UNDERSTANDING THE PERCEPTION OF COVID-19 POLICIES BY MINING A MULTILANGUAGE TWITTER DATASET

Christian E. Lopez[1,2*], **Malolan Vasu**[1] and **Caleb Gallemore**[3]

[1] Computer Science Department, Lafayette College, Easton, PA 18042
[2] Mechanical Engineering Department, Lafayette College, Easton, PA 18042
[3] International Affairs Program, Lafayette College, Easton, PA 18042

Abstract

The objective of this work is to explore popular discourse about the COVID-19 pandemic and policies implemented to manage it. Using Natural Language Processing, Text Mining, and Network Analysis to analyze corpus of tweets that relate to the COVID-19 pandemic, we identify common responses to the pandemic and how these responses differ across time. Moreover, insights as to how information and misinformation were transmitted via Twitter, starting at the early stages of this pandemic, are presented. Finally, this work introduces a dataset of tweets collected from all over the world, in multiple languages, dating back to January 22nd, when the total cases of reported COVID19 were below 600 worldwide. The insights presented in this work could help inform decision makers in the face of future pandemics, and the dataset introduced can be used to acquire valuable knowledge to help mitigate the COVID-19 pandemic.

Link for dataset: https://github.com/lopezbec/COVID19_Tweets_Dataset

1 Introduction

The Coronavirus Disease of 2019, known as COVID-19, is a rapidly spreading disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2). The COVID-19 is now considered a pandemic that has affected countries in all inhabited continents. Since the first cases of COVID-19 reported in Wuhan, China, in December 2019, the number of fatalities worldwide has increased rapidly. Due to its high infection and death rate, governments have implemented a wide range of policies aimed at mitigating the spread of this virus and its impact. Such actions began with the Chinese government order to quarantine Wuhan on January 23rd, 2020, to, most recently, multiple countries declaring state of emergency and implementing strict quarantine and social distancing measurements (e.g., US, Italy, Argentina, Spain).

*Corresponding Author. 569 Rockwell Integrated Science Center, Lafayette College, Easton, PA 18042, lopezbec@lafayette.edu

Most government leaders have implemented measures to incentivize, and in some cases enforce, "social distancing" to reduce the spread of COVID-19. These measures have resulted in the cancelled entertainment events, closures of schools and colleges, and businesses reducing hours of operation, implement telecommuting, or close altogether. There is no doubt the pandemic and the measures set in place to mitigate it have and will continue to drastically impact the lives of millions. As this pandemic and the responses to it are unprecedented, however, we are likely to be surprised by how people respond.

Since the early stages of the disease, people have expressed their opinion and shared information, as well as misinformation, about it via social media platforms, such as Twitter. As COVID-19 spreads to other countries and governments try to mitigate its impact by implementing counter measures, people have also used social media platforms to express their opinion about the measures themselves, the leaders implementing them, and the ways their lives are changing. The use of social media, such as Twitter, as platforms to express opinions and share information about COVID-19, will only continue to grow, precisely because of the "social distancing" measures set in place to mitigate it.

Policymakers could mine this social media data to explore popular discourse about the pandemic and the measures set in place to mitigate it. We plan to analyze a corpus of tweets that relate to COVID-19 with the objective of identifying common responses to the pandemic and how these responses differ across time, countries, and policies. Moreover, insights as to how information and misinformation about this pandemic and the polices are transmitted are presented. Finally, we introduce and share with the research community a dataset of tweets collected from all over the world, in multiple languages, dating back to January 22nd when the total cases of reported COVID-19 were below 600 worldwide. Here, we describe and present descriptive statistics of this dataset, and explain our data collection methods and intended analyses.

2 Dataset Description

The dataset presented is being continuously collected using the Twitter API. The dataset presented here (v1.5) covers Jan 22, 2020 to Apr 18, 2020 and contains 65,388,324 tweets. The keywords used for search tweets are: *virus* and *coronavirus* since January 22, 2020, *ncov19* and *ncov2019* since

February 26, 2020, and *covid* since March 7, 2020. The number of tweets collected for each keyword is presented in Table 1.

The average daily number of tweets collected on dataset v1.5 was 207,581.98 (SD=191,064.93, Mdn=0). In the first few months, the number of tweets collected increased steadily from 724,877 in January and 2,994,768 in February, 27,414,279 in March, to 34,254,400 in April. Table 2 shows the summary statistics for the daily number of tweets collected each month and on the first 18 day(s) of April.

Table 1: Breakdown by keyword

Keyword	Number of Tweets	Percentage
coronavirus	22,823,353	34.90%
virus	22,059,400	33.74%
covid	20,342,766	31.11%
ncov2019	82,870	0.13%
ncov19	79,935	0.12%

Table 2: Statistics on number of daily tweets per month

Month	Mean	SD	Median
Jan	36,243.85	11,699.56	0
Feb	47,536.0	18,136.58	0
Mar	193,058.3	167,347.42	0
Apr	380,604.44	82,262.93	0

Table 3: Distribution of Tweets per Language

Language	Number of Tweets	Percentage
English	38,455,759	59.41%
Spanish; Castilian	11,203,759	17.31%
Portuguese	4,309,676	6.66%
Bahasa	2,684,019	4.15%
French	2,514,900	3.89%
Italian	880,907	1.36%
Thai	829,668	1.28%
Japanese	663,834	1.03%
Hindi	535,609	0.83%
Turkish	483,823	0.75%
Catalan; Valencian	401,929	0.62%
Tagalog	383,925	0.59%
German	371,524	0.57%
Dutch; Flemish	182,799	0.28%
Other	831,355	1.28%

While the dataset contains tweets from 63 languages, only English-language tweets were collected from 22 January to 30 January 2020. English-language tweets remain the most prominent in the dataset accounting for 59.41% of the total. Figure 1 present the distribution of collected tweets by language for the keyword *coronavirus*.

A basic sentiment analysis was performed on the dataset using a bag of words approach, where an existing sentiment

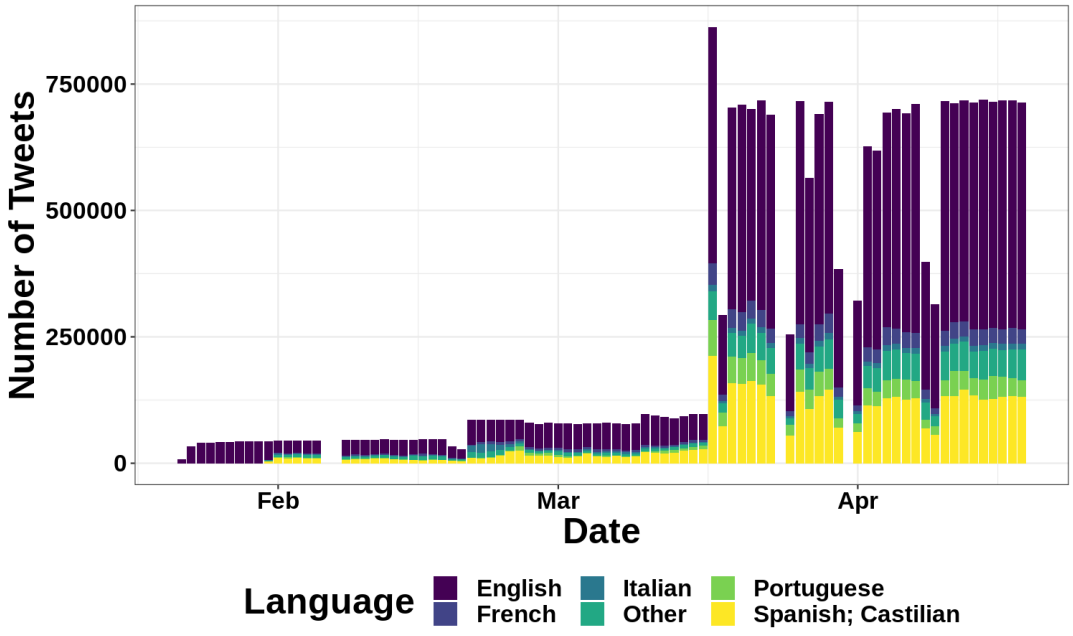

Figure 1: Collected tweets, by language, for the keyword *coronavirus*.

lexicon was used to look up sentiment values for each individual word of a pre-processed tweet. Figure 2 presents the average sentiment for tweets over the observation period, grouped by keyword.

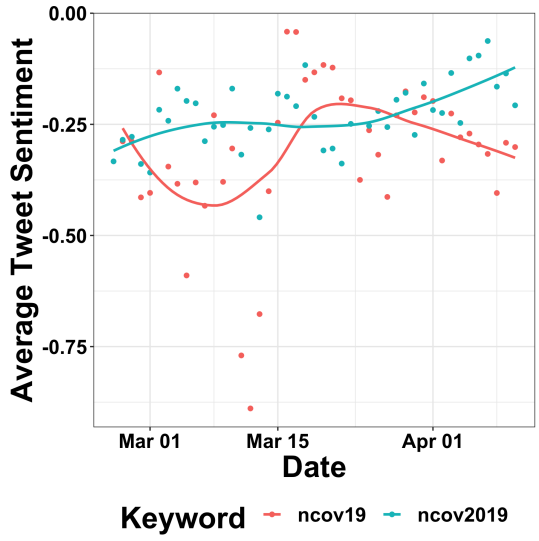

Figure 2: Average sentiment of tweets by keyword.

Information about retweets and likes was also collected. Figure 3 presents the distribution of collected tweets' average number of retweets over the observation period for the keyword *coronavirus*. While the overall level of retweeting appears to have declined in February, retweets rose abruptly as the crisis ramped up in Europe in late February and early March. Since mid-March however, the overall level of retweeting seems to be on a gradual decline.

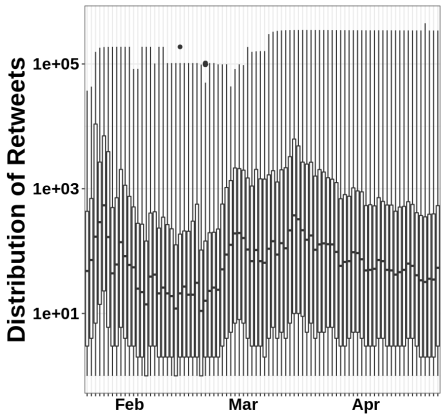

Figure 3: Distribution of retweet numbers, on a log10 scale, for the keyword *coronavirus*.

Only a relatively small percentage of the collected tweets contain geolocations. Figure 6 presents the locations observed as of Apr 18, 2020 for the keyword *coronavirus* . Figures 4 and 5 present information about the number of hashtags and mentions used on average in a tweet by day for

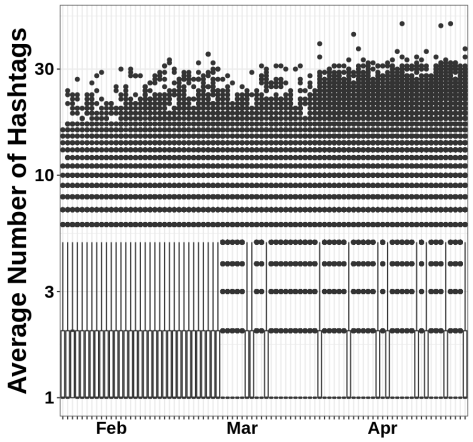

Figure 4: Distribution of the average number of hashtags for each tweet, on a log10 scale, for the keyword *coronavirus*.

the keyword *coronavirus* respectively. It is apparent that the tweets, disregarding the outliers, generally use very few mentions and hashtags. Table 4 provides numerical summaries of the observed tweets number of retweets, likes, mentions, hashtags, and geolocation information.

Table 4: Summary statistics for collected tweets

Number of Geolocated Tweets	23,667
Maximum Observed Retweets	1,064,693
Maximum Observed Likes	8732

3 Dataset Accessibility

The dataset v1.5 was realized on Apr 18, 2020. The dataset is available on Github at: https://github.com/lopezbec/COVID19_Tweets_Dataset. The dataset is released in compli-

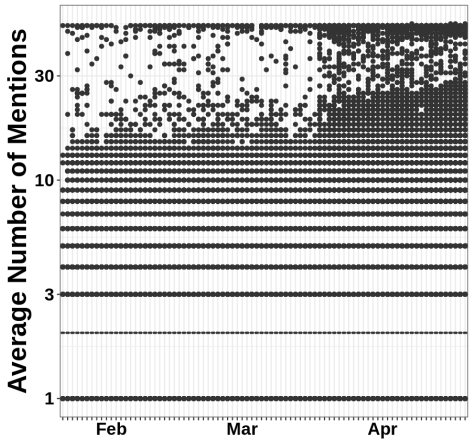

Figure 5: Distribution of the average number of mentions for each tweet, on a log10 scale, for the keyword *coronavirus*.

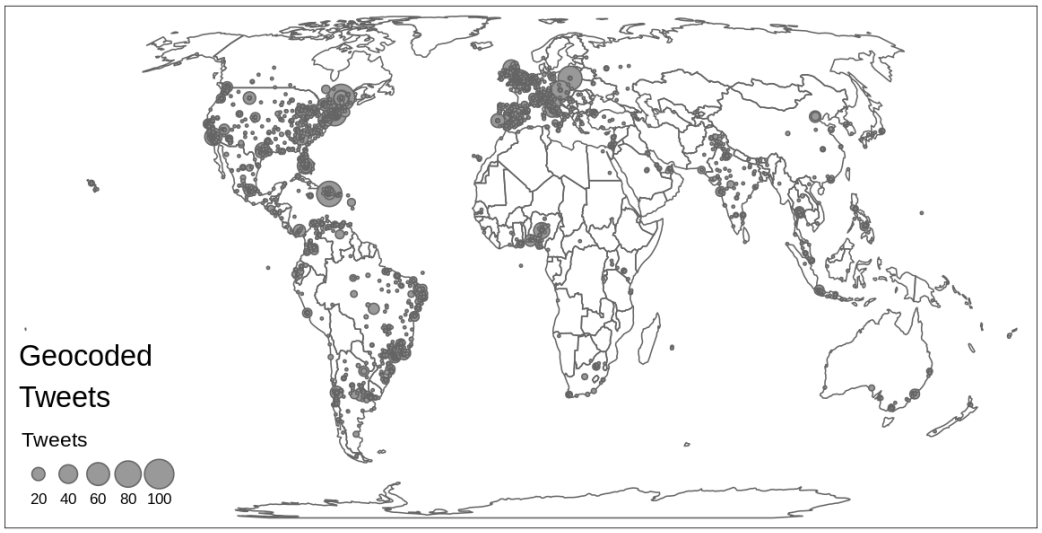

Figure 6: Geolocated tweets as of Apr 18, 2020 for the keyword *coronavirus*.

ance with the Twitter's Terms & Conditions. Hence, only the tweets-IDs are made available to researchers. However, using the Twitter API the tweets can be "rehydrated" and the data of tweets that have not been deleted can be accessed (more details on the GitHub page). This dataset is still being continuously collected and routinely updated. If you have technical questions about the data collection, please contact the corresponding author.

4 Future Work

This research note's main objective was to introduce and share with the research community a dataset of tweets related to the COVID-19. We are continuously collecting and routinely updating the dataset. Similarity, we will be using Natural Language Processing and Text Mining, and Network Analysis to analyze the corpus of tweets to identify common popular responses to the pandemic and how these responses differ across time. Moreover, with this dataset we will explore to how information and misinformation about COVID-19 is transmitted via Twitter.