Chair in
Information and Service Systems

UNIVERSITÄT
DES
SAARLANDES

Covid19 Global Forecasting

Lecture	Data Science summer semester 2020
Participants	Danish Shahzad, Danish (2572862), dani.shahzad87@gmail Hafeez Ullah, Hafeez (2572872), hafizhafeezullah05@gmail
Submission date	07-17-20
Chair	Univ.-Prof. Dr.-Ing. Wolfgang Maaß Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken

Executive Summary

Covid-19 prediction is the new hot topic in datascience these days to see the predicted trajectory of this epidemic. A pandemic caused by the SARS-CoV-2 virus has been in a state of emergency worldwide for some time now, and its impact on everyday life and the world economy in general was barely predictable at the beginning. In this report, We 'll be foscusing on visualzation of past Covid-19 data and upon the results of past data, We are going to have forecasting for 5-chosen countries. We have made a choice of those countries which have most deaths currently. We have made it generic so any country can be in the forecasting. Up to date, Its working on currently (US,Brazil,UK,Mexico,Italy). Objective of this project is to study COVID-19 outbreak with the help of some visualizations techniques. We have made forecasting for confirmed cases possible along death cases. We have also predicted hospital bed needed for their trajectory based on their confirmed cases to be possible in next 60 days.

Results show that COVID-19 doesn't have very high mortality rate as we can see which is the most positive take away. Also, the healthy Recovery Rate implies the disease is curable. The only matter of concern is the exponential growth rate of infection.

Contents

1 Introduction

The current outbreak of the new coronavirus (SARS-CoV-2) have affected virtually every person on Earth, either directly or indirectly. Many people will die of the infectious disease caused by this coronavirus (Covid-19), and others will lose people close to them. Many more will suffer other extreme hardships – psychological, social and financial – due to the extensive physical distancing measures that are reducing the spread of the virus. While there may be some perceived "silver linings", such as temporarily reduced air pollution and CO2 emissions, and for some an opportunity to slow down and contemplate their ways of living, in the balance the effects are already tremendously challenging for the world, and are likely to get much worse before the pandemic is over. Its been almost 4 months that Covid-19 has spread all over globe. In terms of data, there have been huge addition in dataset with every coming day.
We have forecasted and showed analysis of 5 countries as our project requirement. We have made it generic so any country can replace those selected countries, based on top deaths. First, we gave overall analysis and data visualization for all the countries in terms of confirm cases, deaths, recovery cases, current active cases. We have forecasted the future trend up to 60 days. We have implemented and trained our model on machine learning algorithms and it is discussed in analysis section. For the medical equipment needed to fight against the novel virus, it is difficult to predict as it depends on a lot of factors. Country GDP, economic growth, budget for medical care and how expensive the treatment is.
So, it is very difficult to forecast as of now. Maybe some special dataset containing all the dependable factors may come up soon. In our case, we have predicted based on beds availability to their whole population and forecasted upon this feature.

Currently, WHO working on it and gives and authentic research on Covid-19. WHO is bringing the world's scientists and global health professionals together to accelerate the research and development process, and develop new norms and standards to contain the spread of the coronavirus pandemic and health care for those affected. The solidarity of all countries will be essential to ensure equitable access to COVID-19 health products. The current COVID-19 pandemic is unprecedented, but the global response draws on the lessons learned from other disease outbreaks over the past several decades. As part of WHO's response, the R&D Blueprint was activated to accelerate diagnostics, vaccines and therapeutics for this novel coronavirus. The Blueprint aims to improve coordination between scientists and global health professionals, accelerate the research and development process, and develop new norms and standards to learn from and improve upon the global response.

According to WHO for research gap, Governance become a key challenge in many countries, and not that much research is conducted on pandemic governance and needs to be considered as a core research gap. , There seems to be a strong gap of incorporation of biological hazards in disaster response, recovery and long-term preparedness. New research is required in the areas of supply chain management, business continuity planning and short to medium term response and recovery planning. We also did not notice much research on the risk assessment methodologies, which is an integral part of the disaster response and risk reduction. Multi-disciplinary research incorporating public health, disaster risk reduction, economics of pandemics, social psychology, anthropology, sociology, psychology and ecology are required.

2 Data Set

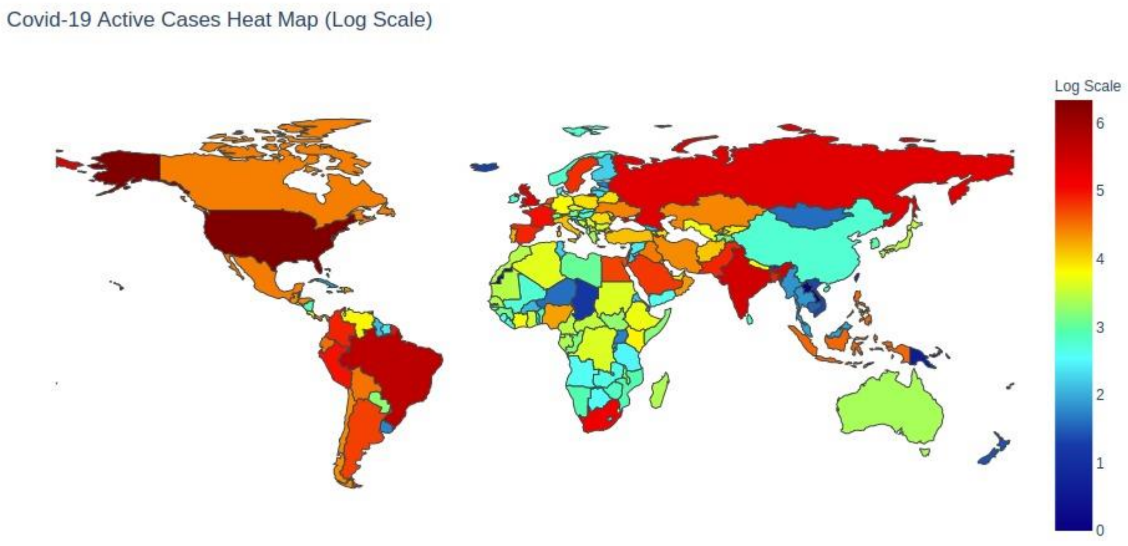
We have taken 4 datasets to perform on our prediction. It was very difficult to fulfill requirement from one dataset due to non-availability of feature all in one dataset. 3-datasets taken from John Hopkin University and 1 Dataset taken from GitHub. The one which we taken from GitHub has beds information which we needed in predicting medical condition needed to accommodate patients get infected from covid-19. While john Hopkin 3 datasets contained each for confirmed cases, confirmed deaths, recovered cases. Recovered cases information was not available in GitHub dataset. So, for that reason we chose multiple dataset to perform prediction.

The datsets we chose had records from the start of pendemic mostly from March and till July beggining. It has a lot of junk information which does not needed so we did drop those in the start of code. A lot of dataset with very few information was available so we had to annotate data first. There were a lot of information that we were indended to aim to use in prediction so we drop those. One of the biggest data set we use was composed of 28k rows contating a lot of infromation for every country.

Data pre-preparation was the most difficult steps in our Data Science project. The reason is that each dataset we taken was different and were short of information and were not highly specific to the project. Nevertheless, there are enough commonalities across predictive modeling. Pre-process provided us a data preparation required for the project, informed and the evaluation of machine learning algorithms performed after pre-processing.

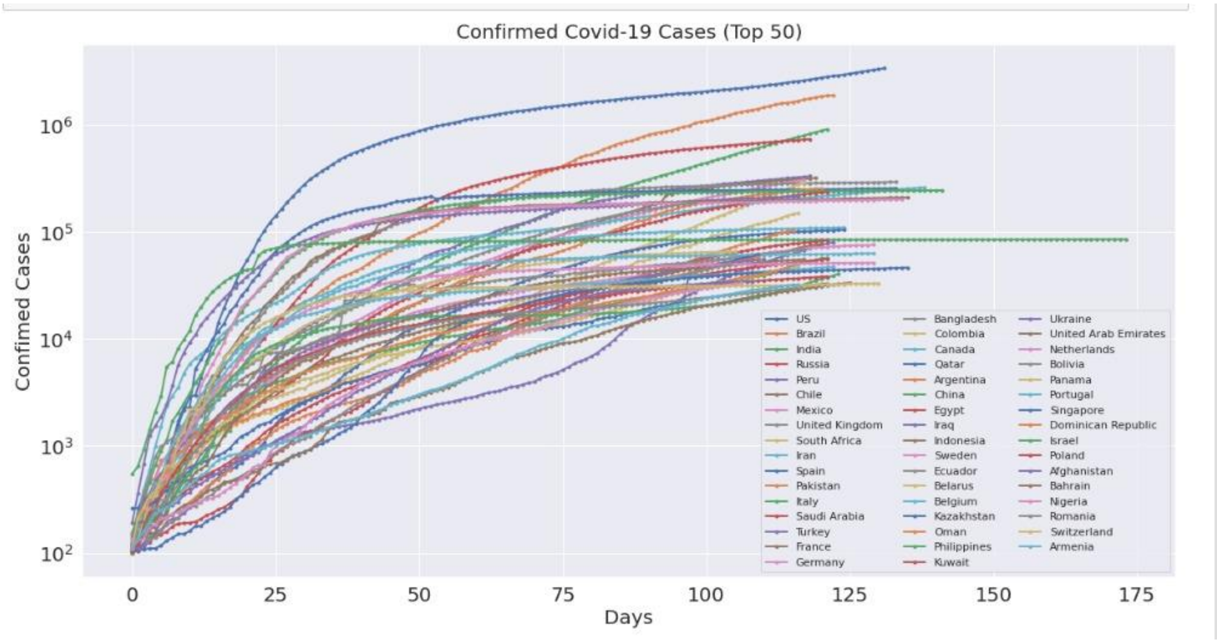
3 Procedure and Analysis

Our major problem was our 3rd member of Group left us in the start of project when we got assigned with project. Then Danish and I(Hafeez) decided to work on project with the model in mind LSTM-RNN after discussion. Out of many machine learning algorithms we chose LSTM-RNN. LSTM-RNN fits best in our process and analysis. LSTM is like a recurrent neural network (RNN) architecture that remember values over arbitrary intervals. LSTM is well-suited to classify, process and predict time series given time lags of unknown duration. Relative insensitivity to gap length gives an advantage to LSTM over alternative RNNs, hidden Markov models and other sequence learning methods. Since recurrent networks possess a certain type of memory, and memory is also part of the human condition, so it works as repeated analogies to memory in the brain.



Visualization of Global Active Cases.

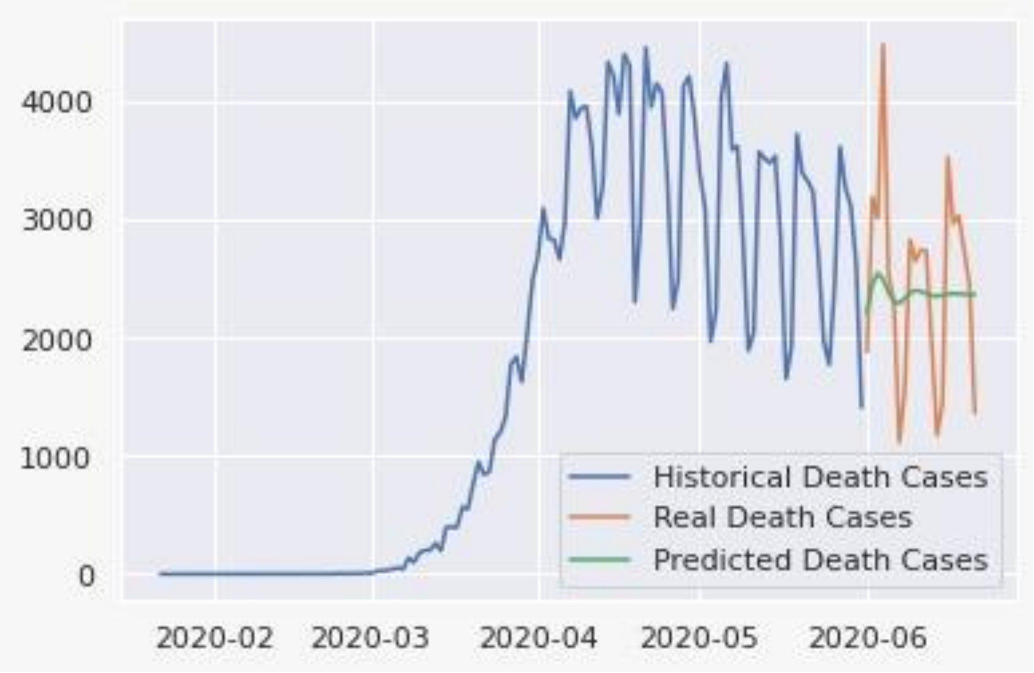
First, we did clean datasets and from those we visualized the current trend analysis of all the globe. Attach pictures represents global trends. Then we applied LSTM-RNN to predict deaths, confirm cases expected trajectory. And we also further performed analysis on medical health facilities that should be met. All the model was trained on it. We trained our model for more than 400 epochs which provides less validation error in prediction in only 1 time series feature. There was one issue that we were only able to use 1 feature at a time because of the complexity issue, we used log scale because in limited time there was huge number of cases that we needed to show. Heat map did a great job for us in visualization. One of the other reasons to choose LSTM-RNN, LSTM easily works in gradient explosion and gradient vanish problem. So, using this model allow us to efficiently train network. We have also implemented Sklearn, data normalizer, which is also a model based on scaling.



4 Results and Discussion

Please write in whole sentences and use the following key points as assistance.
After training our model for over 400 epochs we are finally able to make a progress with prediction for future. Our results are relative comparable with any other prediction done till now. Our results shows prediction up to next 60 days based on confirmed deaths and confirmed cases, and medical conditions. We also overcome on the problem of initialization in convergence.

Data was not consolidated and were missing info, biggest problem was data being not aligned, it should be in trend analysis, we did resolve this issue. one issue is scale of data, e.g no of confirm cases are in millions and no of deaths are in thousands, that’s why we did not considered multi feature training.



Prediction of feature based on deaths.

Bibliography

Pytorch

Github

Stackoverflow

John Hopkins University

Appendix

- The code is formulated in Jupyter notebook to make it, self-explanatory and every task is well described. Apart from that general description of the data organization and methodology is as follows.
- First not a single dataset contains all useful information if have, the naming convention is not generalized. We have used John Hopkins dataset for formulating network prediction and trend analysis, beside that we were also ambitious to study the medical and political ground realties of individual selected country.
- At this point we load confirmed, recovered and death related time series data first. Based on location information we indexed them and formulated active cases for trend analysis.
- Next as a data scientist, we visualized these data trends in a more suitable and efficient manner with the help of interactive tools like heatmap and Altair library. In this stage key challenge was that the data in x-axis was not aligned. So, for some countries trend analysis could not be directly carried out on raw data. This make us to use alignment strategy for data, we performed this using a functional approach by considering a thresholding criterion and drop all data un-aligned data. Next we shifted all data to thresholding index and performed Global Trend analysis, which provides good understanding.
- Next we consider the approach to select top infected/affected countries in a generic way of sorting. Based on that we pick top 5 countries.
- Next approach is to perform trend analysis to these selected countries using interactive tool. Where we understand the role of confirmed cases in trend analysis. At this stage we started deep learning-based **forecasting**.
- In detailed we performed forecasting for confirmed cases and future mortality cases. We have provided these plots in report also. To give you brief overview regarding deep learning approach, we have used here LSTM based RNN model, which is good in two aspects, firstly it caters the gradient vanishing and exploding, hence provide proper backpropagation/smooth learning and secondly, it can memorize the trend based on its gated activation value.
- One key aspect to cover here is that, why we did not combined data for learning more accurate results; the implication mainly is that, our feature here are daily confirm cased, or daily deaths which are not parametric scaled, means the scale of these features vary and hence the model find it difficult to generalize. The other point which is worth mentioning is the use of scaling. In deep learning, one common thing is the normalization prior learning, this is somewhat fundamental and provide quick learning by limiting weights magnitude and provide generalization. In fixed datasets (images etc.) we can easily perform this normalization, in time series data where we could not expect future/maximum value and our scale will vary for every new data value. Thanks for Sklearn-MinMaxScalar model which provide a generic approach to fix data normalization in a parametric way. We employed this strategy for both training and test set. Next thing

is to formulate data in a proper windowed manner so that training and evaluation can be formulated in batched manner.

- Beside LSTM we have used as a loss function MSE (Mean Square Error), minimizing this leads to good approximation of regress feature. As a optimizer we have used Adam, which is state of the art. Our Model implementation is quite generalized, extending it more complex in deep encoder decoder or sequence/attention based was not under the scope of this study.
- At this point I would like to provide basic overview of results which we deduced from forecasting. Firstly, for the case of confirmed cases, our predictions are far away from ground truth, the reason is the scalar method provides you max (1) for the data which rapidly increase exponentially in time. This results in false learning; however, the same approach is effective for second case when we are forecasting deaths.
- Like every machine learning model, there lies some assumption. The task of forecasting exponentially growing predictor gives you the closely associated label if data is normalized properly. However, in death prediction the model predicts very good approximation of Pr[y/x] conditional prediction which in our case is the median y-axis value (regression result) .
- In the last, next unhandled thing was to study the political and ground realities in terms of life standard, medical policies and corresponding trend analysis. For this we tried to fuse a separate dataset into our analyzed dataset. As a limitation we don’t have any generic dataset which is relatable and provide key features for selected countries. The inner and outer joining methods used in dataset creates more implications and hence leads to more challenges dramatically in data sciences.
- In the last we would like to thank you for proving this study as a project.

Declaration of Authorship

I affirm that I have produced the work independently, that I have not used any aids other than those specified and that I have clearly marked all literal or analogous reproductions as such.

Location, Date

Saarbrucken 17-July-20

Hafeez Ullah, Hafeez

Name, first name

Danish Shahzad, Danish

Name, first name