2021 2nd International Conference on Big Data Economy and Information Management (BDEIM) | 978-1-6654-8288-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BDEIM55082.2021.00069

Applications of Big Data in Fighting the COVID-19 and Future Employment

Kejun Li
School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK
gaoming@cas-harbour.org

***Abstract*-Since the coronavirus (SARS-CoV-2) outbreak in 2019, it has infected millions of people and claimed the lives of tens of thousands of people. During the coronavirus pandemic, big data and its applications have become one of the few powerful means to fight the virus. Nowadays, many countries and research institutions are using big data and its applications to track and control the spread of the contagious disease. In the future, people can use big data to fight against such epidemics. For example, comparative genomic research on virus variants, accelerated by big data analysis, can yield important information about virus mutations and evolutionary selection. This article mainly discusses the application of big data during the COVID-19 pandemic and how to fight similar epidemics in the future. Software and applications have been developed based on big data to track and predict infections. IoT-based solutions have been deployed in preliminary diagnosis. Therefore, in similar epidemics in the future, big data can accelerate tracking, prediction, diagnosis, and treatment, which helps the government and experts make more informed decisions to fight the virus and reduce its social impact.**

Keywords-Big data, COVID-19, Epidemics, Social impact, Application, Data model

I. INTRODUCTION

The COVID-19 pandemic has become an unprecedented global public health emergency. This pandemic poses a major challenge to public health, epidemiological planning, and health care systems.

This public health incident that swept the world is not the first in human history; similar incidents have occurred several times in the past hundreds of years. The Black Death caused by the bubonic plague from 1346 to 1353 caused about 200 million deaths. The cholera pandemic from 1852 to 1860 caused about 1 million deaths. The H2N2 flu stroke between the end of the 19th century to the mid-20th century claimed the lives of 1 million people. Viruses such as SARS-Cov (2002-2003) and H1N1 influenza (2009) have spread and affected humans in various ways[1]. This shows that new diseases caused by RNA viruses will continue to appear and affect humans in the future. Nevertheless, the global pandemic caused by the COVID-19 warns people that the world is not prepared for such incidents.

According to the World Health Organization, as of 5:51 pm CEST, 30 August 2021, 216,303,376 cases of COVID-19 have been confirmed globally, and the death toll has reached 4,498,451. Among these, the USA has the highest number of confirmed cases, and India comes next. Although more than 4 billion vaccine doses have been administered, the situation has not been significantly suppressed due to the new variant Delta.

At present, the global response to COVID-19 is mainly focused on: minimizing the rate of serious morbidity and mortality; and minimizing the damage to society[2]. Therefore, many research institutions, government entities, and industries have gathered together and are using advanced technologies such as big data analysis and artificial intelligence to fight this pandemic. For example, Shanghai and the United States are using the advantages of big data analysis to combat the risk of coronavirus. Key information such as the physical condition and travel history of more than 100,000 people was recorded to manage the pandemic[3]. This article aims to discuss the application of big data technology in the COVID-19 pandemic and explore how to use these big data tools to enable mankind to better respond to such outbreaks in the future.

II. BIG DATA APPLICATIONS IN FIGHTING THE COVID-19

To date, big data applications have proved to be effective in many aspects of fighting the COVID-19 pandemic, especially in tracking or diagnosis.

Some researchers and government agencies are using big data to track COVID-19 cases in real-time. The epidemic analysis combines all data, including positive cases, deaths, people who have recovered from the disease, monitoring contacts of positive instances, population movements, travel history, population density, etc. With the use of artificial intelligence and machine learning, the data can be processed to build a disease model, which can predict the infection rate (high or low) and its impact. AarogyaSetu, a mobile tracking application launched by the Indian government, is a great example of how big data can be used to combat the disease. The workflow of AarogyaSetu is shown in the figure below. [5]

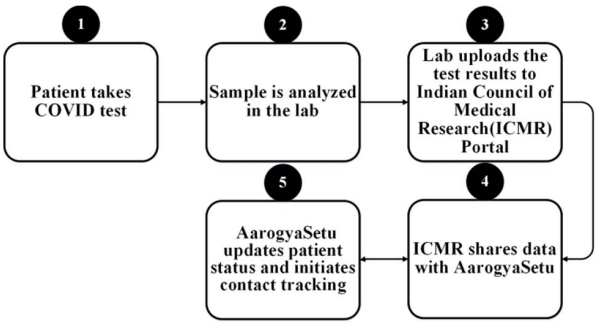

Figure 1. The workflow of the AarogyaSetu app. Adapted from the official website[5]

DOI 10.1109/BDEIM55082.2021.00069

By picking up the location data through GPS and Bluetooth and then comparing it with the existing database, the app can identify people who might have been in contact with confirmed corona cases, hence providing early warning in time and tracking the chain of infection[6].

The application of big data to the genome has played an important role in responding to the coronavirus outbreak. Some researchers have sequenced the coronavirus genome, which helps track the spread of the disease. On January 11, 2020, the open-access virology website (http://virological.org/)[7] released the first SARS-CoV-2 genome. Subsequently, the Chinese Center for Disease Control and Prevention quickly released the other four virus genomes to the Global Influenza Data Sharing Initiative (GISAID; https://www.gisaid.org/). The genomic data set submitted in GISAID can also be used for COVID -19 haplotype network analysis, essential for secondary outbreak prevention.

In addition, big data is also applied to the diagnosis and prediction of the COVID-19. Since the radiological manifestations of COVID-19 infection and viral pneumonia are very similar, distinguishing them in the clinical setting becomes very important. Therefore, the chest CT of patients infected with COVID-19 and patients with viral pneumonia is used as the input data of the deep learning system [8]. The final accuracy rate in distinguishing the two was 86.0%, and the accuracy rate in distinguishing COVID-19 patients from healthy individuals was 94.0%. In addition, machine learning was used in developing a prognostic prediction algorithm. The algorithm is used to predict the mortality risk of people infected with COIVD-19 [9]. This helps identify patients with potential needs for further medical care, allowing medical staff to allocate medical resources rationally.

III. FUTURE APPLICATION OF BIG DATA IN THE EPIDEMIC FIGHT

The fight against COVID-19 is not yet over with the emergence of new virus variants and new cases reported across the world. Researchers have developed tools and methods from past experience, which can be useful in the fight against COVID-19 and other similar diseases in the future.

This section will describe four main fields of application, including tracking, prediction, diagnosis, and treatment. Figure 1 below shows how to use big data to fight against similar diseases.

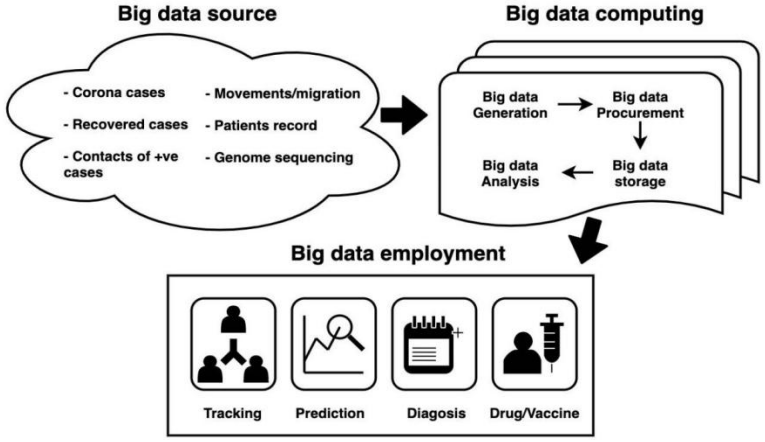

Figure 2. The Epidemic fight using big data: sources, computing process and employments

A. Tracing and Prediction

One of the features of the COVID-19 is that the patients can transmit the virus before they show any symptoms. With the COVID-19 variants becoming more contagious, it is critical for governments to track and contain corona cases and make epidemic management plans in time.

First of all, social big data have significant uses in tracking. Jahanbin et al adopted a fuzzy model called EClass1-multiple-input-multiple-output(EClass1-MIMO) to monitor and track the news about the spread of contagious diseases. The model has four successive phases, namely Filtering, Crawling, Deploying fuzzy rules and Visualization[10].

The framework of this algorithm is shown in the figure below.

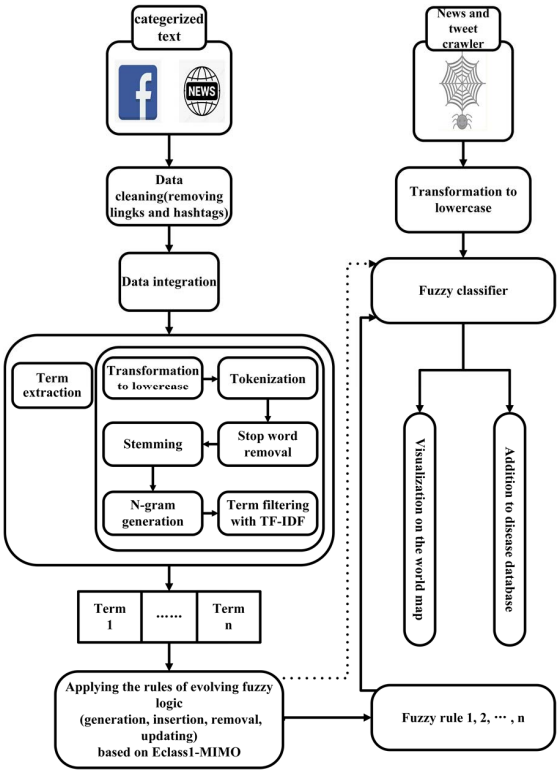

Figure 3. The framework of the fuzzy model. Adapted from Jahanbin et al. [10]

The fuzzy logic rule is deployed for web mining and crawling, managing, storing, and visualizing the latest and trending tweets. These tweets were later classified by the Fuzzy Algorithm for Extraction, Monitoring, and Classification(FAMEC) model and compared to the real dataset. The final findings were highly similar to the incidence cases described by WHO [11].

Moreover, big data can help identify key spreaders, which is crucial in government departments developing action plans to combat the pandemic.

Researchers have proposed a CovidKeySpreader method to identify key spreaders in Indian states based on patient interaction networks. Each network is regarded as an undirected graph G(V,E), where nodes represent individual patients and edges represent interactions between patients. Each node in the network is classified according to its characteristics and then assigned to the relevant community. Meanwhile, nodes and communities are scored. Finally, the random walk algorithm is used to realize the iterative ordering of nodes. The node with the highest ranking is taken as the key spreader in its associated network [12].

The performance is evaluated using the SIR model, where the nodes are each assigned with any of the three statuses: Susceptible(S), Infected(I) and Recovered(R). The model will simulate the spread of virus and generate a reference ranked list. The difference between the ranked list generated by the proposed algorithm and the reference list is quantified by the Kendall correlation coefficient τ, which can be calculated using the equation below:

$$\tau(R, R^E) = \frac{n_c - n_d}{n(n-1)/2} \quad (1)$$

Where n_c and n_d are the concordant pairs and discordant pairs respectively. R represents the list of nodes generated by the proposed algorithm, and R^E is the reference list.

Experimental results show that the proposed method performs better in identifying key spreaders than basic centrality measures. Both the CovidKeySpreader method and the SIR model can be adopted in future tracking applications.

In addition, the application of big data to predictive analysis can make epidemic control more effective. Quick and reliable forecasts help governments take preventive measures.

Many data models can act as valuable tools in predictions. In the Time-dependent dynamic model(TDDM) developed by Tang et al., the rate of contact is regarded as a time-dependent variable, as defined in the following equation.

$$c(t) = (c_0 - c_e)e^{-r_1 t} + c_b \quad (2)$$

where C_0 represents the initial contact rate, C_b is the minimum contact rate considering control measures, and r_1 is the exponential reduction in the rate of contact. [13]

Deep learning or machine learning can also be used in predictive analytics for prediction and recognition. For example, the Random Forest algorithm is capable of data classification and regression, and bagging is applied to prevent data inaccuracies. Another algorithm is the K-means algorithm

based on Spark. It is used to cluster the data according to the common characteristics of the data. As the number of infections increases in a pandemic, the K-means algorithm can divide infection data into different groups to propose new cases of infection to predict transmission and mortality. In the COVID-19 pandemic, there are research institutions using the FBProphet algorithm through training models to predict the number of cases. FBProphet algorithm can be very effective in predicting time series data. Kaggle provides data sets for predicting COVID-19, with attributes such as recovered cases and deaths [11].

B. Diagnosis and treatment

In addition to tracking and predicting the spread of the virus, big data can also be applied to diagnosis and treatment. Reliable, fast and convenient diagnoses are essential to fighting the virus. In the pandemic, a fast and accurate diagnosis of COVID-19 can not only save a large number of lives, but also limit its spread; the medical data collected can act as the training dataset for machine learning.

First of all, the application of big data to virus detection can shorten the detection time while ensuring accuracy. The traditional virus detection for respiratory diseases is through reverse transcription-polymerase chain reaction (RT-PCR) [14]. However, this detection method is time-consuming and expensive. Additionally, it requires specific instruments and equipment. As a result, many countries are unable to conduct such tests in time due to technological limitations. Based on the previous experience of responding to the COVID-19 pandemic, there needs to be a method that can achieve rapid detection of the virus. Smart devices based on big data have proven to be a simpler and faster measure. Imran et al. proposed a mobile application named AI4COVID-19, which can be used for diagnosis based on coughs. This application has been upgraded to minimize the misdiagnosis rate, and its accuracy in distinguishing COVID-19 from other types of cough has exceeded 90% [15].

Besides, as the coronavirus affects various organs and organ systems in the human body, patients require personalized treatments according to their conditions. Therefore, a cloud service centre can be established, which is connected to smart devices or biosensor devices[16]. Sensors have the characteristics of simple operation and diverse functions, which is helpful for diagnostic research, such as immunoassay. They can check patients' health conditions and transmit data to the cloud service centre. The cloud server enables doctors to monitor patient conditions in real-time, making preliminary online diagnosis possible. During COVID-19, because of the low doctor-patient ratio, it is difficult to monitor every patient. But the IoT devices helped doctors provide personalized treatment for patients without being exposed to viruses [17].

Big data has been widely used in drug research. Vaccines and specific drugs developed based on big data can play a decisive role in responding to pandemics similar to COVID-19. Although Google's DeepMind is a company known for its AlphaGo game algorithm, it is now able to use artificial intelligence to predict the structure of SARS-CoV-2 membrane proteins that may be useful for the development of new drugs. The solution based on molecular docking has also been used in drug research. In addition, it launched a big data-driven drug repositioning program, using its machine learning function to combine knowledge maps with literature to develop targeted vaccines.

IV. Conclusion

This article mainly describes the control and prevention measures people have used so far that involves big data and its corresponding tools in the global crisis brought about by the COVID-19 pandemic. It also explores, in the future, when a similar pandemic comes again, how people can use big data to fight the crisis. Firstly, big data can be used to develop applications for tracking and epidemic predictions. Then, a faster and safer clinical diagnosis can be made through artificial intelligence based on big data. Big data analysis is also deployed in vaccine development and drug repositioning.

Although big data is already showing the potential to be used to fight COVID-19 and similar pandemics, the processing methods for big data are still at a relatively early stage. Therefore, despite its obvious advantages, many areas still need to be resolved and upgraded. The recently proposed big data platform based on artificial intelligence algorithms lack a standard data set, which will lead to deviations in test results [18]. Consequently, in this case, it is difficult to judge which algorithm is more suitable for virus detection. Sometimes, researchers do not disclose their unique data sets. Therefore, a standard data set is one of the main challenges faced by big data platforms. The other issue is the privacy and security of people's personal data. During the COVID-19 pandemic, the government requires people to share personal information, such as location, travel history, diagnosis reports, daily activities, etc. However, people are often unwilling to share their personal information because they are worried about personal privacy leakage. To deal with this concern, some existing technologies (for example, blockchain, incentive mechanism) can be applied to solve the problem.

Moreover, big data can also contain misinformation. For example, during the COVID-19 pandemic, much misinformation appeared, which undoubtedly had a negative impact on the fight against the pandemic. Therefore, the World Health Organization has adopted the "Epidemic Information Network" (EPI-WIN) to disseminate data to important partners. Other social platforms have also collected and analyzed data about the coronavirus by scanning keywords in order to curb the spread of misinformation [19].

In terms of tracking and prediction, there are also problems that need to be resolved. Machine learning requires a large amount of data, without which an AI model cannot be constructed [20]. This places high demands on the amount of virus-related data. In addition, there are also problems with big data obtained through social media. Because how to filter out accumulated noise in a large amount of social data can be challenging.

Although there are still many problems in the employment of big data in the pandemic, this does not change the fact that it is a powerful weapon for people to fight the virus. In the future, it is necessary to conduct further research in this field to solve the problems in the source and processing of big data.

REFERENCE

[1] Kaur, Simran, and Hasija, Yasha. Role of Computational Intelligence Against COVID-19. Computational Intelligence Methods in COVID-19: Surveillance, Prevention, Prediction and Diagnosis. Singapore: Springer Singapore, 2020. 19-43. Studies in Computational Intelligence. Web.

[2] Tuite, Ashleigh R, Fisman, David N, & Greer, Amy L. Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. Canadian Medical Association Journal (CMAJ), 2020, 192(19), E497-E505.

[3] Kupferschmidt, Kai. Genome analyses help track coronavirus' moves. Science (American Association for the Advancement of Science), 2020, 367(6483), 1176-1177.

[4] EPhzibah, E. P., & Sujatha, R. Big data management with machine learning inscribed by domain knowledge for health care. International Journal of Engineering & Technology, 2017, 98.

[5] AarogyaSetu, 'how big data can be used to combat the disease?'. https://www.aarogyasetu.gov.in/

[6] Verma, Sandhya, & Gazara, Rajesh Kumar. Big Data Analytics for Understanding and Fighting COVID-19. In Computational Intelligence Methods in COVID-19: Surveillance, Prevention, Prediction and Diagnosis (Studies in Computational Intelligence, pp. 333-348). Singapore: Springer Singapore. 2020.

[7] Zarocostas, John. How to fight an infodemic. The Lancet (British Edition), 2020, 395(10225), 676.

[8] Song, Ying, Zheng, Shuangjia, Li, Liang, Zhang, Xiang, Zhang, Xiaodong, Huang, Ziwang, et. al. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 1.

[9] Shi, Feng, Xia, Liming, Shan, Fei, Song, Bin, Wu, Dijia, Wei, Ying, et al. Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification. Physics in Medicine & Biology, Physics in medicine & biology, 2021-02-19.

[10] Zhang, Y., & Holmes, E. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. Cell, 2020, 181(2), 223-227.

[11] Jiang, X., Coffee, M., Bari, A., & Wang, J. Towards an artifificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials & Continua, 2020, 62(3), 537–551.

[12] Kia Jahanbin, & Vahid Rahmanian. Using twitter and web news mining to predict COVID-19 outbreak. Asian Pacific Journal of Tropical Medicine, 2020, 13(8), 378-380.

[13] Waggoner, Jesse J, Stittleburg, Victoria, Pond, Renee, Saklawi, Youssef, Sahoo, Malaya K, Babiker, Ahmed, et. al. Triplex Real-Time RT-PCR for Severe Acute Respiratory Syndrome Coronavirus 2. Emerging Infectious Diseases, 2020, 26(7), 1633-1635.

[14] Jahanbin, K., Rahmanian, F., Rahmanian, V., & Jahromi, A. S. (2019). Application of Twitter and web news mining in infectious disease surveillance systems and prospects for public health. GMS hygiene and infection control, 14. https://www.egms.de/static/pdf/journals/dgkh/2019-14/dgkh000334.pdf

[15] Ali Imran, Iryna Posokhova, Haneya N. Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, et. al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Informatics in Medicine Unlocked, 2020, 20, 100378.

[16] Yadav, Samir S, & Jadhav, Shivajirao M. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data, 2019, 6(1), 1-18.

[17] Tang, B., Bragazzi, N. L., Li, Q., Tang, S., Xiao, Y., & Wu, J. (2020). An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov). Infectious disease modelling, 5, 248-255. https://pubmed.ncbi.nlm.nih.gov/32099934/

[18] Hasan, A., & Kamal, A. Social Network Analysis for the Identification of Key Spreaders During COVID-19. In Computational Intelligence Methods in COVID-19: Surveillance, Prevention, Prediction and Diagnosis (Studies in Computational Intelligence, pp. 61-77). Singapore: Springer Singapore. 2020.

[19] Mahalle, P. N., Sable, N. P., Mahalle, N. P., & Shinde, G. R. Predictive analytics of COVID-19 using information, communication and technologies. 2020.

[20] Raza, K., & Qazi, S. Nanopore sequencing technology and internet of living things: A big hope for u-healthcare. Sensors for health monitoring (Vol. 5). London, UK: Elsevier, Academic Press. 2019.