Correlation of COVID-19 Vaccination and Confirmed Cases Using Scalable Big Data

Jae Hoon Lee, Timothy Rochester, Riker Santivong, Wesam Farjo, Justin Licea, Jongwook Woo
Department of Information System, California State University Los Angeles
Los Angeles, California, United States of America
[e-mail: {jlee464, troches, rsantiv, wjargee2, jlicea7, jwoo5}@calstatela.edu]
*Corresponding author: Jongwook Woo

Abstract

The paper conducts an analysis of United States COVID-19 sentiment and the correlation of COVID-19 cases and vaccination rates, using a Big Data platform that is linearly scalable when the data size of the number of cases, the vaccination, and the tweets increase. Sentiment analysis was conducted using Twitter's data and shows that the overall sentiment in the United States about COVID-19 has remained largely negative over the past year. In addition, the analysis reveals that the number of new cases decreases in countries as vaccinations increase.

Keywords: Hadoop, HiveQL, COVID-19, vaccine, sentiment, Tweeter

1. Introduction

COVID-19 has become the most severe global health emergency in recent history with more than 147 million confirmed cases and over three million deaths worldwide. As a result, there is no shortage of data about the pandemic. In addition, the advent of Twitter allows researchers to gather information on the sentiments of users to various issues. This paper analyzes the data of new COVID-19 cases and vaccination rates in United States (US), Israel, and United Kingdom (UK) and compares the data with the tweet sentiment in the US over time.

2. Related Work

Boon-Itt et al gauges the public's sentiment of COVID-19 by building an application programming interface (API) to collect tweet data between December 13 and March 9, 2020, using Python and RStudio [1]. We use Hive in Hadoop with the data from January 28, 2020, to January 1, 2021, which allow large scale data to store and analyze. Shradha et al built time series prediction of the confirmed and fatality cases using the legacy and Big Data methods. Our paper is not for predictive analysis [5].

Monika et al collect COVID-19 data set to illustrate the confirmed and fatality cases using the legacy BI tool and forecast the three days cases [7]. The CDC's COVID Data Tracker contains maps and charts that track cases, deaths, and other trends of COVID-19 in the US [2]. Using that data, our paper looks at the relationship between the rate of vaccinations and new COVID-19 cases per million in each country.

Rossman et al studied the effects of the Pfizer vaccine rollout in Israel and sought to determine the effectiveness of the vaccine deployment. The study results show a decline in hospitalizations due to COVID-19 three to four weeks after the start of the national vaccination campaign [6]. Our study explores the temporal changes that occur after vaccination and additionally expands the previous study to examine the changes in multiple countries.

3. Background

The large-scale data set was stored using HDFS and the data was analyzed using HiveQL. Sentiment analysis was performed using Twitter data. The data engineering for this was to clean the Twitter data and build an Excel 3D map for tempo-spatial visualization.

4. Experimental Result

The datasets include global coronavirus cases, global vaccination, and Twitter data of USA. The datasets were in CSV file format with a total 5.58 GB. The dataset is collected and analyzed using Hadoop 3.1.4 Cluster illustrated in Table 1. Hadoop cluster is linearly scalable as the data set increases by adding additional servers. It can be extended for data prediction using Spark Machine Learning services.

Version	Hadoop 3.1.4
Total Nodes	3 (x 16 cores)
Total Node Memory Size	192 GB
RAID	24TB/18TB (RAID)
PySpark	Python 2.7.5, Spark 2.3.2.3.1.4.0-315

Table 1. Specifications of SCU Hadoop Server

The workflow architecture is composed of Data Collection and Transformation, Analysis, and Visualization as shown in **Fig. 1.**

5. Data Analysis

Raw data was uploaded and stored in HDFS and then loaded into tables with HiveQL. A sample size table was created to test codes and then the codes were used to analyze each dataset. The tables downloaded into CVS files then used Excel 3D Map and Tableau for visualization.

5.1 Sentiment Analysis from Twitter

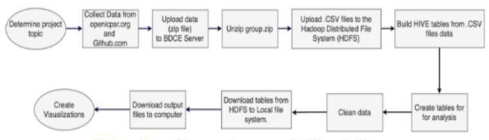

Fig. 1. - Experimental Workflow.

Twitter is a popular social media application with 192 million users and is therefore a rich source of data to conduct sentiment analysis to determine the public's general opinion on different topics. This is commonly done by assigning a value to terms that have been classified as "negative," "neutral," or "positive" and then giving a value to each classification. An average of each classification is then calculated to determine the public's general sentiment.

We filtered tweets from the US to analyze the overall sentiment of users on COVID-19. using Hadoop cluster, and collected the data into HDFS and developed Hive tables and QL codes. For example, the Hive table 'tweetid_sentiment' stores the tweets from USA with the format in **Fig. 2.**

country	date	sentiment_category	cnt
United States	1/28/2020	negative	2255
United States	1/28/2020	neutral	485
United States	1/28/2020	positive	487
United States	1/28/2020	very negative	150
United States	1/28/2020	very positive	11

Fig. 2. tweetid_sentiment

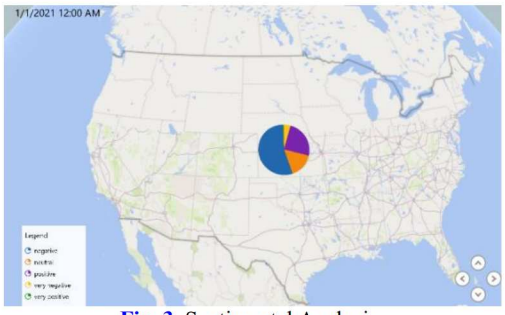

Fig. 3. Sentimental Analysis

Fig. 3 illustrates the overall sentiment in the US from 1/28/2020 – 1/1/2021. It shows the 5 sentiment categories below:

- 54% Negative
- 6% Neutral
- 25% Positive
- 4% very negative
- 1% very positive

5.2 Vaccination & New Cases

Vaccination started in USA and Europe in early 2021. Thus, we can find out the co-relationship between the vaccinations and the new cases of COVID-19 [4]. We measure the new cases per

million and the number of fully vaccinated people per hundred between January 1 and April 23, 2021 in USA, UK, and Israel.

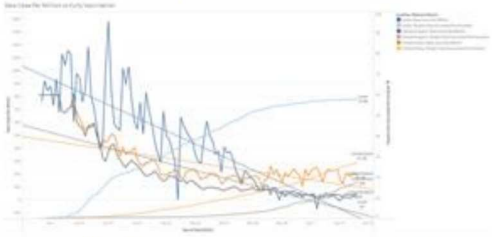

Fig. 4. Trends in Israel, UK, and the US

As shown in **Fig. 4**, we observe that the number of new cases decreases as vaccinations increases in those three countries.

5.3 Correlation in California (CA), US

Delta variant was first detected in India in October 2020, and then in the US in May 2021.

We find out the co-relationship between the new cases, new fatality cases, and vaccination in CA between December 15, 2020, and August 24, 2021.

Data from California Health And Human Services Open Data Portal [2, 3, 4, 8] was used to compute the new cases, new fatality cases, and full-vaccinations per 1k population as shown in **Fig. 5.**

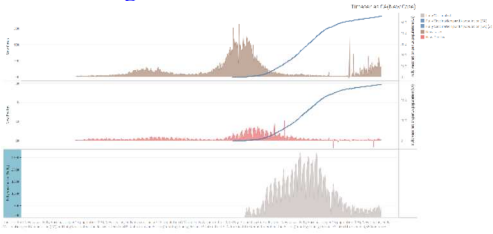

Fig. 5. The relationship of the new cases, the new fatality cases, and full-vaccinations in California

Conclusions

The analysis resulted determine the following conclusions:

1. The majority of Twitter sentiments for COVID-19 are negative.
2. Before the Delta variant, when the vaccination rate increased, COVID-19 new cases decreased.
3. Before the Delta variant, higher full-vaccination rates tended to result in fewer new cases in CA.
4. Since July 2021, there has been an increase in cases in CA due to the Delta variant however, as vaccination rates rise, the number of new deaths has declined.

In general, the more fully vaccinated the population, the lower the fatality rate due to COVID-19 even the Delta variants dominate.

References

[1] Boon-Itt, S., & Skunkan, Y. "Public perception of the covid-19 pandemic on twitter: Sentiment analysis and topic modeling study," *JMIRPublic Health and Surveillance, 6*(4). doi:10.2196/21978, 2020

[2] CDC COVID Data Tracker, 2021 Retrieved from https://covid.cdc.gov/covid-data-tracker/#/datatracker-home

[3] COVID-19 World Case Dataset, 2021. Retrieved from https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv

[4] COVID-19 World Vaccination Dataset, 2021. Retrieved from https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinations/vaccinations.csv

[5] Shradha Shinde, Jay Joshi, Sowmya Mareedu, Yeon Pyo Kim, Jongwook Woo, "Scalable Predictive Time-Series Analysis of COVID-19: Cases and Fatalities," arXiv:2104.11349, 2021

[6] Rossman, H., Shilo, S., Meir, T., Gorfine, M.,Shalit, U., & Segal, E. COVID-19 dynamics after a National Immunization program in Israel. *Nature Medicine*. doi:10.1038/s41591-021-01337-2 , 2021

[7] Monika Mishra, Dalya Manatova, Jongwook Woo, "COVID-19 Data Statistics and Forecasting," KAUPA News Letter, vol. 20, no. 2, pp. 8-9, March 32 2020

[8] Statewide COVID-19 Vaccines Administered By County, 2021. Retrieved from https://data.chhs.ca.gov/dataset/vaccine-progress-dashboar