Guest editorial

Using data science to understand the COVID-19 pandemic

Data science in pandemic

The coronavirus disease, a novel severe acute respiratory syndrome (SARS COVID-19), has become a severe global health crisis due to its unpredictable nature and lack of adequate treatment. The COVID-19 pandemic has generated a strong demand for using technologies such as data science to understand or mitigate the adverse effects of the COVID-19 on public health, society and the economy (He *et al.*, 2021).

In the current era of big data, data science and data analytics have become increasingly crucial in academia, healthcare, public relationships and business operations. Machine learning (ML) models could be effective in identifying the most critical factors responsible for the overall fatalities caused by the COVID-19. However, the functional capabilities of ML models in conducting epidemiological research, especially for the COVID-19, have not been substantially explored. There are several related research methodologies regarding the COVID-19 data analytics. For instance, adopted ML models and Random Forest (RF) have been used to perform the regression modeling and provide useful information to identify the relevant critical explanatory variables and evaluate interconnections between and among the key explanatory variables and the COVID-19 case and death counts (Gupta *et al.*, 2021). Time-series analyses have been used to examine the rate of incidences of the COVID-19 cases and deaths (Khayyat *et al.*, 2021). Social network analysis (SNA) has been used to track cases and simulations for modeling the COVID-19 outbreaks (Bahja and Safdar, 2020). Researchers have built models to interpret patterns of public sentiment on disseminating health-related information and assess the political and economic influence of the pandemic.

Data visualization

During the pandemic, data visualization plays a crucial role in helping the public understand the COVID-19 cases and trends. Numerous government healthcare departments or agencies have launched the data visualization portal on their websites and social media platforms. The United States Centers for Disease Control and Prevention (CDC) has created a COVID-19 data visualization webpage ((www.cdc.gov/coronavirus/2019-ncov/covid-data/data-visualization.htm) which is updated frequently. An essential thing for data visualization is to keep the data up to date and shows the latest update date and time. CDC provides the weekly COVID-19 hospitalization rates since the beginning of the pandemic. CDC also offers COVID-19 View surveillance data about US COVID-19 activities. Chen *et al.* (2020) report how to use data visualization to track the virus spreading. A study by Khanam *et al.* (2020) found that males are more prone to this disease, and older people are more at risk by comparing the infections of people through data visualization. The data visualization tool also helps users get a better understanding of information about the confirmed cases of the COVID-19 (Leung *et al.*, 2020), such as the common symptoms for this novel coronavirus and the differences from other SARS, Middle East respiratory syndrome (MERS) and swine flu.

Predictive modeling

The COVID-19 pandemic daily data can also be used to predict key aspects of the case growth. Predictive modeling can assist the proper planning for healthcare resources and related socioeconomic decision-making (Marmarelis, 2020). Predictive modeling can provide reliable estimates of key parameters of the unfolding infectious process. Useful insights can be generated to assist policy and planning makers in the dynamic structure of the infectious process (Marmarelis, 2020). For example, modeling can show the maximum number of total confirmed cases and further help the proper clinical management. For example, the health-care facilities need to prepare more resources to accommodate the possible surge of the confirmed cases and higher hospitality rates. Based on the public health data, institutions can obtain the predicted model to interpret the dynamic characteristics and help policy planning and operational implementation. And another case study in Pakistan shows the rate of mortality would decrease by the point of date. The total number of deaths will reach its maximum point; then, it will gradually decrease (Daniyal *et al.*, 2020). This research applied three regression models and used the data from the National Institute of Health of Pakistan. With the predicted model, the government and policymakers should create strict rules and regulations and follow the prevention guidelines.

Sentiment analysis

Social media plays an essential role during the COVID-19 pandemic. When a pandemic breaks out, many people express opinions on social media and seek to interact with others using social media. People often obtain the news of the COVID-19 and the information on preventing coronavirus from social media. There was negativity, fear, disgust and sadness about the lockdown, positive and negative sentiments about vaccine and treatment drugs. The sentiment changes from the beginning of the lockdown to the middle of the pandemic are worth investigating. Policymakers, public health-care departments, and the government should be aware of the changing sentiment of the public on social media platforms. The two major social media platforms are Facebook and Twitter. Many researchers did sentiment analysis about the COVID-19 during the pandemic. There may be different sentiment trends on the pandemic in other countries and cultures. As there are many aspects influencing public sentiment, one work uses a Recurrent Neural Network (RNN) to classify emotions on Twitter (Nemes and Kiss, 2021). The research identified positive and negative tweets on social media platforms during the pandemic. Alamoodi *et al.* (2020) conducted a systematic review on sentiment analysis

The current issue and full text archive of this journal is available on Emerald Insight at: **https://www.emerald.com/insight/2398-6247.htm**

Information Discovery and Delivery
49/3 (2021) 189–192
 [ISSN 2398-6247]
[DOI 10.1108/IDD-08-2021-161]

Information Discovery and Delivery
Volume 49 · Number 3 · 2021 · 189–192

and its applications in fighting the COVID-19 and infectious diseases. As a result, they emphasized the current standpoint and opportunities for research in this area.

Social network analysis

Social networks can amplify the spread of both harmful and beneficial behaviors during the pandemic. The coronavirus can quickly spread from person to person. People in the center of social networks tend to have connections with more people, and thus they are more likely to among the first to be infected. But the same central people may be instrumental in slowing the disease spread by promoting positive interventions like hand washing, facemask and physical distancing to a wide range of people (Shi *et al.*, 2020). SNA is the measurement of social entities and their relationships (Oliveira and Gama, 2012). Using SNA as an analysis method can support emergency agencies in understanding the network structure of people's interactions as situations rapidly emerge. In a social network, centrality measures network players who are most connected and hold influential positions in the network (Craig *et al.*, 2020). Yum (2020) uses SNA to assess twitter data and explores how public key players play their roles in social networks for COVID-19. The results show that the presidents, the World Health Organization (WHO) and its regional offices, the Centers for Disease Control, and news channels play a crucial role in the news of COVID-19 for people.

Big data analytics

Big data analytics can be used to identify people that need quarantine based on their travel history, predict contagious disease diffusion, forecast future disease outbreaks (Elkin *et al.*, 2017) and speed up the development of anti-viral drugs and vaccines. In Taiwan, big data analytics has been leveraged to help identify COVID-19 cases and generate real-time alerts through analyzing clinical visits, travel history and clinical symptoms (Wang *et al.*, 2020). Machine learning (ML) methods can play vital roles in identifying COVID-19 patients by visually analyzing their chest x-ray images, predicting the trends of spreading COVID-19 and prioritizing testing for COVID-19 when testing resources are limited (Elaziz *et al.*, 2020; Zoabi *et al.*, 2021). As artificial intelligence and machine learning scientists have been eagerly searching and waiting for real-time data generated by this pandemic around the world, timely delivery of COVID-19 patient data, such as physiological characteristics and therapeutic outcome of COVID-19 patients, followed by subsequent data transformation for easy access, is extremely important, but challenging (Alimadadi *et al.*, 2020). Randhawa *et al.* (2020) identifies an intrinsic COVID-19 virus genomic signature and uses it together with a machine learning-based alignment-free approach for an ultra-fast, scalable and highly accurate classification of whole COVID-19 virus genomes. This research also suggests that the alignment-free whole-genome machine-learning approach can provide a reliable real-time option for taxonomic classification for novel viral and pathogen genome sequences.

Assessing the impact and consequences of the pandemic

Coronavirus is still spreading worldwide and has caused severe consequences on people, the economy and society. Governments are under a lot of pressure to combat the outbreak and reduce its negative implications on people and the economy. Any mismanagement in addressing the pandemic could carry political costs, as their legitimacy and competency will be called into question by the people. The public has asked the governments to provide transparency and access to timely information to battle the COVID-19 outbreak. Clear direction in mitigating the consequences of the pandemic is necessary to win the battle against the pandemic. Accurate and reliable information is needed to alert the public in preventing the spread of the coronavirus. In the absence of facts and trust, rumors and panic are more likely to spread as people became emotional. These emotions created anger and fear that posed a threat to the government. Besides, the harmful effects of the pandemic are not distributed equally, and some minority groups are hit especially hard based on recent data analysis of publicly available data (Fairlie *et al.*, 2020; Jahromi and Hamidianjahromi, 2020). Further data science research and analysis is needed to increase our understanding of the impact and consequences of the pandemic for different populations and various sectors of society.

Papers in this special issue

The idea for this special issue began in June 2020 after the World Health Organization declared a global pandemic as the coronavirus rapidly spreads across the world. As of March 6, 2020, the global death toll from the coronavirus pandemic has topped 70,000. In light of the emergence and spread of the COVID-19 across the world, we launched a special issue call for papers that help understand and mitigate the rapid spread of this disease from the data and information science perspective. We are particularly interested in papers that explore how to track, model, understand and predict the spread of the COVID-19 through various data science and visualization methods. The goal is to provide a platform for sharing timely research to help societies address this unprecedented challenge.

In this special issue, we have accepted seven articles. These articles cover simulations, sentiment analysis, social media behaviors and mental health in the COVID-19 pandemic. The article "Twitter users exhibited coping behaviors during the COVID-19 lockdown: An analysis of tweets using mixed methods" by Mittal *et al.* follows the quasi-inductive approach to conduct the research. Data were extracted using relevant keywords from Twitter, and a sample was drawn from the Twitter data set to ensure the data is more manageable from a qualitative research standpoint and that meaningful interpretations can be drawn from the data analysis results. This study found that during the lockdown, most users on Twitter shared positive opinions toward it because of its potential to halt the spread of the COVID-19 and prevent further deaths. Many people were still able to keep themselves engaged and entertained although some users reported negative sentiments.

The article "A comparative study of modified SIR and Logistic predictors using local level database of COVID-19 in India" by Bajaj *et al.* deployed the modified SIR and the

Information Discovery and Delivery
Volume 49 · Number 3 · 2021 · 189–192

Logistic model to analyze the COVID-19 patients' database of India and three Municipal Corporations, namely, Akola, Kalyan-Dombivli and Mira-Bhayander. This study provides evidence to show the superiority of the modified SIR over the Logistic model. The models give accurate predictions for a period of up to 14 days. The prediction accuracy of the models is limited due to changes in government policies. This can be observed by the drastic increase in the COVID-19 cases after Unlock 1.0 in India. The models have proven to effectively predict for both the National and Municipal Corporation (local level) databases.

The article "An agent-based model for simulating COVID-19 transmissions on university campus and its implications on mitigation interventions: A case study" by Zhou *et al.* developed an agent-based model to mimic the virus transmission dynamics on campus. Scenario-based experiments are conducted to evaluate various interventions, including course modality shift (from face-to-face to online), social distancing, mask use and vaccination. A case study is performed for a typical US university. With 10%, 30%, 50%, 70% and 90% course modality shift, the number of total cases can be reduced to 3.9%, 20.9%, 35.6%, 60.9% and 96.8%, respectively, comparing against the baseline scenario (no interventions).

The article "An empirical investigation of precursors influencing social media health information behaviors and personal healthcare habits during coronavirus (COVID-19) pandemic" by Muhammad *et al.* collected data through an online survey conducted in two different universities situated in highly COVID-19-affected cities – Wuhan and Zhengzhou, China. The valid data consists of 230 useful responses from WeChat users. To analyze the final data set, structural equation modeling (SEM) is used. The results indicate that perceived health information credibility (PIC), trust on the medium (TRM) and peer influence (PI) significantly affect health ISI which further affects health information re-sharing behaviors (IRB) and personal healthcare habits (PHH). Besides, the results also identify that PI has a direct, positive and significant effect on health IRB via social media during the COVID-19 pandemic.

The article "An analysis of attitude of general public toward COVID-19 crises – sentimental analysis and a topic modeling study" by SV and Ittamalla analyzed 433,195 tweets collected from February 1, 2020, to June 27, 2020, regarding the COVID-19 pandemic. Natural language processing (NLP) was used to analyze the tweets for this study. NLP was used to track the changes in the general public's sentiment toward COVID-19 crises and Latent Dirichlet Allocation (LDA) was used to understand the issues that shape the general public's sentiments during the crisis time. Using Python library Word Cloud, the authors further derived how the primary concerns regarding COVID-19 crises vary from February to June of 2020.

The article "COVID-19 and India: what next?" by Behl *et al.* leverages the susceptible, infected, recovered and dead (SIRD) epidemiological framework for predictive modeling. The basic reproduction number R0 is derived by an exponential growth method using RStudio package R0. The differential equations reflecting the SIRD model have been solved using Python 3.7.4 on the Jupyter Notebook platform. Python Matplotlib 3.2.1 package is used for visualization. The study offers insights on the peak-date, peak number of COVID-19 infections and end-date about India and five of its states. The results subtly indicate the amount of effort required to eliminate the infection. It could be leveraged by the political leadership and industry doyens for economic policy planning and execution.

The article "Delivery Forecasting mental health and emotions based on social media expressions during the COVID-19 pandemic" by Tommasel *et al.* presents an approach for forecasting mental health conditions and emotions of a given population during the COVID-19 pandemic in Argentina based on social media contents. Mental health conditions and emotions are captured via markers, which link social media contents with lexicons. They built time series models to describe the evolution of markers and their correlation with crisis events. They also used the time series for forecasting markers and identifying high prevalence points for the estimated markers. The authors evaluated different forecasting strategies that yielded different performance and capabilities. In the best scenario, high prevalence periods of emotions and mental health issues can be satisfactorily predicted with a neural network strategy, even at the early stages of a crisis (e.g. a training period of seven days).

Xin Tian
Department of Information Technology, Kennesaw State University, Kennesaw, Georgia, USA

Wu He
Old Dominion University, Norfolk, Virginia, USA, and

Yunfei Xing
Central China Normal University, Wuhan, China

References

Alamoodi, A., Zaidan, B., Zaidan, A., Albahri, O., Mohammed, K., Malik, R., . . . Alaa, M. (2020), "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: a systematic review", *Expert Systems with Applications*, p. 114155.

Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P.B., Joe, B. and Cheng, X. (2020), "Artificial intelligence and machine learning to fight COVID-19".

Bahja, M. and Safdar, G.A. (2020), "Unlink the link between COVID-19 and 5G networks: an NLP and SNA based approach", *IEEE Access*, Vol. 8, pp. 209127-209137.

Chen, B., Shi, M., Ni, X., Ruan, L., Jiang, H., Yao, H., . . . Ge, T. (2020), "Data visualization analysis and simulation prediction for covid-19. arXiv preprint arXiv:2002.07096".

Craig, K., Humburg, M., Danish, J.A., Szostalo, M., Hmelo-Silver, C.E. and McCranie, A. (2020), "Increasing students' social engagement during COVID-19 with net. Create: collaborative social network analysis to map historical pandemics during a pandemic", *Information and Learning Sciences*, Vol. 121 Nos 7/8, pp. 533-547.

Daniyal, M., Ogundokun, R.O., Abid, K., Khan, M.D. and Ogundokun, O.E. (2020), "Predictive modeling of COVID-19 death cases in Pakistan", *Infectious Disease Modelling*, Vol. 5, pp. 897-904.

Elaziz, M.A., Hosny, K.M., Salah, A., Darwish, M.M., Lu, S. and Sahlol, A.T. (2020), "New machine learning method for image-based diagnosis of COVID-19", *Plos One*, Vol. 15 No. 6, p. e0235187.

Elkin, L.S., Topal, K. and Bebek, G. (2017), "Network based model of social media big data predicts contagious disease diffusion", *Information Discovery and Delivery*, Vol, Vol. 45 No. 3, pp. 110-120.

Fairlie, R.W., Couch, K. and Xu, H. (2020), The impacts of COVID-19 on minority unemployment: first evidence from April 2020 CPS microdata (No. w27246), National Bureau of Economic Research.

Gupta, M., Jain, R., Taneja, S., Chaudhary, G., Khari, M. and Verdú, E. (2021), "Real-time measurement of the uncertain epidemiological appearances of COVID-19 infections", *Applied Soft Computing*, Vol. 101, p. 107039.

Jahromi, A.H. and Hamidianjahromi, A. (2020), "Why African Americans are a potential target for COVID-19 infection in the United States", *Journal of Medical Internet Research*, Vol. 22 No. 6, p. e19934.

Khanam, F., Nowrin, I. and Mondal, M.R.H. (2020), "Data visualization and analyzation of COVID-19", *Journal of Scientific Research and Reports*, Vol. 26 No. 3, pp. 42-52.

Khayyat, M., Laabidi, K., Almalki, N. and Al-Zahrani, M. (2021), "Time series Facebook prophet model and python for COVID-19 outbreak prediction", *Computers, Materials & Continua*, Vol. 67 No. 3, pp. 3781-3793.

Leung, C.K., Chen, Y., Hoi, C.S., Shang, S., Wen, Y. and Cuzzocrea, A. (2020), September). "Big data visualization and visual analytics of COVID-19 data", *2020 24th International Conference Information Visualisation (IV)*, *IEEE*, pp. 415-420.

Marmarelis, V.Z. (2020), "Predictive modeling of Covid-19 data in the US: adaptive phase-space approach", *IEEE Open Journal of Engineering in Medicine and Biology*, Vol. 1, pp. 207-213.

Nemes, L. and Kiss, A. (2021), "Social media sentiment analysis based on COVID-19", *Journal of Information and Telecommunication*, Vol. 5 No. 1, pp. 1-15.

Oliveira, M. and Gama, J. (2012), "An overview of social network analysis", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 2 No. 2, pp. 99-115.

Randhawa, G.S., Soltysiak, M.P., El Roz, H., de Souza, C.P., Hill, K.A. and Kari, L. (2020), "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study", *Plos One*, Vol. 15 No. 4, p. e0232391.

Yum, S. (2020), "Social network analysis for coronavirus (COVID-19) in the United States", *Social Science Quarterly*, Vol. 101 No. 4, pp. 1642-1647.

Wang, C.J., Ng, C.Y. and Brook, R.H. (2020), "Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing", *JAMA*, Vol. 323 No. 14, pp. 1341-1342.

Zoabi, Y., Deri-Rozov, S. and Shomron, N. (2021), "Machine learning-based prediction of COVID-19 diagnosis based on symptoms", *NPJ Digital Medicine*, Vol. 4 No. 1, pp. 1-5.

Further reading

Barkur, G. and Vibha, G.B.K. (2020), "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: evidence from India", *Asian Journal of Psychiatry*, Vol. 51, p. 102089.

Chakraborti, S., Maiti, A., Pramanik, S., Sannigrahi, S., Pilla, F., Banerjee, A. and Das, D.N. (2021), "Evaluating the plausible application of advanced machine learnings in exploring determinant factors of present pandemic: a case for continent specific COVID-19 analysis", *Science of the Total Environment*, Vol. 765, p. 142723.

He, W., Zhang, Z.J. and Li, W. (2021), "Information technology solutions, challenges, and suggestions for tackling the COVID-19 pandemic", *International Journal of Information Management*, Vol. 57, p. 102287.

Muthusami, R. and Saritha, K. (2020), "Statistical analysis and visualization of the potential cases of pandemic coronavirus", *VirusDisease*, Vol. 31 No. 2, pp. 204-208.

Yin, S., Zhang, N. and Dong, H. (2020), "Preventing COVID-19 from the perspective of industrial information integration: evaluation and continuous improvement of information networks for sustainable epidemic prevention", *Journal of Industrial Information Integration*, Vol. 19, p. 100157.