Analyzing the Relationship Between COVID-19 and Human Behavior

Scott Chu, Weijian Feng, Junru He, Jingyi Jiang
{zc2396,wf2099,jh7948,jz4725}@nyu.edu
New York University
NY, USA

ABSTRACT

Final Project of Realtime and Big Data Analytic, Spring 2023

KEYWORDS

Covid-19, Hadoop HDFS, MapReduce, Hive

1 INTRODUCTION

The COVID-19 pandemic has had a significant impact on human behavior, including changes in social interactions, mobility patterns, and healthcare-seeking behaviors. In this project proposal, we aim to analyze the relationship between COVID-19 and human behavior using multiple datasets. We will analyze four datasets, including Twitter dataset, COVID-19 Cases, Google search trends, and Google Map Mobility Reports, to explore how the pandemic has affected people's behavior. By understanding how COVID-19 has affected people's behavior, we can better prepare for future pandemics and improve public health policies.

2 METHODS

In this section, we first present the overall framework of our system and how the big data tools are incorporated into our system. Then we will also show our data source and describe the datasets.

2.1 Overall framework

We utilized MapReduce to clean and profile our data. We then used Hive and Presto to join different datasets based on a common key (i.e. Date). Finally, we utilized Tableau and Python to perform analysis and generate visualizations using Seaborn and Matplotlib.

2.2 Data source

Dataset 1: COVID-19 Case Surveillance Public Use Data
Description: This dataset[2] contains 96.6 million rows of case surveillance cases from 01/01/2020 to 03/03/2023 over the US. Each record has 12 features, including earliest date, report date, positive specimen date, symptom onset date, current status, sex, age group, race and ethnicity, hospitalization, ICU admission status, death status, presence of underlying comorbidity or disease. We will use date related features: earliest date, positive specimen date and symptom onset date to analyze average duration from onset and laboratory confirmation. Analyzing the distribution of cases uses age, sex, race and ethnicity features.
Size of data: 10.8GB

Dataset 2: COVID-19 Twitter dataset.
Description: It is a dataset[1] of tweets acquired from the Twitter Stream related to COVID-19 chatter. It started from March 11th 2020 yielding over 4 million tweets a day. The data collected from the stream captures all languages, but the higher prevalence are: English, Spanish, and French. For NLP tasks the dataset provide the top 1000 frequent terms, the top 1000 bigrams, and the top 1000 trigrams.
Size of data: 5GB.

Dataset 3: COVID-19 Symptoms Search Trends
Description: The dataset[4] is about the search trends of symptoms in Google starting from 2020 to 2022 in different countries. It includes date,country code, and normalized search volumes of a wide range of symptoms. For search volumes, the range is from 0 to 100. The larger value means more search. Each row represents the daily search volume of different symptoms in a country. We can use this dataset and also the dataset about daily cases to analyze how case changes can impact people's behavior on google search. We will focus on some common symptoms (both physical and mental), such as anxiety, depression, dry eye syndromes, fever, gastroparesis, fatigue, insomnia, neck pain,skin condition. Then analyzing the statistics of these features in different time periods and regions.
Size of data: 1.99 GB

Dataset 4: Covid Community Mobility Reports by Google
Description: The Google Community Mobility Reports for COVID-19 provide insights into how people's movements have changed in response to the pandemic. These reports use aggregated, anonymous data from Google Maps to show how people are moving around in different regions and countries. [3]
The reports track changes in movement patterns in six different categories: retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and residential areas. By comparing current data to a baseline period, typically the period from January 3 to February 6, 2020, the reports show how much people are staying at home, avoiding public places, and using public transportation, among other things.
Size of data: Regional (1.05 GB)

3 DATA HANDLING

In this section, we discuss how did we do the experiments. Our programs are run on NYU HPC. The codes are able to run in a Linux system with Hadoop. The specific configuration can be found at our repository.

3.1 Profile and Filter

Covid-19 Twitter:
The Twitter dataset is composed of four distinct parts, each providing a unique perspective on the tweets gathered.

The first part is the tweets record. Comprising 5 columns, the Twitter dataset offers a comprehensive view of each tweet. The first column is the unique Twitter ID, which serves as a reference to the corresponding tweet using Twitter Dev Tools. However, the approval for access to this tool has not yet been granted. The second and third columns indicate the date and time of the tweet, respectively. The fourth column denotes the language used in the tweet, while the last column displays the country code. We filter the country code = US for future analysts. And then calculate tweets number per day.

The second part is the single, bi, and tri-gram data which are statistical results derived from Twitter's raw data. These datasets provide a count of how often words appeared on a given day, and are organized by the corresponding n-gram (single, bi, or tri-gram). The process of creating these datasets involved splitting the words used in each tweet and tallying the frequency of each word's appearance. Then we calculate the top 100 popular words during these years.

Covid-19 Case Surveillance:

The case dataset contains 12 features. We want to determine the number of cases per day to trace the trend of infection over time. There are four features about the case date, which are the earliest date, report date, positive specimen date, and symptom onset date. We select the earliest date as it is the earlier clinical date or date received by CDC. Additionally, we aim to investigate the association between age, sex, race and ethnicity with death status. So we retain these columns and remove hospital status, medical condition from the analysis.

To implement the filter, we used a map-only job to extract the earliest date, age, sex, race and ethnicity, and death status and drop the rest features. Then, we extracted the earliest date as key and set the value as 1 in another mapper. In the corresponding reducer, we counted the case number on each date.

Google Search Trend:

The dataset is composed of date, location, and over 400 symptoms. It is about the symptom search trends all over the world. Since our analysis focuses on the data in the US, we first extract all the data about search trends in the US by checking if location_key includes US in the mapper. Also, we pick 9 common symptoms for the analysis, and in the mapper, we extract the columns about anxiety, depression, dry eye syndromes, fever, gastroparesis, fatigue, insomnia, neck pain,skin condition. Then in the reducer, we take the average value of data in different areas in the US in one day. Also, since some data is missing, so in the mapper phase, we ignore all the rows that have missing values.

Google Mobility Report:

The Mobility Report contains data from different countries, not all of them provides full data. So We firstly did a sampling on 1000 data and found over 70 percents were partial valid. So we cannot just neglect them but extract the useful elements in them.

We use mappers to extract the country name and year-month as keys, then filter the valid data which contains at least one following column's data: Retail and recreation, Grocery and pharmacy, Parks, Transit stations, Workplaces, and Residential. Then in the reducer, group the same country's monthly data together, then calculate the average change percentage of each data partition, which is the monthly change rate of each country. We have modified the Java code for the mapper and the reducer to handle missing values in the row. If a value is missing, it will be ignored for that specific column.

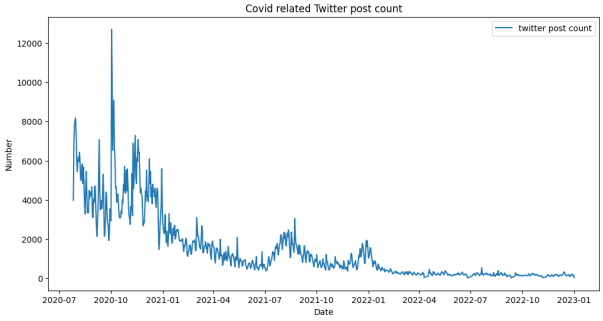
3.2 Process and Analyze

Covid-19 Twitter:

We only have Twitter data from July 2020 on wards, as data prior to that was not tagged with the country label.

As can be seen from the figure 1, the popularity of Covid 19 related tweets has been declining since January 2021, despite a few spikes caused by outbreaks during that period. Covid 19 is gradually becoming less discussed on Twitter. Therefore, the analysis that follows will focus on data from July 2020 to January 2022.

Figure 1: Tweets count



Then from Figure 2, we see a strong consistency between the number of Covid-related tweets and the number of confirmed cases of Covid. So we calculate their correlation coefficient and find that it is up to 0.46, which means strongly positive relative.

We discovered that the peak in the number of tweets typically occurs before the peak in COVID cases. To determine the maximum correlation, we experimented with various lag times. The highest correlation coefficient of 0.55 was achieved when the tweet number lagged by seven days. Another interesting discovery is Here's a hypothesis: After people become infected with COVID and develop symptoms, they first search for related information on platforms such as Twitter. They only get tested or seek medical help when their condition becomes more serious. The spread of information on the internet is faster than the spread of COVID. When one person is infected, those around them also search for information related to COVID, and the speed of information dissemination is faster than that of COVID transmission. However, existing data cannot prove these two hypotheses.

We analyzed the most popular grams in COVID-related tweets and found that people tend to post such tweets after being infected. The content of these tweets is not just about personal health but also addresses social topics such as lockdown and government policies. Our analysis suggests that COVID is gradually becoming a social topic.

Covid-19 Case Surveillance:

To accurately analyze the correlation between COVID-19 and variables such as sex, age, and race, relying solely on changes in daily case numbers is insufficient. The reason for this is that the number of individuals belonging to each group within the population varies.

Figure 2: Covid confirmed case and Tweets number

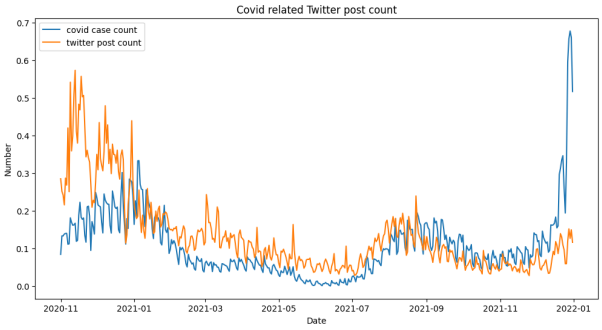
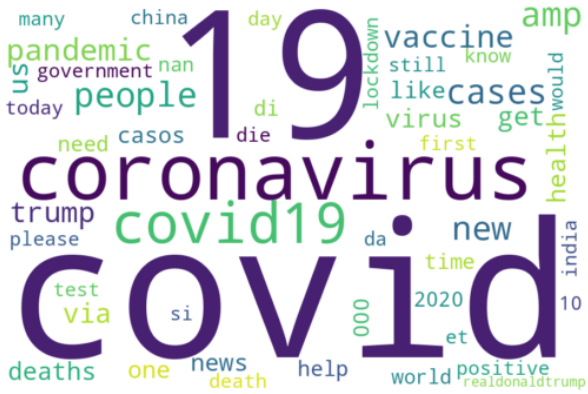


Figure 3: Covid related tweets world cloud



Thus, a more appropriate method for analysis is to use probability-based measurements such as infection rate, mortality rate, and infection fatality rate. To determine the infection and mortality rates, it is necessary to obtain population data for the United States, which is not currently included in our analysis. However, we were able to calculate the infection fatality rate(IFR) for different demographic groups based on data from this pandemic. The infection fatality rate is calculated using the following formula:

$$IFR\,in\,\% = \frac{Number\,of\,deaths\,from\,disease}{Number\,of\,infected\,individuals} \quad (1)$$

Figure 4: Covid Cases Number and Death Number

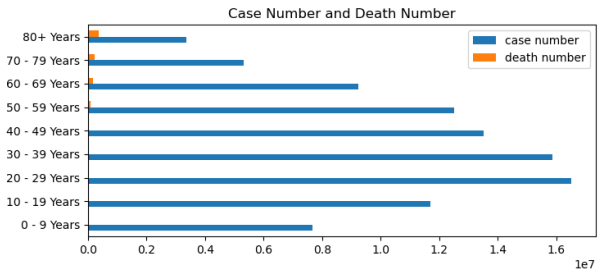
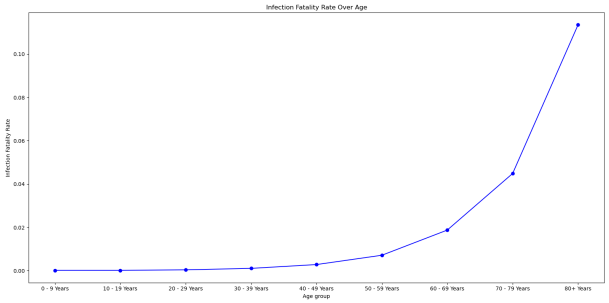


Figure 5: Infection Fatality Rate Over Age

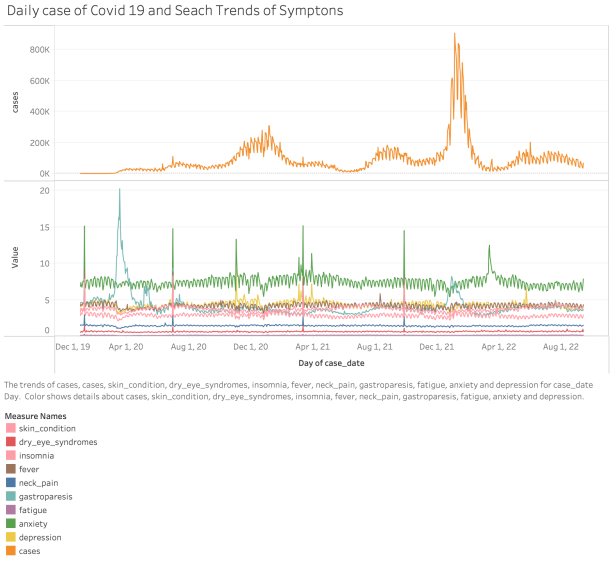


Based on the data presented in Figure 4 and Figure 5, it is evident that the infection fatality rate rises as age increases. This highlights the fact that older individuals face a greater risk of dying from the virus and should take additional measures to protect themselves during the pandemic. However, the relationship between sex, race, and death is not clear from the given information and requires further investigation.

Google Search Trend:

We use Hive to join the google search trends dataset with the daily cases in the US to discover the relationship between covid-19 and people's concern about the symptoms. I used Tableau to draw the plots. From the figure 6, we can see that the search trends of symptoms change with the cases. For example, at time 12/2020, there is an increase in cases, and we can see an obvious increase in anxiety and depression. Also, at around early 2022, which is the outbreak of omicron, we can see an increase in the search trends of gastroparesis.

Figure 6: Covid Cases and Symptom Search Trends



Comparing the correlation efficiently between each symptom and cases, we can see that insomnia has the highest correlation

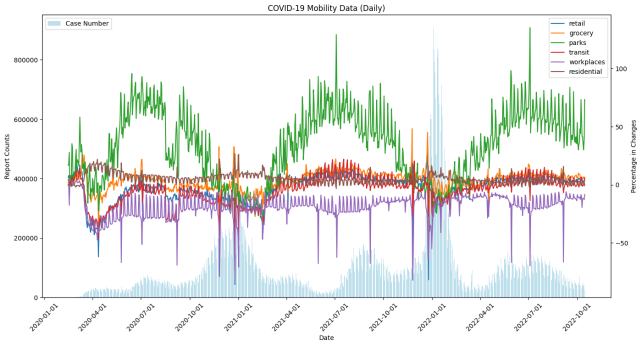
coefficient (0.22), and anxiety, depression, and skin condition has negative correlation coefficients.

Since before 04/2022, covid-19 has not been a major contagious disease in the US, I chose a shorter time period to focus more on the time when covid has an impact on people's lives. Therefore, I chose 04/2021 to 12/2022, which is the outbreak of covid-19. During this time period, the correlation between cases and symptoms are more obvious, we can see that insomnia: 0.34, anxiety: 0.18, depression: 0.21. This tells us that the search volume of mental issues are more correlated with covid-19 cases.

Google Mobility Report:

For the mobility reports, we want to find its relationship with the Covid cases changes. We got the daily count of the covid cases from the Covid-19 Case Surveillance profiled dataset. And we can join this two tables to find more info. We use Hbase to handle those two CSV files since they are both fit in SQL database. After loading them into tables of Hbase, We used the inner-join sql command to create a new table by using date as the shared keys. Then we export those data and draw them in a graph: (x axis – date, y left axis – covid cases, y right axis – percentage changes, column – covid cases, line – mobility of locations)

Figure 7: COVID-19 Mobility Data (Daily)



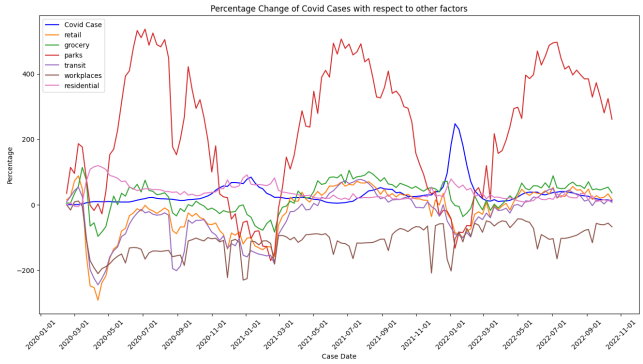
Before processing it, we had a overview from this graph that people intended to stay at home when the Covid waves come through. We want to verify this hypothesis, so we have to get the correlation of the mobility change rates and covid change rates. The mobility data are the increase rate of the daily location visited counts by the first week's counts(which means pre-covid level). So we also calculates the increase rate of covid cases compared to the first week's covid data. To normalize those data, we want to use weekly rate instead of daily rate to see the trends.

And then we got this graph: (Blue line–Covid cases)

From this graph, we can have a basic view of the mobility weekly changes relating to the covid cases weekly changes. But it's not trivial to see the magnitude of the relation. So we also calculate the correlation and conduct several p-value tests.

Based on the corrected results of the correlation analysis and hypothesis test, there is a statistically significant relationship between the weekly percentage change in covid cases and several mobility indicators. The strongest correlation is observed with parks, which have a negative correlation of -0.41 and a p-value of 0.0001(very significant), indicating that as the number of COVID-19 reports

Figure 8: Normalized Percentage Changes of Mobility Data vs Covid (Weekly)



Location	Correlation	P-value
Parks	-0.41	0.0001
Retails	-0.17	0.0102
Grocery and Pharmacy	-0.17	0.0392
Transit stations	-0.18	0.0334
Work places	-0.13	0.1350
Residential	0.22	0.0102

Table 1: Ratio of parallel region in the whole program

increases, the percentage change in park visits decreases. Similarly, retail and recreation, grocery and pharmacy, and transit stations have negative correlations respectively, with p-values smaller than 0.05 suggesting that as COVID-19 reports increase, the percentage change in visits to these locations also decreases. The workplace indicator has a weak negative correlation of -0.13 with a p-value of 0.1350, which is not statistically significant(bigger than 0.05). This indicates that there may not be a strong relationship between the number of COVID-19 reports and the percentage change in workplace visits. Finally, the residential indicator has a positive correlation, with a p-value statistically significant, suggesting that as the number of COVID-19 reports increases, the percentage change in residential visits also increases. Overall, the results suggest that as COVID-19 reports increase, people tend to visit public spaces less frequently, and they are more likely to stay at home.

4 RESULTS

In this section we present the results obtained from the data handling. We will analyze the measurements obtained with the profiling tools and discuss the bottlenecks reflected by them.

Covid-19 Twitter:

The correlation between COVID-related tweets and confirmed cases was found to be strongly positive, with a correlation coefficient of 0.46. The peak in the number of tweets was found to occur before the peak in COVID cases, and the maximum correlation coefficient of 0.55 was achieved when the tweet number lagged by seven days. Our analysis also suggests that COVID is gradually becoming a social topic, with tweets addressing social issues such as lockdown and government policies.

Based on our analysis, we suggest that monitoring Twitter-related tweet data can provide an early warning of the outbreak of epidemic diseases such as Covid and other large-scale pandemics. Furthermore, analyzing the content of tweets can also provide insight into the social evolution of the disease, such as its transformation from a physical illness to a social topic. This approach can provide useful information for public health professionals and decision-makers to develop more timely and effective prevention and response strategies.

Covid-19 Case Surveillance:

COVID-19 infection fatality rate rises as age increases. Adults aged 80 and over have a COVID-19 death rate 2.5 times higher than those aged 70-79, and 6 times higher than those aged 60-69. It is crucial to prioritize the care of elderly COVID-19 patients due to the significant differences in symptoms and prognosis compared to other age groups[5]. Strict prevention measures, early diagnosis, and aggressive care are vital to reducing the mortality of older adults during the epidemic.

Google Search Trend:

From the google symptom search trends, we can see that mental symptoms has higher positive correlation with the cases from 01/2020 to 12/2022, and this correlation is more obvious during the period of covid-19 outbreak. Therefore, we can see can covid-19 is not only a "coronavirus disease" that will impact physical health, it will also bring some mental issues to people because because covid-19 will also cause many social problems such as the higher employment. Also, this implies that many people have no ideal with how to manage mental health concerns.

From the report, we suggest that people can pay more attention to their mental health. Also, some public institution, such as schools, can have some mental health services to help people maintain a good mental health.

Google Mobility Report:

For the mobility report, we find that people began working from home since the beginning of the Covid no matter how the pandemic changed. But other places, especially parks,are very sensitive to the case wave. The People stayed at home immediately in a covid outbreak and rushed outdoors when the wave was over. This result is according to the expectation except for the workplace, which means people keep work from home even after the covid waves.

5 CONCLUSION

From this study, we have found the following trend in this Covid-19 pandemic. The death rate due to Covid-19 is closely linked to age. Individuals who contract the virus often engage in online discussions and searches about it on social media platforms. Then we can see that Covid-19 might brings some other social problems, and the outbreak has heightened mental health concerns, leading to a greater prevalence of related symptoms. With regard to mobility patterns, people have consistently worked from home since the pandemic's onset, regardless of its progression. However, visits to public spaces, particularly parks, have decreased as the number of Covid-19 reports has risen.

REFERENCES

[1] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalin, and Gerardo Chowell. 2023. *A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.* https://doi.org/10.5281/zenodo.3723939 This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates. Release: We have standardized the name of the resource to match our pre-print manuscript and to not have to update it every week..

[2] CDC. 2023. *COVID-19 Case Surveillance Public Use Data.*

[3] Google. 2022. *Google Covid-19 Community Mobility Report.* Technical Report. Google Map.

[4] Google. 2022. *Google COVID-19 Search Trends symptoms dataset.*

[5] Dadras Omid and et al. 2022. COVID-19 mortality and its predictors in the elderly: A systematic review. *Health science reports* (2022).