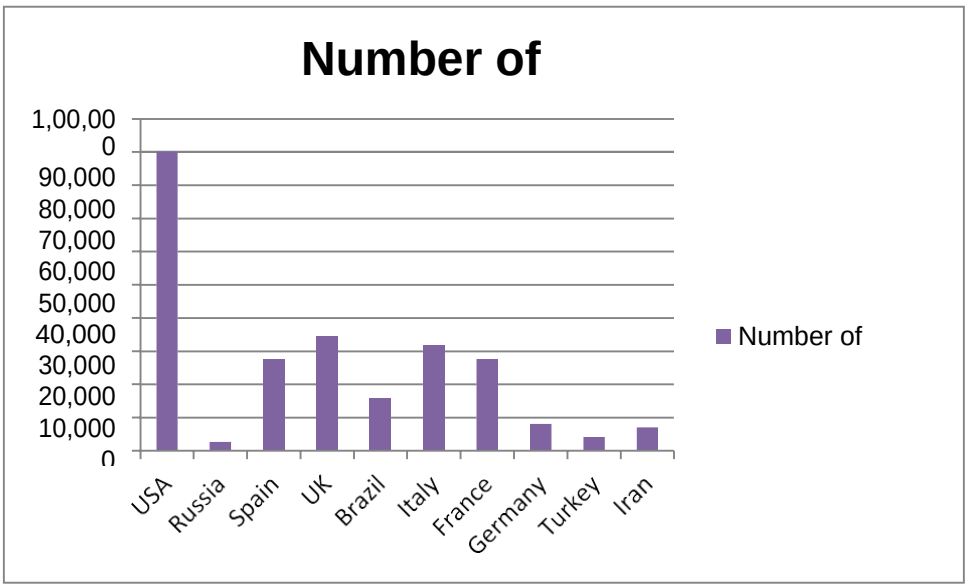

Fig. 6: Most affected countries showing number of recovered COVID-19 cases.

III. CONCLUSION

COVID-19 outbreak which took place in China was recorded and visualized through different online platforms. Data analysis was done through several methods. Exploratory Data Analysis was used to analyze data and visualize the dataset provided by different sources regarding the emergence of the disease. Visual Exploratory Data Analysis (EDA) was used as a method to analyze the rate at which COVID-19 was spreading throughout the globe. In this method, the data was analyzed through a map that helps an individual to understand the epidemiological nature of COVID-19. Susceptible-Exposed-Infectious-Recovered (SEIR) model was used to predict the time and the rate taken for the spreading up of disease throughout the globe. In this modeling method, real-time data was collected and visualized to forecast the rate of increasing cases for COVID-19 and was also used to forecast the analysis of infection. The results from the SEIR model were further used to analyze data for sentiment analysis among the community regarding the outbreak of COVID-19. The COVID-19 outbreak spreads not only through the country's policy but also through the social responsibility of each individual. The different online platform, updates the viewers with the situation of the disease including the number of confirmed cases, number of recoveries and number of deaths taking place throughout the world. Data analysis for COVID-19 is done to make aware humans against the infection caused by Corona.