COVID-19 DISEASE ANALYSIS, PREDICTION AND SOLUTION PROPOSITION

NANDHINI V
NITHYA KAMAL
NIVEDHA N
PRAVEEN KUMAR V

1. ABSTRACT

The pandemic of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is spreading all over the world and has become the most pressing issue for mankind. The COVID-19 disease has changed the global landscape completely. A high reproduction rate and a higher chance of complications have led to border closures, empty streets, rampant stockpiling, mass self-isolation policies, and an economic recession. The Supply Chain Management Systems that have been in practice in industries till now have not been able to meet the demand and supply surge amongst the industries across the globe. With so much information and expert opinions, to see different nations adopting different strategies, from complete lockdown to social distancing to herd immunity, one is left thinking as to what the right strategy is for them. Is there any basis to these opinions and advice? An attempt of data modelling and analyzing Coronavirus (COVID-19) spread with the help of data science and data analytics in python code is necessary. This analysis will help us to find the basis behind common notions about the virus spread from purely a dataset perspective. This paper has thoroughly explored the present covid - 19 situation across the world in order to get a clarity towards predicting the future. The paper then concentrates on bringing an almost accurate prediction of the spread of the covid - 19 disease using two different models, a machine learning model and a deep learning model. The reason for using two different models for prediction is to bring an analogy between the predictions and come out with the most accurate one. Having analyzed the main problem, the paper concludes by throwing some light on focused testing, a solution that's proposed to control the effect of Covid - 19 pandemic.

2. INTRODUCTION

Even after the distribution of vaccines in full swing in our country, the disease cannot be totally eradicated and there are ample chances for the rise in covid cases and hence the emergence of third wave. The insufficiency of vaccines makes the situation worse for the already overstretched Indian health care system. The purpose of complete or partial lockdown is to slow down the spread of the novel coronavirus. We're unsure of the emergence of a third wave in our country. Thus, it is difficult for the Government to pass an order to completely open schools, colleges, theatres and other public places. Even people who have completed both the doses of vaccination are instructed to follow the COVID -19 protocols. The important question for Indian health officials is how many new confirmed cases will be seen in the future and by what time? It is not economically sustainable to have another lockdown, as the large workforce in India is employed in the informal sector as a daily wage laborer. Keeping this in my mind we have taken a three-way approach. One is to understand the severity of the ground situation; and the second is the prediction, which will help the health officials to make the plans accordingly; Third is the solution proposition which will definitely not eradicate the disease completely but would tide us over in the future.

3. RELATED WORK

Various types of research have been carried in order to detect the onset of Covid-19. For prediction, a variety of machine learning approaches are utilized, each with a distinct level of accuracy.

Random forest classifier and Extra tree classifier algorithm was used by Deepak Painuli et al. to forecast Covid 19. They used autoregressive integrated moving average (ARIMA) time series analysis for forecasting. The accuracy of Random Forest classifier is 91.63% and Extra tree classifier algorithm is 93.62%.

Kolla Bhanu Prakash et al. used supervised machine learning techniques such as SM, KNN+NCA, Decision Tree classifier, Gaussian Naïve Bayes Classifier, and XGBoost Classifier to analyze and prediction of Covid 19. They able to obtain an average accuracy of 91.4%.

Epidemiological SIR model and Statistical Machine learning model was used by Sourish Das et al. for prediction of Covid 19 disease. They estimated the basic reproduction number and predicted the cases ahead of time using the same. Using evaluation of R0 they predicted the progression.

J. N. Dhanwant et al. forecasted Covid 19 growth in India using Susceptible-Infected-Recovered (S.I.R) model. They used a separate algorithm which was hardcoded in python using SciPy that learns the social-contact structure and gives a suitable value for \b{eta}.

Burhanuddin Bhopalwala et al. gave a potential Machine Learning approach that can help control COVID-19. They developed Priority based Covid 19 testing model and that helps in determining the priority of giving limited healthcare services to highly populated countries. They used some Natural Language Processing techniques and KMeans algorithm to obtain desired results.

4. PROPOSED WORK

4.1 EXPLORATORY DATA ANALYSIS

Up to date time series covid dataset is obtained from John Hopkins University and is used throughout the project. Confirmed, recovered and death cases for 279 countries across the world are tabulated to a data frame using the Pandas library. In the most confirmed cases list, the US has topped the table. Italy has the most death cases while China is far ahead in recovery count of cases.

4.2 POLYNOMIAL REGRESSION

Covid cases with restricted timeline rather than in entirety were found to give comparatively better results. Hence the covid cases for a restricted time series is chosen for training and testing where test size is chosen to be 25%. The Linear regression model is trained and tested to come with an accuracy of 99.993 % which is pretty good and the graph lines plotted for actual and predicted cases also have closer proximity. For readability, the covid cases for upcoming 20 days are forecasted for both global cases and India and tabulated in the form of a data frame.

4.3 LSTM MODEL

The deep learning model serves better for a large pool of time series data. LSTM is especially chosen because it works better remembering past information. Data is preprocessed using min-max scaler of scikit learn. Test size is chosen to be 20%. The model has one input layer followed by three LSTM layers. LSTM layers have a dropout of 0.5 to avoid overfitting. For model compilation, loss is taken as “mean-squared error” and optimizer as adam optimizer.

Learning rate of our model is reduced by ReduceLROnPlateau. Model is trained with 200 epochs and batch size of 100. Trained model has a loss of 0.0021 and mean squared error of 0.0021 which is almost zero and obviously a better performance metric. The graph lines plotted for actual and predicted cases also have closer proximity. For readability, the covid cases for upcoming 20 days are forecasted for both global cases and India and tabulated in the form of a data frame.

4.4 A POTENTIAL APPROACH

A priority-based testing is achieved by pbATS mechanism which is again an AI/ML approach. For pbATS, population is categorized based on features(age/sex), travel history and symptoms. As a part of data preprocessing, feature engineering is done. All the standard NLP techniques for vectorizing the master-symptom are performed. BoW+W2V (Word2Vec-gensim) is used. Clustering is done using KMeans++. Thus, from the obtained word cloud of master symptoms, people can be categorized and tested respectively.

5. EXPERIMENTAL ANALYSIS

5.1 MODEL SELECTION FOR PREDICTION

We've tried out different regression models and found polynomial regression to serve good and give better accuracy of our model. Since the data pool is increasing day by day, a deep learning model serves better to handle a large amount of data. The reason why we went with RNN for Time Series prediction instead of other models is, LSTMs are better at identifying complex pattern logics from data by remembering what's useful and discarding what's not. The LSTM model which is being used for forecasting has an exponential trend in the number of Covid-19 cases which is quite similar to the Real number of cases. This model also gives better results if it is trained with more epochs.

5.2 DATASET RESTRICTION

Confirmed, death and recovered cases from Jan 1 2020 to till date are not found to be feasible for our polynomial regression model. So, we've planned to restrict our dataset to a certain timeline in such a way that the accuracy is increased. Same technique is applied to predicting India cases using LSTM in order to get better performance.

5.3 A POTENTIAL APPROACH

We planned to categorize the population based on the features (age, sex), symptoms, and past travel history. We have performed all the standard NLP techniques for vectorizing the master-symptom. BoW+W2V (Word2Vec-gensim) is used. W2V is used because we need to cluster the symptoms based on the relationship (not similarity & counts), which helps in the clustering process. BoW is used instead of TF-IDF because our Dataset does not have many rare, occurring words that need more importance.

6. CONCLUSION

After meticulously understanding the ground situation, an almost accurate prediction of the future is attempted. As per the results the confirmed cases seem to have been increasing in the same rate without sudden fall or rise. We've also proposed a solution that implements focused and prioritized testing which tides us over but it would definitely not eradicate the disease. In order to prevent the emergence of a third wave, at least partial lockdown and safety measures have to be followed. Increased distribution of vaccines is also luckily in favor of us.

REFERENCE

- **https://www.frontiersin.org/articles/10.3389/fpubh.2020.00357/full**
- **https://www.researchgate.net/publication/341821120**
- **https://arxiv.org/abs/2004.03147**
- **https://www.semanticscholar.org/paper/Forecasting-COVID-19-growth-in-India-using-(S.I.R)-Dhanwant Ramanathan/3b2e57cabc0b994cc47d2efeac73f29e73810a21**