Guolin Wang. Relationship between Google Trend and Daily Trend of New COVID-19 cases in the United States. A Master's Paper for the M.S. in I.S degree. April, 2021. 47 pages. Advisor: Jaime Arguello

This study is a statistical analysis examining the relationship between Google Trend and Daily Trend of the number of new COVID-19 cases in the United States using single term versus selecting 5 terms per day and taking time-lag into account. Google Trend was used to measure the interest of users in search terms over time during the pandemic of COVID-19. To make up for the lack of open datasets of search queries, tweets filtered by "COVID-19" related hashtags were used to generate top-5 terms per day according to term frequency on a 7-day moving average. Spearman correlation analysis was applied to examine the correlations between the two trends and examine the impact of time-lag. The results showed high correlation between the two trends by selecting 5 terms per day. Besides, taking time-lag into account improved the correlation between the difference of search interest and the number of new COVID-19 cases. The results indicate that it is important to account for the fact that covid-related language may change over time and time-lag should be considered as an important factor when conducting trend analysis. This study demonstrates the possibility of using Google Trend to predict future case trend of COVID-19 in areas with a large population of Web search users.

Headings:

Trend analysis

Time series analysis

Correlation (Statistics)

Text mining

RELATIONSHIP BETWEEN GOOGLE TREND AND DAILY TREND OF NEW COVID-19 CASES IN THE UNITED STATES

by

Guolin Wang

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2021

Approved by

Jaime Arguello

Table of Contents

Introduction

Corona Virus Disease 2019 (COVID-19), a new strain of coronavirus against which no previous immunity exists and shows a pattern of human-to-human transmission, has evolved from an isolated diseases in a region of China to a global pandemic that has brought countries to a standstill, pushed hospital systems to the brink, and dragged the global economy into a recession (“A Timeline of COVID-19 Developments in 2020 | AJMC,” n.d.). On September 30, 2020, the Centers for Disease Control (CDC) reported 7,213,419 cases in the United States with 2.8% fatality.

The median incubation period of COVID-19 is estimated to 5.1 days and 97.5% of those who develop symptoms will do so within 11.5 days (Lauer et al., 2020). In addition, there exists delay of CDC-reported data, which makes it difficult for traditional surveillance system to provide timely reflection of the pandemic of COVID-19. Internet searches has been reported to correlate with traditional surveillance data and can even predict the outbreak of disease epidemics several days or weeks earlier (Li et al., 2020). This research is a statistical analysis designed to measure, understand, and assess the statistical relationship between Google Trend and Daily Trend of new COVID-19 cases and take time-lag into account. Google Trend was used to measure the search interest of users or relative search volumes (RSV) on terms over time during the pandemic of COVID-19. Google does not provide open dataset of search queries. Therefore, this research leveraged tweets filtered by “COVID-19” related hashtags to select top-5 terms per day based on term frequency on a 7-day moving average. This alternative method is formed on a hypothesis that the terms used in social media posts and search queries related to COVID-19 are same or similar. Daily Trend of new COVID-19 cases was obtained from CDC COVID Data Tracker. Spearman correlation analysis and time series analysis were applied to examine the correlations between the two trends. The anticipated impacts of this study were to examine the correlation and find stable time-lag between Google Trend and Daily Trend of new COVID-19 cases, so that provides help to internet-based surveillance systems for COVID-19.

2 Literature Review

This section presents a review of the spread of COVID-19 as well as prior research and studies about pandemics of infectious diseases. It starts with an introduction to the timeline of COVID-19 development in the United States from January to July. Secondly, it summarizes prior research on using social media data to predict the prevalence of specific infectious diseases. Finally, it gives a brief review of current trend analysis studies on COVID-19 and prior research applying trend analysis to other infectious diseases.

2.1 Timeline of COVID-19 Development in the United States

The objective of the research is to examine the relationship between Google Trend and Daily Trend of new laboratory confirmed cases of COVID -19 in the United States. It is important to develop an accurate timeline of COVID-19 in the United States for the convenience of research design and implementation. On January 9, the World Health Organization (WHO) announced that the spate of pneumonia-like cases in Wuhan could have stemmed from a new coronavirus ("A Timeline of COVID-19 Developments in 2020 | AJMC," n.d.). At this point, there were 59 laboratory confirmed cases globally and none of them were found in the United States. On January 20, WHO confirmed human-to-human transmission and reported the first case of COVID-19 in the United States ("Timeline of the COVID-19 pandemic in the United States - Wikipedia," n.d.). On January 31, WHO issued Global Health Emergency and another case of a person return from Wuhan was confirmed in California, which marked the seventh known cases in the U.S (“CDC officials confirm 7th US case of coronavirus, in California man,” n.d.).

The CDC COVID data tracker recorded first case on January 22 (“CDC COVID Data Tracker,” n.d.). However, according to report from *The Washington Post* (“While CDC coronavirus tests stalled for six weeks, a German lab made 1.4 million tests - The Washington Post,” n.d.), CDC revised its fault test for COVID-19 on February 28, resulting in missing or inaccurate data prior to February 29, while another Global COVID-19 tracker developed by 1Point3Acres (“Global COVID-19 Tracker & Interactive Charts | Real Time Updates & Digestable Information for Everyone | 1Point3Acres,” n.d.), whose data have been referenced by CDC and JHU COVID-19 World Map, recorded the first case on March 13. Due to different report and data collection mechanisms, the daily cases and peaks varies between different COVID-19 cases trackers. Since CDC COVID data tracker and 1Point3Acre Global COVID-19 tracker apply 7-day moving average to compute daily new COVID-19 cases, this research mainly depend on CDC COVID data tracker and use 1Point3Acres Global COVID-19 tracker as complement to divide the stages of the pandemic on a weekly basis. In the following section, the CDC COVID data track was denoted as CDC tracker and 1Point3Acre Global COVID-19 tracker was denoted as 1Point3Acre Tracker.

Based on the CDC tracker shown in Figure 1, this research divided Daily Trend in number of new COVID-19 cases into three stages and the time rage of each stage was shown in table 1. The first stage runs from 02/09/2020 to 04/12/2020 which is the date of first peak. The second stage runs from 04/13/2020 to 06/14/2020. The turning point is

June 13 by CDC tracker and June 14 by 1Point3Acres tracker. In order to divide the stages into weeks, this research chose June 14 as the end date. The third stage runs from 06/15/2020 to 07/26/2020 which is two days after the highest peak from 02/09/2020 to 10/01/2020.

Table 1: Stages of the pandemic of COVID-19 in the United States

Stages	Start Date	End Date	Span (Week)
1	2/29/2020	4/12/2020	6
2	4/13/2020	6/14/2020	10
3	6/15/2020	7/26/2020	6

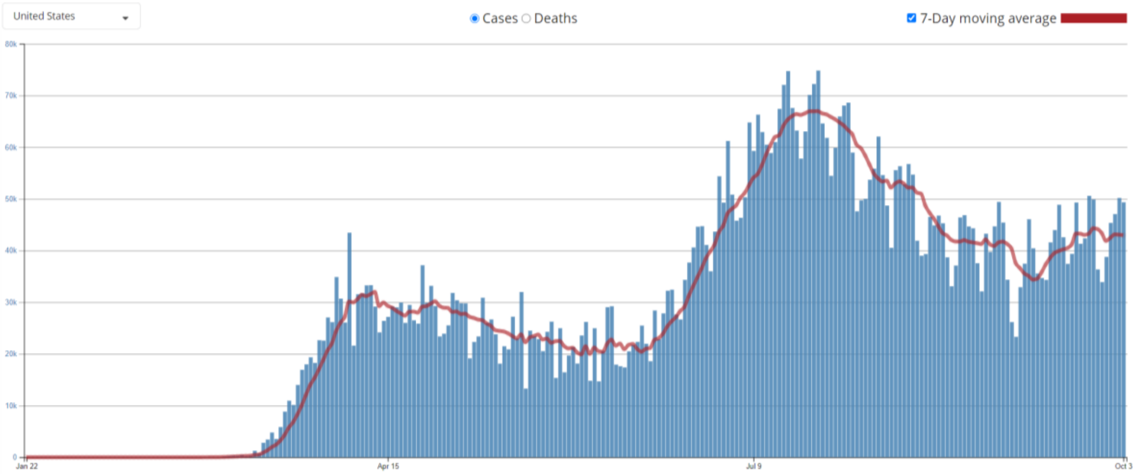


Figure 1: Daily Trend in number of new laboratory-confirmed COVID-19 cases in the United States reported to CDC

2.2 Mining on Social Media about Infectious Diseases Prediction

This research raised two hypotheses: 1) The terms used in social media posts and search queries related to COVID-19 are same or similar; 2) There exists a stable time-lag between Google Trend and Daily Trend of new COVID-19 cases. This research was not going to prove the hypothesis 1. Instead, this subsection uses previous literature to provide evidence to support the hypothesis. This section also contributes to the hypothesis 2 by quoting the research examining time-lag between tweets term frequency and Daily Trend of infectious diseases.

In the following months after the pandemic of COVID-19 began, lots of research that contribute to developing predictive models of COVID-19 have been published. One of the most important methods is mining on social media posts such as Twitter and Weibo.

Social media posts can be used for real-time content analysis, sentiment analysis and knowledge translation research, allowing health authorities to analyze users' perspectives and reactions to infectious diseases and then respond to public concerns. In the early stage of pandemic of COVID-19, (Gao, Yada, Wakamiya, & Aramaki, 2020) built a multilingual COVID-19 Dataset consisting of tweets in English and Japanese and microblogs of Weibo in Chinese. It suggested that the trend of social media posts is related to the outbreak of pandemic and changed according to precautionary measures. In addition, it presented possible utilizations of this dataset to develop sentiment analysis. (Abd-Alrazaq, Alhuwail, Househ, Hamdi, & Shah, 2020) conducted a study that leveraged Twitter data to identify four main themes consisting of 12 main topics related to COVID-19. In combine with sentiment analysis, this study illustrated the public attitudes toward each topic, which indicated health authorities may build real-time surveillance system and allocate resources to mitigate public concern by monitoring social media posts. Another word frequency and sentiment analysis conducted by (Rajput, Grover, & Rathi, 2020) with expanded time period from the end of January 2020 to April 2020 showed a similar result of public sentiments. This study did not differentiate sentiments by topics or themes but by tweets from WHO or from general public. Its results showed that about 85% percent tweets related to COVID-19 are positive while 15% are negative, which shared common characteristics to the sentiment analysis results towards the topics examined by (Abd-Alrazaq et al., 2020) with positive on ten topics and negative on the rest two topics.

The data sources and research on COVID-19 are relatively less than that of previous global pandemic such as Dengue and H1N1which give inspirations for further research on COVID-19. (Marques-Toledo et al., 2017) examined the association between Dengue cases and tweets, Dengue cases and Google Trend, as well as Dengue cases and Wikipedia access logs which served as a complement. This study used data from 2012 to 2016 (to 2015 for Wikipedia since 2016 is non-epidemic period) and the result indicated that tweets, Google Trend and Wikipedia data were associated with Dengue cases at country level. In this study, Google Trend containing the term “Dengue” in Brazil was obtained from 2012 to 2016 and aggregated per week. The result showed stronger linear association between Google Trend and Dengue cases ($r = 0.92$, $p < 0.001$) in comparison to that of tweets and Dengue cases ($r = 0.87$, $p < 0.001$). In particular, the results showed similar distribution for normalized Google Trends and Tweets Trend. This finding supports the hypothesis 1 though the gap still exists between terms used by tweets and search queries.

During the last global pandemic of H1N1, there were many research built predictive model based on tweets. For example, (Chew & Eysenbach, 2010) conducted a content analysis on tweets finding that several Twitter activity peaks coincided with major news stories and its results showed statistically significant correlation between Tweets Trend and H1N1 incidence data. Besides, it showed a gradual adoption of WHO recommended terminology, which indicated the terms frequently used by tweets related to COVID-19 may change over time. This is one of the reasons why dynamically selecting top-5 terms related to COVID-19 per day may generate inspiring results. (Signorini, Segre, & Polgreen, 2011) examined the use of tweets information to track rapidly evolving public sentiments in respect to H1N1 and developed a model to monitor influenza-related traffic within the United States. This research developed a fairly accurate estimate model to make real-time estimation of national and reginal weekly influenza-like illness (ILI) based on support-vector regression. Its estimation of disease activity in real time was 1-2 weeks faster than CDC-reported data which was in line with the fact that CDC-reported data has a common delay. This fact supports the hypothesis 2 that stable time-lag exists between Google Trend and Daily Trend of new COVID-19 cases on the basis that the hypothesis 1 holds true.

2.3 Trend Analysis about Infectious Diseases

There are plenty of research using Google Trend to predict infectious diseases' growth trends, mortalities, etc. This subsection has two main objectives: 1) summarizing trend analysis on COVID-19; 2) summarizing trend analysis applied to previous pandemic such as H1N1.

There are some studies used search terms related to COVID-19 to examine the relationship between Google Trend and Daily Trend of new COVID-19 cases. (Walker, Hopkins, & Surda, 2020) picked loss-of-smell-related searches. Its results showed strong correlation between Daily Trend of new confirmed cases, deaths, and relative search volume (RSV) which is as same as the Google search interest defined in this study. But when the loss-of-smell related searches were applied to H1N1, no such correlation was observed. Though the authors did not provide details about why they chose loss-of-smell related search terms, it might be the case that loss-of-smell is a typical symptom for COVID-19 but not for H1N1. According to BBC report on August 19 (“Coronavirus smell loss ‘different from cold and flu’ - BBC News,” n.d.), smell loss was much more profound in the COVID-19 patients and leaded investigator Prof Carl Philpott, from the University of East Anglia, stated that smell and taste tests could be used to discriminate between COVID-19 patients and people with a regular cold or flu.

(Lin, Liu, & Chiu, 2020) applied trend analysis to Google Trend of terms “wash hands” and “face mask” and the Daily Trend of new COVID-19 cases. They chose the two terms for the objective of examining whether Google searches of the two terms would protect from increased number of confirmed cases of COVID-19. Google Trend of “wash hand” showed negative association with COVID-19 cases while that of “face mask” had no significant correlation. The authors attributed the increase of search trend of “face mask” to the shortage of mask but needed further evidence to support it.

(Ayyoubzadeh, Ayyoubzadeh, Zahedi, Ahmadi, & R Niakan Kalhori, 2020) developed multiple linear regression and long short-term memory (LSTM) models to predicting the number of COVID-19 cases. The results showed that the most effective factors of the linear regression model besides the previous day incidence included the search interest of handwashing, hand sanitizer, and antiseptic topics. The predictions of the linear regression model were not precise enough and the LSTM model results indicated overfitting. In this study, it covered many terms such as “Coronavirus”, “COVID-19”, “Antiseptic selling”, “Handwashing”. However, the authors were not explicitly listing the reasons about why they chose such terms except for saying that the related concepts were suggested by one of the authors.

There are some related studies examined the relationship between Google Trend and other topics. For example, (Hong, Lawrence, Williams, & Mainous III, 2020) used Google Trend search interest of telehealth and telemedicine to describe population-level interest and examined its relationship with cumulative numbers of COVID-19 cases and Telehealth Capacity of US Hospitals in response to COVID-19 by Pearson correlation. With clear target, coming with specific terms related to telehealth and it found strong correlation between population-level interest and cumulative numbers of COVID-19 cases.

Another interesting type of research is applying time series analysis to describe the association of Google Trend and Daily Trend of infectious diseases. (Effenberger et al., 2020) conducted a research examining the relationship of the Google Trend of term “Coronavirus” and Daily Trend in number of new COVID-19 cases across the world. It found the time-lags between the peak of search interest and the peak of newly infected cases in Europe and U.S were 11.5 days, 7 days in Brazil and Australia. (Shin et al., 2016) examined the correlation between the Middle East respiratory syndrome coronavirus (MERS-CoV), Google Trend, and Twitter Trends using the term “MERS”, “MERS symptoms” and “MERS hospital” in Korea. The results showed the peak of new laboratory confirmed cases was 5 days later than the peak of Google Trend and Twitter Trend. And the overall graph patterns were similar, which further supports the hypothesis 1. The two research shared a common term selecting strategy by selecting terms directly related to the name of infectious diseases.

(Ginsberg et al., 2009) applied time-lag correlation analysis to detecting seasonal influenza epidemics. The authors illustrated the reason for choosing queries about health-seeking behavior. Traditional surveillance system used by CDC and European Influenza Surveillance Scheme rely on clinical data including ILI physician visits. Therefore, monitoring heath-seeking behavior in the form of queries to online search engine was helpful to improve early detection of disease activity. Besides, this research provided a query selection process with scoring and evaluation of how many top-scoring queries to be included in the ILI-related query fraction coming with maximal performance. Leveraging plenty of influenza data and Google Trend between 2003 and 2008, the results showed the RSV of certain queries was highly correlated with the percentage of physician visits, and the final model was able to accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting time-lag of about 1 day. (Wilson et al., 2009) conducted similar research in New Zealand for the ILI cases of H1N1 using existing Google Flu Trends (GFT) web services which includes millions of queries. However, the time-lag was unstable before and after the peaks of flu cases. Google shut down GFT on August 20, 2015 and there were lots of research and data showed the GFT was not working well when comparing its results to real world (Lazer, Kennedy, King, & Vespignani, 2014). GFT includes millions of queries, however, the evaluation of how many top-scoring queries to be included in the research of (Wilson et al., 2009) showed the best performance was obtained by summing the top 45 search queries. This comparison between these two research shows the importance of selecting suitable queries.

Based on literature review, some current research of trend analysis on COVID-19 selected terms without providing adequate evidence, or used terms directly related to the disease such as “Coronavirus” and “COVID-19”, and most of them using fixed terms within their study period. As the terms frequently used by tweets related to COVID-19 changed over time, this research dynamically picked top-5 terms related to “COVID-19” per day to conduct correlation and time series analysis to describe the association between Google Trend and Daily Trend of new COVID-19 cases. Besides, this research leveraged Twitter data, which was used to select candidate terms, to address the lack of open data source of search queries. Instead of using the scoring method in (Wilson et al., 2009), this research used term frequency of tweets related to COVID-19 as the term selection criteria.

3 Research Questions and Hypotheses

Based on the literature review, we can conclude that internet searches and social media posts are correlated with traditional surveillance data, which can be helpful to predict the future trends or outbreaks of infectious diseases several days or weeks earlier in the countries or areas with a large volume of internet users. There are some research applying trend analysis to the early stages of the pandemic of COVID-19, which used one to several pre-selected terms. However, covid-related language may change over time and the report process of laboratory testing data to CDC may cause time-lag which should be considered. The purpose of this study is to examine the relationship between Google Trend and Daily Trend of new COVID-19 cases in the United States. The independent variable Google Trend is measured by interest of terms over time, which is defined as numbers represent search interest relative to the highest point on the chart for the given region and time ("Google Trends," n.d.). The terms were selected from tweets with "COVID-19" related hashtag based on term frequency on a 7-day moving average. The dependent variable Daily Trend of new COVID-19 cases is defined as new laboratory confirmed COVID-19 cases per day on a 7-day moving average provided by CDC.

This research developed based on two hypotheses:

1) The terms used in social media posts and search queries related to COVID-19 are same or similar.

2) There exists a stable time-lag between Google Trend and Daily Trend of new COVID-19 cases.

On the basis that the two hypotheses hold true, this research was developed to answer two questions:

1) How important is it to account for the fact that covid-related language may change over time?

2) How important is it to account for a time-lag?

4 Methodology

In this section, I will provide a detailed description of data and research methods used by this study to examine the relationship between Google Trend and Daily Trend of new COVID-19 cases in the United States.

4.1 Data Description

The data used by this research consists of three parts: 1) tweets with "COVID-19" related hashtag posted between stage 1 to stage 3 defined by the Timeline, 2) Google Trend of candidate terms, 3) Daily Trend of the number of new COVID-19 cases in the United Sates.

4.1.1 Tweets

For privacy and security issues, Google does not provide open datasets of search queries for public research. In literature review, (Marques-Toledo et al., 2017) and (Shin et al., 2016) provided evidence supporting hypothesis 1, the terms used in social media posts and search queries related to COVID-19 are same or similar. On the basis that this hypothesis holds true, tweets filtered by "COVID-19" related hashtags posted within stage 1 to stage 3 served as the datasets for extracting candidate terms for Google Trend.

A repository containing an ongoing collection of tweets IDs associated with the novel coronavirus COVID-19, which commenced on January 28, 2020, was leveraged by this research ("COVID-19-TweetIDs/2020-01 at master · echen102/COVID-19-TweetIDs · GitHub," n.d.). There were roughly 1,000,000 tweets per day. Due to the rate limit of retrieving tweets through tweet id by Twitter API, this research randomly selected 10,000 tweets IDs per day.

The retrieved tweets underwent data preprocessing before generating candidate terms. Firstly, tweets in English were selected for the scope and convenience of the experiment because the retrieved tweets in the repository came from users across different countries using different languages. Secondly, all texts were converted to lowercase, the texts in square brackets, words containing numbers, links, punctuations, emoji, and stopwords were removed from the contents. Finally, the left contents were tokenized and used to generate top-5 candidate terms per day based on term frequency on a 7-day moving average. The form containing all candidate terms was shown in Appendix 1.

4.1.2 Google Trend

Google Trend is a website published by Google that analyzes the popularity of top search queries in Google Search across various regions and languages ("Google Trends - Wikipedia," n.d.). In the scope of this research, the Google Trend data in the language of English within the United States between February and July in 2020 were used.

4.1.3 Daily Trend of New COVID-19 Cases

CDC provides a tracker of COVID-19 cases in the United States. In case of drastic fluctuations caused by report delay or mechanism changes, 7-day moving average were applied to calculate the Daily Trend in number of new laboratory-confirmed COVID-19 cases per day in the United Sates.

4.2 Data Analysis Methods

4.2.1 Difference of Search Interest

The candidate terms of Google Trend were generated from tweets based on term frequency on a 7-day moving average. More specifically, this research picked 5 terms with highest term frequency per day, then calculating the difference of Search Interest (SI) of each term according to function (1). SI was the interest of a term overtime retrieved by Google Trend on a day. The baseline of a term was calculated by the average SI of the term from 12/1/2019 to 12/31/2019 which is roughly a month before the pandemic started.

$$D(t) = SI(t) - B(t) \quad (1)$$

D: Difference, SI: Search Interest, B: Baseline, t: term

To find out how important is it to account for the fact that covid-related language may change over time, the Difference of Search Interest (D_SI) of the term “Coronavirus” and the D_SI of the top-5 terms selected per day were calculated to describe the public interest over COVID-19 of each day. The sum of the D_SI of the top-5 terms after min-max normalization was used by spearman correlation analysis.

4.2.2 Spearman Correlation Analyses

After drawing the trend of D_SI of “Coronavirus” and that of the top-5 terms per day and Daily Trend of COVID-19 Cases, spearman correlation analysis was applied to examine the correlation between the trends. The Spearman's rank-order correlation is the nonparametric version of the Pearson product moment correlation. The Spearman correlation coefficient ρ can take value from +1 to -1, which measures the strength and direction of association between two ranked variables. A ρ of +1 indicates a perfect positive association of ranks, a ρ of 0 indicates no association between ranks, and a ρ of -1 indicates a perfect negative association of ranks. The closer ρ is to |1|, the stronger the association between the ranks. This research further conducted time series analyses to assess the temporal relationships ranging from 1 to 14 days before and after, which was in line with to the common delay of CDC-reported data.

5 Result and Discussion

This section consists of the results of the Daily Trend of new laboratory-confirmed COVID-19 cases, the trend of D_SI of the term “Coronavirus” and that of the top-5 terms per day selected based on term frequency, the results of spearman correlation analysis and the results of time series analysis.

5.1 Daily Trend of COVID-19 Cases

After min-max normalization of the new laboratory-confirmed COVID-19 cases from stage 1 to stage 3, which covered 2/29/2020 to 7/26/2020, the Daily Trend of new laboratory-confirmed cases was shown in Figure 2.

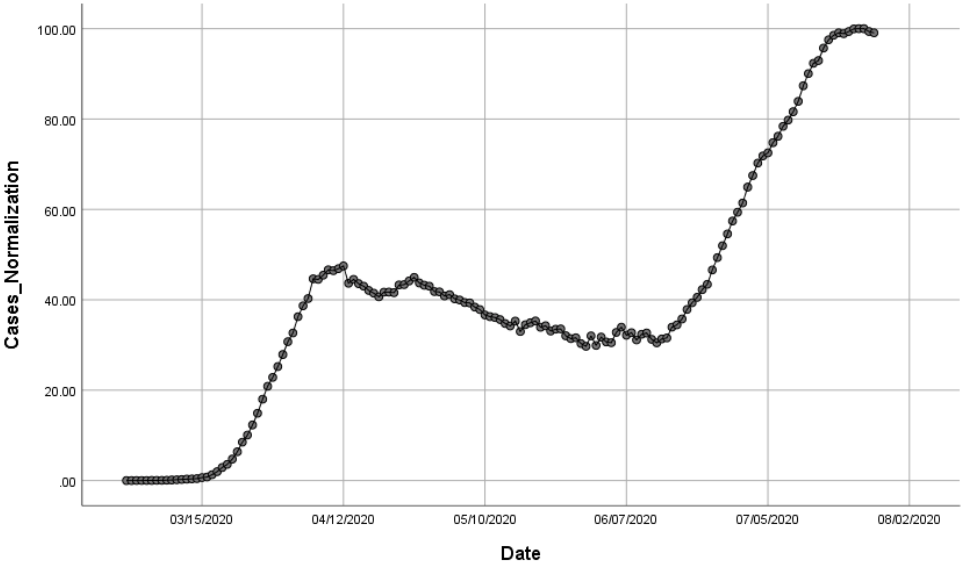


Figure 2: Daily Trend of normalized new laboratory-confirmed COVID-19 cases in the United States

5.2 Trends of D_SI of "Coronavirus" and Top-5 terms

D_SI stands for the Difference of Search Interest defined in section 4.3.1. Figure 3 and 4 showed the trend of the D_SI of the term "Coronavirus" and the top-5 terms per day selected based on term frequency from tweets after min-max normalization accordingly.

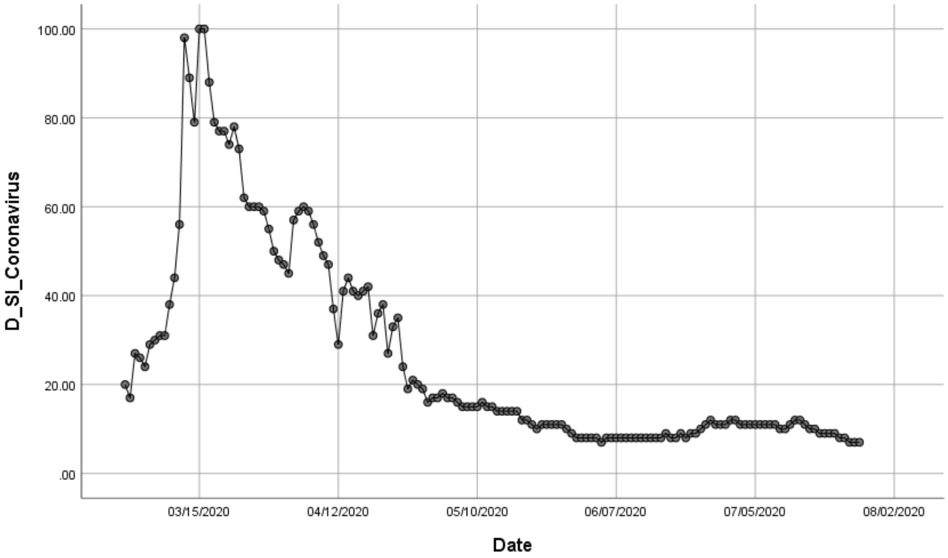


Figure 3: Trend of Difference of Search Interest of term "Coronavirus"

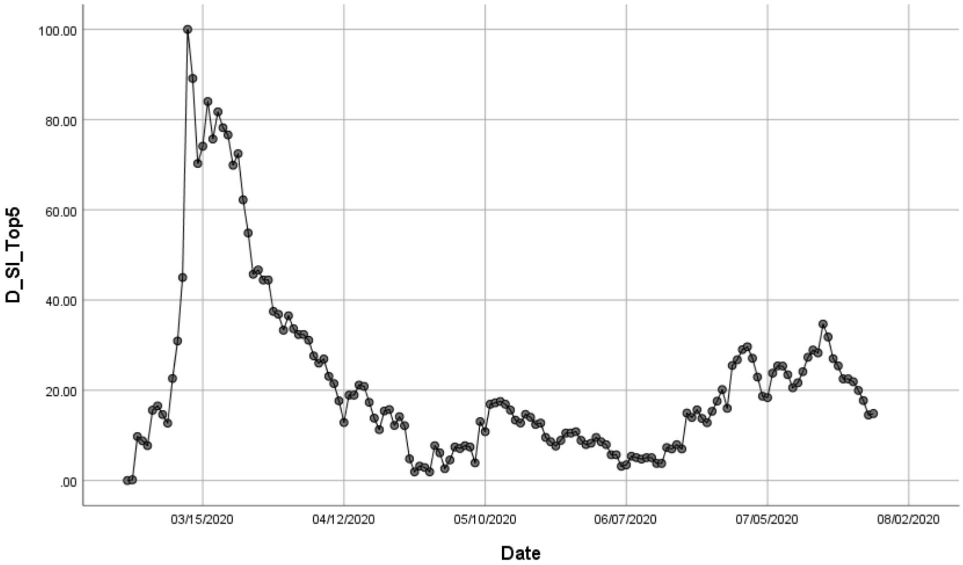


Figure 4: Trend of Difference of Search Interest of Top-5 terms selected from tweets based on Term Frequency

5.3 The results of Spearman Correlation Analysis

Spearman correlation analysis was applied to examining the correlation between D_SI of term “Coronavirus” and new laboratory-confirmed COVID-19 cases per day, D_SI of top-5 terms selected from tweets based on term frequency and new laboratory-confirmed COVID-19 cases per day.

5.3.1 Correlation Analysis Containing Stage 1 to Stage 3

The results of correlation analysis between D_SI of the term “Coronavirus” and new COVID-19 cases per day were shown in Figure 5 and Table 2. Figure 5 showed Dual Y Axes with Scale X Axis of Cases_Normalization in blue, Value of D_SI_Coronavirus in red by Date. The result of spearman correlation analysis corresponding to Figure 5 was shown in Table 2. The correlation coefficient of Cases_Normalization and D_SI of term “Coronavirus” was -0.349 (ρ(df) = -0.349, df = N-2=147) and it was significant at 0.01 level, which indicated that D_SI_Coronavirus tended to decrease when Cases_Normalization increased. In other words, there existed negative monotonic association between D_SI of term “Coronavirus” and new laboratory-confirmed COVID-19 cases per day considering all three stages.

Table 2: Spearman Correlation Analysis of Cases_Normalization and D_SI_Coronavirus

Correlations

			D_SI_Coronavirus
Spearman's rho	Cases_Normalization	Correlation Coefficient	-.349**
		Sig. (2-tailed)	0
		N	149

**. Correlation is significant at the 0.01 level (2-tailed).

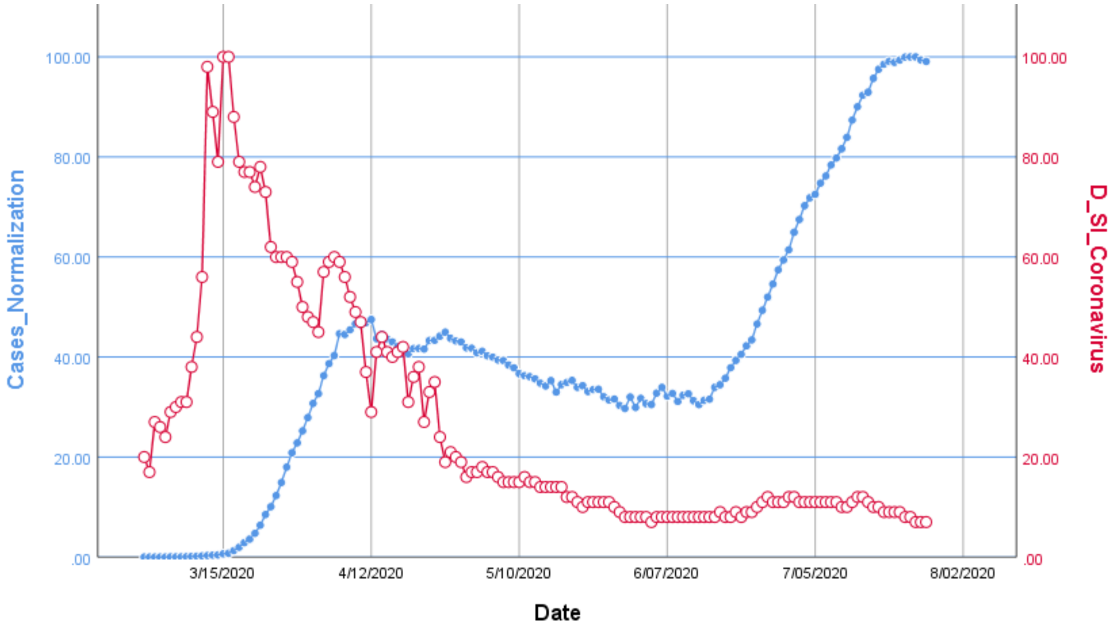


Figure 5: Dual Y Axes with Scale X Axis of Cases_Normalization, Value of D_SI_Coronavirus by Date

The results of correlation analysis between D_SI of Top-5 terms and new COVID-19 cases per day were shown in Figure 6 and Table 3. Figure 6 showed Dual Y Axes with Scale X Axis of Cases_Normalization in blue, Value of D_SI_Top5 in red by Date. The result of spearman correlation analysis corresponding to Figure 6 was shown in Table 3. The correlation coefficient of Cases_Normalization and D_SI_Top5 was 0.056 (ρ(df) = 0.056, df = N-2=147) and it was statistically insignificant, which indicated that there was no monotonic association found between D_SI of top-5 terms selected from tweets based on term frequency and new laboratory-confirmed COVID-19 cases per day considering all three stages.

Table 3: Spearman Correlation Analysis of Cases_Normalization and D_SI_Top5

Correlations

			D_SI_Top5
Spearman's rho	Cases_Normalization	Correlation Coefficient	0.056
		Sig. (2-tailed)	0.498
		N	149

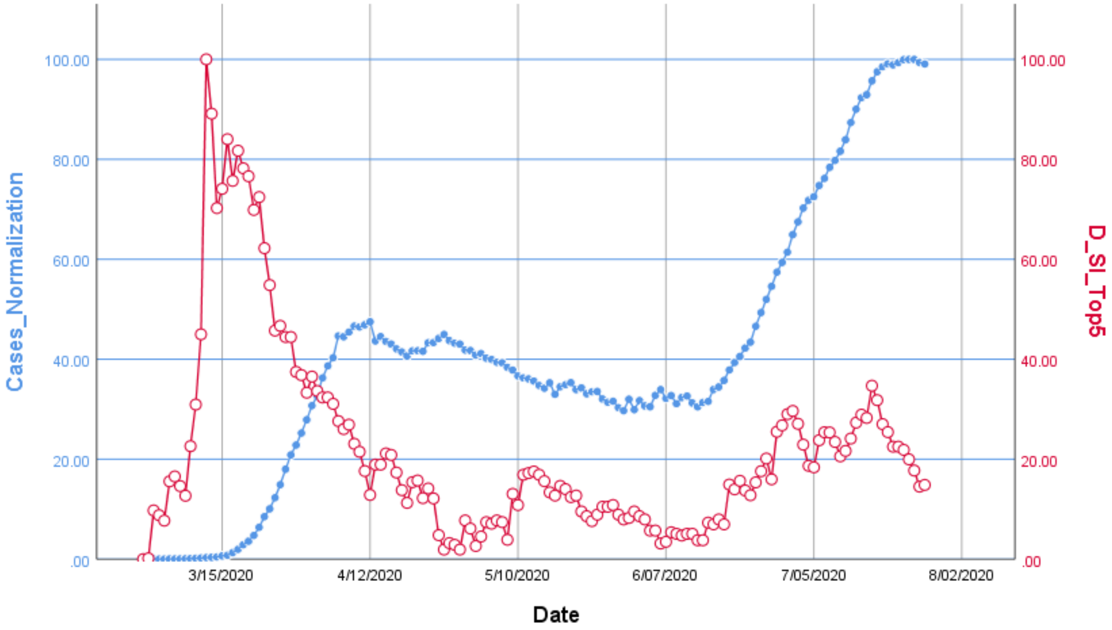


Figure 6: Dual Y Axes with Scale X Axis of Cases_Normalization, Value of D_SI_Top5 by Date

5.3.2 Correlation Analysis excluding Stage 1

By comparing table 2 and table 3, the correlation coefficient of Cases_Normalization and D_SI_Coronavirus was negative while that of Cases_Normalization and D_SI_Top5 was positive though not statistically significant. In combine with Figure 7 Dual Y Axes with Scale X Axis of D_SI_Coronavirus, Value of D_SI_Top5 by Date, the early stage of the two lines shared similar trend. Therefore, this research applied correlation analysis excluding the early stage, the stage 1, running from February 29 to April 12.

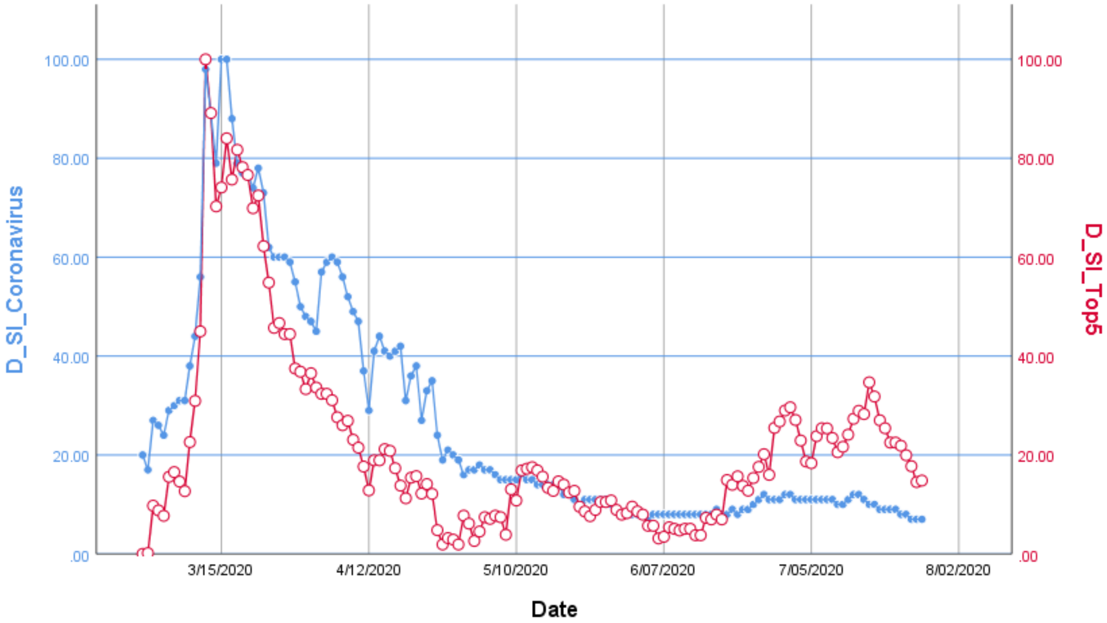


Figure 7: Dual Y Axes with Scale X Axis of D_SI_Coronavirus, Value of D_SI_Top5 by Date

The results of correlation analysis between D_SI of term "Coronavirus" and new COVID-19 cases per day excluding stage 1 were shown in Figure 8 and Table 4. Figure 8 showed Dual Y Axes with Scale X Axis of Cases_Normalization in blue, Value of D_SI_Coronavirus in red by Date excluding Stage 1. The result of spearman correlation analysis corresponding to Figure 8 was shown in Table 4. The correlation coefficient of Cases_Normalization and D_SI of term "Coronavirus" was 0.179 (ρ(df) = 0.179, df = N-2=103) and it was statistically insignificant, which indicated there was no monotonic association between D_SI of term "Coronavirus" and new laboratory-confirmed COVID-19 cases per day considering Stage 2 and Stage 3.

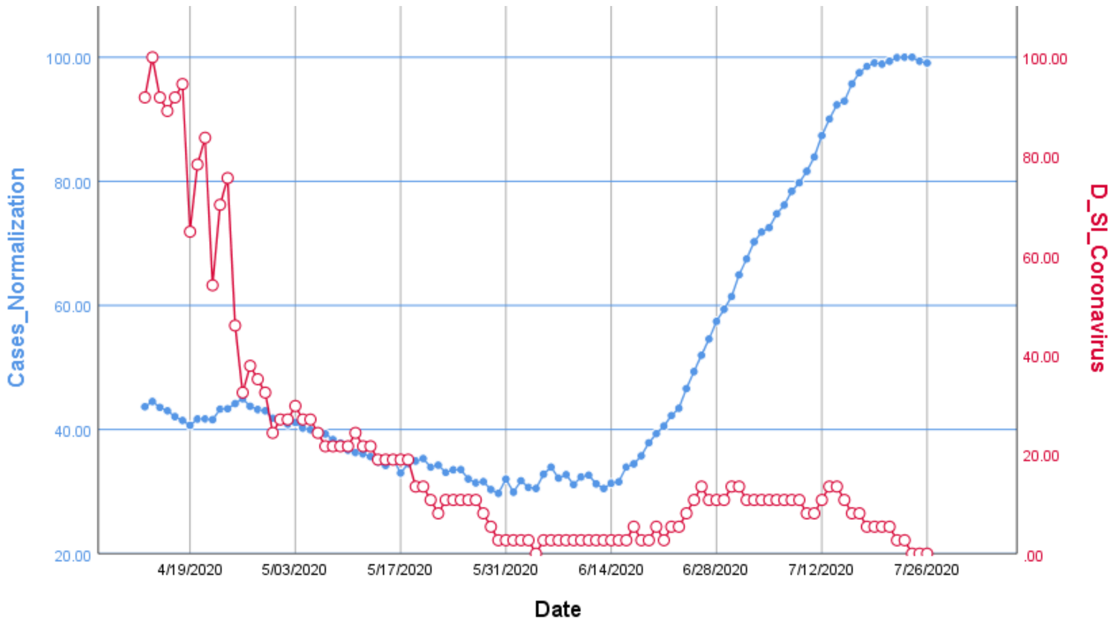


Figure 8: Dual Y Axes with Scale X Axis of Cases_Normalization, Value of D_SI_Coronavirus by Date excluding Stage 1

Table 4: Spearman Correlation Analysis of Cases_Normalization and D_SI_Coronavirus excluding Stage 1

Correlations

			Cases_Normalization
Spearman's rho	D_SI_Coronavirus	Correlation Coefficient	.179
		Sig. (2-tailed)	.068
		N	105

The results of correlation analysis between D_SI of Top-5 terms and new COVID-19 cases per day excluding Stage 1 were shown in Figure 9 and Table 5. Figure 9 showed Dual Y Axes with Scale X Axis of Cases_Normalization in blue, Value of D_SI_Top5 in red by Date excluding stage 1. The result of spearman correlation analysis corresponding to Figure 9 was shown in Table 5. The correlation coefficient of

Cases_Normalization and D_SI_Top5 was 0.702 (ρ(df) = 0.702, df = N-2=103) and it was significant at 0.01 level, which indicated that D_SI_Top5 tended to increase when Cases_Normalization increased. In other words, there existed positive monotonic association between D_SI of top-5 terms selected from tweets based on term frequency and new laboratory-confirmed COVID-19 cases per day considering stage 2 and stage 3.

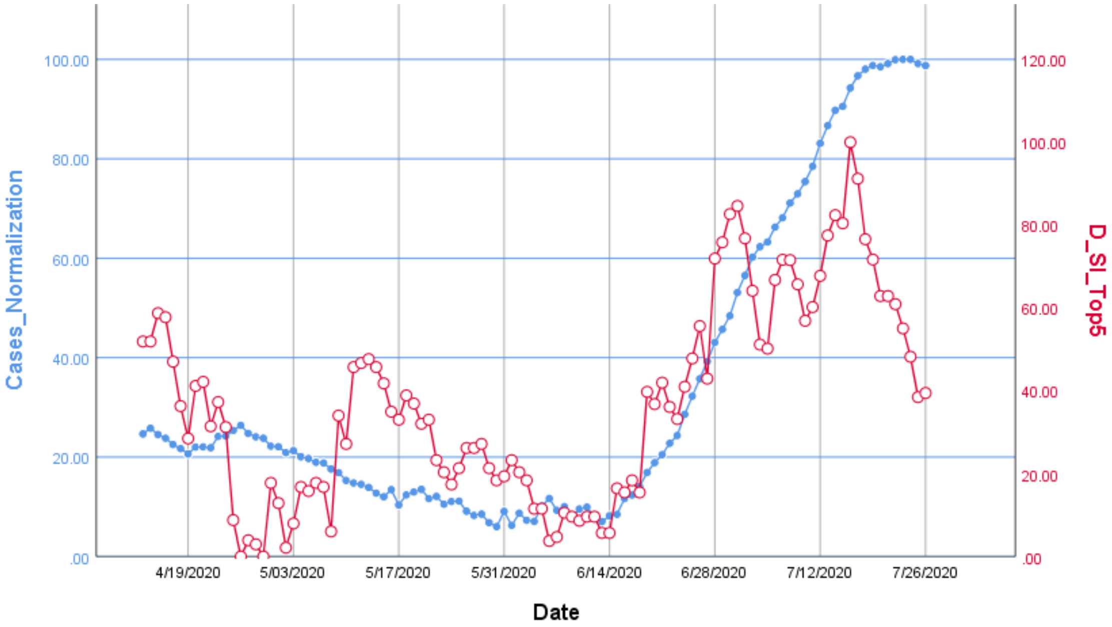
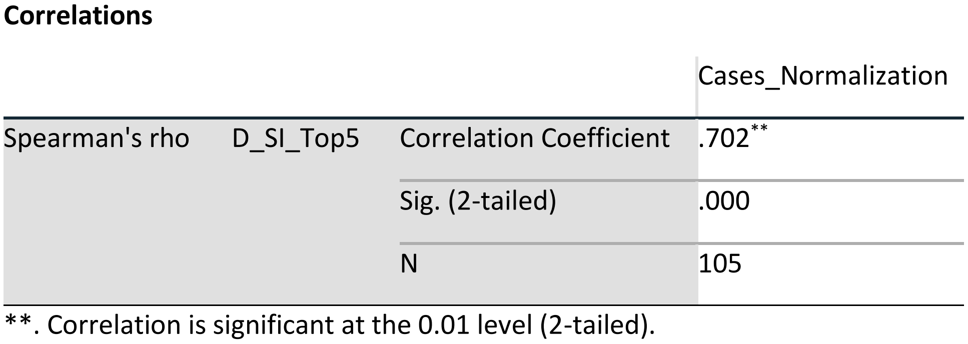


Figure 9: Dual Y Axes with Scale X Axis of Cases_Normalization, Value of D_SI_Top5 by Date excluding Stage 1

Table 5: Spearman Correlation Analysis of Cases_Normalization and D_SI_Top5 excluding Stage 1



Correlations

			Cases_Normalization
Spearman's rho	D_SI_Top5	Correlation Coefficient	.702**
		Sig. (2-tailed)	.000
		N	105

**. Correlation is significant at the 0.01 level (2-tailed).

5.4 The result of Time Series Analysis

The result of correlation analysis between D_SI of Top-5 terms and new COVID-19 cases per day excluding Stage 1 indicated statistically significant monotonic association and the corresponding ρ was 0.702 which indicated relatively strong association. Based on this correlation, this research applied spearman correlation analysis to examine how important is it to account for a time-lag between D_SI of Top-5 terms selected from tweets based on term frequency and Cases_Normalization ranging from -14 days to +14 days.

Table 6 and Table 7 showed the correlation coefficient of spearman correlation analysis of Cases_Normalization and D_SI_Top5 with time-lag from -14 to +14. “-N” means to move forward the Daily Trend of new laboratory-confirmed COVID-19 case of each day by N days, while “+N” means to move backward the Daily Trend of new laboratory-confirmed COVID-19 case of each day by N days.

According to Table 6, the ρ of D_SI_Top5 and Cases went up from “-1” to “-3” and went down after “-3” to “-14”. In contrast, Table 7 showed the ρ of D_SI_Top5 and Cases went down from “+1” to “+14”. The highest correlation (ρ = 0.711) was attained at “-3”. All the result in Table 6 and 7 are statistically significant at 0.01 level. Corresponding line charts are shown as Figure 10 and Figure 11.

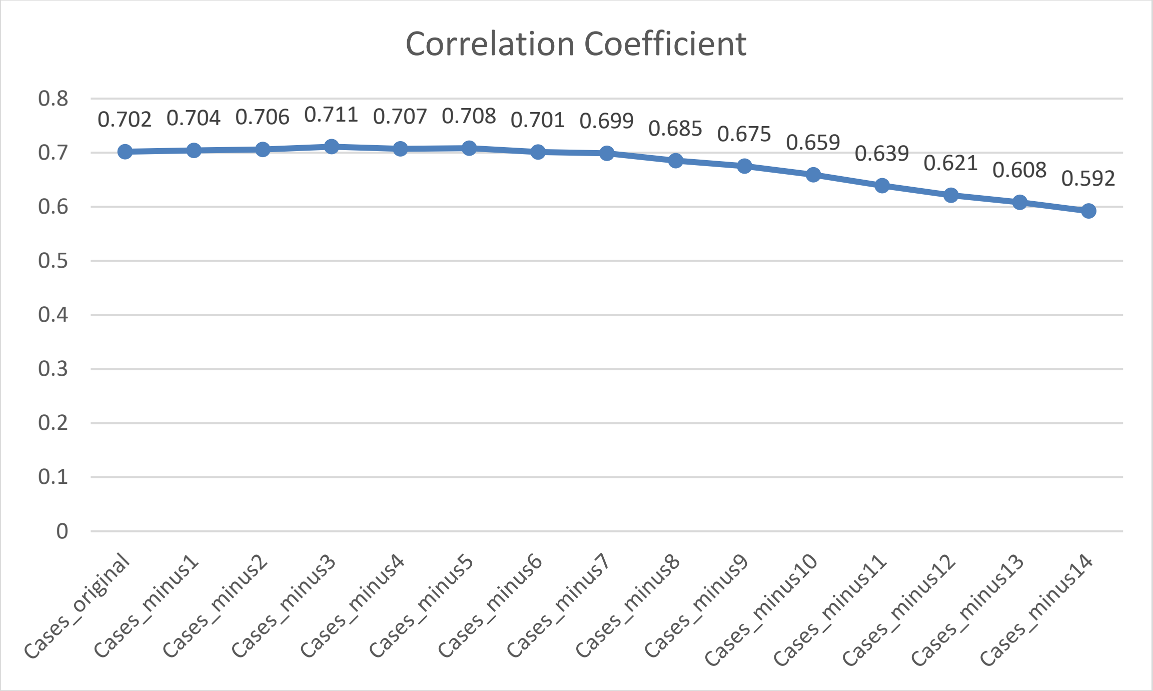


Figure 10: Correlation Coefficient by time-lag from “-1” to “-14”

Table 6: Spearman Correlation Analysis of Cases_Normalization and D_SI_Top5 with time-lag from -1 to -14

		Cases_original	Cases_minus1	Cases_minus2
D_SI_Top5	Correlation Coefficient	.702**	.704**	.706**
		Cases_minus3	Cases_minus4	Cases_minus5
D_SI_Top5	Correlation Coefficient	.711**	.707**	.708**
		Cases_minus6	Cases_minus7	Cases_minus8
D_SI_Top5	Correlation Coefficient	.701**	.699**	.685**
		Cases_minus9	Cases_minus10	Cases_minus11
D_SI_Top5	Correlation Coefficient	.675**	.659**	.639**
		Cases_minus12	Cases_minus13	Cases_minus14
D_SI_Top5	Correlation Coefficient	.621**	.608**	.592**

** Correlation is significant at the 0.01 level (2-tailed).

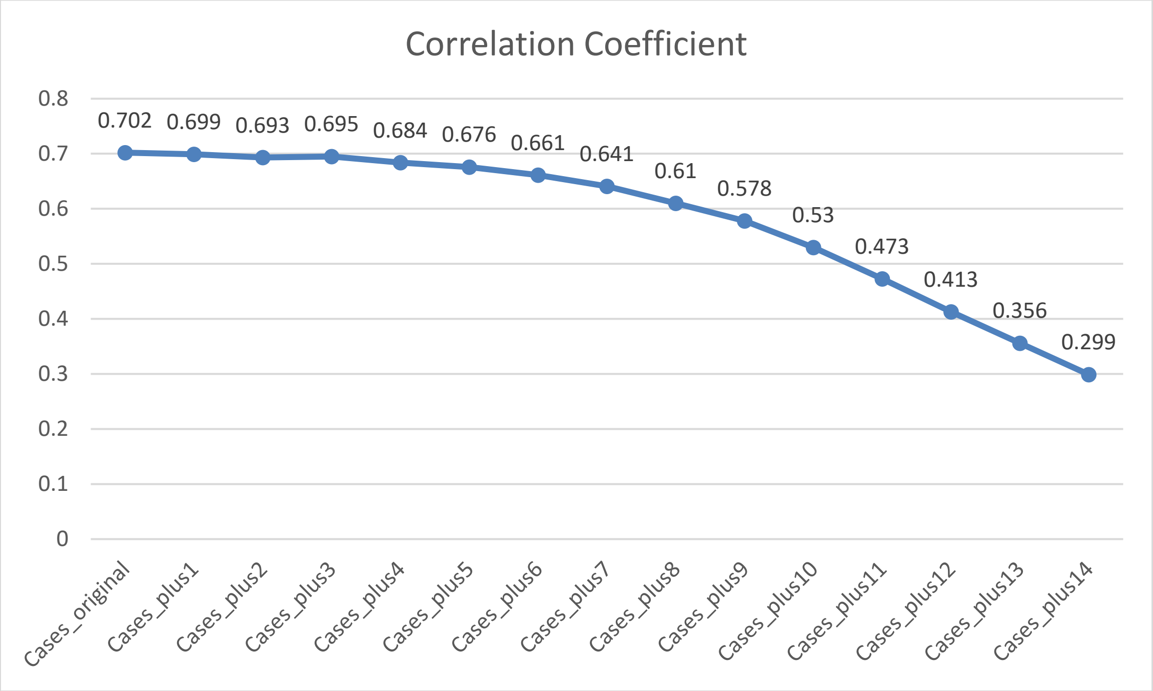


Figure 11: Correlation Coefficient by time-lag from “+1” to “+14”

Table 7: Spearman Correlation Analysis of Cases_Normalization and D_SI_Top5 with time-lag from +1 to +14

		Cases_original	Cases_plus1	Cases_plus2
D_SI_Top5	Correlation Coefficient	.702**	.699**	.693**
		Cases_plus3	Cases_plus4	Cases_plus5
D_SI_Top5	Correlation Coefficient	.695**	.684**	.676**
		Cases_plus6	Cases_plus7	Cases_plus8
D_SI_Top5	Correlation Coefficient	.661**	.641**	.610**
		Cases_plus9	Cases_plus10	Cases_plus11
D_SI_Top5	Correlation Coefficient	.578**	.530**	.473**
		Cases_plus12	Cases_plus13	Cases_plus14
D_SI_Top5	Correlation Coefficient	.413**	.356**	.299**

** Correlation is significant at the 0.01 level (2-tailed).

5.5 Discussion

The results of correlation analysis between D_SI of term “Coronavirus” and new COVID-19 cases per day in section 5.3.1 suggested that there existed negative monotonic association between D_SI of term “Coronavirus” and new laboratory-confirmed COVID-19 cases per day considering all three stages with ρ = -0.349. And if excluding stage 1, as shown in section 5.3.2, there were no statistically monotonic association being found.

One possible explanation for this phenomenon was because stage 1 covered the early stage of the pandemic of COVID-19 when it existed an unusual surge on searches for information about COVID-19. A lot of people searched for all kinds of information about COVID-19 due to panic, lack of knowledge and official news sources for COVID-19 in the early stage of the pandemic. This kind of surge caused high rank of D_SI of term “Coronavirus” in the early stages while only a few laboratory-confirmed cases were found in the United States at the meantime, which contributed to a negative correlation coefficient. Another explanation was the testing capacity in the early stages was so insufficient that the laboratory-confirmed cases could not reflect the real time COVID-19 cases.

Similar phenomenon was found in the results of correlation analysis between D_SI of Top-5 terms selected from tweets based on term frequency and new COVID-19 cases per day. According to section 5.3.1, there were no statistically significant monotonic association found between D_SI_Top5 and Cases_Normalization. However, if excluding stage 1, as shown in 5.3.2, there existed relatively strong positive monotonic association between D_SI _Top-5 and Cases_Normalization with ρ = 0.702.

There were some possible reasons that caused such positive monotonic association between D_SI of Top-5 terms selected from tweets based on term frequency and new COVID-19 cases per day in stage 2 and stage 3. Firstly, lots of people with suspicious symptoms of COVID-19 such as fever, cough, shortness of breath or difficulty breathing, chills, muscle pain, headache, sore throat, and new loss of taste or smell will search for COVID-19 related information, though the idealized pattern that every people with symptoms of COVID-19 will search for related information may not hold true. If the correlation is derived from this reason, the SI of each day should reflect the COVID-19 cases N days after. Secondly, people may search for COVID-19 related information as the number of COVID-19 cases shot up. In this case, moving backward the Daily Trend of new laboratory-confirmed COVID-19 case of each day by N days may help to reveal the correlation. Thirdly, COVID-19 related social events, news, editorials, national pandemic strategies and so on may draw public attention which drives people search for more information. Looking at the Appendix 1. Candidate terms from 2/29/2020 – 07/26/2020, some candidate terms such as “trump”, “lockdown” and “mask” are related to such social events. This kind of events and activities affected people’s behavior which may contribute to the correlation between Google Trend and Daily Trend of new COVID-19 cases. The reasons listed above, and other factors may individually or collectively cause such correlation.

According to Section 5.4 the result of time series analysis, the highest correlation coefficient was attained at “-3”, which suggested that an increase / decrease in covid-related searches on day T suggests an increase / decrease of new COVID-19 cases on day T + 3. That is to say Google Trend has the ability to predict Daily Trend of new COVID-

19 cases three days ahead. This time-lag can be caused by many different factors such as the common delay of CDC-reported data, and people with suspicious COVID-19 symptoms searching for related information before taking COVID-19 tests.

Based on the results of present study, it might be insufficient to monitor or predict trend of COVID-19 using single or a few terms, such as “Coronavirus”. However, inputting too many terms or queries like the GFT may not get good results either, as there were lots of research and data shows that GFT was not working well when comparing its results to real world (Lazer, Kennedy, King, & Vespignani, 2014), by which time series were computed for about 50 million common queries entered weekly from 2003 to 2008. This research provided a substitute way for selecting candidate terms, which leveraged the data of social media such as Twitter to select terms based on term frequency.

The correlation between Google Trend and Daily Trend of new COVID-19 cases as well as the time-lag found by the time series analysis can serve as supplementary to traditional surveillance for COVID-19 and be applied to predict COVID-19 Trend. However, as the testing capacity and reporting procedures changed over time, the correlation and time-lag may change as well. Therefore, it is important to extend the time span of the experiment to test whether the correlation and time-lag are stable on a larger time range.

6 Conclusion

This study started with organizing the timeline of pandemic of COVID-19, dividing it into three stages based on the peak of new COVID-19 cases per day covering 02/09/2020 to 07/26/2020. Then, this study summarized previous works on the topic of mining on social media about infectious diseases including COVID-19, H1N1, Dengue, which indicated that health authorities may leverage social media data to improve traditional surveillance system during pandemic of COVID-19 and provided support to the two hypotheses that the terms used in social media posts and search queries related to COVID-19 are same or similar and time-lag may exists between Google Trend and Daily Trend of COVID-19. In addition, this study went through research in the topic of search trend analysis about infectious diseases and summarized research applying time series analysis to describe the association of Google Trend and Daily Trend of infectious diseases which indicated stable time-lag may exist while it is important to select suitable queries.

This study leveraged tweets data to generate top-5 terms per day based on term frequency on a 7-day moving average from 10,000 randomly selected tweets per day related to COVID-19. Google Search Interest of candidate terms were used to calculate the Difference of Search Interest for each day. Then this study applied spearman correlation analysis to examine the relationship between Google Trend and Daily Trend of new COVID-19 cases in the United States. This study applied spearman correlation analysis using single term “Coronavirus” and using the top-5 selected terms per day with the Daily Trend of new COVID-19 cases. The results indicated that using the selected top-5 terms per day resulted in much stronger correlation if excluded the early stage of the pandemic of COVID-19, when an unusual surge of searches may exist and contribute to a negative correlation coefficient. The result of time series analysis suggested that Google Trend has the ability to predict Daily Trend of new COVID-19 cases three days ahead. Based on the results, it is important to account for the fact that covid-related language may change over time and using single terms such as “Coronavirus” may not provide adequate information on a larger time scale. In addition, excluding the early stage of a pandemic from trend analysis will be helpful to reduce the impact of data anomalies due to panic. However, this study did not provide a good way to divide the early stages of pandemic from the others which may result in excluding time that reflects common search activities. As the pandemic of COVID-19 continues, there are more data can be used for trend analysis. It is important for future work to test whether the correlation and time-lag are stable on a larger time scale and leverage the correlation and time-lag to help traditional surveillance system.

Appendix 1. Candidate terms from 2/29/2020 – 07/26/2020 (in alphabetical order)

Date	1	2	3	4	5
2/29/2020	cases	china	coronavirus	people	trump
3/1/2020	china	coronavirus	people	trump	us
3/2/2020	china	coronavirus	people	trump	us
3/3/2020	china	coronavirus	people	trump	us
3/4/2020	china	corona	coronavirus	people	trump
3/5/2020	corona	coronavirus	people	trump	virus
3/6/2020	corona	coronavirus	people	trump	virus
3/7/2020	corona	coronavirus	people	trump	virus
3/8/2020	corona	coronavirus	people	trump	virus
3/9/2020	corona	coronavirus	people	trump	virus
3/10/2020	corona	coronavirus	people	trump	virus
3/11/2020	corona	coronavirus	people	trump	virus
3/12/2020	corona	coronavirus	people	trump	virus
3/13/2020	corona	coronavirus	people	trump	virus
3/14/2020	corona	coronavirus	people	trump	virus
3/15/2020	corona	coronavirus	people	trump	virus
3/16/2020	corona	coronavirus	people	trump	virus
3/17/2020	corona	coronavirus	people	trump	virus
3/18/2020	corona	coronavirus	pandemic	people	virus
3/19/2020	corona	coronavirus	pandemic	people	virus
3/20/2020	corona	coronavirus	pandemic	people	virus
3/21/2020	corona	coronavirus	pandemic	people	virus
3/22/2020	corona	coronavirus	pandemic	people	virus
3/23/2020	corona	coronavirus	pandemic	people	virus
3/24/2020	corona	coronavirus	pandemic	people	virus
3/25/2020	coronavirus	home	pandemic	people	us
3/26/2020	coronavirus	home	pandemic	people	us
3/27/2020	coronavirus	home	pandemic	people	us
3/28/2020	coronavirus	home	pandemic	people	us
3/29/2020	coronavirus	lockdown	pandemic	people	us
3/30/2020	coronavirus	lockdown	pandemic	people	us
3/31/2020	coronavirus	lockdown	pandemic	people	us
4/1/2020	coronavirus	lockdown	pandemic	people	us
4/2/2020	coronavirus	lockdown	pandemic	people	us
4/3/2020	coronavirus	lockdown	pandemic	people	us
4/4/2020	coronavirus	lockdown	pandemic	people	us
4/5/2020	coronavirus	lockdown	pandemic	people	us
4/6/2020	coronavirus	pandemic	people	trump	us
4/7/2020	coronavirus	pandemic	people	trump	us

4/8/2020	coronavirus	lockdown	pandemic	people	us
4/9/2020	coronavirus	lockdown	pandemic	people	us
4/10/2020	coronavirus	lockdown	pandemic	people	us
4/11/2020	coronavirus	lockdown	pandemic	people	us
4/12/2020	coronavirus	lockdown	pandemic	people	us
4/13/2020	coronavirus	lockdown	pandemic	people	us
4/14/2020	coronavirus	lockdown	pandemic	people	us
4/15/2020	coronavirus	lockdown	pandemic	people	us
4/16/2020	coronavirus	lockdown	pandemic	people	us
4/17/2020	coronavirus	lockdown	pandemic	people	us
4/18/2020	coronavirus	lockdown	pandemic	people	us
4/19/2020	coronavirus	lockdown	pandemic	people	us
4/20/2020	coronavirus	lockdown	pandemic	people	us
4/21/2020	coronavirus	lockdown	pandemic	people	us
4/22/2020	coronavirus	lockdown	pandemic	people	us
4/23/2020	coronavirus	lockdown	pandemic	people	us
4/24/2020	coronavirus	lockdown	pandemic	people	trump
4/25/2020	coronavirus	lockdown	pandemic	people	trump
4/26/2020	coronavirus	lockdown	pandemic	people	trump
4/27/2020	coronavirus	lockdown	pandemic	people	trump
4/28/2020	coronavirus	lockdown	pandemic	people	trump
4/29/2020	coronavirus	lockdown	pandemic	people	trump
4/30/2020	coronavirus	lockdown	pandemic	people	us
5/1/2020	coronavirus	lockdown	pandemic	people	us
5/2/2020	coronavirus	lockdown	pandemic	people	us
5/3/2020	coronavirus	lockdown	pandemic	people	us
5/4/2020	coronavirus	lockdown	pandemic	people	us
5/5/2020	coronavirus	lockdown	pandemic	people	us
5/6/2020	coronavirus	lockdown	pandemic	people	us
5/7/2020	coronavirus	lockdown	pandemic	people	us
5/8/2020	coronavirus	lockdown	pandemic	people	us
5/9/2020	coronavirus	covid	lockdown	pandemic	people
5/10/2020	coronavirus	covid	lockdown	pandemic	people
5/11/2020	coronavirus	covid	lockdown	pandemic	people
5/12/2020	coronavirus	covid	lockdown	pandemic	people
5/13/2020	coronavirus	covid	lockdown	pandemic	people
5/14/2020	coronavirus	covid	lockdown	pandemic	people
5/15/2020	coronavirus	covid	lockdown	pandemic	people
5/16/2020	coronavirus	covid	lockdown	pandemic	people
5/17/2020	coronavirus	covid	lockdown	pandemic	people
5/18/2020	coronavirus	covid	lockdown	pandemic	people
5/19/2020	coronavirus	covid	lockdown	pandemic	people
5/20/2020	coronavirus	covid	lockdown	pandemic	people
5/21/2020	coronavirus	covid	lockdown	pandemic	people

5/22/2020	coronavirus	covid	lockdown	pandemic	people
5/23/2020	coronavirus	covid	lockdown	pandemic	people
5/24/2020	coronavirus	covid	lockdown	pandemic	people
5/25/2020	coronavirus	covid	lockdown	pandemic	people
5/26/2020	coronavirus	covid	lockdown	pandemic	people
5/27/2020	coronavirus	covid	lockdown	pandemic	people
5/28/2020	coronavirus	covid	lockdown	pandemic	people
5/29/2020	coronavirus	covid	lockdown	pandemic	people
5/30/2020	coronavirus	covid	lockdown	pandemic	people
5/31/2020	coronavirus	covid	lockdown	pandemic	people
6/1/2020	coronavirus	covid	lockdown	pandemic	people
6/2/2020	coronavirus	covid	lockdown	pandemic	people
6/3/2020	coronavirus	covid	lockdown	pandemic	people
6/4/2020	coronavirus	covid	lockdown	pandemic	people
6/5/2020	coronavirus	covid	lockdown	pandemic	people
6/6/2020	coronavirus	covid	lockdown	pandemic	people
6/7/2020	coronavirus	covid	lockdown	pandemic	people
6/8/2020	coronavirus	covid	lockdown	pandemic	people
6/9/2020	coronavirus	covid	lockdown	pandemic	people
6/10/2020	coronavirus	covid	lockdown	pandemic	people
6/11/2020	coronavirus	covid	lockdown	pandemic	people
6/12/2020	coronavirus	covid	lockdown	pandemic	people
6/13/2020	coronavirus	covid	lockdown	pandemic	people
6/14/2020	coronavirus	covid	lockdown	pandemic	people
6/15/2020	coronavirus	covid	lockdown	pandemic	people
6/16/2020	coronavirus	covid	lockdown	pandemic	people
6/17/2020	coronavirus	covid	lockdown	pandemic	people
6/18/2020	coronavirus	covid	lockdown	pandemic	people
6/19/2020	cases	coronavirus	covid	pandemic	people
6/20/2020	cases	coronavirus	covid	pandemic	people
6/21/2020	coronavirus	covid	pandemic	people	trump
6/22/2020	coronavirus	covid	pandemic	people	trump
6/23/2020	coronavirus	covid	pandemic	people	trump
6/24/2020	coronavirus	covid	pandemic	people	trump
6/25/2020	coronavirus	covid	pandemic	people	trump
6/26/2020	coronavirus	covid	pandemic	people	trump
6/27/2020	coronavirus	covid	pandemic	people	trump
6/28/2020	cases	coronavirus	covid	pandemic	people
6/29/2020	cases	coronavirus	covid	pandemic	people
6/30/2020	cases	coronavirus	covid	pandemic	people
7/1/2020	cases	coronavirus	covid	pandemic	people
7/2/2020	cases	coronavirus	covid	pandemic	people
7/3/2020	cases	coronavirus	covid	pandemic	people
7/4/2020	coronavirus	covid	mask	pandemic	people

7/5/2020	coronavirus	covid	mask	pandemic	people
7/6/2020	coronavirus	covid	mask	pandemic	people
7/7/2020	coronavirus	covid	mask	pandemic	people
7/8/2020	coronavirus	covid	pandemic	people	wear
7/9/2020	coronavirus	covid	pandemic	people	wear
7/10/2020	coronavirus	covid	pandemic	people	wear
7/11/2020	cases	coronavirus	covid	pandemic	people
7/12/2020	coronavirus	covid	mask	pandemic	people
7/13/2020	coronavirus	covid	mask	pandemic	people
7/14/2020	coronavirus	covid	mask	pandemic	people
7/15/2020	coronavirus	covid	mask	pandemic	people
7/16/2020	coronavirus	covid	mask	pandemic	wear
7/17/2020	coronavirus	covid	mask	pandemic	wear
7/18/2020	coronavirus	covid	mask	pandemic	wear
7/19/2020	coronavirus	covid	mask	pandemic	wear
7/20/2020	coronavirus	covid	mask	pandemic	people
7/21/2020	coronavirus	covid	mask	pandemic	people
7/22/2020	coronavirus	covid	mask	pandemic	people
7/23/2020	coronavirus	covid	mask	pandemic	people
7/24/2020	coronavirus	covid	mask	pandemic	people
7/25/2020	coronavirus	covid	mask	pandemic	people
7/26/2020	coronavirus	covid	mask	pandemic	people

References

Marques-Toledo, C. de A., Degener, C. M., Vinhal, L., Coelho, G., Meira, W., Codeço, C. T., & Teixeira, M. M. (2017). Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. *PLoS Neglected Tropical Diseases*, *11*(7), e0005729. doi:10.1371/journal.pntd.0005729

Rajput, N. K., Grover, B. A., & Rathi, V. K. (2020). Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic. *arXiv*.

Shin, S.-Y., Seo, D.-W., An, J., Kwak, H., Kim, S.-H., Gwack, J., & Jo, M.-W. (2016). High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Scientific Reports*, *6*, 32920. doi:10.1038/srep32920

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *Plos One*, *6*(5), e19467. doi:10.1371/journal.pone.0019467

Spearman's Rank-Order Correlation - A guide to how to calculate it and interpret the output. (n.d.). Retrieved March 30, 2021, from https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide-2.php

A Timeline of COVID-19 Developments in 2020 | AJMC. (n.d.). Retrieved September 4, 2020, from https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *Journal of Medical Internet Research*, *22*(4), e19016. doi:10.2196/19016

Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & R Niakan Kalhori, S. (2020). Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR public health and surveillance*, *6*(2), e18828. doi:10.2196/18828

CDC COVID Data Tracker. (n.d.). Retrieved October 4, 2020, from https://covid.cdc.gov/covid-data-tracker/#trends

CDC officials confirm 7th US case of coronavirus, in California man. (n.d.). Retrieved October 4, 2020, from https://www.cnbc.com/2020/01/31/california-and-cdc-officials-confirm-7th-case-of-coronavirus-in-the-us.html

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *Plos One*, *5*(11), e14118. doi:10.1371/journal.pone.0014118

Coronavirus smell loss “different from cold and flu” - BBC News. (n.d.). Retrieved October 5, 2020, from https://www.bbc.com/news/health-53810610

COVID-19-TweetIDs/2020-01 at master · echen102/COVID-19-TweetIDs · GitHub. (n.d.). Retrieved March 17, 2021, from https://github.com/echen102/COVID-19-TweetIDs/tree/master/2020-01

Effenberger, M., Kronbichler, A., Shin, J. I., Mayer, G., Tilg, H., & Perco, P. (2020). Association of the COVID-19 pandemic with Internet Search Volumes: A Google

TrendsTM Analysis. *International Journal of Infectious Diseases, 95,* 192–197. doi:10.1016/j.ijid.2020.04.033

Gao, Z., Yada, S., Wakamiya, S., & Aramaki, E. (2020). NAIST COVID: Multilingual COVID-19 Twitter and Weibo Dataset. *arXiv*.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. doi:10.1038/nature07634

Global COVID-19 Tracker & Interactive Charts | Real Time Updates & Digestable Information for Everyone | 1Point3Acres. (n.d.). Retrieved September 4, 2020, from https://coronavirus.1point3acres.com/

Google Trends. (n.d.). Retrieved September 30, 2020, from https://trends.google.com/trends/?geo=US

Google Trends - Wikipedia. (n.d.). Retrieved October 5, 2020, from https://en.wikipedia.org/wiki/Google_Trends

Hong, Y.-R., Lawrence, J., Williams, D., & Mainous III, A. (2020). Population-Level Interest and Telehealth Capacity of US Hospitals in Response to COVID-19: Cross-Sectional Analysis of Google Search and National Hospital Survey Data. *JMIR public health and surveillance*, *6*(2), e18961. doi:10.2196/18961

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., … Lessler, J. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine, 172*(9), 577–582. doi:10.7326/M20-0504

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205. doi:10.1126/science.1248506

Li, C., Chen, L. J., Chen, X., Zhang, M., Pang, C. P., & Chen, H. (2020). Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Euro Surveillance, 25*(10). doi:10.2807/1560-7917.ES.2020.25.10.2000199

Lin, Y.-H., Liu, C.-H., & Chiu, Y.-C. (2020). Google searches for the keywords of "wash hands" predict the speed of national spread of COVID-19 outbreak among 21 countries. *Brain, Behavior, and Immunity*, *87*, 30–32. doi:10.1016/j.bbi.2020.04.020

Marques-Toledo, C. de A., Degener, C. M., Vinhal, L., Coelho, G., Meira, W., Codeço, C. T., & Teixeira, M. M. (2017). Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. *PLoS Neglected Tropical Diseases*, *11*(7), e0005729. doi:10.1371/journal.pntd.0005729

Rajput, N. K., Grover, B. A., & Rathi, V. K. (2020). Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic. *arXiv*.

Shin, S.-Y., Seo, D.-W., An, J., Kwak, H., Kim, S.-H., Gwack, J., & Jo, M.-W. (2016). High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Scientific Reports*, *6*, 32920. doi:10.1038/srep32920

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *Plos One*, *6*(5), e19467. doi:10.1371/journal.pone.0019467

Timeline of the COVID-19 pandemic in the United States - Wikipedia. (n.d.). Retrieved October 4, 2020, from https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_the_United_States

Walker, A., Hopkins, C., & Surda, P. (2020). Use of Google Trends to investigate loss-of-smell-related searches during the COVID-19 outbreak. *International forum of allergy & rhinology*, *10*(7), 839–847. doi:10.1002/alr.22580

While CDC coronavirus tests stalled for six weeks, a German lab made 1.4 million tests - The Washington Post. (n.d.). Retrieved October 4, 2020, from https://www.washingtonpost.com/business/2020/03/16/cdc-who-coronavirus-tests/

Wilson, N., Mason, K., Tobias, M., Peacey, M., Huang, Q. S., & Baker, M. (2009). Interpreting "Google Flu Trends" data for pandemic H1N1 influenza: The New Zealand experience. *Eurosurveillance*.