# Characteristics of online social network during a pandemic

Muthualagesan Suryavarathan*1, Sanjay Kumar*1

## Abstract

Social media, such as Twitter and Facebook, plays a critical role in pandemic management by propagating emergency information to a pandemic-affected community. It ranks as the fourth most popular source for accessing emergency information. Emergency agencies and organizations are on the periphery of the social network, connecting a community with other communities. Here, we intend to study and analyze the Twitter dataset generated during the COVID-19 pandemic using natural language processing and social network analysis. The results of this study will help emergency agencies develop their social media operation strategies for a pandemic mitigation plan.

**Keywords:** Social Network Analysis, Natural Language Processing, COVID-19

## 1. INTRODUCTION

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus was first identified in December 2019 in Wuhan, China. The World Health Organization declared a Public Health Emergency of International Concern regarding COVID-19 on 30 January 2020 and later declared a pandemic on 11 March 2020. As of 17 April 2021, more than 140 million cases have been confirmed, with more than 3 million deaths attributed to COVID-19, making it one of the deadliest pandemics in history.

Social media platforms can provide rich and useful information to predict and explain the characteristics and status of disease outbreaks. Text mining can be used to extract health information from social media platforms such as Twitter [1]. Twitter data enable researchers to obtain large samples of user-generated content, thereby garnering insights to inform early response strategies. Social media data text mining has been used to track diseases and assess public awareness concerning health issues, enabling disease forecasting [2]. Text analysis of Twitter data is one of the most important areas of focus in medical informatics research [3].

COVID-19 is a scientifically and medically novel disease that is not fully understood, as it has yet to be consistently and deeply studied. The use of social media information to analyze syndromic surveillance, focusing on public health–related concerns using web-based information and content is essential [4]. One important reason is that during an outbreak, social media plays a critical role, as these platforms reflect real-time public panic through comments. Twitter, one of these social media platforms, has often served as a communication

* Equal Contribution, [1]Department of Applied Mathematics and Computational Sciences, PSG College of Technology, India

modality during disease outbreaks [5]. Twitter provides rich information to increase public awareness and inform people about outbreak locations. This is very useful to provide insight regarding the issues related to infectious disease outbreaks.

## 2. RELATED WORK

### 2.1. COVID-19 Pandemic on Twitter

Regarding COVID-19, there is a lack of social media data–based research studying the spread of the disease, the behavioral awareness of the public, and emergent conversations on COVID-19. Taking research published in 2020 as examples, Shen et al [6] studied mentions of symptoms and diseases on social media to predict COVID-19 case counts, and Huang et al [7] analyzed social media posts to study the characteristics of COVID-19 patients in China. However, both these studies focused primarily on China. Moreover, Park et al [8] addressed information transmission networks and news-sharing behaviors on Twitter regarding COVID-19 in Korea only. Abd-Alrazaq et al [9] conducted an infoveillance study on aspects of the COVID-19 pandemic, aiming to study the main topics of discussion related to the disease. Chen et al [10] presented basic statistics that tracked only Twitter activity responding and reacting to COVID-19–related events. However, these previous studies did not use Twitter data to address conclusive themes, nor did they perform sentiment analysis during the timeline of COVID-19 in its initial stage. These missing data are important because themes and sentiment analysis can provide a wider overview of public awareness [11]. The evolution of sentiment analysis of Twitter data since the early stages of the COVID-19 pandemic has not yet been fully presented. Greater understanding and public awareness of the pandemic are still needed.

### 2.2. Social Media as a Diagnostic Tool and Referral System

Social media should be used to disseminate reliable information about when to get tested, what to do with the results, and where to receive care [12]. If a vaccine becomes available, the same platforms could be used to encourage uptake and address challenges associated with vaccine hesitancy. These targeted efforts can occur in response to what people search for or in a more personalized approach based on an individual's online profile, posts, and likes. Health systems may become overwhelmed as testing becomes more available and as more mildly ill yet concerned individuals seek care; yet, social media platforms are well poised to enable users to remotely assess symptoms and determine their most appropriate course of action. For those whose test results are positive for COVID-19, the platform could enable users to inform their contacts about the potential exposure and how to follow up for testing.

## 3. DATA COLLECTION AND PREPROCESSING

On Twitter, thousands of updates and opinions regarding this pandemic keep springing up every other day. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment and work with other data scientists and machine learning engineers. One such dataset [99] containing the Tweets from 29th March to 30th April 2020 of users who have applied the following hashtags: #coronavirusoutbreak, #covid19, #coronavirusPandemic, #coronavirus, #covid_19, etc. has been

collected and hosted on Kaggle. The dataset contains variables associated with Twitter: the text of various tweets and the accounts that tweeted them, the hashtags used, and the locations of the accounts. In this dataset, there are around 1,67,000 tweets posted in the English language in the US Region.

As a first step towards cleaning and preprocessing the dataset, the columns that are irrelevant to the current research objective have been dropped. To start data processing, the tweet texts were subjected to a series of functions to remove URLs, emojis, special characters, retweets, hash symbols, and hyperlinks pointing to websites; this process also enabled us, as much as possible, to exclude mentions of related diseases that would contaminate the results. Stop words in English (eg, for, the, is) were also removed. Additionally, the hashtags on whose basis the dataset was generated have also been added to the list of stopwords, as they do not add any new information during the text analysis. Further, the tweet texts were converted to lowercase, and words were changed to their root forms (eg, viruses to virus). The tweets were then converted into a corpus (text mining structure). The final tokenized tweet data was used for all further text analysis methods explained in the following sections.

## 4. TREND ANALYSIS

### 4.1. Google Trends

Google Trends provides access to a largely unfiltered sample of actual search requests made to Google. It's anonymized, categorized, and aggregated. This allows us to display interest in a particular topic from around the globe or down to city-level geography. While only a sample of Google searches is used in Google Trends. Providing access to the entire data set would be too large to process quickly. By sampling data, we can look at a dataset representative of all Google searches, while finding insights that can be processed within minutes of an event happening in the real world. Google Trends normalizes search data to make comparisons between terms easier. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics.

### 4.2. Top Hashtags and Mentions

Almost every social media site is known for the topic it represents in the form of hashtags. Particularly for our case, Hashtags played an important part since we were interested in #Covid19,#Coronavirus, #StayHome, #InThisTogether, etc. Hence, the first step was forming a separate feature based on the hashtag values and mentions. After the pre-processing of data, We excluded all the stop words and links in the tweets because it acted as a leakage variable. Then the words have been tokenized so that we can find which are all hashtags and mentions in the tweets by using Regex. Then we performed a frequency distribution of the most occurring hashtags and mentions.

## 5. SENTIMENT ANALYSIS

Sentiment analysis, a natural language processing (NLP) approach, was used to categorize the sentiments appearing in Twitter messages. TextBlob [100] is a Python library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more. TextBlob maintains a dictionary of polarity

scores associated with every word in the English language. For a set of tokens, the TextBlob API calculates a polarity score for the given set of words based on the words present in the set and the order of occurrence in the set.

In this work, every tweet in the dataset has been tokenized to calculate the polarity score ranging from -1 to +1. A negative 1 for a tweet with the most negative sentiment and a positive 1 for a tweet with the most positive sentiment. Neutral sentiment is represented by a zero.

The tweets with an associated polarity have been analyzed to produce various results. Based on the sentiment polarity scores, hashtags and mentions in various sentiments have also been analyzed. Most frequent hashtags and mentions present in the tweets with positive, negative, and neutral sentiments have been collected and listed individually for each sentiment. Besides, the most frequent words used by Twitter users in various sentiments have also been collected and analyzed.

## 6. RESULTS

The results indicate aspects of public awareness and concern regarding the COVID-19 pandemic. Trends during the time of the outbreak of Covid-19, the results of the sentiment analysis showed that people had both positive and negative outlook toward COVID-19, and based on top hashtags we come to know about the themes relating to COVID-19 that is spread and symptoms of COVID-19, the COVID-19 pandemic emergency, how to control COVID-19, and reports on COVID-19. Also with top mentions in tweets, we observe the people and organizations who were influential during the outbreak.

### 6.1 Google Trend Analysis

We explored Google Trends to identify the relation between search trends of people during the period of COVID-19 outbreak in India over the year 2020. Observed the search-term trend (or interest over time) of Netflix and Indian Premier League with comparison to COVID-19 these results show that people were more interested in COVID and Netflix, when the lockdown was enforced in India and interest towards Indian Premier League, was very low during the outbreak later part of the year it went high. Interest in the sport did not have any effect on the pandemic.

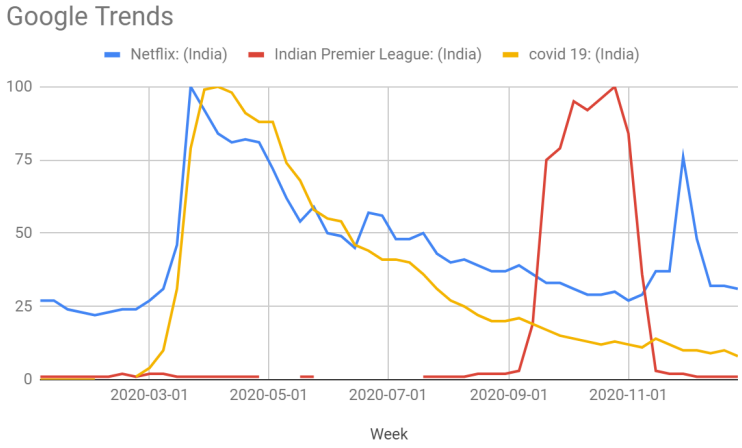


**Figure 6.1. Google Trend Graph of Netflix, IPL and COVID-19 in India**

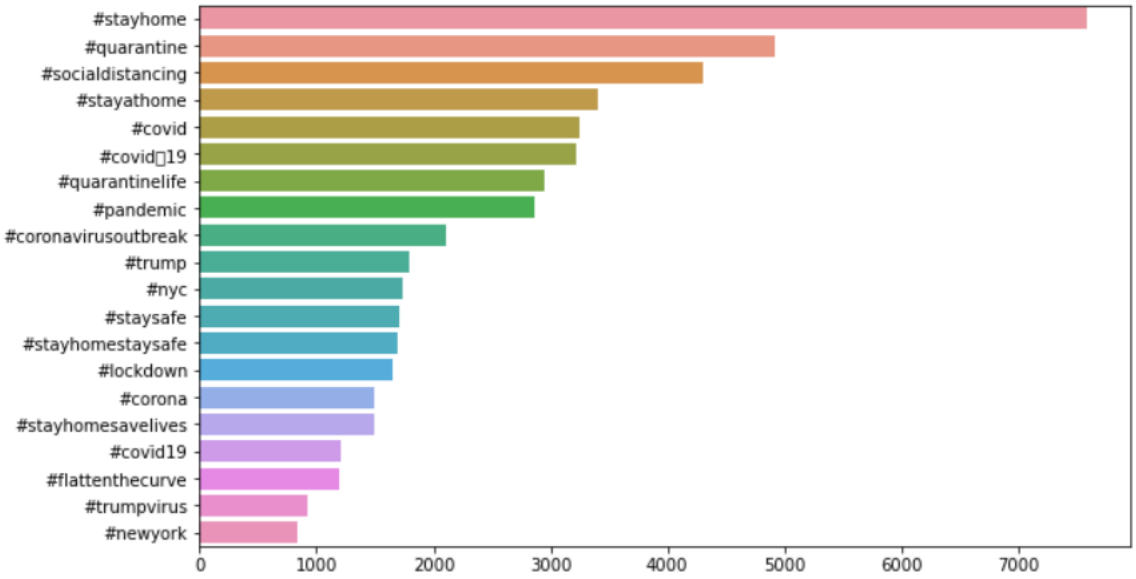


**Figure 6.2. Bar Graph of frequent hashtag**

### 6.2 Top Hashtags and Mentions

This analysis used word clouds, which can provide a visual representation of text appearing in tweets. Word clouds highlight words according to frequency. In this study, the word clouds of frequently appearing words provided deeper insights into tweets related to COVID-19 posted by Twitter users.


**Figure 6.3. Word Cloud of Top hashtags**

According to Figure 6.4, the most frequently appearing hashtags were related to #stayhome, #quarantine, #socialdistancing, etc. This shows that people are spreading the precautions to be followed during the pandemic.


**Figure 6.4. Word Cloud of Top Mentions**

When it comes to mentions in the tweets we observe that people have mostly mentioned elected representatives or government officials, media, and some health organizations those are @realdonaldtrump, @cnn, @who, etc.

### 6.3 Sentiments of Tweets

The sentiment-level analysis further enriched the findings through the clear identification of negative and positive topics on COVID-19. The sentiment of tweets is positive every day in the given period. However, towards mid-April Twitter users were much less active and the percentage

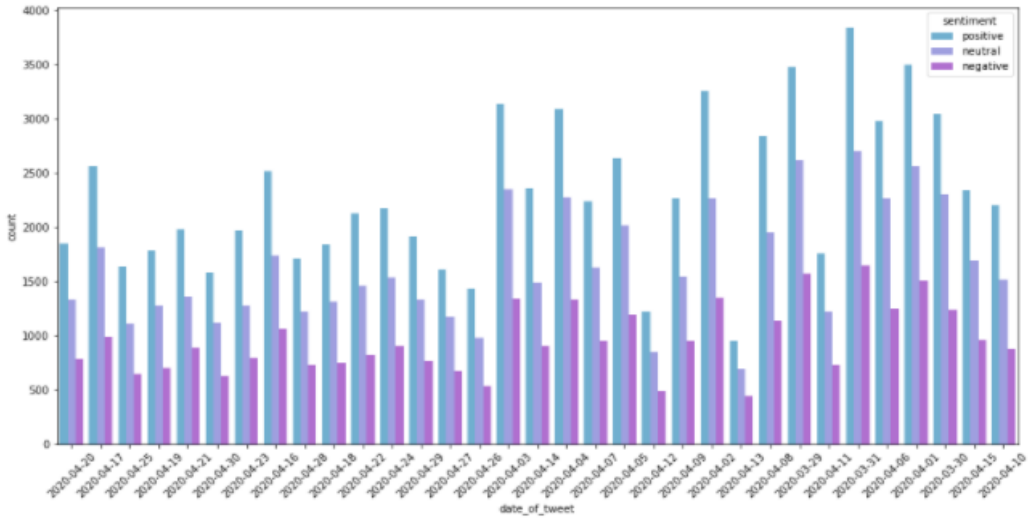


**Figure 6.5. Tweet statistics in different sentiments for chosen period**

of positive tweets diminished as well. This signifies that Twitter users had a positive outlook toward COVID-19. In Figure 6.6 we can see a fatter tail of the polarity in the positive direction.

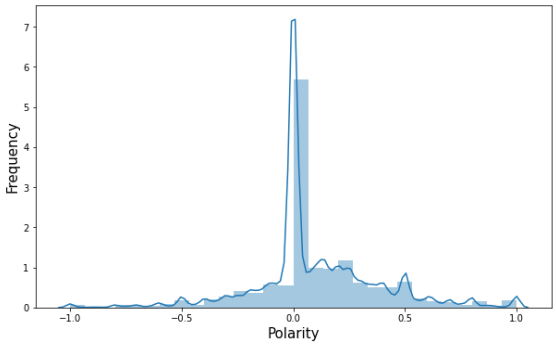


**Figure 6.6 Polarity vs Frequency**

### 6.4 Sentiments of Top Hashtags and Mentions

For the next phase of the sentiment analysis, a word cloud was created using the most frequently used Hashtags and Mentions, which were categorized by different sentiments. The word cloud was developed based on the whole corpus of tweets and also visualization with a bar chart for better comparison on top hashtags and mentions.

With the results of the top hashtag in tweets with the different sentiment, we understood that the top 50 in negative, positive, neutral sentiments all were the same because these hashtags were used in large numbers so it was present in both negative and positive hashtags. To remove this we found the second top 50 and while checking the result we found that the results were different for each sentiment. Some of the positive hashtags were #easter, #hope, this shows the positive things which people were talking about. Negative hashtags were mostly on criticizing the government and also about the COVID-19 those are #trumpownseverydeath, #fakenews, etc. and some of the neutral hashtags observed are #breakingnow, #2020, etc.


**Figure 6.7. Positive Top Mentions**

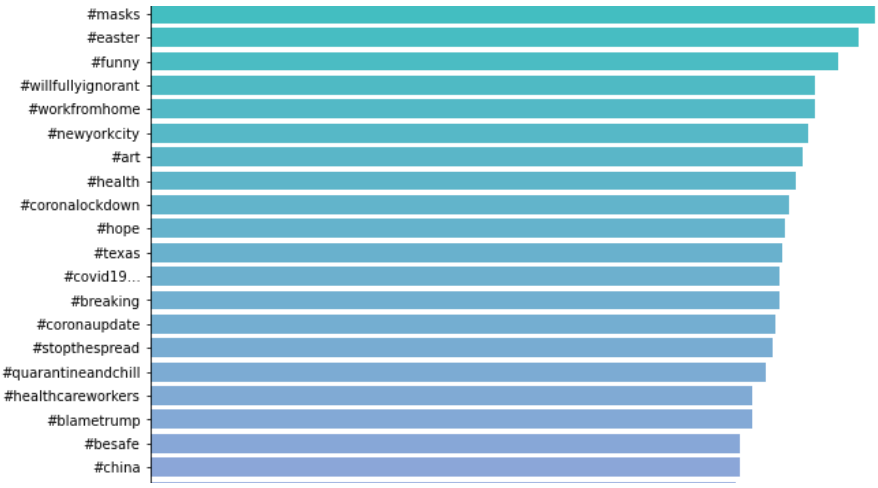


**Figure 6.8. Top hashtags in positive sentiment**

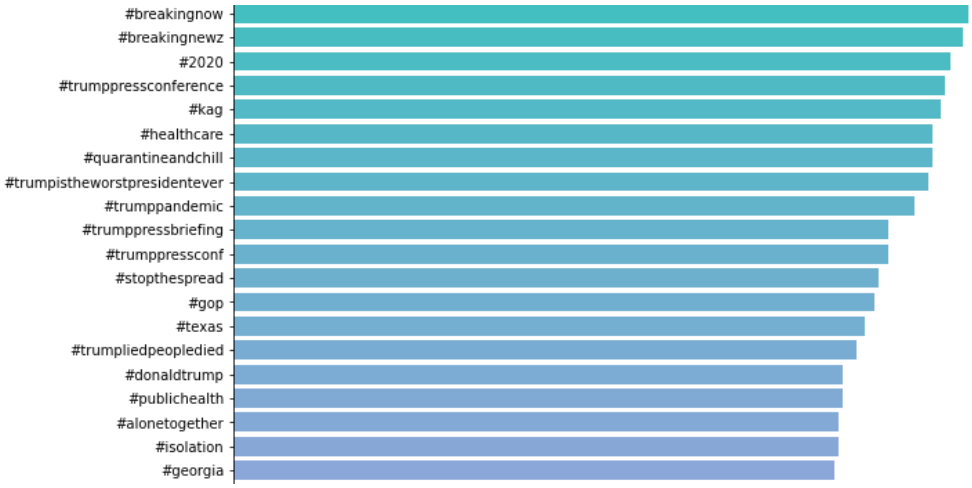


**Figure 6.9. Top hashtags in neutral sentiment**

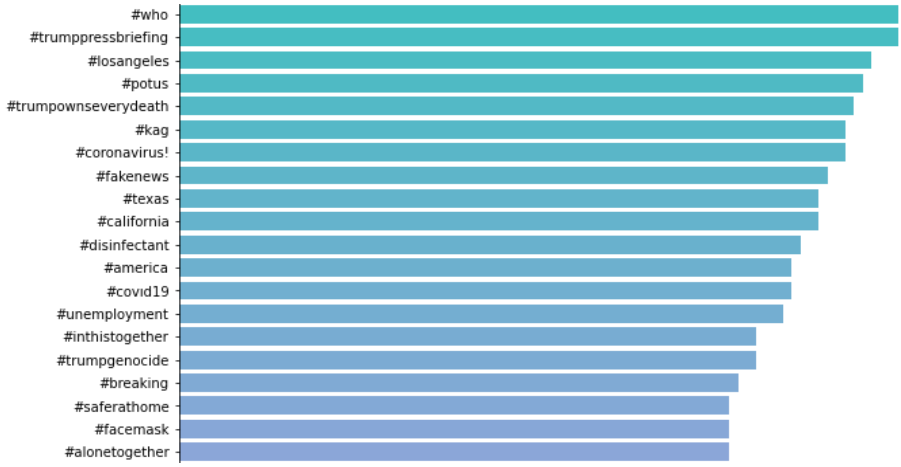


**Figure 6.10. Top hashtags in negative sentiment**

After analyzing the results of the top mentions in tweets with a different sentiment with the word cloud we know people have mentioned common political persons and organisations in their tweets for all the kind of positive, negative and neutral tweets. These political people and organisations who have been mentioned in tweets act as a major source of information about the covid19 and also they have appreciation and criticism for their decisions during these pandemics.

### 6.5 Visualize similarity of Hashtags and Mentions

Trained a Word2Vec model for finding the similarity between the tokenized hashtags and mentions. It is a recent model that embeds words in a lower-dimensional vector space using a shallow neural network. We have used this model with t-SNE for data exploration and visualizing high-dimensional data. It gives you a feel or intuition of how the data is arranged in a high-dimensional space.


**Figure 6.11. Similar hashtags in t-SNE**

## 7. CONCLUSION AND FUTURE WORK

Policymakers should recognize that Twitter data can be used to explore levels of public awareness and emotions about the COVID-19 pandemic. It is important to note that the levels of public awareness are dynamic, which can be observed from the two or three awareness peaks in a period of just a few months in this study.

In this work, a large Twitter dataset has been analyzed using various methods including Natural Language Processing. The sentiment of the people using Twitter towards the impact of COVID-19 on humanity has been analyzed and presented. The search trends and behavior of people during the pandemic have also been analyzed. These results can be interpreted by government institutions and emergency agencies to better understand the population. The results of this study will help emergency agencies develop their social media operation strategies for a pandemic response plan.

Though this study has produced useful results, it contains a few limitations. First, it is worth mentioning that this study used keywords related to COVID-19 to investigate trends and frequencies of keywords. The list of selected keywords may have been incomplete. The keywords used in this study can be extended to cover the search of tweets by combining keywords related to COVID-19 and its symptoms. This work considers the tweets that are posted in the US region and the English language for analysis. Considering regional languages and more regions across the world shall help us build a more efficient model.

In future work, we aim to also perform social network analysis on the Twitter data. By modeling the users as social actors and drawing an edge between them based on their social media interaction (such as retweet, mention, etc.), we aim to produce a social network on which patterns can be explored which in turn provide insights to understand the critical role of social media use for emergency information propagation.

## REFERENCES

1. Jahanbin K, Rahmanian V. Using Twitter and web news mining to predict COVID-19 outbreak. *Asian Pac J of Trop Med 2020; 13.*2020;13(8):378–380. https://www.apjtm.org/text.asp?2020/13/8/378/279651. [Google Scholar]

2. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM, Pavlin JA, Shigematsu M, Streichert LC, Suda KJ, Corley CD. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. *PLoS One.* 2015 Oct 5;10(10):e0139701. doi: 10.1371/journal.pone.0139701. https://dx.plos.org/10.1371/journal.pone.0139701. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

3. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS One.* 2010 Nov 29;5(11):e14118. doi: 10.1371/journal.pone.0014118. https://dx.plos.org/10.1371/journal.pone.0014118. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

4. Werron T, Ringel L. Pandemic Practices,

Part One: How to Turn “Living Through the COVID-19 Pandemic” into a Heuristic Tool for Sociological Theorizing. *Sociologica.* 2020;14(2):55–72. doi: 10.6092/issn.1971-8853/11172. [CrossRef] [Google Scholar]

5. Bartlett C, Wurtz R. Twitter and Public Health. *J Public Health Manag Pract.* 2015;21(4):375–383. doi: 10.1097/phh.0000000000000041. [PubMed] [CrossRef] [Google Scholar]

6. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. *J Med Internet Res.* 2020 May 28;22(5):e19421. doi: 10.2196/19421. https://www.jmir.org/2020/5/e19421/ [PMC free article] [PubMed] [CrossRef] [Google Scholar]

7. Huang C, Xu X, Cai Y, Ge Q, Zeng G, Li X, Zhang W, Ji C, Yang L. Mining the Characteristics of COVID-19 Patients in China: Analysis of Social Media Posts. *J Med Internet Res.* 2020 May 17;22(5):e19087. doi: 10.2196/19087. https://www.jmir.org/2020/5/e19087/ [PMC free article] [PubMed] [CrossRef] [Google Scholar]

8. Park HW, Park S, Chong M. Conversations and Medical News Frames on Twitter: Infodemiological Study on COVID-19 in South Korea. *J Med Internet Res.* 2020 May 05;22(5):e18897. doi: 10.2196/18897. https://www.jmir.org/2020/5/e18897/ [PMC free article] [PubMed] [CrossRef] [Google Scholar]

9. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *J Med Internet Res.* 2020 Apr 21;22(4):e19016. doi: 10.2196/19016. https://www.jmir.org/2020/4/e19016/ [PMC free article] [PubMed] [CrossRef] [Google Scholar]

10. Chen E, Lerman K, Ferrara E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill.* 2020 May 29;6(2):e19273. doi: 10.2196/19273. https://publichealth.jmir.org/2020/2/e19273/ [PMC free article] [PubMed] [CrossRef] [Google Scholar]

11. Samuel J, Ali G, Rahman M, Esawi E, Samuel Y. COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. *Information.* 2020 Jun 11;11(6):314. doi: 10.3390/info11060314. [CrossRef] [Google Scholar]

12. Merchant RM. Evaluating the potential role of social media in preventive health care.https://jamanetwork.com/journals/jama/article-abstract/2758937?utm_campaign=articlePDF&utm_medium=articlePDFlink&utm_source=articlePDF&utm_content=jama.2020.4469