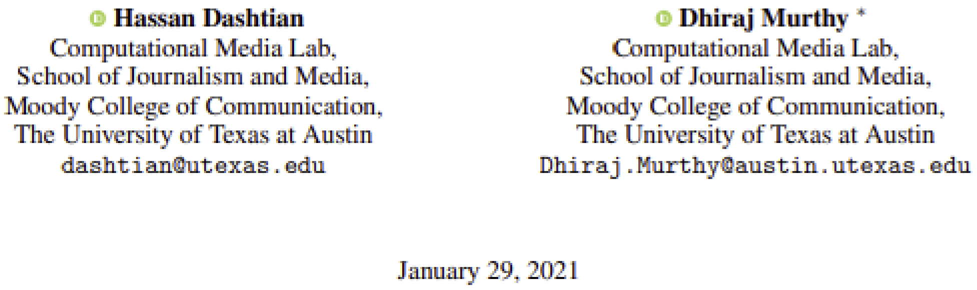
COVID-19 Paper Review

By: Alana R. McGraw

Title Revision: Understanding Public Opinion: Extracting COVID-19 Sentiments Through Twitter Posts



Hassan Dashtian
Computational Media Lab,
School of Journalism and Media,
Moody College of Communication,
The University of Texas at Austin
dashtian@utexas.edu

Dhiraj Murthy *
Computational Media Lab,
School of Journalism and Media,
Moody College of Communication,
The University of Texas at Austin
Dhiraj.Murthy@austin.utexas.edu

January 29, 2021

Abstract

Twitter and other social media platforms represent important sites of engagement and discussion regarding the COVID-19 pandemic. Analysis of the sentiment and information presented on platforms like Twitter can be used for social, health, and political research. In this report, we present a COVID-19 Twitter data set of 19,298,967 million tweets from 5,977,653 unique individuals and summarize some of the attributes of these data. We use topic modeling, sentiment analysis, and descriptive statistics to describe the tweets related to COVID-19 we collected and the location of tweets.

1 Introduction

The COVID-19 pandemic is an unparalleled global health emergency. As such, it has led to an exceptional response on social media platforms, with posts related to social, political, and economic life. Many people rely on social media to stay informed, with 68% of Americans reporting they use social media to access information, and one third of people stating that Twitter is the most important source of scientific information and news [2,3]. In these uncertain times, high volumes of COVID-19-related misinformation on social networks like Twitter is a significant problem [1].

The Ebola outbreak in 2014 [5] and the spread of Zika in 2016 [6] highlighted the importance of studying pandemics through the content of social networks [7,8]. There is a new urgency in monitoring social media content related to COVID-19 [3]. One particular concern is that Twitter can be a source of misinformation about health issues such as COVID-19 vaccination [4]. Given that Twitter is such a vital source of information for the public, is critical to understand the attitudes, perceptions, and responses to COVID-19 present in social media data.

In this paper, we explore the frequency of tweet activity related to COVID-19 and we make our data and source code publicly available for others to use. We collected tweets real time using the Twitter API from March - July 2020 with the following COVID-19-related query terms (‘coronavirus’, ‘covid’ and ‘mask’). Here, we describe our data collection methods, present basic statistics of the dataset, and provide information about how to obtain and use the data.

2 Methods

2.1 Data Collection

The University of Texas Austin Computational Media Lab (CML) collected 19,298,967 million tweets from 5,977,653 individuals between March – June 2020 and called the data set CML-COVID. All data were collected from Twitter through Netlytic 2, a text and social network analyzer [11], which queried Twitter’s REST API for COVID-19 related tweets. The dataset is roughly 15 GB of raw data. To comply with Twitter’s Terms & Conditions (T&C), we have not publicly released the full text/API-derived information from the collected tweets. Our released data set includes a list of the tweet IDs that others can use to retrieve the full tweet objects directly from Twitter using their own API calls. There are a variety of tools that can be used to recover the full tweets objects such as Hydrator3. Twitter also provides documentation in their Developer site4 on how to recover (hydrate) 100 tweets per API request.

2.2 Preprocessing

We pre-processed each raw tweet by concatenating (linking together in a series) and converting csv files of full tweet objects into Python DataFrames and lists to optimize data processing. We removed characters such as “\, /, ∗ and etc.”, filtered out stop words (like the most rare and frequent words), and performed text tokenization (the process of breaking up text into subunits of text called tokens) [12]. These steps are essential for topic modeling and sentiment analysis.

2.3 Data Summary

In order to summarize the data, we began by retrieving the location of users based on what is reported in their profiles, not GPS coordinates. Locations such as “USA” and “United States” are considered the same and merged as a singular location (i.e., ‘United States’). For each state in the United States with identifiable state-level location, we counted the number of tweets and calculated the frequency of tweets per day. We then conducted a frequency analysis by time. We identified the date and time of each tweet and counted the frequencies of tweets for each day.

2.4 Topic Modeling and Sentiment Analysis

We then calculated the sentiment of each tweet. We used TextBlob5 to perform sentiment analysis, the act of determining the author’s opinions based on what they write [15]. To extract information related to sentiment in our collected tweets, we used Textblob to extract the sentiment and scores. We divide tweet sentiment into three main categories - ‘Negative’, ‘Neutral’ and ‘Positive’. For each day we count the number of tweets with one of these three categories.

For topic modeling (a method that exposes the themes in a collection of information and aids in information retrieval), we applied an unsupervised topic clustering technique called Latent Dirichlet Allocation (LDA), which identifies groups of words that are often together in a collection of text [13, 14]. We sampled 20% of the tweets in our dataset and trained an LDA model that was used to estimate the most representative words in each topic. We found extraneous terms (e.g., ‘amp’, ‘dan’, and ‘na’) in our derived topic models that interfered with our topic modeling results. Therefore, we re-ran LDA and removed these terms (see table 3).

3. Results

Users in our dataset posted an average of 3 tweets. A preliminary analysis of the data shows that English is the dominant language in the tweets we collected (65.4%), but other languages were present including Spanish (12.2%). Table 1 summarizes the top 10 languages, the frequency of associated tweets, and the percentage of each language in our dataset. The top 10 locations (by country and city) of users is summarized in Table 2. The frequency of tweets per day in US states is illustrated in Figure 2. The United States has the highest frequency of tweets during the period that we collected these data. The number of tweets is low for most regions and countries. The frequency of COVID-19 related tweets per day is illustrated in Figure 3. Tweet frequency is relatively consistent during our data collection period.

Figure 4 depicts the sentiment by category over time. As figure 4 indicates, neutral tweets were the most numerous, followed by positive tweets. In the first two weeks of April 2020, the gap in frequency between the three sentiment categories is initially reasonably large, but it closes after the second week of April. We obtained 10 topic clusters from the trained LDA-based topic model summarized in Table 3. Three topics are illustrated in Table 3, and topic 1 is Spanish language. Since tweets can be in any of 64 different languages, the topics and the top words may contain words and symbols that are from different languages. As we found, cleaning the data based on stop words in one language is not enough to solve these issues.

Table 1: Top ten most popular languages, the number of associated tweets and their percentage.

ISO Language Code	Language	Number of Tweets	Percentage
en	English	12488955	65.4%
es	Spanish	2333241	12.2%
pt	Portuguese	728483	3.8%
und	undefined	651141	3.4%
fr	French	536100	2.8%
in	Indonesian	483566	2.5%
ja	Japanese	419953	2.2%
it	Italian	262602	1.4%
tl	Tagalog	183694	1.0%
hi	Hindi	155204	0.8%

Table 2: Top ten locations of tweets based on the user profiles.

Location	Number of Tweets
‘ ’ (undefined)	5483327
United States	330563
India	121037
New York, USA	85236
London, England	156034
Washington, D.C., USA	79412
Los Angeles, USA	79335
California, USA	73098
México	54689
United Kingdom	53773

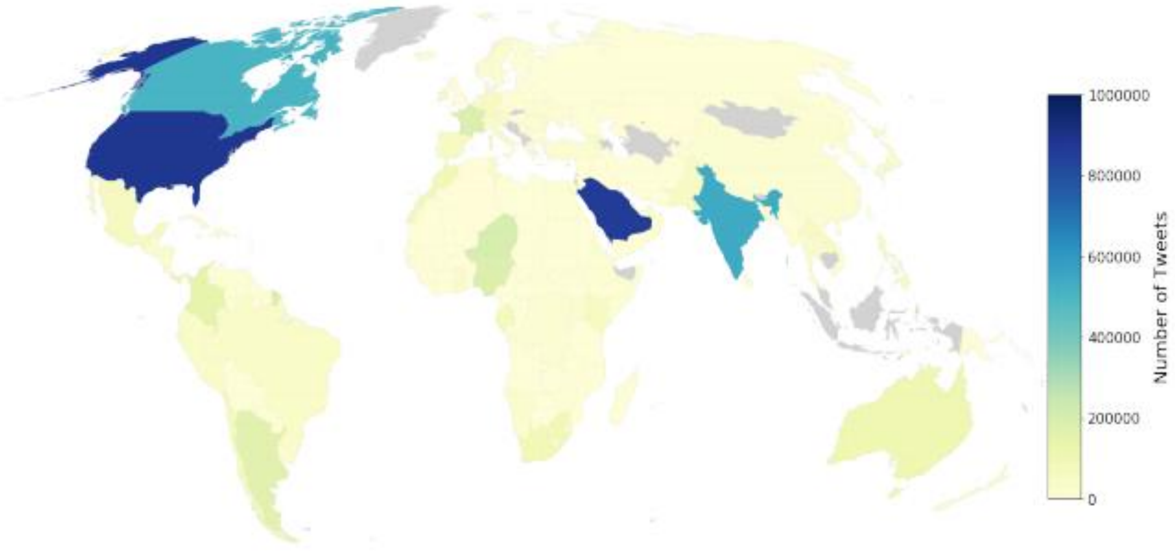


Figure 1: Global distribution, frequency and geographical coverage of the tweets.

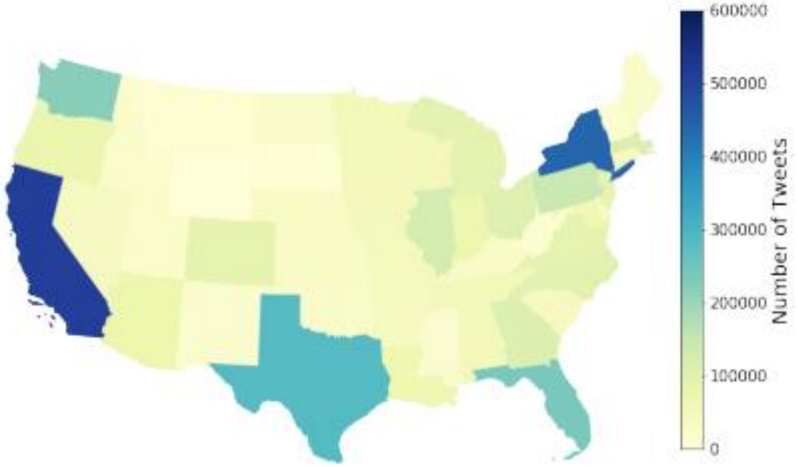


Figure 2: Distribution, frequency and geographical coverage of the tweets in the mainland of US.

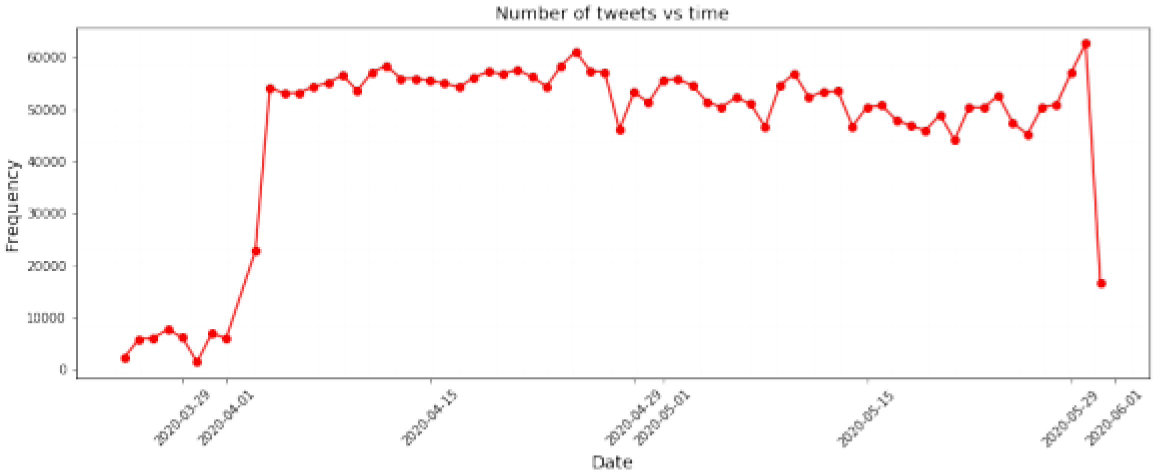


Figure 3: Frequency of tweets related to COVID-19 per day from March to June 2020.

Table 3: Examples of three topics, the top ten most representative words and their weights.

Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights
covid	62819.6	covid	82813.9	covid	86130.0
case	13357.3	coronaviru	21099.4	peopl	18270.8
new	10122.1	#covid	12967.9	coronavir	14549.7
coronaviru	10029.4	do	10428.6	get	13465.0
test	6716.8	caso	8984.3	like	11891.6
#covid	6676.0	di	8884.3	death	11731.4
death	6579.8	da	8110.9	go	10632.7
updat	4990.1	si	7257.3	test	10174.4
via	4887.8	#coronavirus	6097.0	one	9492.1
report	4828.7	com	5698.3	us	9401.0

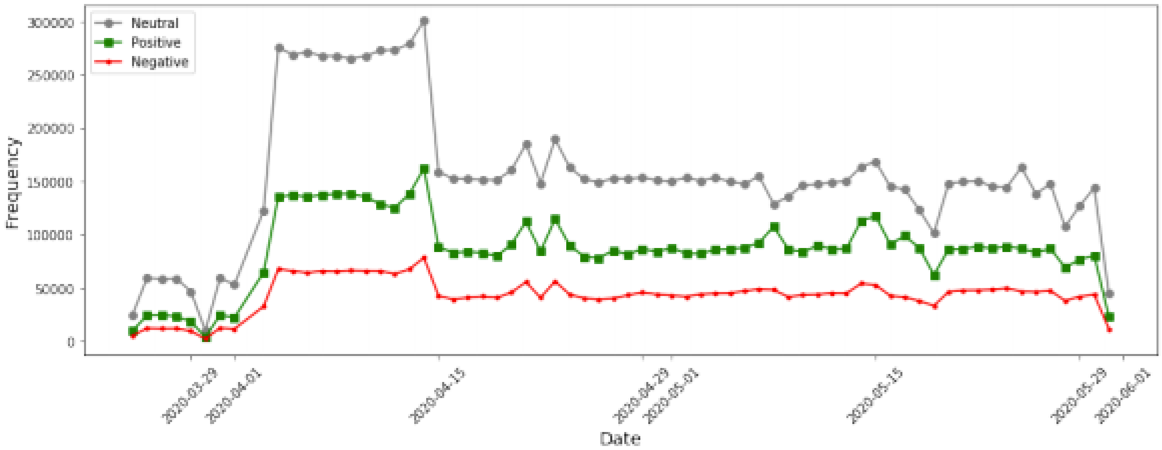


Figure 4: Daily evolution of sentiment of tweets by frequency.

4. Conclusions

Most COVID-19 related tweets sampled are positive or neutral, reflecting the public attitude towards the pandemic and pandemic related topics.

Though sentiment analysis has its limitations with large tweet corpora, we do believe, like others, that there is some utility in understanding top-level sentiment of these data [10].

References

[1] Brennen, J. Scott, et al. "Types, sources, and claims of COVID-19 misinformation." Reuters Institute 7 (2020): 3-1.

[2] Ortiz-Ospina, Esteban. "The rise of social media." Our World in Data 18 (2019).

[3] Singh, Lisa, et al. "A first look at COVID-19 information and misinformation sharing on Twitter." arXiv preprint arXiv:2003.13907 (2020).

[4] Broniatowski, David A., et al. "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate." American journal of public health 108.10 (2018): 1378-1384.

[5] Gomes, Marcelo FC, et al. "Assessing the international spreading risk associated with the 2014 West African Ebola outbreak." PLoS currents 6 (2014).

[6] Petersen, Eskild, et al. "Rapid spread of Zika virus in the Americas-implications for public health preparedness for mass gatherings at the 2016 Brazil Olympic Games." International Journal of Infectious Diseases 44 (2016): 11-15. 5CML-COVID A PREPRINT

[7] Crook, Brittani, et al. "Content analysis of a live CDC Twitter chat during the 2014 Ebola outbreak." Communication Research Reports 33.4 (2016): 349-355.

[8] Fu, King-Wa, et al. "How people react to Zika virus outbreaks on Twitter? A computational content analysis." American journal of infection control 44.12 (2016): 1700-1702.

[9] Cinelli, Matteo, et al. "The covid-19 social media infodemic." arXiv preprint arXiv:2003.05004 (2020).

[10] Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M. Mohammad. "Sentiment analysis of short informal texts." Journal of Artificial Intelligence Research 50 (2014): 723-762.

[11] Gruzd, Anatoliy

[12] Grefenstette, Gregory. "Tokenization." In *Syntactic Wordclass Tagging*, pp. 117-133. Springer, Dordrecht, 1999.

[13] Hu, Yuening, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. "Interactive topic modeling." *Machine learning* 95, no. 3 (2014): 423-469.

[14] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

[15] Feldman, Ronen. "Techniques and applications for sentiment analysis." *Communications of the ACM* 56, no. 4 (2013): 82-89.

Appendix

Note to the Editor:

This work is not ready to be published. While the idea is interesting, it fails to effectively communicate those ideas. The structure of the report defies convention, making it unnecessarily inaccessible for readers. The main ideas need to be developed and introduced effectively in order to give this report more of a focus.

The abstract is incomplete; it does not describe the main results or conclusions. The introduction provides reasonable background and presents the need for this study well, but there is no clear objective stated. The methods are reasonably described and reproducible within the confines of Twitter's Terms & Conditions, but some of the languages needed to be changed in order to make them more accessible for readers without a background in this area.

Once the results were separated from the methods, it was clear that the information is not very focused, likely due to the lack of clear objectives. There is unnecessary information about tweet location and language that could be made shorter or in some cases removed altogether to improve clarity. There is no conclusion section and there is no interpretation of the results. There is some discussion of the limitations of the methods, but it is very brief. Overall, it is clear that this work is not fully developed. It lacks a clear main idea and fails to communicate in an effective manner and therefore would be unlikely to benefit your audience.

Notes for the Authors

1) Title
   a) The title is not very informative, particularly to a broad audience that is likely unaware of the terms used. Change it to something that will grab the attention of the audience and accurately describe what the work is about.
2) Abstract
   a) The abstract section should contain a summary of the main results of this study and the interpretation of these results.
   b) It should communicate the main objective of the report as well.
3) Introduction
   a) This first paragraph of the introduction may benefit from focusing on talking about COVID-19. Providing more background about the disease and the severity of the pandemic will be helpful for providing future readers with context.
   b) Is the sentence "Social media data related to the COVID-19 pandemic can be used to study: the impact of social networks on health info-/mis-information, (2) how misinformation diffusion and spreading can influence behavior and beliefs and (3) the effectiveness of COVID-19-related actions and campaigns deployed by agencies and governments at global and local scales " necessary? Does it relate to the main objective of this paper? If it does it needs to be made more clear why and made more concise.
   c) The most essential change needed is the addition of an objective or aims for the analysis. Readers expect to be able to find the objective and reason for it at the end of the introduction. In this case, the aims need to be added to help create a central focus for the paper, but failing to meet the audience's expectations also indicates a major communication problem.

4) Methods
    a) The methods section is combined with the results and the limited conclusions. This is a problem because it impacts how well you will be able to communicate with your audience. The failure to follow the conventional style will make it more difficult for readers to find the information they are looking for.
    b) Some of the terms in the methods need to be defined in order to make the report is more accessible to a wider audience. These terms include concatenating, text tokenization, topic modeling, sentiment analysis, and Latent Dirichlet Allocation.
    c) For reproducibility, describe how you sampled the tweets used in the topic modeling. Was it random sampling?
5) Results
    a) The results lack focus, likely due to the absence of a clear objective. The topic modeling and sentiment analysis seem like they should be the main focus of the paper but they are perceived as an afterthought because of how the data is presented and the focus on less relevant figures.
    b) Among the tables and figures, there is a large focus on language and location. Describing language and location does not seem to be the purpose of this report but they take up a lot of space and may distract from what is more important.
    c) In table 1, the language column should be before the language code to improve clarity.
    d) In table 2, you should remove the undefined category from the table. It distracts from what this table is trying to indicate. It is also unclear whether the report benefits from this table
    e) Figure 1 is not necessary; it does not show much information and it is unclear how it is important to this report
6) Discussion
    a) Any conclusions and interpretations of the results are largely absent in this report. The conclusion section should restate the main idea and main results. It should reflect on what those findings mean in the context of the current literature, and discuss the limitations of the study. It should end by discussing the future directions for research and very briefly summarize the main takeaways.
    b) There is not much that I could change in the conclusion section since there was so little author interpretation. This should be written considering what the main objective was and how the authors interpreted their tweet analysis
7) References
    a) Be sure to label the references.