The applications of Data Science in the context of COVID-19 outbreak

Chandupraveen Gudi, Reg. 100329132

Abstract

Epidemics or pandemics such as COVID-19 are unexpected and highly hazardous. To control these pandemics, governments worldwide have to take some preventive measures and raise awareness by taking optimal decisions, that can be achieved if reliable data is available. It is evident that, we are now living in the data age and data is available abundantly all over the internet in various platforms. By applying the techniques available in Data Science, valuable information can be extracted that can help the organizations to monitor and control the wide spread of pandemics. This essay focusses on various dimensions and applications of Data Science used in fighting the current COVID-19 pandemic, which includes the types and characteristics of data, various tasks and algorithms applied, success stories and the lessons learnt.

1 Introduction

Due to the rapid growth in the transportation infrastructure, population mobility across the globe has seen a significant rise which could lead to a pandemic in a very short span in case of any disease outbreak (Jia et al., 2020). Presently, the world is experiencing a pandemic with the newly discovered novel corona virus COVID-19. First case was discovered in China in December 2019. As per the latest epidemiological update by WHO as of 31st January 2021, 102 million corona virus positive cases and close to 2.2 million deaths were reported from 222 countries and territories (WHO, 2021). To combat with COVID-19 pandemic, both government and private organizations are developing and promoting global actions and strategies together. Applying Data Science techniques in health sector makes it possible to be more reliable in controlling, monitoring and studying patients, diseases. A recent study in intensive care units conducted by the Massachusetts Institute of Technology found that Data science technologies can accurately forecast vital details such as the hospital stay duration, total patients in need of surgery, and the risk of infecting and spreading the disease in patients (Corsi et al., 2020).

2 Literature Review

Corsi et al. (2020) aims to do a literature review on the usage of Data Analytics in health sector during epidemics and pandemics with a wide range of articles from various journals. The paper focused on the types of data used for analysis and how data analytics assist public and private organizations to fight pandemics. Arora et al. (2020) describes how South Korea was able to reduce the impact of COVID-19 by employing Data Science techniques using exploratory data analysis.

3 Data used and its Characteristics

Internet and social media searches have become important data sources. These platforms collect data when the users verify their health condition and concerns about the risk of a pandemic. Even though the data from official sources is accurate and comply to ethical and privacy concerns, it is a very cumbersome, costly and bureaucratic process which will prevent in making decisions at early stages of the outbreak. Data from search engines, on the other hand, helps policymakers make decisions and adapt quickly to potential epidemics, allowing them to monitor epidemics faster and more accurately. Social networking can be used as a method to study behavioral changes during outbreak of any pandemic. Social media like Facebook, Twitter, Blogs and websites also collect data which is a thought-provoking way to get public opinion in realtime. Twitter is the most widely used media for data analysis during COVID-19 pandemic with fewer errors and more stability, Twitter has an advantage of providing minute-by-minute analysis over search engine logs, and is used effectively to increase public awareness and thereby reduce the prevalence of COVID-19 (Corsi et al., 2020).

GPS data generated by mobile phones can be used to easily track the direct contact of a diagnosed patient without jeopardizing personal privacy, there by alerting the contacts of the infected person to isolate. IoT refers to connection of electronic gadgets to the internet. During COVID-19, IoT played a very crucial role in collecting real-time data. Data collected from unmanned supermarkets, online ticket booking, hotel itinerary and other public security systems were used to assess the personalized behavior and customer's habits during the pandemic. Genome sequencing is the process of analyzing a DNA sample to help determine the origin of COVID-19, vaccine and drug development. Could computing platforms like nextstrain.org helps researchers across the globe to share and analyze gnome sequences in real-time, analyzing this data helps in finding the various mutants of COVID19 and evolutionary relationships of COVID19 due to other viruses (Jia et al., 2020).

Data characteristics were classified as unstructured, structured, and semi-structured. Unstructured data can be text without any specific format, images, audio recordings, and videos. Complex processing systems are required to process Unstructured data. Processing of unstructured data have gained momentum due to the advent of various Big data technologies. Structured data is the data with specific rules generally in the form of tables, this can be data in any relational database. Some of the properties of structured data include consistency, non-redundancy, data integrity and data security. Semi-structured data is stored along with labels or identifiers to denote the data type. It is generally XML, JSON, HTML..., it supports different data types, such as text, image and video (Corsi et al., 2020).

4 Tasks used and Algorithms applied

Real-time visualization played a key role in controlling COVID-19, Google presented Geographical Information System using interactive maps for manifesting pandemic distribution, hotspot reports which gives a clear indication of the spread of COVID-19. Rumor-mining, opinion-mining, sentiment analysis and public appeasement were extracted using Natural Language Processing with Deep Learning by classifying sensitive information from reports, news, documents and all sorts of social media. Key methods used for such analysis include window-based neural networks, Word2Vec's vector representation, recurrent neural networks, convolutional neural networks. Genome sequence of COVID-19 can be expressed as a thermal matrix to find root cause using image classification algorithm and convolutional neural networks, which revealed the origin of virus is from Bats. To estimate the level of COVID-19 transmission due to direct contacts, Markov chain Monte Carlo method found to be very useful with a predicted scale to be 95% confident. Deep Learning methods like Deep-Auto-Encoders(DAE) were used in China to predict the spread of COVID-19, which gave a reference to the policy makers to take decisions on economic recovery. Data science community has helped the agencies to fight the pandemic with various machine learning algorithms like susceptible–infectious–susceptible, susceptible–infectious–recovered, complex host–vector propagation for analyzing social media data (Jia et al., 2020).

Mathematical modeling techniques are quite often used for forecasting and monitoring pandemics like COVID-19. ARGO(AutoRegression Google) is a regression tool by Goggle used to predict any outbreak of diseases around the globe. ARGO have the ability to capture the changes in the research behavior of people and real-time tracking. Another algorithm for predicting the disease outbreaks is Support Vector Machine Regression algorithm(SVR). Machine Learning algorithms along with Artificial Neural Networks can predict the possible occurrence of deaths due to COVID-19 (Corsi et al., 2020). A survey shows that Decision Tree algorithms performs better than Naive-Bayes, logistic regression, linear regression, support vector machine, KNN, and random forest models with an accuracy of 99.85% to predict the recovery rate of COVID-19 patients (Muhammad et al., 2020).

5 Success Stories

Using social media platforms like wechat and alipay, to stop the spread of COVID-19 China has implemented an innovative method using Quick Response code in healthcare which tracks the travel history and health status of the people. People must scan the QR code while accessing any facility or purchasing or travelling, which in turn allow the person if there is no threat else barred from the entry. With this method China was able to stop the spread of COVID-19 much earlier than rest of the world (Jia et al., 2020). South Korea has set an ample to the rest of the world by controlling the spread of COVID-19 by using best medical practices and implementing state of the art technologies like Data Mining and AI. South Korea's majority success lies in collecting the data, the government was able to collect the personal data like CCTV, mobile device data, creditcard data. By analyzing the data, government alerted people with emergency messages incase if they enter infected hotspots (Arora et al., 2020).

6 Problems

Extracting data from social media is difficult as these sources possess heterogeneous characteristics, un-structuredness and dynamic behavior. Besides, such data can have other problems like large quantities of data, lack of accuracy, lack of continuity in the data collection, lack of specific location of consumers, heterogeneity of information, and sampling bias (Corsi et al., 2020). Data Reliability has always been a disputed issue due to the high availability of information, data collected for COVID-19 is no exception. Poor quality of data will mislead the machine learning and deep learning algorithms to predict the wrong results. Organizations had to invest more in preprocessing the data to extract the hidden information by eliminating noisy data. Since the mathematical models had to rely on the data available from various sources which may be trusted and untrusted sometimes, there is a high chance of possibility to have outliers in the data even after many pre-processing steps are applied. Due to this there will be a possibility of False positives and False negatives (Jia et al., 2020).

7 Lessons Learned

HealthMap, which includes multiple data sources to monitor the spread of epidemics around the world, has limited with English language configurations, loosing initial notifications of diseases reported in other languages leading the epidemic being detected at a later date, this is a major lesson learned for any future disease outbreaks (Corsi et al., 2020). Predicting the outbreaks of diseases using the historical data is assumed to be a reliable means of data, but the results show that, with the use of data from social media platforms like twitter prediction models perform better by reducing the error of 17-30%. Also, forecasts with the data from twitter are almost four weeks advanced when compared with the historical data predictions. Collaboration is the one among the major lessons learnt during COVID-19 pandemic. During the initial phases of COVID-19, all the countries didn't work collaboratively by sharing the medical data. Countries globally reacted to share the data available with them after COVID-19 disease has become a Pandemic and a global crisis (Jia et al., 2020).

8 Conclusion

Necessary steps should be taken to avoid misinterpretation of data due to its high availability which will eventually lead to taking erroneous decision making, it is always advisable to use the latest Data Science techniques along with the traditional methods to get the most out of data.

References

Arora, A. S., Rajput, H., and Changotra, R. (2020). Current perspective of covid-19 spread across south korea: exploratory data analysis and containment of the pandemic. *Environment, development and sustainability*, pages 1–11. 32837283[pmid].

Corsi, A., de Souza, F. F., Pagani, R. N., and Kovaleski, J. L. (2020). Big data analytics as a tool for fighting pandemics: a systematic review of literature. *Journal of Ambient Intelligence and Humanized Computing*.

Jia, Q., Guo, Y., Wang, G., and Barnes, S. J. (2020). Big data analytics in the fight against major public health incidents (including covid-19): A conceptual framework. *International journal of environmental research and public health*, 17(17):6161. 32854265[pmid].

Muhammad, L. J., Islam, M. M., Usman, S. S., and Ayon, S. I. (2020). Predictive data mining models for novel coronavirus (covid-19) infected patients' recovery. *SN computer science*, 1(4):206–206. 33063049[pmid].

WHO (2021). Weekly epidemiological update - 2 february 2021.