Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 10 October 2021

Review Article

A Real Time Analytics for COVID-19 Tracking Based on Big Data Extracted on a Free Online Provider

Abderrahmane Ez-zahout[1]*, Slimane El Ouafi[2] and Omar Aitoulghazi[2]

[1]IPSS Team, Faculty of Science, Mohamed V University, Morocco

[2]School of Science and Engineering, Al Akhawayn University, Morocco

***Corresponding Author:** Abderrahmane Ez-zahout, IPSS Team, Faculty of Science, Mohamed V University, Morocco.

Received: June 25, 2021

Published: September 07, 2021

Abstract

The COVID-19 pandemic has caused a large number of human losses and impacted different sectors; economic, social, societal, and health systems around the world. Controlling such pandemic requires understanding its characteristics, causes and consequences, which can be done using data. Big data analytics tools play a vital role in building knowledge required in making decisions and precautionary measures. However, to do that, many challenges are faced whether it is about privacy and security, data sharing, information correctness, or patient cooperation. To better understand how this pandemic is affecting our world and the countries that got the most affected, we did some analyzed data that we extracted from a free provider and visualized.

Keywords: COVID-19; Big Data Analytics; Making Decision; Patterns; Knowledge

Introduction

The spread of the COVID-19 global pandemic lead to an exponentially mounting and extraordinary volume of data that can be used to improve our understanding of big data management research as well as a deeper understanding of a range of analytical tools that could be utilized to better anticipate and respond to such critical and risky events that impact the peaceful life of people. Many governments found themselves forced to use data driven solutions to find effective solutions to this pandemic. Throughout this paper we will cover the different solutions in different areas that were presented by researchers using big data analytics tools. Even if these solutions turned out to be effective, many challenges came to the surface while modeling them.

Related work

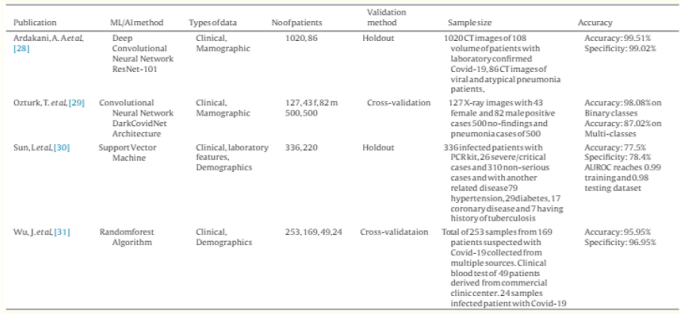
This study [2,12] categorizes the kind of analytics in big data, giving examples of data, web and text mining that could be used. First, the authors explain the descriptive and diagnostic analytics. This type of analytics aims to present data in an easy, understandable format to analyze and showcase the cause and effect relationships. For descriptive data analysis, it does nothing but summarizes previous data to give an overall overview of probable patterns and trends; this process relies on data visualization and statistical computations (e.g. correlation). On the other hand, diagnostic analytics provides a historical account that aims to detect problems or opportunities from the given, existing operations. For this purpose, a causal-explanatory statistical modeling is the approach to adopt. Then, the authors move to discuss another type of analytics which is the predictive analytics. As its name refers to, this type is all about predicting what may happen in the future, and this where most of machine learning and data mining techniques are used; this type is classified into three predictive analytics techniques: statistical inference, machine learning and other techniques that tackle unstructured data. Statistical inference consists of using statistical approaches to analyze large amounts and high dimensional structured data. Moreover, machine learning (e.g. regression, artificial neural networks, classification, clustering, association, etc.) makes use of different complex algorithms to build predictive models on very large datasets. The last class tackles unstructured data; in this category, we find text mining, image and sentiment analysis, web

mining and social networks analysis. Finally, the last type of analytics that is discussed in this paper, is the prescriptive analytics. Those analytics help in making decisions and quantify errors. They are categorized into three: optimization, simulation and logic-based models. Optimization uses linear programming, non-linear programming, stochastic optimization, Bayesian optimization and evolutionary computation. On the other hand, simulation consists on modeling a real-life or hypothetical operation to predict how a system would behave. Finally, the logic-based models give a hypothesized representation of theory of change about how an intervention leads to a specified output.

The following paper [2] shows an example of such application of machine learning in the detection of covid19 cases. It compares different adopted algorithms using different datasets and sample sizes of data. The two top ones are based on artificial neural networks, and most specifically, transfer learning. Transfer learning, in the world of ANNs, is about making use of already trained models and their optimized parameters. The first model is a deep convolutional neural network based on ResNet-101 network; it was able to reach about 99.5% in accuracy. Followed by another convolutional neural network that is based on the DarkCovidNet Architecture, and it has been 98% accurate in detecting covid19 cases. Other included machine learning techniques are a support vector machine-based model, and it achieved an accuracy of 77.5%. The last model is a random forest algorithm, with an accuracy of almost 96% (Figure 1).

Figure 1: Comparative analysis of the adopted models [2,4,9,10,13,14].

The spread of the global pandemic, COVID-19, has generated a huge and varied amount of data that can be used by applying the following big data analytics techniques in various areas.



Publication	ML/AImethod	Typesofdata	Noofpatients	Validation method	Samplesize	Accuracy
Ardakani, A. A*et al.* [28]	Deep Convolutional Neural Network ResNet-101	Clinical, Mamographic	1020,86	Holdout	1020CTimagesof108 volumeofpatientswith laboratoryconfirmed Covid-19,86CTimagesof viralandatypical pneumonia patients,	Accuracy:99.51% Specificity:99.02%
Ozturk, T. *et al.*[29]	Convolutional Neural Network DarkCovidNet Architecture	Clinical, Mamographic	127,43 f,82 m 500,500	Cross-validation	127X-ray imageswith43 female and82 malepositive cases500no-findingsand pneumoniacasesof500	Accuracy:98.08%on Binaryclasses Accuracy:87.02%on Multi-classes
Sun, L*et al.*[30]	Support Vector Machine	Clinical, laboratory features, Demographics	336,220	Holdout	336infectedpatientswith PCRkit,26severe/critical casesand310non-serious casesandwithanother relateddisease79 hypertension,29diabetes,17 coronarydiseaseand7having historyoftuberculosis	Accuracy:77.5% Specificity:78.4% AUROC reaches0.99 trainingand0.98 testing dataset
Wu, J.*et al.*[31]	Randomforest Algorithm	Clinical, Demographics	253,169,49,24	Cross-validataion	Total of253samplesfrom169 patientssuspectedwith Covid-19collectedfrom multiplesources.Clinical bloodtestsof 49patients derivedfromcommercial cliniccenter.24samples infectedpatientwithCovid-19	Accuracy:95.95% Specificity:96.95%

Figure 2: Areas on which big data analytics techniques are applied for Covid-19 [3,11].

Based on [3], In order to try to control the Covid-19 pandemic, several studies came up with solutions that are related to one of these three areas: diagnosis, estimate or predict risk score, and healthcare decision-making.

When it comes to the diagnosis area, the suspected Covid-19 cases are diagnosed using the RT-PCR test which stands for: Reverse Transcription-Polymerase Chain Reaction.

However, due to the increased demand for diagnosing suspected Covid-19 cases, a high number of researchers have proposed alternative diagnosis test. But, before coming up with the alternatives, the researchers had to identify the symptoms associated with positive results of the Covid-19 cases. The study discovered that the most common symptoms of Covid-19 cases are fever, myalga, and anosmia, while, on the other hand, negative cases had few symptoms limited to sore throat or no symptoms at all. Another study from [6] tried to separate a covid-19 cough sound from other respiratory sounds using a website and Android app based of crowdsourcing data from thousands unique users. To do that, they used logistic regression, gradient boosting trees, and support vector machines classifiers as methods to distinguish the differentiate between users on different levels: asthmatics parents, smokers. Based on these methods, they were able to reach a 82% distinction accuracy from the two coughs; however, they would still need more studies to increase the characteristics of a Covid-19 cough sound.

Other researchers [8], tried to use machine learning techniques, spark-based linear models, multilayer perceptron, and Long Short-Term Memory with a two-stage cascading platform to increase the predicted accuracy of the test on different datasets. They tried these models on two datasets for cardiac arrhythmia and resource locator, which made them perform with higher accuracy and lower computation time.

The second area they worked on is the estimation or prediction of risk score. Here, researchers wanted to estimate the risk score, that helps professionals in the medical field, determine the priority and urgency of each case. From there, they are able to come up with solutions to control more the spread of the pandemic. One of these solutions came from the hypothesis that Covid-19 infection could lead to serious cardiovascular diseases or even worse. To confirm this hypothesis, researches in [7] had to use statistical analysis by employing multi-factorial logistic regression model to study Covid-19 related causes. From that study, they found out that patients with diabetic conditions, hypotension patients, and elder males have higher chances to develop serious heart-related conditions.

Another solution that was referred to control and estimate the spread of the pandemic was done by [5] using Internet of Things (IoT). The goal of this research was to find unregistered Covid-10 patients and infectious places to help authorities to control these places by disinfecting them and quarantine the person that had contact with patients that turned out to be Covid-19 positive.

Some researchers on their side, tried to provide a model that predicts the successive steps that Covid-19 patients go through to help plan an efficient prevention process. The stages are referred to as SIDARTHE: susceptible, infected, diagnosed, ailing, recognized, threatened, healed, and extinct. After combining the model stages and the available Covid-19 data, it was found that its use is of urgent necessity.

The third and last area of study was the healthcare decision-making. After the world witnessed a tremendous increase in the number of Covid-19 cases, medical tools and models that help in making decision started to be highly demanded. One of the models that was developed was the Conscious-based Susceptible-Exposed-Infective-Recovered(C-SEIR) model. This model helps in analyzing the usefulness of the lockdown and protective countermeasures to decrease the influence of Covid-19.

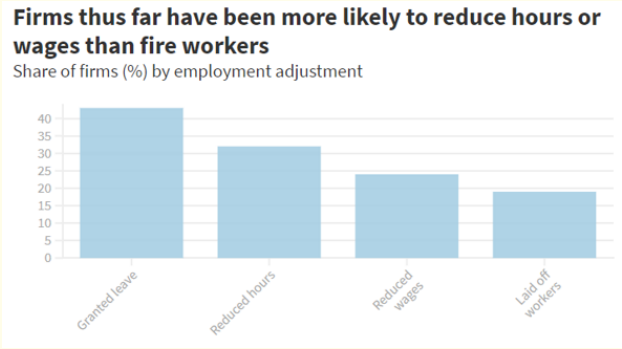


Figure 3: Impact of Covid-19 on businesses and jobs [1].

Some researchers on their side, worked on a model that monitor patients and keeps track of symptoms daily. The model then provides advice and reminders via text messages or by phone.

Although the application of big data analysis tools to rise solutions to the Covid-19 pandemic came out with multiple beneficial outcomes. Several challenges were faced while designing these solutions, some of them are security and privacy, information correctness, and patient cooperation.

Medical data is known to be very critical which makes its access for studies complicated. Therefore, it is necessary to define the mechanisms and strategies that facilitate access to these data while respecting patients' privacy.

Information correctness became a serious topic of discussion after the wild spread of fake information on social media, especially when it comes to medical information such as diseases, viruses, vaccines, and other related topics. The previous may not only cause the damage or government efforts to control the situation but also negatively impact the psychological behavior of people. Nevertheless, many AI and big data analytics tools can be used to filter social media information and alert people.

Finally, for researchers to come up with models and solutions, they need data, which cannot be collected without the cooperation of patients. However, many patients are not willing to share their private health information with others, making data collection difficult.

Most of the other studies [2,10,13,14] focus on comparing different adopted, developed models in tackling such analytics problematics, discussing many of the challenges and advantages of such models.

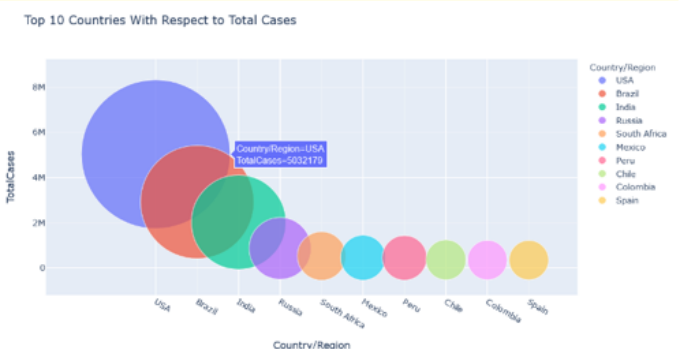


Figure 4: Closed classrooms around the world [1].

Impact of covid-19 on different sectors

Covid19 has impacted many sectors, causing the world global economy to reach level it did not for decades, medical conditions that are horrible and an education far below average because of online classes.

We can see from (Figure 3) how Covid-19 impacted micro, small and medium enterprises. Most of these enterprises found their sales drop by half during the pandemic, forcing them to reduce hours and wages; which lead to a reduce in family incomes.

In (Figure 4), we can see the number of schools that were closed during the pandemic. The closed schools lead to an approximate number of 1.5 billion of students out of school.

Data anaylsis and visualization

The data that is used in this project is extracted from the worldometers website. It is then imported to the Hadoop distributed filesystem (HDFS) and analyzed. The following figures, implemented in python using the plotly library, illustrate some of the visualizations performed on the data for the sake of analysis and understandability:

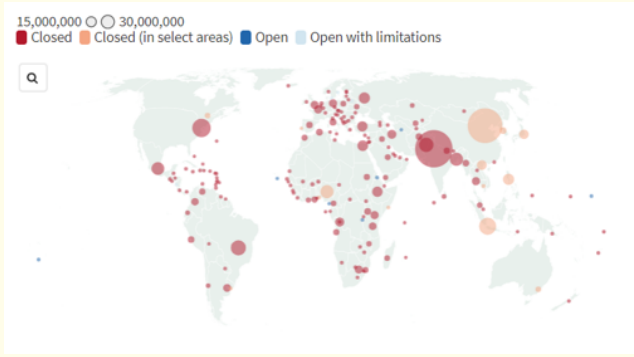


Figure 5: Top 10 countries with the highest total cases.

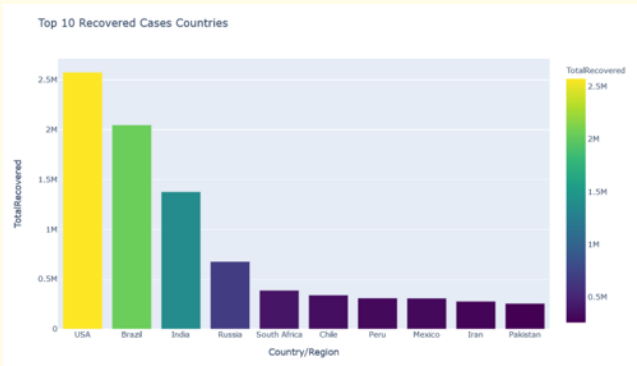


Figure 6: Top 10 countries with the highest total deaths.

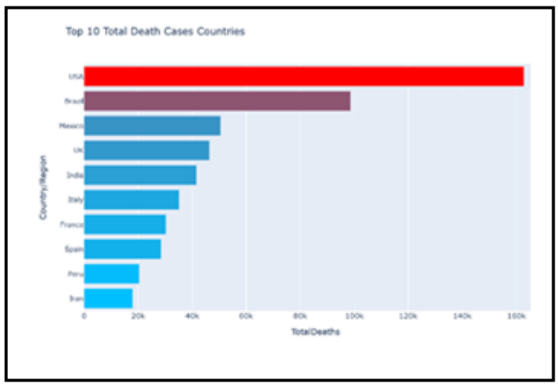


Figure 7: Top 10 Countries with the highest recovered cases.

Since the outbreak of the Covid-19 crisis in early 2020, all governments have been at the forefront of managing the COVID-19 health pandemic and its economic and social impacts. Together with central governments and social security bodies, they have significant responsibilities in the different areas affected by the COVID-19 crisis. In many countries, subnational governments are responsible for critical aspects of health care, from primary care to secondary care, including hospital management, accounting for 25% of total public health expenditure, on average (Figure 8).

Conclusion

Our work aims to analyze the real time data provided in the worldometers dataset about covid19. Our study tackles a batch analysis that is based on stored data once in a timestamp (e.g. a day) using CRON scheduling. We believe that such analysis of this data could be approached in such manner because the data itself gets updated once in a while; hence, we could treat it as batch at once, nevertheless, its huge size actually requires a distributed system such as HDFS.

However, and for the sake of improvement, we think that making use of Spark Streaming would be suitable for such issue. Thanks to its notion of treating real time data as mini batches of data, we could avoid implementing actual scripts to run our analysis automatically; that would save us a lot of effort and resources.

For more advanced treatment, Kafka could be adopted to collect the data in actual real time resulting in sequentially updated analysis that do not go beyond a number seconds using Spark MLlib or simply other Python libraries for analysis.

To conclude, both the state of art and industry insist on how important big data analytics, using different machine learning and data mining techniques, are in today's world. The problem of covid19 pandemic is just one example of how these fields contributed to humanity and helped in providing a variety of solutions that led to efficient decision making. Indeed, there are a lot of improvements to be made, however, research has already been giving birth to many new ideas that would make these components the pillars of computer science in general. Nowadays, data is everything. However, it has become characterized by three main Vs that are well known in the world of big data, volume, velocity and variety. This data requires new type of processing and storing in a distributed, parallel fashion. Volume represents how huge the size of data is today, while velocity represents how fast data is being collected and received, and finally variety refers to different kinds of data one can gather (e.g. structured, semi-structured and unstructured). These new problems pushed us today to invent many new tools to process and analyze big data such as Apache Hadoop (HDFS for storing and other tools in the ecosystem to process and collect data such as Flume) and Spark Engine to process data and apply some machine learning techniques on the collected data.

Bibliography

1. Paul Blake Divyanshi Wadhwa. "2020 Year in Review: The impact of COVID-19 in 12 charts". World Bank Blogs (2020).

2. Agbehadji., *et al*. "Review of Big Data Analytics, Artificial Intelligence and Nature-Inspired Computing Models towards Accurate Detection of COVID-19 Pandemic Cases and Contact Tracing". *International Journal of Environmental Research and Public Health* 17.15 (2020): 5330.

3. Alsunaidi S J., *et al*. "Applications of Big Data Analytics to Control COVID-19 Pandemic". *Sensors* 21.7 (2021): 2282.

4. Ardakani AA., *et al*. "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks". *Computers in Biology and Medicine* 121 (2020): 103795.

5. Benreguia B., *et al*. "Tracking COVID-19 by Tracking Infectious Trajectories". *IEEE Access* 8 (2020): 145242-145255.

6. Brown C., *et al*. "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data". In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data (ACM), New York, NY, USA, 25-27 August (2020): 3474-3484.

7. Chen Q., *et al*. "Cardiovascular Manifestations in Severe and Critical Patients with COVID -19". *Clinical Cardiology* 43 (2020): 796-802.

8. Khan MA., *et al*. "A Two-Stage Big Data Analytics Framework with Real World Applications Using Spark Machine Learning and Long Short-Term Memory Network". *Symmetry* 10 (2018): 485.

9. Lalmuanawma S., *et al*. "Applications of Machine Learning and Artificial Intelligence for Covid-19 (SARS-CoV-2) pandemic: A review". *Chaos, Solitons and Fractals* (2020): 110059.

10. Ozturk T., *et al*. "Automated detection of COVID-19 cases using deep neural networks with X-ray images". *Computers in Biology and Medicine* 121 (2020): 103792.

11. PEX Process Excellence Network 6 Ways Pharmaceutical Companies are Using Big Data to Drive Innovation and Value (2020).

12. Sheng Jie., *et al*. "COVID-19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions". *British Journal of Management* (2021).

13. Sun L., *et al*. "Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19". *Journal of Clinical Virology* (2020): 104431.

14. Wu J., *et al*. "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results". *medRxiv* (2020).

Volume 3 Issue 10 october 2021