Literature Review

Data Analytics and Visualization (COMP09014)

Ronan Mc Cormack S00144576

Institute of Technology Sligo

In December 2019, Coronavirus 2 (SARS-CoV-2) was first discovered in Wuhan, China. This novel coronavirus soon spread to every country worldwide and has caused millions of deaths since (Yadav, 2020). With this tragedy, comes a collection of real time data related to Covid-19 cases, deaths, vaccinations, and other variables from each country worldwide. This collection of data was a contribution of many organizations and institutes.

Through-out 2020 and 2021, there have been many publish articles analysing the data received from many different organizations and institutions across the World. Many different industries have contributed to this research from Biology to Technology and Engineering. The collective effort to analyse and investigate the data has given way to break throughs in vaccine development and techniques to combat the virus.

This literature review examines the trends in real time Covid-19 data. Additionally, it examines how data can be used to forecast the Covid-19 outbreak in different countries. The topics I will cover in this literature review are Covid-19, Data Analysis, Data Visualization, Challenges, and my project proposal and objectives. All topics were formed from the literature cited in this literature review. The resulting dashboard gives an insight into the basic metrics of this virus.

1. Covid-19

In 2016, the International Heath Regulations Health Committee released a report stating the importance of sharing data related to public health emergencies (Dye, et al., 2016). When the Covid-19 pandemic hit in 2020, from across the Globe, data related to Covid-19 poured in, from the US Centers for Disease Control and Prevention (CDC), The European Center for Disease Prevention and Control and the Chinese Center for Disease Control and Prevention (China CDC), to name just a few.

With this enablement of real-time data collection at scale, the next steps for WHO were to analyse and detect new cases of the Covid-19 variant and mitigate the impact Covid-19 patients were having on the World's healthcare system (Ting, et al., 2020). The effects of Covid-19 are not only felt within the healthcare system. The wider effects on the World economy are spreading at an astounding speed with Governments absorbing the high costs of large public outbreaks (Moorthy, et al., 2020).

Using digital technologies such as big-data analysis, artificial intelligence (AI) and machine learning, we can analyse and understand healthcare trends related to Covid-19 on a large scale, using real-time data collection. This is evident on dashboards such as "Worldometer"[1] and John Hopkins University[2].

2. Data Analysis

As evident in the dashboards, the analysis of Covid-19 has been broken down into various phases. The most basic level being the time series analysis of the confirmed cases of Covid-19.

[1] https://www.worldometers.info/coronavirus/

[2] https://coronavirus.jhu.edu/map.html

Prior to any deaths related to the virus, this basic information was the basis to any further reporting and analysis. The simplistic approach to analysis these basis features has been the topic of investigation for trends and patterns (AL-Rousan & AL-Najjar, 2020) (Gupta & Pal, 2020).

When comparing the state of two countries, the confirmed Covid-19 cases data is used as the primary variable with other variables such as deaths and vaccinated populated being the second variables. This basic information is also used in forecasting methods to predict trends are possible scenarios related to the virus (Gupta & Pal, 2020).

Due to the large collection of data, there is the need to normalize the data into common factors and variables. As noted previously, there are many different organizations from around the World all contributing their own data which is then compiled by different institutions for further investigation. Two common situations occur when performing data analysis, as identified by (Qin & Chiang, 2019):

1. Most of the data is normal but contains many outliers.
2. Using the data to uncover trends or variables to decide of the operation on the data is normal or not.

2.1. Data Analysis Techniques

(Tiwari, 2017) identifies three techniques used to mine big data. These three techniques are: Clustering, Classification and Association rule mining. In relation to the Covid-19 data, the application of all three techniques is relevant.

Clustering allows features e.g., confirmed cases, to be grouped together to indicate their resemblance. As the data is gathered and reported from a central source, it avoids clustering issues such as noisy data.

Classification allows for target categories to be the variable of interest. For any variable e.g., confirmed deaths, we can easily classify the range of the variable into categories such as high, medium, or low, depending on the parameters specified.

Association rule mining can give insight into trends in Covid-19 data, such as high confirmed cases in specific regions within countries. When trying to identify high infection areas and areas of concern, governments can turn to this technique to help mitigate public heath exposure.

3. Data Visualization

As discussed earlier, there are several resources available to visualize the impact of Covid-19. Using visual data, we can increase our observation and understanding of the pandemic, while trying to pinpoint treads related to the extreme outbreak (Dey, et al., 2020). Across all industries, the emergence and use of visual dashboard has gained popularity.

With freely available data in relation to Covid-19, we can capture the timing of the first reported cases of Covid-19 and how it has spread globally (Dong, et al., 2020). (Gupta & Pal, 2020) presents, in Figure 1, the trend of cumulative infected cases (Natural Log) in India for a 60-day period following the 22$^{nd\ of}$ January 2020. These graphs provide simple, easy to visualize, analysis of the trend in India for that 60-day period. For public health and government officials, visualizations such as the one presented by (Gupta & Pal, 2020) , allow them to make key decisions on how to be proactive against the Covid-19 pandemic.

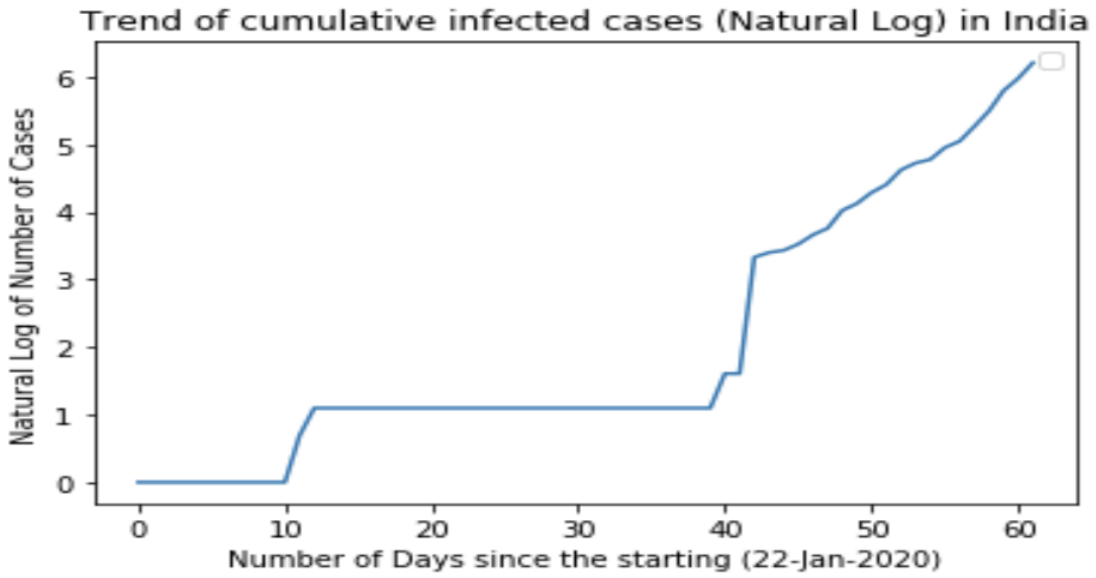

Figure 1: Trend of cumulative infected cases (Natural Log) in India for a 60-day period following January 22nd, 2020.

While the most common method of analysis is the time series analysis of confirmed cases, other researchers dig deeper into the information to try and identify trend bases on variables such as gender and region within a country (AL-Rousan & AL-Najjar, 2020).

In Figure 2, (AL-Rousan & AL-Najjar, 2020) plot the distribution of recovered. Unrecovered, deceased, and un-deceased cases based on gender. The X^2 test (observed value minus the expected value, squared, divided by the expected value) performed by (AL-Rousan & AL-Najjar, 2020) confirms that the gender variable is statistically significant in the number of both the recovered and deceased cases.

The X^2 formula:

$$\sum X_{i-j}^2 = \frac{(O-E)^2}{E}$$

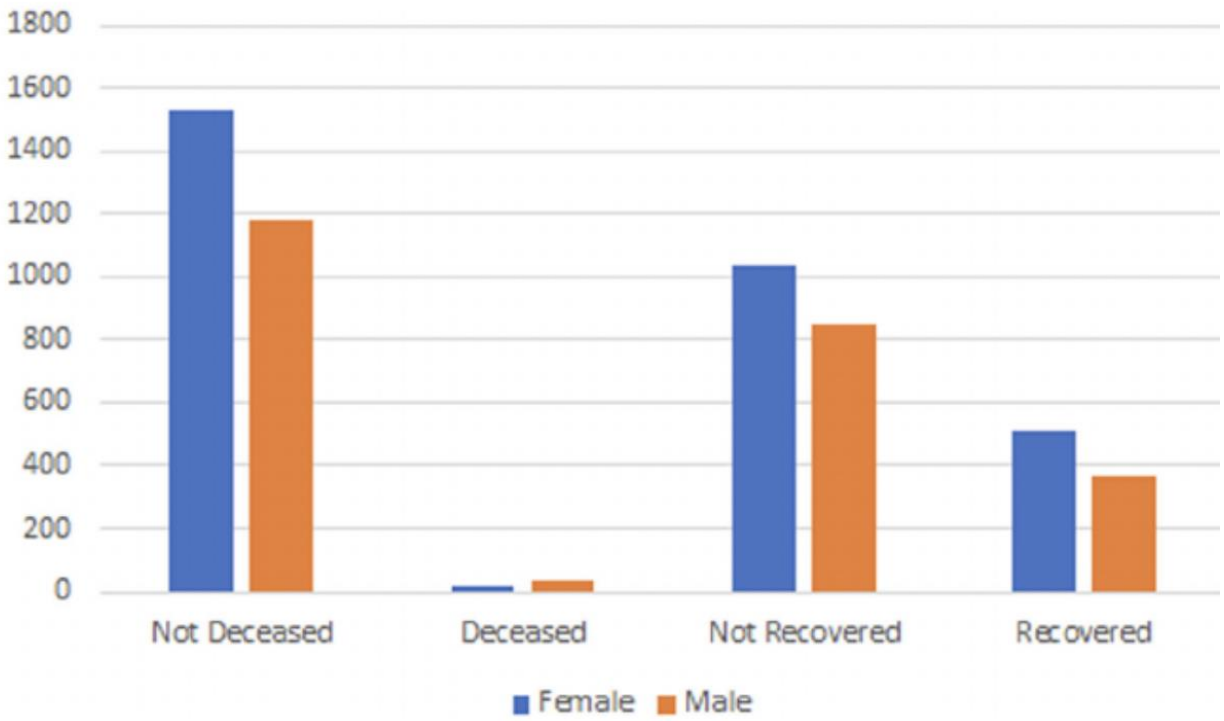

Figure 2: Distribution of recovered, unrecovered, deceased, and un-deceased cases based on gender in South Korea.

4. Challenges

Due to the sheer amount of data available, we often run into challenges. The "truth" of the data has been called into question many times throughout the pandemic which seeks to discredit the life-saving research and analysis done by world class researchers. Digital technology can enhance public health education but can also be cause of mis-leading information.

In Singapore, the social-media platform, WhatsApp (owned by Facebook), has partnered with the government to allow users to access accurate information and clarify uncertainties the public may have regarding this information (Ting, et al., 2020). (Garrett, 2020) calls for actions to counteract false information, such as funding of their messages on an unprecedented scale with genuine urgency.

(Lowe & Matthee, 2020) describe other challenges such as scalability and readability. Other researchers also expand on this by describing the difficulty trying to fit visualizations onto a limited number of pixels (Molina-Solana, et al., 2017).

(Li, et al., 2016) introduce the topic of 'over plotting', which can be applied to certain dashboard that visualize Covid-19 data. This can be ineffective at plotting the trends and patterns in the data as the dataset is not summarized efficiently. (Li, et al., 2016) also re-enforce this point by discussing the changing technology, with mobile device screens becoming lower in pixel count which means it is important to scale the data to fit the lower pixel count.

5. Project Proposal and Objective

The objective of my project is to address the question, could new digital technology be used for Covid-19? Using data analysis on real-time data collected from around the World, I will present

a dashboard visualizing the fundamental variables of Covid-19. As discussed earlier, the reliance of public health services on accurate real-time information can be pivotal to plan and mitigate the impart to healthcare indirectly related to Covid-19.

I propose for a simpler presentation of the data as larger dashboards can have an overwhelming number of visuals. The focus of not 'over-plotting' is one of the objectives.

The aim with the dashboard is for the ease of applications if areas such as healthcare where quick information is critical is making decisions on how to deal with escalating events linked to Covid-19. There have been numerous researchers who agree that the data should be easy to retrieve and accurate (Li, et al., 2016) (Molina-Solana, et al., 2017).

Bibliography

AL-Rousan, N. & AL-Najjar, H., 2020. Data analysis of coronavirus COVID-19 epidemic in South Korea based on recovered and death cases. *Journal of Medical Virology,* 92(9), pp. 1603-1608.

Dey, S. K., Rahman, M. M., Siddiqi, U. R. & Howlader, A., 2020. Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. *Journal of Medical Virology,* 92(6), pp. 632-638.

Dong, E., Du, H. & Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases,* 20(5), pp. 533-534.

Dye, C., Bartolomeos, K., Moorthy, V. & Kieny, M. P., 2016. Data sharing in public health emergencies: A call to researchers. *Bulletin of the World Health Organization,* 94(3), p. 158.

Garrett, L., 2020. COVID-19: the medium is the message. *The Lancet,* 395(10228), pp. 942-943.

Gupta, R. & Pal, S. K., 2020. Trend Analysis and Forecasting of COVID-19 outbreak in India. *medRxiv.*

Gupta, R. & Pal, S. K., 2020. Trend Analysis and Forecasting of COVID-19 outbreak in India. *MedRxiv.*

Li, X., Kuroda, A. & Matsuzaki, H., 2016. an interactive visualization platform optimized for visual analysis of big data. pp. 109-111.

Lowe, J. & Matthee, M., 2020. Requirements of Data Visualisation Tools to Analyse Big Data: A Structured Literature Review. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* pp. 469-480.

Molina-Solana, M., Birch, D. & Guo, Y. k., 2017. Improving data exploration in graphs with fuzzy logic and large-scale visualisation. *Applied Soft Computing Journal,* Volume 53, pp. 227-235.

Moorthy, V., Restrepo, A. M. H., Preziosi, M. P. & Swaminathan, S., 2020. Data sharing for novel coronavirus (COVID-19). *Bulletin of the World Health Organization,* 98(3), p. 150.

Qin, S. J. & Chiang, L. H., 2019. Advances and opportunities in machine learning for process data analytics. *Computers and Chemical Engineering,* Volume 126, pp. 465-473.

Ting, D. S. W., Carin, L., Dzau, V. & Wong, T. Y., 2020. Digital technology and COVID-19. *Nature Medicine,* 26(4), pp. 456-461.

Tiwari, P., 2017. Accident Analysis by using Data Mining Techniques Quantum Inspired Machine Learning and Information Retrieval View project Implementation of Machine Learning and Deep Learning View project.

Yadav, R. S., 2020. Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. *International Journal of Information Technology (Singapore),* 12(4), pp. 1321-1330.