Data Analysis on Covid-19

Ritvik Gautam, Mayank Raj, Abhishek Kumar and Kavita Saini

February 27, 2021

Data Analysis on Covid-19

Ritvik Gautam
School of Computer Science And Engineering
Galgotias University
Uttar Pradesh,Indiaritvik_gautam.scsebtech
@galgotiasuniversity.edu.in

MayankRaj
School of Computer Science And Engineering
Galgotias University
Uttar Pradesh,Indiamayank_raj1.scsebtech
@galgotiasuniversity.edu.in

Abhishek Kumar
School of Computer Science And Engineering
Galgotias University
Uttar Pradesh,India
abhishek_kumar4.scsebtech
@galgotias university.edu.in

Dr. Kavita Saini (Associate Professor)
School of Computer Science And Engineering
Galgotias University
Uttar Pradesh,India
kavita@galgotiasuniversity.edu.in

Abstract- The spread of COVID-19 across many parts of the world is an issue of concern for all government units of the world. India is also facing this very dark time and trying to control the spread and has managed to control its growth rate through some strict precautions. The source of data has been gathered from multiple sources and many other certified websites. The need of humanity in this dark time is to predict the future accurately by telling when number will reach its peak and when it will decrease. The help has been provided to the public welfare professionals to accommodate preventive measures by keeping the economy of the country in balance. Identities such as sex, longitudes and latitudes, age factor, etc. have been represented using R, data visualization techniques.Covid-19 analysis is the process of investigating number of people suffering from this on the basis of collection of data on growth rate through the use ofnetworks.

What we are trying to do in this project is that we are going to import an already given data in the form of a CSV file and then visualize it constantly at every step using an R package known as covid19.analytics which is a library collection for covid19 data function and it will show the deaths recovered latitude and longitude and various details of that place. We also use different libraries in this like dplyrand prophet and many more We will have time series Forecasting techniques and we will analyzing the summary reports of different countries in same time period which includes death rate, growth rate, recovered and comparison and analysis of this data has been done in order to understand the best suitable model data analysis.

Keywords- COVID-19 Analysis, Time series, forecasting techniques and regressions, R, covid19.analytics, dplyr, prophet.

I. Introduction

Novel Corona Virus, or more simply it's known as the Covid-19 virus, is transmitted to the "Nidovirus" family, or "Nidovirales". COVID-19 is associated with a respiratory disorder which has been declared a worldly pandemic in the early's of 2020 by the World Health Organization(WHO). The most common signs or symptoms of COVID-19 are uneasiness, fever, weakness, dry cough, pain, no taste in food and soreness, congestion in nose, thin fluid in nose or sore throat. Corona virus is a "hazardous" disease which can be passed through one person to other in the form of droplets through nose or mouth of an infected person when he coughs or sneeze and is a major concern for maintaining a distance of 2.5m (8 ft) from the infected person.

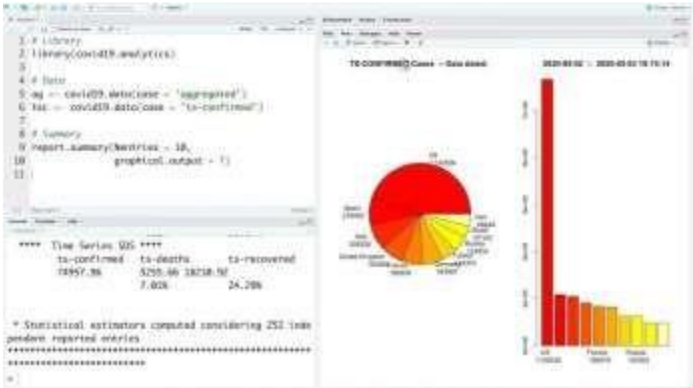
In accordance with the latest data, currently there are more than 48.8 million peoples are infected with Novel Corona virus and about 1.42M people from different parts of the Earth are reported. On January 31st, 2020, India reported the first ever case of Covid-19 in Kerala where a student has just returned from Wuhan and by then the number of case starts increasing in the whole country. No vaccine or medicine has been available in current times, especially from the recovery of COVID-19. This paper analyzes the current practice of COVID-19.

II. Scope/Objective

This analysis on different case studies is very useful for the government organization and in different areas of India, administration sections of India, Indian health workers, scholars and scientists. This study will be very good for foreign governing sections to look into various things related to COVID-19 in their areas.

Our main objective is to create a system that can provide a summary report, output (pie chart and bar graph), totals per location, growth rate, totals plot, world map. of cases in each country, SIR model (susceptible infectedrecovered).

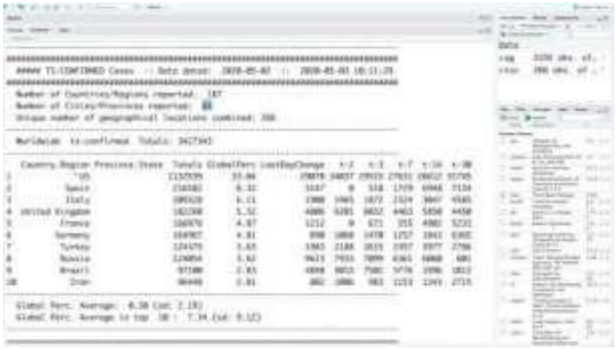
This analysis tries to work on a wider scale
Analyze the prevalence of COVID19 in India.
To answer a wide variety of research questions, specific methods were used that included various data sets and sources, modelling methods and result variability



```
# Library
library(covid19.analytics)

# Data
ag <- covid19.data(case = 'aggregated')
tsc <- covid19.data(case = "ts-confirmed")

# Summary
report.summary(Nentries = 10,
               graphical.output = T)

```

```
****  Time Series SDS ****
      ts-confirmed   ts-deaths       ts-recovered
      78957.96       5255.66 18218.92
                     7.02%           24.28%

 * Statistical estimators computed considering 252 independent reported entries
*****************************************************************
************************
>
```

CONFIRED CASES TOP 10



```
################################################################################
  ##### TS-CONFIRMED Cases  -- Data dated:  2020-05-02  ::  2020-05-03 10:11:25
################################################################################
  Number of Countries/Regions reported:  187
  Number of Cities/Provinces reported:  83
  Unique number of geographical locations combined: 266
--------------------------------------------------------------------------------
  Worldwide  ts-confirmed  Totals: 3427343
--------------------------------------------------------------------------------
   Country.Region Province.State  Totals GlobalPerc LastDayChange   t-2   t-3   t-7  t-14  t-30
1             *US                1132539      33.04         29078 34037 29515 27631 28622 31795
2           Spain                 216582       6.32          3147     0   518  1729  6944  7134
3           Italy                 209328       6.11          1900  1965  1872  2324  3047  4565
4  United Kingdom                 182260       5.32          4806  6201  6032  4463  5850  4450
5          France                 168870       4.87          1212     0   671   355  4902  5233
6         Germany                 164967       4.81           890  1068  1478  1257  1843  6301
7          Turkey                 124375       3.63          1983  2188  2615  2357  3977  2786
8          Russia                 124054       3.62          9623  7933  7099  6361  4068   601
9          Brazil                  97100       2.83          4898  8023  7502  5776  1996  1812
10           Iran                  96448       2.81           802  1006   983  1153  1343  2715
--------------------------------------------------------------------------------
  Global Perc. Average:  0.38 (sd: 2.19)
  Global Perc. Average in top  10 :  7.34 (sd: 9.12)
--------------------------------------------------------------------------------
```

```
Data
agg     3250 obs. of ...
tsc     266 obs. of ...
```

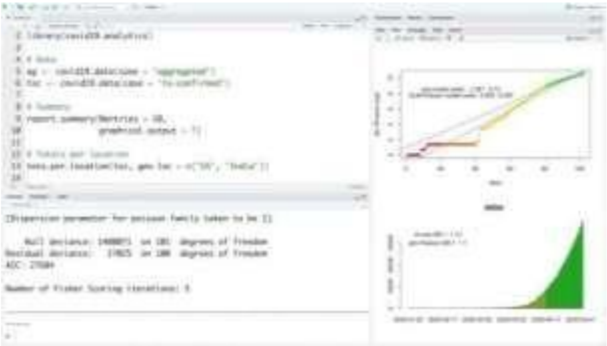
SUMMARY REPORT

III. LITERATURE REVIEW

According to the various papers present in other's work, there exist some study that keep an eye on trend insights and predicting for the Indian portion. Study on Indian Territory presents both short and long timeline trends.

These study reports use TS data from the JHU database and put before us forecasts with the help of the ARIMA model, exponential plaining methods, the SER model, and the regression model. In addition, studies in the India portion from the past time are more focusing on presenting TSA on the basis of aggregate data for the India portion.

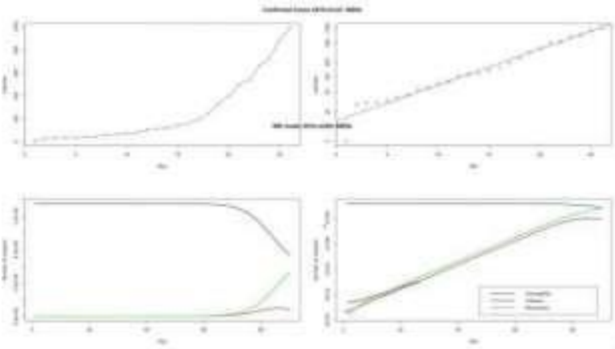
TOTALS PER LOCATION



```
library(covid19.analytics)

# Data
ag <- covid19.data(case = "aggregated")
tsc <- covid19.data(case = "ts-confirmed")

# Summary
report.summary(Nentries = 10,
               graphical.output = T)

# Totals per location
tots.per.location(tsc, geo.loc = c("US", "India"))
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1408071  on 101  degrees of freedom
Residual deviance:   17025  on 100  degrees of freedom
AIC: 27884

Number of Fisher Scoring iterations: 5
```

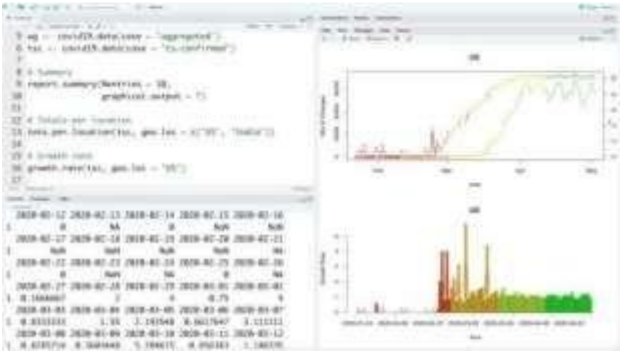
SIR MODEL-INDIA



Similar to others, there exist other mathematics based models planned that were made to analyze COVID-19 outbreak trends in India. A model was presented to study the effect of social distinction on the basis of age and sex of sick people in Indian subcontinent. It compared the demographics of the countries between Indian, Italiano and Chinese and gave suggestions on the weakest age group and gender groups among every on of the countries. The study also forecasted an increase infection cases with varying locked down times in India. In the same manner, a interconnected structure way was used by a study to watch if a particular node cluster was coming into existence. Only the travelling info node were taken into consideration by the writers to examine what major areas are affecting the return of Indian tourists to Indian region.The study also put before us the SER models to look at the rate with which the spread of virus in patients in the Indian region. Analyzing on laboratories where testing takes place and infra was also put before us by previous authors.

The work of doctors and health man force was also put before us by some study. In India, the role of health employees were less focused upon as the stages of the expansion of virus were in 2 or more stages, as compared to other countries such as Italian ,Spanish and the US.

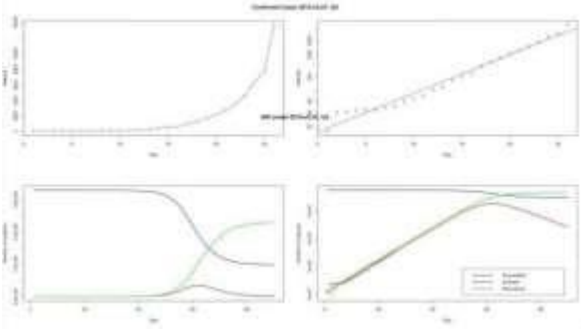
TOTALS PER LOCATION



```
ag <- covid19.data(case = "aggregated")
tsc <- covid19.data(case = "ts-confirmed")

# Summary
report.summary(Nentries = 10,
               graphical.output = T)

# Totals per location
tots.per.location(tsc, geo.loc = c("US", "India"))

# Growth rate
growth.rate(tsc, geo.loc = "US")
```

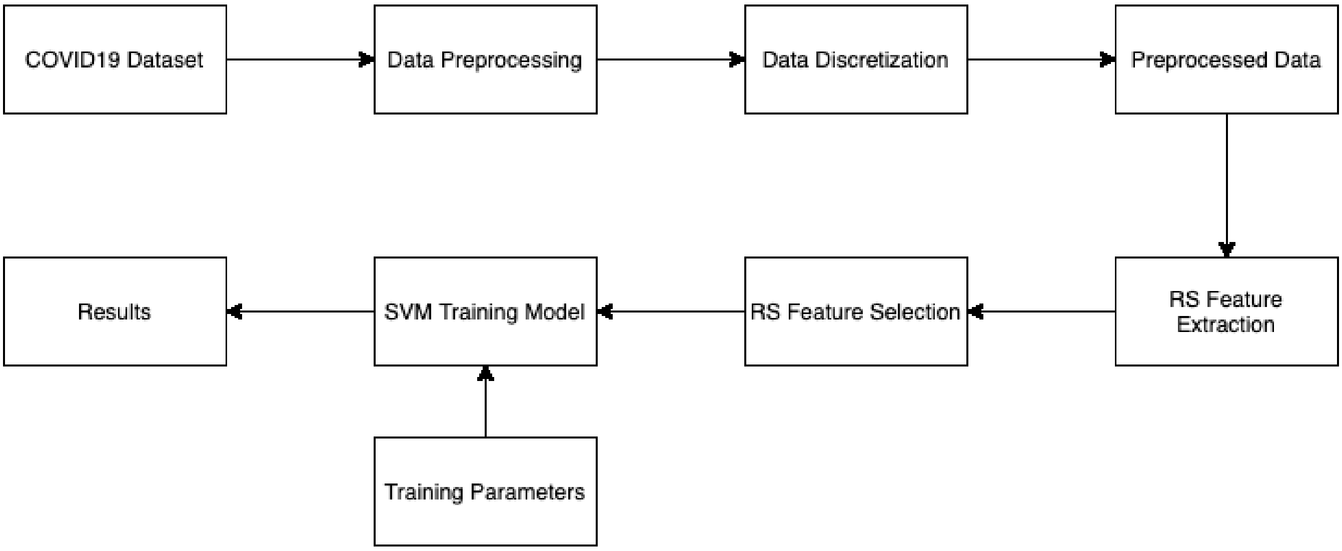
```
  2020-02-12 2020-02-13 2020-02-14 2020-02-15 2020-02-16
1          0         NA          0        NaN        NaN
  2020-02-17 2020-02-18 2020-02-19 2020-02-20 2020-02-21
1        NaN        NaN        NaN        NaN         NA
  2020-02-22 2020-02-23 2020-02-24 2020-02-25 2020-02-26
1          0        NaN         NA          0         NA
  2020-02-27 2020-02-28 2020-02-29 2020-03-01 2020-03-02
1  0.1666667          2          0       0.75          5
  2020-03-03 2020-03-04 2020-03-05 2020-03-06 2020-03-07
1  0.8333333       1.55   2.193548  0.6617647   3.111111
  2020-03-08 2020-03-09 2020-03-10 2020-03-11 2020-03-12
1  0.4235714  0.7604444   5.194675   0.050383   1.166370
```

SIR MODEL US



Keeping India aside, models are present for other country too mainly for Chinese, Italian and US as the no. of infected patients was really big. The study has worked on different mathematical forms to gauge the spreading of the virus, predicted the no. of infected people, commented on each country's readiness to deal with the spread of COVID-19, and flattened curves under varying conditions. Discovered a lot of patterns. Research for the World Forum is still in its first phase.

In relation to the research actions that took place in India, the study remains to change on the effect of various policy making efforts towards the control of this deadly virus. This study tries to work extensively for the analysis of COVID spread in India.



Architecture of RS Based SVM System

IV. METHODOLOGY

We have given description of our model in detail. Firstly, we summarize the dataset obtained from the Johns Hopkins Center for System Science and Engineering and tested pre-processing methods. Then we have given description of the machine learning process to predict the spread of the epidemic.

A. Dataset preparation and datapre-processing

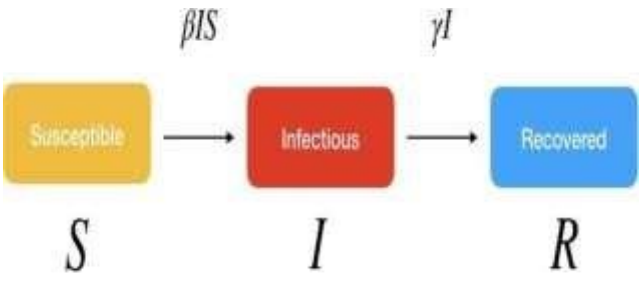
- Real-time data from the Johns Hopkins Center for Systems Science and Engineering was usedin this project. The following steps for data processing are as follows:

- Data is cleaned up and null values are replaced by column values.

- The data were converted using a standard scalar tool in Python to facilitate Gaussian distribution to predict the spread of the epidemic.

- Optimized using a logarithmic scale to exclude data

B. Mathematical modeling usingSIRModel



The SIR model is a mathematical model that computes the theoretical number of people infected and is used to predict infectious diseases. The aim of the model is to find the relation between dependent and independent variables.

- 't'is the Independent variance

One group looks at individuals and the other group looks at a small group of people..

- We see a group made up of a small group People for better accuracy.

- Is a parameter that measures transmission from one person to another. The probability of communication and transmission of the disease are determined..

$$-\beta IS=ds/dt \quad (1)$$

$$\beta IS-\gamma I=di/dt \quad (2)$$

$$\gamma I= dr/dt \quad (3)$$

$$D = 1/\gamma \quad (4)$$

By multiplying the original reproductive number by the percentage of non-immunized peopleThe balanced state of the disease is obtained and is equal to 1.

$$R_0=\beta/\gamma \quad (5)$$

Let people who are immune to disease isp. The steady state of the epidemic can then be represented

$$R_0(1-p)=1\rightarrow 1-p=1/R_0\rightarrow p_c=1-(1/R_0)(6)$$

C. *Prediction using RS-SVM.*

SVM is a classical machine learning algorithm that is used to make forecasts. Rough Set (RS) attribute reduction SVM is used. This method increases the accuracy of forecasting.

This algorithm works as follows:

- The deduction set is considered empty.

- Separation conditions are determined according to test.

- quality of best distinguishing feature is selected.

- test is repeated several times until a set of U defined cuts is available

D. Prediction using polynomial regression

Polynomial regression is a quadratic method of linear regression. The great feature of using polynomial regression is to remove dependencies between variables that may be unequal.

.

E. Prediction using RNN

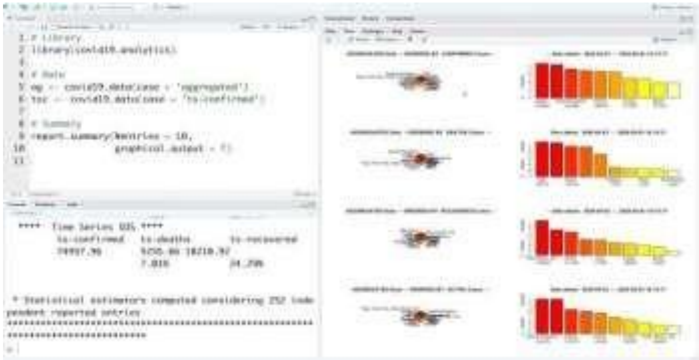
Neural networks (RNNs) regularly form a neural network algorithm section with internal memory in each hidden layer. Maintains information previously calculated.

Maintains the information previously calculated. RNN repeats because they set the same parameters for each input to each hidden layer. Inputs are processed by weight and bias in each hidden layer. Once the product is produced, it is copied and returned to the iteration network.

We used the LSTM (Long Short-Term Memory) procedure to estimate epidemic mortality. The great advantage of using LSTM is that it is the right way to classify, process and forecast.

V. Problem Formulation

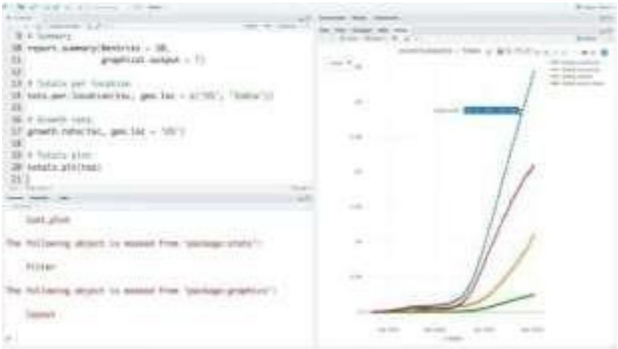
The COVID-19 cases in India is increasing continiously. National and local authorities have a difficult time compiling a pattern, analyzing and predicting the spread of COVID19 in India. The objective of this paper is to create a statistical model to better understandthe COVID-19spread in India by a careful study of the reportedcases.



```
# Library
library(covid19.analytics)

# Data
ag <- covid19.data(case = 'aggregated')
tsc <- covid19.data(case = 'ts-confirmed')

# Summary
report.summary(Nentries = 10,
               graphical.output = T)
```

```
****  Time Series [illegible] ****
       ts-confirmed    ts-deaths         ts-recovered
       79957.96        5250.86 18210.92
                       7.01%             24.29%

 * Statistical estimators computed considering 252 inde
pendent reported entries
**************************************************************
*************************
>
```

As we know some efforts have been made for the analysis of the spread of Covid-19. We aim to create a system that can provide a summary report, output (pie chart and bar graph), totals per location, growth rate, totals plot, incoming world map. of cases in each country, the SIR model (susceptible infected recovered) and forecasts appropriate number of cases in the future. Therefore, we can prepare and take steps to protect ourselves.



```
# Summary
report.summary(Nentries = 10,
               graphical.output = T)

# Totals per location
tots.per.location(tsc, geo.loc = c('US', 'India'))

# Growth rate
growth.rate(tsc, geo.loc = 'US')

# Totals plot
totals.plt(tsc)
```

```
    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
>
```

TOTALS PLOT

VI. Tools Required For Implementation

- R
- RStudio
- Libraries/Packages: i) covid19.analytics ii) dplyrlub
  iii)prophet iv) ridate
  v) ggplot2 vi) tidyr vi) gganimate
- Google Developer APIkey

VII. FeasibilityAnalysis

The corona poses a number of challenges to ongoing clinical trials, including new and non-standard causes of disappearance, including essentially high rates of missing outcome data. The International Drug Trial Guidelines advise examiners to review plans to deal with missing data in conduct and statistical analysis, butclear recommendations are lacking. Since the virus is new to us and little is known about it, there may be discrepancies in predictions because the data is not very accurate and therefore not giving an approximate number. Minor to major changes may occur in infected people in thefuture.

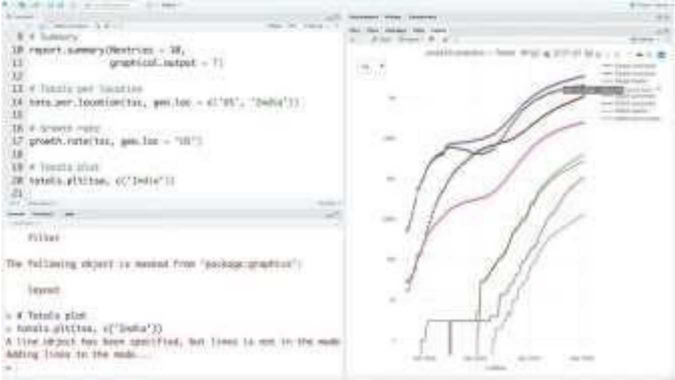
VIII. Complete WorkPlanLayout

A. FirstPhase

We will start by writing the programming part on R studio on covid-19 analysis in which first we will install some libraries(covid19.analytics,prophet,dplyr,lubridate,ggplot2 ,etc) from packages which will help in understanding the analysis of the covid-19.In this analysis first we will be storing inbuilt data functions of particular libraries in function by saving it and it will gave information about the confirmed cases, recovered case, total per location, growth rate, totals plot , deaths of people through covid-19 pandemic along with names of cities and countries with their latitudes and longitudes giving us the exact location. Then we will be producing a timeseries summary report of top 10 or 100 countries which has the following details: For Confirmed Cases and For Death Cases we willrepresent:

- Countries and Citiesreported
- Graphical outputs showing in the form of graphsand pie-chart
- Cases in comparison with globalpercentage

We also have a data function in this analysis through which we can see a picturization of world-map just like in google earth and it will show the total number of cases in each country.



```
# Summary
report.summary(Nentries = 10,
               graphical.output = T)

# Totals per location
tots.per.location(tsc, geo.loc = c('US', 'India'))

# Growth rate
growth.rate(tsc, geo.loc = 'US')

# Totals plot
totals.plt(tsa, c('India'))
```

```
    filter

The following object is masked from 'package:graphics':

    layout

> # Totals plot
> totals.plt(tsa, c('India'))
A line object has been specified, but lines is not in the mode
Adding lines to the mode...
```

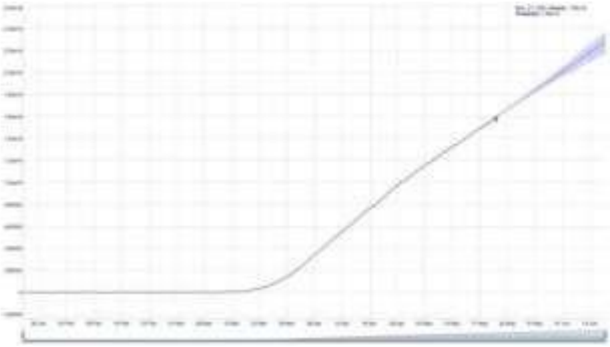
TotalsPlot

B. SecondPhase

Gathering multiple review papers i.eabout 50 and then we will be skimming through them, picking important points and connecting the gaps between them and then drafting a review/survey paper on our own and try publishing it in a journal.

IX. Merits Of The ProposedSystem

- Forecast of growth in cases infuture.



- Help in Better Prevention of upcomingsituation.
- Getting the exact idea of current scenariousing different representations (pie charts,graphs).
- Understandingtheoverallspreadintheworldby latitude and longitude.

World Map showing No. of Active Cases

X. Limitations

“Data is not reality.it only tries to approximatereality.”

- Every COVID cases may not be reported,checked and confirmed.
- Lag due to reporting or dataentry.
- Actual v/s predicted plot may notbe linear.(underestimate or overestimate)

Covid-19 data limitations are a challenging factor in Predicting time series data. The extension of the Recurrent Neural Network (RNN) as a short-term memory (LSTM) cell and its variants such as Stacked LSTM, Bi-Directional LSTM and Convolutional LSTM are used to measure predictions

XI. Future Scope Of theObject

This paperwork can be extended to higher levels in the future.

In the future, we can study the loss of the total economy at the end of Kovid-19 in various regions and prepare a reasonable plan to recover it, to help countries recover the economy rate.And the aerosol transmission of Kovid-19 can be verified

In addition the analysis of future prediction can be extended resulting in more exact prediction to estimate a more accurate total number of cases in India.

XII. Conclusion

The main objective of the paper is to study and analyze COVID-19 prevalence. The spread of the disease in India compared to other countries, state-wise trend of the epidemic to know how it is spreading and to analyze the health sector of India and finally to predict the future of the epidemic in India.

XIII. Acknowledgement

I would like to express my deep appreciation to all team member **(Ritvik , Mayank , Abhishek)** who provided me the possibility to complete this report. A special gratitude I give to my Guide, **Dr. KAVITA Ma’am**, whose contribution in stimulating suggestions and encouragement, helped me to coordinate and writing this review paper.

XIV. References

❖ https://www.mohfw.gov.in/

❖ YouTube

❖ https://www.covid19india.org/

❖ HoseinpourDehkordi

❖ Andrea Apolloni, Chiara Poletto, and Vittoria Colizza. “ Age-specific contacts and travel patterns in the spatial spread of 2009h1n1 influenza Pandemic”, BMC infectious diseases, 13(1):176, 2013

❖ ., “ Disease prediction by machine learning over big data from healthcare communities”,

❖ PahulpreetSingh Kohli and ShriyaArora, “ Application ofmachine learningindis ease prediction”. In 2018 4th International Conference on Computing Communication and Automation (ICCCA), pages 1 –4. IEEE, 2018