International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-12, Issue-4 (August 2022)
www.ijemr.net https://doi.org/10.31033/ijemr.12.4.19

Big Data Analysis on COVID-19

Vinay Kailash Upadhyay[1], Dr. Dinesh Dattatray Patil[2] and Prof. Yogesh Patil[3]
[1]Department of Computer science & Engineering, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, INDIA
[2]Associate Professor and Head, Department of Computer science & Engineering, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, INDIA
[3]Assistant Professor, Department of Computer science & Engineering, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, INDIA

[1]Corresponding Author: vinayupadhyay158@gmail.com

ABSTRACT

Over the past 2 years, the Coronavirus has rapidly spread to all parts of the world. Scientist and researchers are continuing their research to find a permanent cure. As the number of cases are increasing, so the tests are for Coronavirus is increasing rapidly, it is impossible to maintain data of test due to the time and cost factors. Big data is very helpful to maintain the track record of the COVID-19 infected patients in a very systematic way and will reduce the time delay for the results of the medical tests and modulate doctors to give proper medical treatment to the infected person. Big data analytics play an important role in building knowledge, studies required in making decisions and precautionary measures. However, due to the vast amount of data available on COVID-19 from various sources, there is a need to review the roles of big data analysis in controlling and tracking the spread of COVID-19, presenting the main challenges and directions of COVID-19 data analysis, as well as providing a framework on the related existing applications and studies to facilitate future research on COVID-19 analysis. Keywords- big data analytics, 2019 novel coronavirus disease (COVID-19).

Keywords-- Big Data Analytics, COVID-19, Health Science

I. INTRODUCTION

Coronaviruses are a large family of viruses, and it can cause illness ranging from the common cold to more severe diseases such as the Middle East Respiratory Syndrome (MERS) and severe acute respiratory syndrome (SARS). The two member of this family are responsible to be spread the coronaviruses named as MERS-CoV and SARS-CoV. First case of SARS was detected in 2002 in China and MERS was first in 2012 in Saudi Arabia. The latest virus seen in Wuhan, China is called SARS-COV-2, and it causes coronavirus. So, to track Covid-19, testing become an important part of every person daily life in past 2 years, like RT-PCR and antigen testing, due to which it becomes difficult for government and hospitals to maintain data of infected people. To maintain such huge data big data plays a very helpful hand to reduce complexity and replication of data.

II. LITERATURE SURVEY

Nanshan Chen and others [1], performed a retrospective, single-centre study of various patients' data from Jinyintan Hospital in Wuhan, China. In this research they described the epidemiological data (short term) or long-term exposure to virus epicentres, signs and symptoms, laboratory results, CT Findings and clinical outcomes. Though this research does not directly focus on the prediction of COVID-19, it gives us a better understanding of the clinical outcomes.

A. COVID-19 Research

Partially because of the COVID-19 pandemic, many researchers have explored on different aspects of the COVID-19 disease. These led to numerous works on COVID-19 in different disciplines or areas:

- For medical and health sciences, there have been systematic reviews on literature about medical research on COVID-19 [7][8], (b) clinical and treatment information, as well as (c) drug discovery and vaccine development.
- For social sciences, there have been studies on crisis management for the COVID-19 outbreak.

B. Confirmed Cases and Mortality

Many existing works on the COVID-19 epidemiological data focused on reporting simply the numbers of confirmed cases and mortality spatially and/or temporally. In other words, they highlight (a) spatial differences among different continents, countries, or sovereignties and/or (b) temporal trends, which both may demonstrate how effective different public health strategies and mitigation techniques—such as social/physical distancing, stay-at-home orders, and/or lockdown—help in "flattening the (epidemic) curve".

While these overall numbers of confirmed cases and mortality are important in showing the severity of the

disease at a specific time or time interval. However, it is equally important to:

- explore the breakdown of these numbers among different gender and/or age groups, and
- discover other useful knowledge (e.g., symptoms, clinical course and outcomes, transmission methods) from the epidemiological data.

A reason is that the discovered knowledge can reveal useful information (e.g., some characteristics of COVID-19 cases) associated with the disease. This, in turn, helps users to get a better understanding on characteristics of the confirmed cases of COVID- 19 (rather than just the numbers of cases).

Abdel-Basst, Mohamed [2] developed a diagnosis model for COVID-19 detection and diagnosis of symptoms to define appropriate care measures, Using Best Worst Method (BWM) by Symptoms and CT scans data type, used Body sensors as data source The model can differentiate COVID-19 from four other viral chest diseases with 98% accuracy.

Stojanovic, R.; Skraba, A.; Lutovac [3], Design a medical device to detect and track respiratory symptoms of COVID-19, Symptoms could be use as data type, Headsets and mobile phone use to gather information The approach provided good and stable results and can be expanded to include more sensors to detect other COVID-19 symptoms. Jeong, H.; Rogers, J.A.; Xu [4], Discuss the importance of developing complementary technologies to diagnose and monitor COVID-19 infections, using data type like activity data, medical data from Sensors Recommend deploying advanced wearable technologies configured to directly address needs in COVID-19 monitoring and noticing the symptoms [4].

III. OUR BIG DATA SOLUTION

In this section, we describe our big data solution on COVID-19 epidemiological data.

A. Data Collection and Integration

Recall from Section I that big COVID-19 epidemiological data can be characterized by their variety in two aspects. First the data can be generated and collected from a wide variety of data sources. As a concrete example, in India, healthcare is a responsibility of state governments. So, Indian COVID19 epidemiological data are gathered from each state (or territory), and state wise data are obtained from health regions (which are also known as Primary Health Care Units) within the states every district.

Second, the big COVID-19 epidemiological data can contain a wide variety of information, which usually includes:

- administrative information—such as (a) an unique privacy-preserving identifier for each case, (b) its location, and (c) episode day (i.e., symptom onset day or its closest day).
- case details—such as (a) gender, (b) age, and (c) specific occupation of the cases.
- symptom-related data—such as a Boolean indicator to indicate whether the case is asymptomatic or not. If not (i.e., symptomatic case), additional information is captured, which include:
- onset day of symptoms, and
- a collection of symptoms (including cough, fever, chills, sore throat, runny nose, shortness of breath, nausea, headache, weakness, pain, irritability, diarrhea, and other symptoms).
- clinical course and outcomes—such as:
- hospital status (e.g., hospitalized in the intensive care unit (ICU), non- ICU hospitalized, not hospitalized), and
- a Boolean indicator to indicate whether the patients recovered from the disease or not. If so (i.e., recovered case), additional information (e.g., recovery day) is captured.

B. Data Pre-Processing

After collecting and integrating data from heterogeneous sources, we pre-process the collected and integrated data. Recall from Section I that big COVID-19 epidemiological data can be characterized by their veracity. Specifically, we observe that there are some missing, unstated or unknown information (i.e., NULL values). Given the nature of these COVID-19 cases (e.g., for timely reporting of cases, privacy-preservation of the identity of cases), it is not unusual to have NULL values because values may not be available or recorded. For some other attributes related to case details (e.g., personal information like gender, age), patients may prefer not to report it due the privacy concerns. As there are many cases with NULL values for some attributes, ignoring them may lead to inaccurate or incomplete analysis of the data. Instead, our solution keeps all these cases for big data.

IV. RESULT

A. Real-Life COVID-19 Data Collection

1) Data Collection, Integration and Preprocessing

To evaluate and demonstrate the usefulness of our big data solution, we tested it with different COVID-19 epidemiological data including the Indian cases from Statistics Indian [68, 69]. With this dataset, data have been collected and integrated from State and Union territories public health authorities by the Primary Health Care of India (PHC). We pre-process data and generalize some attributes to obtain a dataset with the following attributes:

1. A unique privacy-preserving identifier for each

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-12, Issue-4 (August 2022)
https://doi.org/10.31033/ijemr.12.4.19

case
2. A generalized region/location
3. Episode week (or onset week of symptoms).
4. Gender (cf. sex at birth, which consists of male and female), including (a) male, (b) female, (c) others including unstated gender and non-binary gender (e.g., lesbian, gay, bisexual, transgender, queer/questioning, two-spirited (LGBTQ2+)).
5. Age group: ≤ 19, 20s, 30s, 40s, 50s, 60s, 70s,and ≥ 80s.
6. Occupation group, including: a) healthcare worker, b) school or day care worker (or attendee), c) long-term care resident, and d) other occupation.
7. Asymptomatic: Yes and No.
8. Set of 13 symptoms, including cough, fever,chills, sore throat, runny nose, shortness of breath, nausea, headache, weakness, pain, irritability, diarrhea, and other symptoms.
9. Hospital status, including: a) hospitalized in the ICU, b) hospitalized but not in the ICU, and c) not hospitalized.
10. Transmission method, including: a) community exposures, and b) travel exposures.
11. Clinical outcome: Recovered and death.
12. Recovery week.

As of September 01, 2022, the dataset has captured 62748 COVID-19 cases in India. Among them, 190,108 cases with stated episode week. Moreover, although the first Indian case occurred in Week 3, there were not more than 1000 new daily cases for following few weeks.

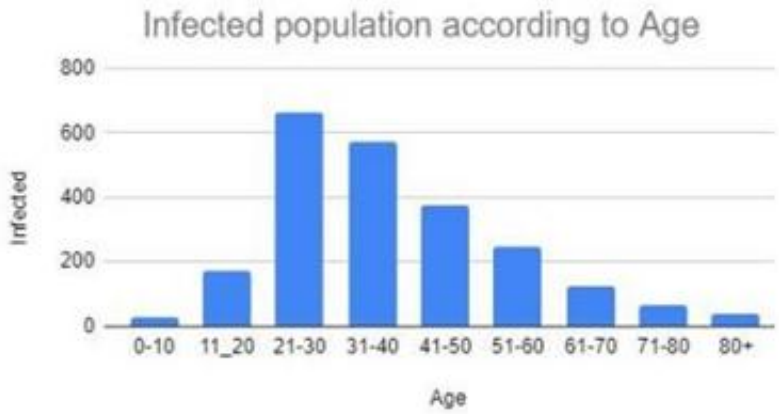

Figure 1: Bar Diagram showing the variation of infectedcases with Age

2) Big data on Cases

Once the data are pre-processed, our big data solution first analyses and mines the national data. With 201,341 COVID-19 cases with stated gender and age (out of an estimated Indian population of 1.4 Billion), the solution reveals that about 0.53% of the population contacted the disease.

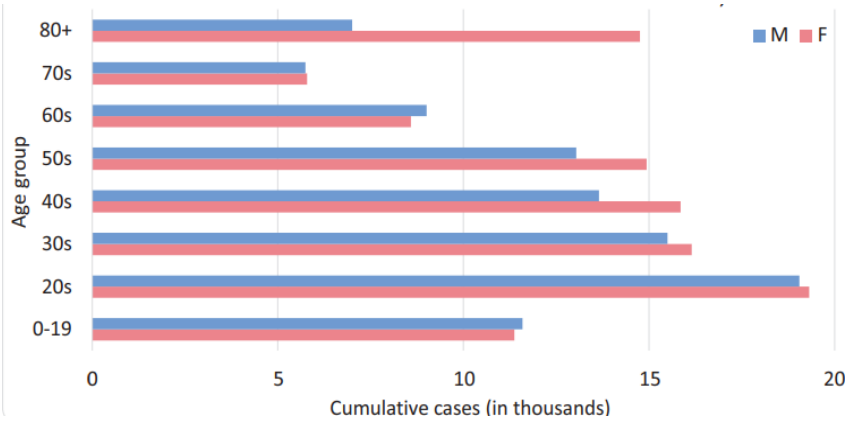

Figure 2: Distribution of cumulative COVID-19 cases

Then, our big data solution analyses and mines all 16 gender, age group-combinations. The resulting distribution of COVID-19 cases is shown in Fig. 2 a. The bar chart reveals that (a) despite being the most populated age groups, youth of 0-19 does not have the highest number of cases. Instead, (b) youth of 20shave the highest number. In contrast, (c) seniors in their 70s have the lowest number of cases. Moreover, (d) female in their 80s have more cases than their male counterparts.

3) Big data on Hospital Status

In addition to examining the cumulative cases, our solution also examines the hospital status among the 16 combinations. Table II reveals that, (a) as the ageincreases, the absolute number of hospitalized cases also increases. When combined with Table I, we observe that (b) despite the number of cases decreases from age groups 20s to 70s, the number of hospitalization increases. This means that, when young people catch COVID-19, a majority of them do not need to be hospitalized. When people age, their chance of requiring hospitalization once they catch COVID-19 increases. (c) Between the two genders, more male in their 30+ are admitted into the ICU than female.

Table 1: Cumulative Number Of Hospitalization As Of November 12, 2020

	MALE		*FEMALE*		*Age Group*
	ICU admitted	*Non-ICU hospitalized*	*ICU admitted*	*Non-ICU hospitalized*	*Total hospitalized*
0-19s	*11*	*79*	*11*	*95*	*196*
20s	*38*	*147*	*54*	*193*	*432*
30s	*74*	*288*	*62*	*292*	*716*
40s	*159*	*438*	*99*	*372*	*1068*
50s	*421*	*777*	*211*	*557*	*1966*
60s	*548*	*957*	*269*	*690*	*2464*
70s	*489*	*1150*	*268*	*1042*	*2949*
80s+	*204*	*1660*	*194*	*2346*	*4404*
Total	*1944*	*5496*	*1168*	*5587*	*14195*

4) Functionality Check with Related Works

After demonstrating the features and usefulness of our big data solution in analysing real-life COVID-19 data, let us evaluate its functionality when compared with related works. First, most of therelated works are observed to report mostly the numbers of cases and deaths. They do not provide privacy preserving details and epidemiological characteristics of those COVID-19 cases, which are provided by our solution. Second, our solution also provides details for each gender, age group-combination, which are unavailable in the related works.

V. CONCLUSION AND FUTURE WORK

In this paper, the effectiveness of big data analytics and its usefulness in latest Covid-19 is proposed with its application. It helps to reduce replication of data and ignoring null values the solution provides users with flexibility of including or excluding these values. This paper provides generalise solution by using some simple data and attributes (eg: age into age group). It also provides users with flexibility to express their preference (e.g., "must include symptoms") in mining of frequent patterns. It discovers frequent patterns from each of the 16 (gender, age group)-combinations. Moreover, it compares and contrasts the discovered frequent patterns among these combinations. As ongoing and future work, we transfer knowledge learned from thecurrent work to data science on other big data in many real-life applications and services.

REFERENCES

[1] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu & Yuan Wei, et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, china: a descriptive study. *The Lancet, 395*(10223), 507–513.

[2] Abdel-Basst, Mohamed, R. & Elhoseny, M. (2020). A model for the effective covid-19 identification in uncertainty environment using primary symptoms and ct scans. *Heath Inform. J.*, 1–18.

[3] Stojanovic, R., Skraba, A. & Lutovac, B. (2020). A headset like wearable device to track covid-19 symptoms. In: *Proceedings of the 2020 9th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro*.

[4] Jeong, H., Rogers, J.A. & Xu, S. (2020). Continuous on-body sensing for the covid-19 pandemic: Gaps and opportunities. *Sci. Adv*.

[5] A.K. Chanda, et al. (2017). A new framework for mining weighted periodic patterns in time series databases. *ESWA, 79*, 207-224. .
[6] C.K. Leung, et al. (2014). Fast algorithms for frequent itemset mining from uncertain data. In: *IEEE ICDM*, pp. 893-898.
[6] Timmers, T. Janssen, L. Stohr, J., Murk, J.L. & Berrevoets, M. (2020). Using eHealth to support covid-19 education, self-assessment, and symptom monitoring in the Netherlands: Observational Study. *JMIR mHealth, 8*, e19822.
[7] Epstein, R.H. & Dexter, F. (2020). A predictive model for patient census and ventilator requirements at individual hospitals during the coronavirus disease 2019 (COVID-19) pandemic: A preliminary technical report. *Cureus, 12*, e8501.
[8] https://covid19.who.int/region/searo/country/in.
[9] https://www.mohfw.gov.in/.
[10] Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H. & Saadi, M. (2017). Big data security and privacy in healthcare: A review. *Procedia Comput. Sci*.
[11] Schmidt, B.-M., Colvin, C.J., Hohlfeld, A. & Leon, N. (2020). Definitions, components and processes of data harmonisation in healthcare: A scoping review. *BMC Med. Inform. Decis. Mak*.