Exploration of COVID-19 Twitter Data

Megan Parsons

Introduction

Coronavirus disease 2019 (COVID-19), caused by the novel coronavirus SARS-CoV-2, is an acute respiratory infection first reported in Wuhan, China, in late 2019.[1] In the United States, public perception of the resulting pandemic has been shaped by a myriad of factors, including the federal response; official communications from government agencies; socio-demographic variables; state-level policies; partisan news media; and social media.[2], [3] The proliferation of misinformation in particular is detrimental to efforts in containing the spread of the virus, and it has been linked to effects such as vaccine hesitancy and reduced likelihood of compliance with public health guidelines.[4] Better understanding the nature and etiology of COVID-19 public perception can aid in the development of more effective strategies to produce, target, and disseminate health information to diverse populations. Therefore, the goal of this project is to analyze publicly available Twitter data using sentiment analysis techniques and network analysis methods to generate insights into public perception of COVID-19 in the United States.

In this paper, we discuss observations from the raw dataset; detail the steps of the preprocessing pipeline; characterize the final dataset; and begin to probe questions regarding public perception of COVID-19 misinformation. Because misinformation is somewhat imprecisely defined in the academic literature, we chose to focus our analysis on hydroxychloroquine, which was introduced as a twitter hashtag during initial data collection and was widely touted as a potential COVID-19 therapeutic despite myriad study limitations and methodological concerns surrounding the initial supporting studies.[5] We conclude with future directions for this analysis.

Understanding the possible relationship between microblogging communications and public perception may provide insight into the types of communications that increase compliance with government recommendations and public health guidelines. Furthermore, characterizing “misinformation” and tracking its propagation through social networks can inform strategies to prevent the spread of inaccurate information during public health crises.

Characterization of Raw Data

This dataset is composed of publicly-available Twitter data (Tweet IDs) associated with SARS-CoV-2 from late January through June 2020.[6] The Tweet IDs are sourced from the GitHub repository *COVID-19 Tweet IDs*, which used both the Twitter streaming API and search API to gather historic and real-time Tweet IDs according to specific COVID-19 keywords and user accounts of interest. Although this dataset is volume-limited at the time of stream/search and subject to fluctuations in internet quality, it provides almost 200 million COVID-19 Tweets to analyze.

Tracked keywords and accounts used to build the dataset evolved over time. Every keyword or account addition was retroactive for up to one week using the Twitter search API. Keywords included ‘coronavirus,’ ‘wuhanlockdown,’ ‘PPEshortage,’ and

'DuringMy14DayQuarantine,' and tracked accounts included 'CoronavirusInfo,' 'CDCemergency,' and 'WHO.'[6], [7] We observe a slight delay in the addition of keywords after nomenclature is formally introduced, which suggests that the earliest Tweets captured in the dataset under a particular keyword might not be representative of some of the earliest topical Tweets. For instance, the World Health Organization (WHO) released a statement on 11 February 2020 declaring official names for the virus that causes the novel coronavirus disease as well as the disease it causes.[8] These terms (COVID-19 and SARS-CoV-2) were not added to the keyword list until 16 February 2020 and 06 March 2020, respectively, and some derivatives of these terms (COVID‒19, COVID__19) were not added until later.[7] For a list of the most up-to-date keywords and accounts included in this dataset, please reference Appendix A.

This dataset was housed on a remote server and transferred using Secure/SSH File Transfer Protocol (SFTP).[1] The file transfer was facilitated by FileZilla Client, which supports simultaneous multi-file transfer and resume transfer.[9] Basic preprocessing to access the Tweet data was completed to analyze the entire "raw" dataset, which ranges from 21 January 2020 to 30 June 2020, and is representative of all original locations and languages in the data.

Characteristic	Total
Number of Tweets	192,140,934
Keywords Used for Data Collection	75
Accounts Followed for Data Collection	9
Dates Tracked	21 Jan 2020 - 30 Jun 2020
Languages Represented	65

Table 01. Characteristics of Raw COVID-19 Dataset

The Tweet IDs were hydrated using Twarc, resulting in Tweet objects stored in gzipped JSON Lines files.[10] The files were organized into folders by month and the naming convention encoded the hour (in UTC) that the Tweet was created (COVID-19-TweetIDs-master / {year}-{month} / coronavirus-tweet-id-{year}-{month}-{day}-{hour}). There are a total of 192,140,934 Tweets in this dataset.

Plotting the total Tweets per month from January to June 2020 reveals that the most Tweets were posted in June of 2020 (over 80 million of the 192 million total). The fewest Tweets were posted in January (only 7,708,262), but this is expected because data collection did not occur until later in the month and the virus hadn't spread widely yet.

[1] We would like to thank the Stringhini lab for facilitating access to this dataset.

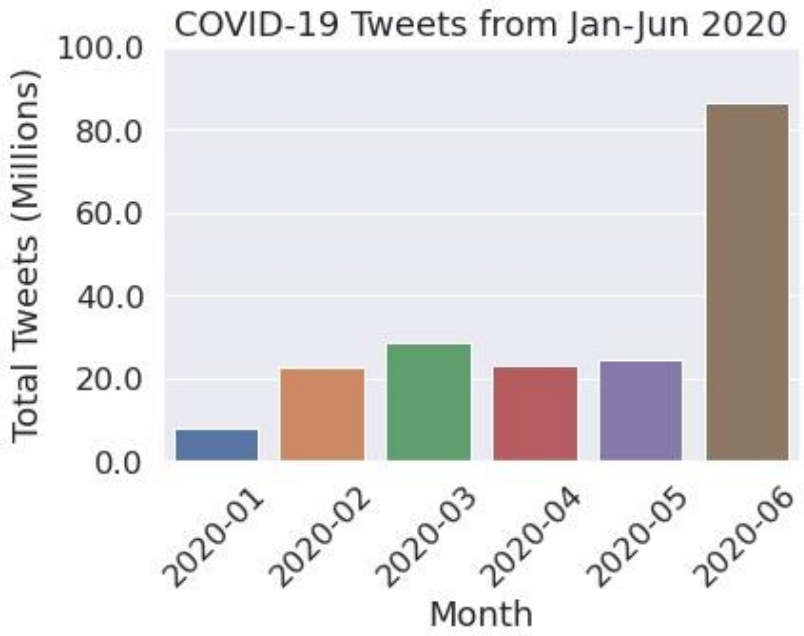

Figure 01. Tweet count per month from 21 January 2020 until 30 June 2020.

Tweets in this dataset are classified as one of 65 languages or 'undefined'. The most common Tweet language is English (63.56%), which suggests that COVID-related events that unfolded in countries where the primary language is not English may be underrepresented in these Tweets.

Language	Language Code	Total	(%)
English	en	122131168	63.56%
Spanish	es	27832348	14.49%
Portuguese	pt	7316162	3.81%
Indonesian	in	5185312	2.70%
French	fr	4705012	2.45%
Undefined	und	4486085	2.33%
Japanese	ja	4042086	2.10%
Thai	th	2490747	1.30%
Hindi	hi	2333008	1.21%
Italian	it	2078577	1.08%

Table 02. Most common Tweet languages represented in the raw COVID-19 Twitter dataset. This table displays the language; the ISO code; the number and percentage of associated Tweets.

Figure 02. The top 100 locations in this dataset based on parsing of user-provided profile location information.

There are several ways to elucidate a Tweet's location of origin: (1) geo-tagging; (2) profile location; and (3) location data given in the Tweet text itself. Twitter allows users to geo-tag Tweets at the time of posting, which adds latitude and longitude data to the Tweet object, but this option is only implemented by an estimated 0.85% to 2.0% of users.[11], [12] Figure 02 was generated by decoding the approximate coordinates of the user profile location using GeoPy, then visualizing the results on an interactive map using Folium.[13], [14]

These may not necessarily be representative of the top locations where COVID-19 Tweets have originated. This is often due to the flexibility of user inputs resulting in variability in the raw location string (e.g., 'United States' and 'USA' are considered separate locations in the count). It was observed, however, that variability also arose though subtle variations in character or diacritic usage ('Mexico' vs. 'México'), the inclusion of differing combinations of city, state, and/or country information ('San Francisco, CA' vs. 'San Francisco, CA, USA'), and the inclusion or exclusive use of flag emoji ('Australia' vs. 'Australia 🇦🇺' vs. '🇦🇺'). We use this information to develop a location processing pipeline used in later visualizations.

Preprocessing Pipeline

Due to Twitter's Terms of Service (TOS), the dataset from Chen, et. al., is provided as a collection of Tweet IDs.[7] It is important to note that the process of hydrating Tweets, or extracting the Tweet object from the Tweet ID, is lossy, so deleted or protected Tweets are no longer accessible through a Twitter API call. In version 1 of this dataset, it was estimated that ~6% of the Tweets were inaccessible at the time of hydration.[7]

Our aim was to create a dataset of English tweets originating from the United States. To achieve this, we captured all tweets with the English language tag ('lang' = 'en') and used that dataset to identify the best way to efficiently filter the Tweets based on location.

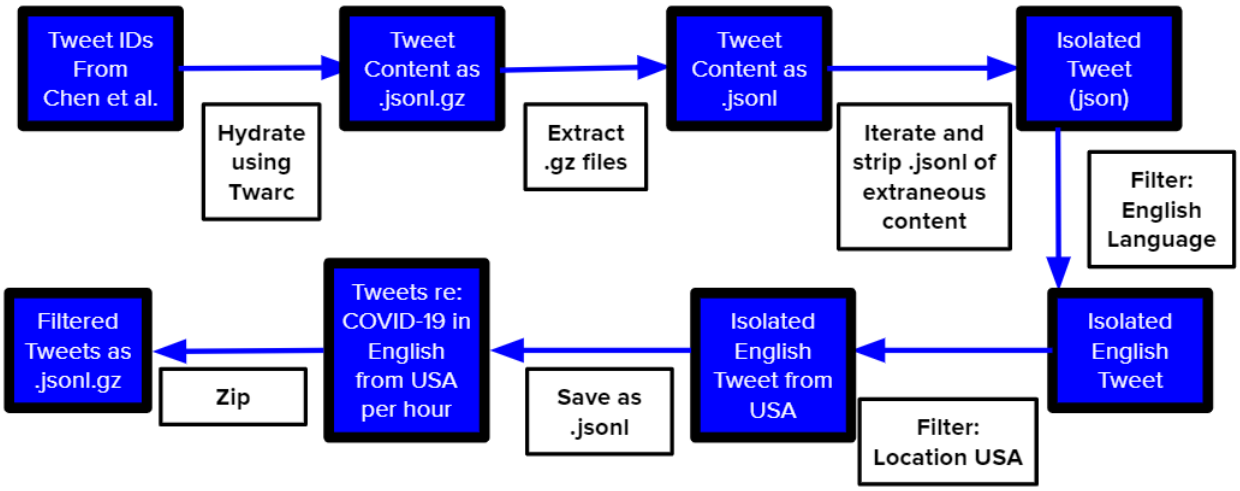

Figure 03. Preprocessing Pipeline. To create the finalized dataset of English Tweets likely originating in the USA, we formatted the data and applied several filters.

As was mentioned, there are several types of location information encoded in Tweet object. Ideally, we would assign the highest probability location per Tweet by integrating results from several methods. Some Tweets, for instance, have geotagged coordinates (longitude and latitude) associated with the Tweet object. There are ways to convert these coordinates to country, state, county, city, or increasingly granular data; however, the most sophisticated methods require the use of APIs like the Google Maps API, which provide unparalleled functionality and accuracy at a monetary cost. Unfortunately, even processing a small subset of the Twitter dataset would quickly exceed the monthly credit offer provided per billing account.[15] Additionally, although geotagged Tweets provide the most accurate geolocation available in the Tweet object, they only account for approximately 0.85% to 2.0% of users.[11] To circumvent this issue, we filtered the data using the Carmen Library.

Carmen Library

Carmen is a library developed at Johns Hopkins University for geolocating tweets for public health analysis.[16], [17] Given a tweet, Carmen will return Location objects that represent a physical location. Carmen uses both coordinates along with other information in a tweet to make geolocation decisions. This library is designed to infer locations from place, coordinate, and user profile information along an Earth → Country → State → County → City hierarchy with the use of frequency statistics and known aliases.[17] This library allowed us to isolate more Tweets for our USA dataset than filtering based on explicitly available Tweet JSON information alone.

Parallelization Strategy
Using the Carmen library would allow us to create a more meaningful dataset by integrating more information into the location filter; however, the program would take weeks to run. To expedite this process, we ran a series of experiments using the Shared Computing Cluster (SCC) on one day's worth of data using 1, 4, and 8 cores. We noted the timing of these tests and set up a job array to process the data more efficiently.

Logging & Troubleshooting
To better troubleshoot issues with the code, we implemented a logging module. In addition, efficiency of the pipeline was enhanced through the automatic creation of datetime formatted results subdirectories; security was elevated by following API key storage best practices; and exception handling was formalized throughout the code.

Finalized Dataset

The finalized dataset consisted of 17,989,581 English Tweets likely originating from the United States. These data represent approximately 8% of the original dataset. We have characterized this dataset prior to further analysis.

Characteristic	Total
Number of Tweets	17,989,581
Keywords Used for Data Collection	75
Accounts Followed for Data Collection	9
Dates Tracked	21 Jan 2020 - 30 Jun 2020
Languages Represented	English ('en')

Table 03. Characteristics of final COVID-19 dataset

We hypothesize that January and February 2020 will have the lowest volume of Tweet data per unit time containing COVID-19 keywords. English is the most prominent language captured in this dataset (representing over 60% of all Tweets as of version 9), but non-English speaking countries were most affected in the early months of the pandemic.[6] In addition, the WHO hesitated to declare COVID-19 a pandemic until almost mid-March, so we predict that Tweet volume will begin to rise in March 2020.

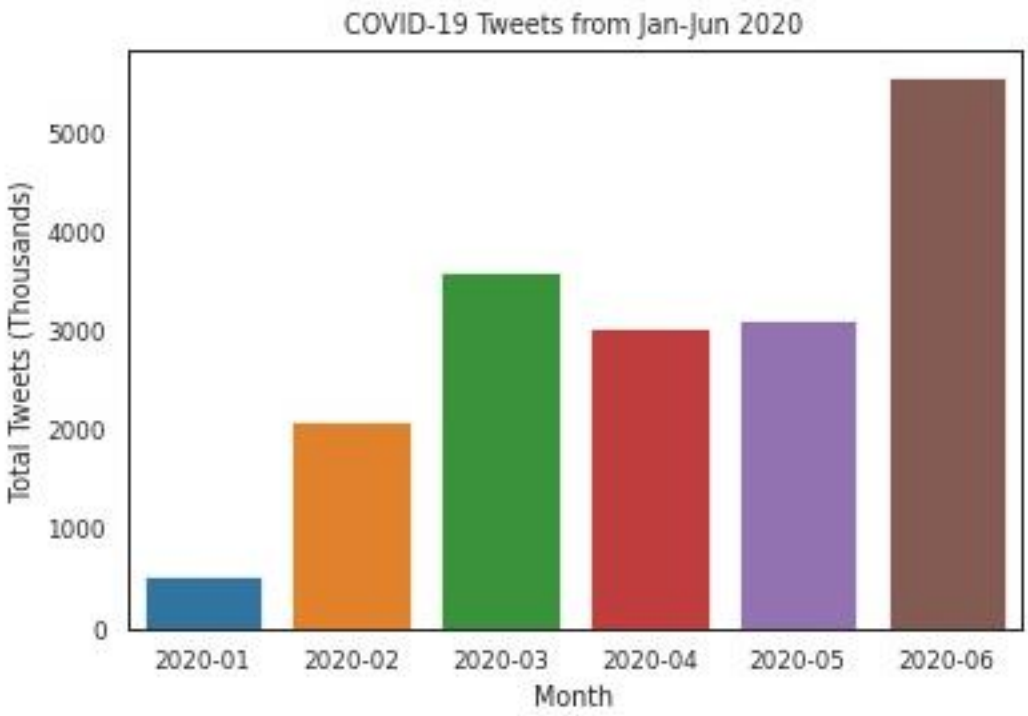

Figure 04. This barplot shows the total number of COVID-19 Tweets per month originating from the USA once non-English Tweets were excluded.

We see in Figure 04 the expected rise in Tweet activity in March of 2020, as well as the relatively lower levels of Tweet activity regarding COVID-19 prior.

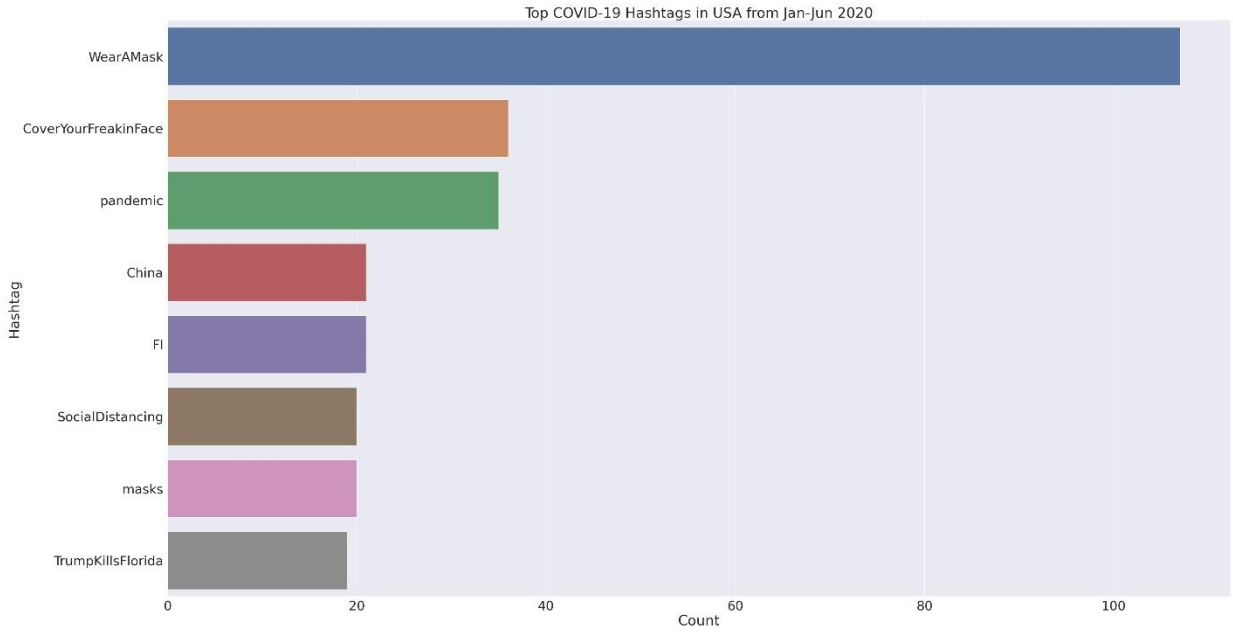

Figure 05. Most common Tweet hashtags in the USA from January to June 2020. This list from the top 15 hashtags excludes all hashtags containing strings ‘covid’ or ‘coronavirus’.

Hashtags are user-generated entities that give context or provide keywords for a Tweet. From our COVID-19 dataset, we determined the top 15 hashtags used from January – June 2020, then excluded hashtags containing the strings ‘covid’ or ‘coronavirus’ (because that is the basis of the dataset). The resulting eight hashtags were:

#WearAMask
#CoverYourFreakinFace
#pandemic
#China
#Fl
#SocialDistancing
#masks
#TrumpKillsFlorida

If we were to organize these hashtags into topics, half of them pertain to public health efforts to stop the spread of the virus, two of them are neutral, and two specifically reference Florida. Florida was notably an ‘epicenter of COVID-19 cases.’[18] We anticipate that the counts on these hashtags would be much higher with further consolidation and processing of the hashtag text.

Using what we observed from characterizing the raw COVID-19 data, we know that the heterogeneity with which location data is reported in Tweet objects can make classification difficult. We developed a location processing pipeline to standardize user profile location data, which was abundant in our dataset.

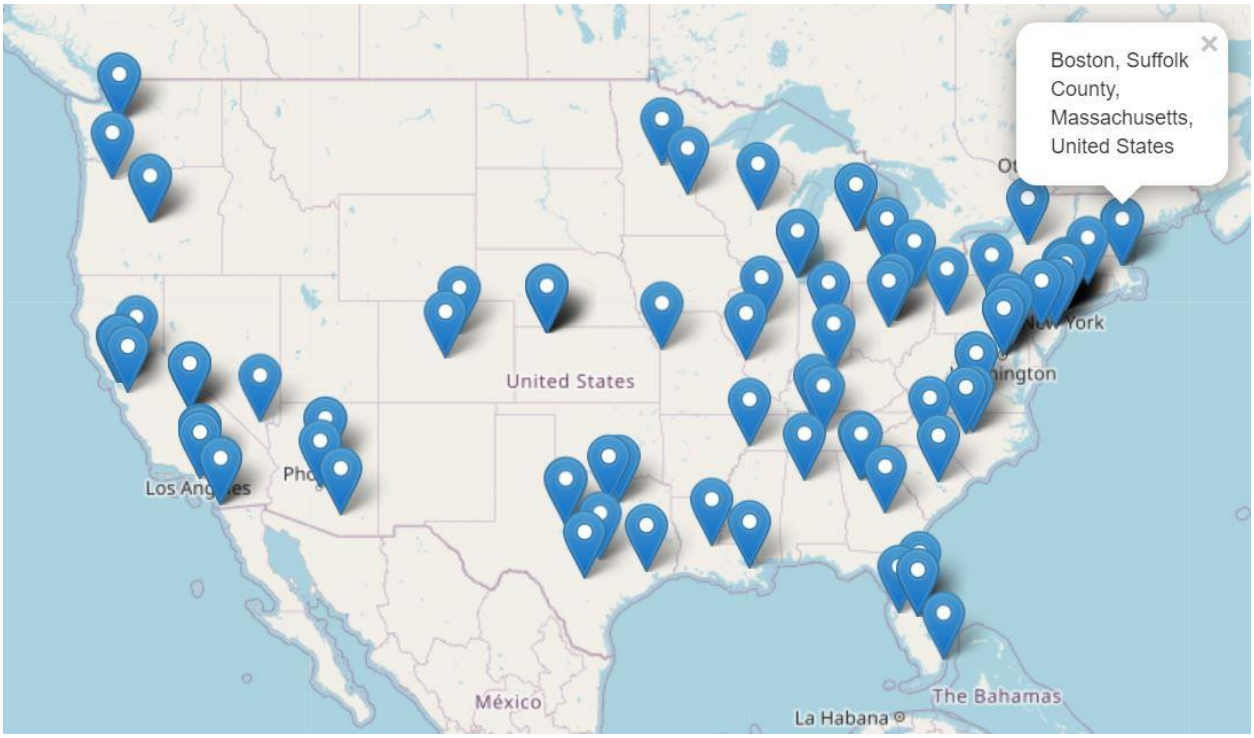

Figure 06. Top 100 user profile locations in the USA mapped on an interactive map**.** Example of user click shown on the Boston, MA pin.

Prior to ranking the locations in this data, we first removed all emoji; made the text lowercase; replaced all accented characters using Unidecode (which has manual character mappings and more robust/intuitive handling of diacritical marks); and removal of extraneous whitespace.[19] A count of the preprocessed locations resulted in fewer misclassifications despite the heterogeneity in expression of user profile location data. We then used GeoPy to classify this processed data into standardized geolocations and store it using a Python dictionary.

From our work, the resulting dictionary of (key, value) = (standardized geolocation information, [processed user profile locations]) pairs characterizing the data prompted the question: If we were to flip the key, value pairs and consolidate the dictionary, would it reveal the number of user locations of the same origin that are not classified as such? We ran this mini-experiment, which revealed that our top 100 list of locations from the Twitter data was actually a top 72 list: almost 30 locations were still classified as distinct from their correct geolocation classifications, even despite the preprocessing and standardization.

In the future, because the use of flags to indicate countries is so common in the user profile data, we could write a small function to map flags to geolocation data and incorporate those data points into our analysis. Accurately geolocating Tweets is an active area of investigation, especially because of its applications to public health and early warning systems.[20]

Sentiment Analysis using the Google Natural Language API

Data Cleaning

In preparation for sentiment analysis, we define ‘cleaning’ the data as the process of removing all Tweet information that is not likely to enrich our understanding or interpretation of the subjective thoughts and feelings expressed in the Tweet. This includes the handling of Twitter tokens, such as @username references, URLs, and #hashtags. To achieve this, we used the Python standard library Regular Expression Operations module.[21]

Properly removing emoji was a nontrivial task using the Regular Expression Operations module. Because emoji patterns are difficult to generalize and capture while reducing false positives, we use the emoji module to handle the most up-to-date listing of emojis and their aliases.[22]

Tokenization of the input text, removal of stopwords, and lemmatization was done using using the Python Natural Language Toolkit (NLTK).[23], [24] Although the goal of both stemming and lemmatization is to return the common base form of words by reducing inflected forms and derivations, we chose to implement lemmatization in our pipeline because it is based on morphological analysis of the words rather than just applying a heuristic.[25] For this reason, it is necessary to include the part of speech associated with each token in the lemmatization function.

Discussion: Data Cleaning

Our pipeline for data cleaning has a few limitations. There are often syntactic and semantic differences between Standard Edited American English (SEAE) and microblogging content due to limitations of the platform (like Twitter's enforced character limits); informal communication preferences of individual users; or an intentionally approachable or casual brand image and communication strategy. The resulting use of abbreviations or slang can be difficult to account for in existing natural language processing algorithms.

For instance, the word "lit" – the past tense and past participle of *light* – is both a common abbreviation for the word "literature" (often referring to the scientific literature in this context) and a slang term meaning "excellent" or "exciting."[26] Even though this slang definition of *lit* was added to the Merriam-Webster dictionary in January 2021, it is predated by an alternative slang definition meaning "intoxicated," which is recorded in the English lexicon as early as 1910.[27]

Original Tweet Text	**"Lit" Usage**	**Sentiment (Score, Magnitude)**	**Interpretation**
Incredibly well done **lit** review about **#hydroxychloroquine** use for COVID. Also really really cool information on how HCQ works for people like me, with #lupus, or #rheumatoidarthritis. Obv tougher to read if you're not savvy with science lingo, though!	[Abbreviation] Scientific Literature	(0.5, 2.2)	Strongly Positive
The #hydroxychloroquine black market is about to be **lit**.	[Slang] Excellent, Exciting	(0, 0)	Neutral
At my health system we havent seen enough covid folks to have anecdotal evidence, but most of the **lit** I have encountered doesn't endorse **Hydroxychloroquine** as an efficacious therapy. Nor tocilizumab. Which hopefully means my rheumatology patients will still be able to function.	[Abbreviation] Scientific Literature	(0, 0.6)	Neutral
After weeks of promoting it,Trump backed away from hydroxychloroquine faster than a kid who just **lit** a firework.	[Past Tense] Light	(-0.5, 0.5)	Slightly Negative

Dr. Anthony Cardillo **lit** up the internet with anecdotal reports of success treating #COVID19 with hydroxychloroquine + zinc Learn more about how HCQ and zinc may synergize @medmastery2 or visit http://clinicaltrials.gov to see the 50+ studies of HCQ combos https://youtube.com/watch?v=BIymfznD7YA	[Past Tense] Light	(0.4, 0.4)	Slightly Positive

Table 04. Using five tweets to illustrate the varied usage of a single term, as well as the output of my sentiment analysis code (sentiment score and magnitude) with the associated interpretation.

The sentiment analysis thresholding function could be further refined given annotated examples of interpretation based on sentiment score and magnitude, i.e., which score and magnitude combinations correspond to strongly positive, weakly positive, neutral, weakly negative, and strongly negative sentiments. These thresholds are often subjective and vary based on use case. However, because this API relies on a pre-trained model that cannot be fine-tuned, our options are limited to enhance performance. Furthermore, this API incurs a monetary cost after the first 5,000 units of analysis, so we had to be selective about which types of questions we tried to address.

In the next section, we detail how we plan to integrate sentiment analysis with Twitter network information to learn more about the spread of misinformation.

Investigating Misinformation: Analyzing Hydroxychloroquine Tweets

Misinformation is somewhat imprecisely defined in the academic literature; however, we have simplified this definition by restricting our analysis to Tweets containing the #hydroxychloroquine hashtag. We first perform some basic characterization of our data.

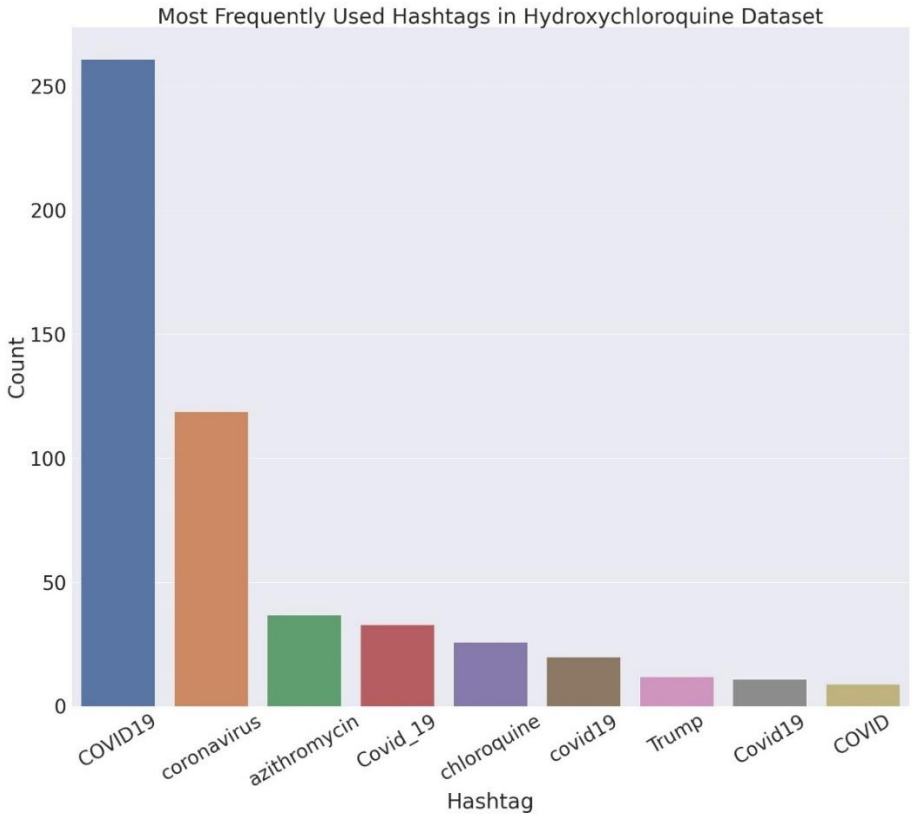

Figure 07. Top 10 hashtags associated with #hydroxychloroquine Tweets. The #hydroxychloroquine tag was removed for visualization purposes.

Hydroxychloroquine Tweets are, perhaps unsurprisingly, associated with the #COVID19 hashtag and its derivatives. This list also contains two associated drugs (chloroquine and azithromycin), as well as the hashtag #Trump. If we condense the top 20 hashtags by topic and provide updated counts from the entire #hydroxychloroquine dataset based on those topics, we can better characterize the hashtags. This process is somewhat subjective, but it is more representative of the hashtags by eliminating the distinction between terms like ‘COVID19’ and ‘coronavirus’. Topic determination from the top 20 hashtags and subsequent categorization was done manually to account for the nuance in the hashtags. Misspellings that did not confer any implied meaning (suggesting that the misspelling was intentional) were categorized as the original writer likely intended. The ‘Trump’ category represents the neutral term without any additional context to avoid ambiguities in the classifications (e.g., ‘TrumpCrimeFamily’ was not included in the count).

Topic	Count
COVID-19	475
azithromycin	41
chloroquine	27
Trump	13
Texas	7
HCQ	7
Cornyn	6
NYC	5
lupus	9
UnitedKingdom	4
zinc	3

Table 05. Top 20 hashtags were manually reduced to categories ('topics') and the counts were updated to reflect themes from the entire set of hashtags.

We suspect that this erroneous idea that hydroxychloroquine might be – alone or in combination with #azithromycin, #zinc, etc. – effective against SARS-CoV-2 might be something that largely propagates through social networks through Re-Tweets. That is, perhaps this information is generated by a relatively small number of users but widely propagated or shared through social media networks, thereby gaining popularity. We can begin to investigate this idea by distinguishing between these hashtags that are shared through original Tweets vs. Re-Tweets.

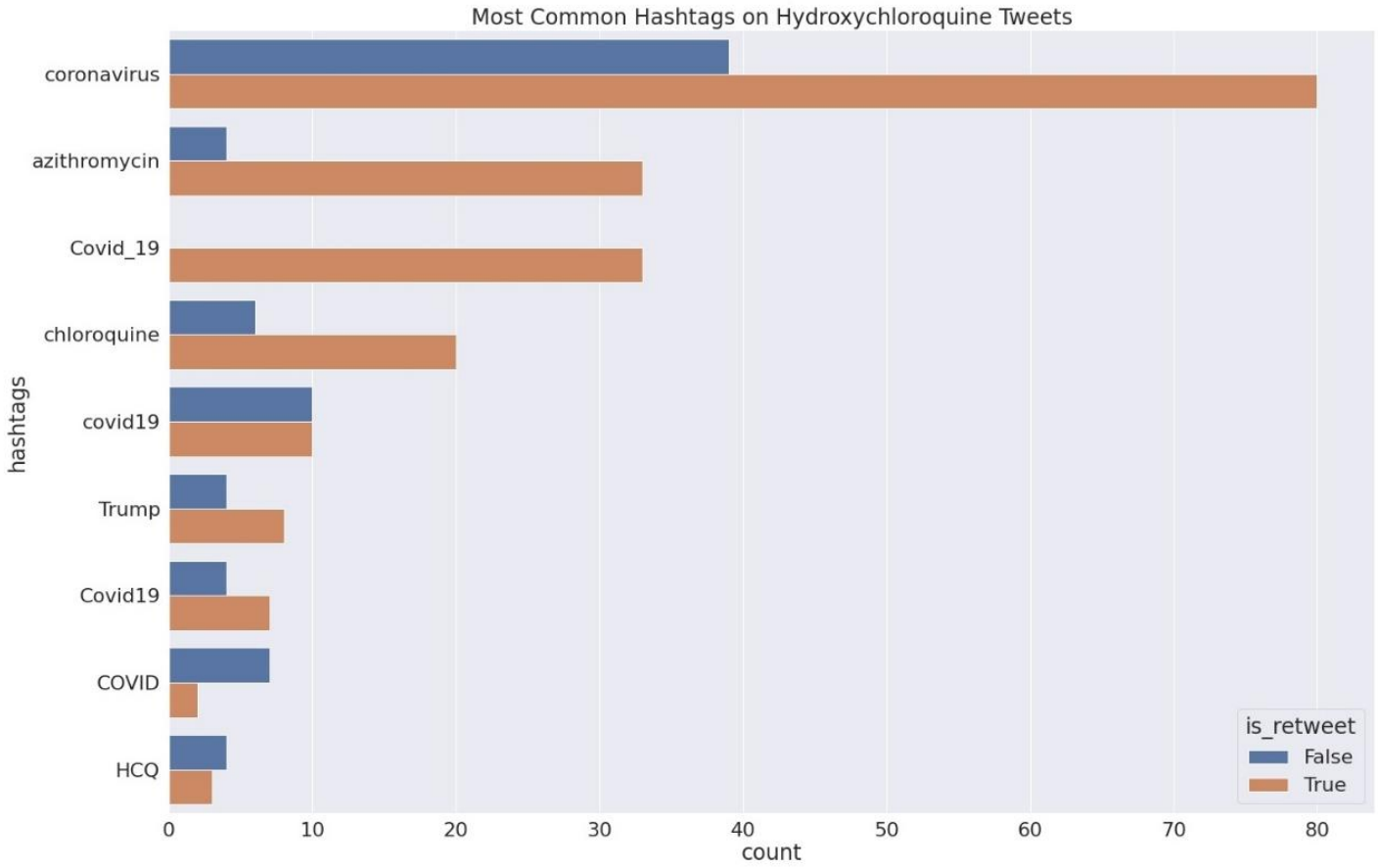

Figure 08. Hashtag and associated counts from the #hydroxychloroquine dataset. These data distinguish between hashtags shared in original tweets (blue) vs. re-tweets (orange).

We see in Figure 08 that #hydroxychloroquine tweets referencing #azithromycin or #chloroquine are much more likely to be Re-Tweets rather than original Tweets. Interestingly, the hashtag #HCQ, a common abbreviation for hydroxychloroquine, is more likely to be used on original Tweets. Perhaps this is due to the ease of typing an abbreviation, but in the future, it would be interesting to investigate whether the scientific validity of the Tweet content can be associated or somehow correlated with hashtag selection. Regardless, I would like to repeat this analysis and reintroduce the #hydroxychloroquine hashtag and perhaps delve a bit deeper into the nuance in this plot.

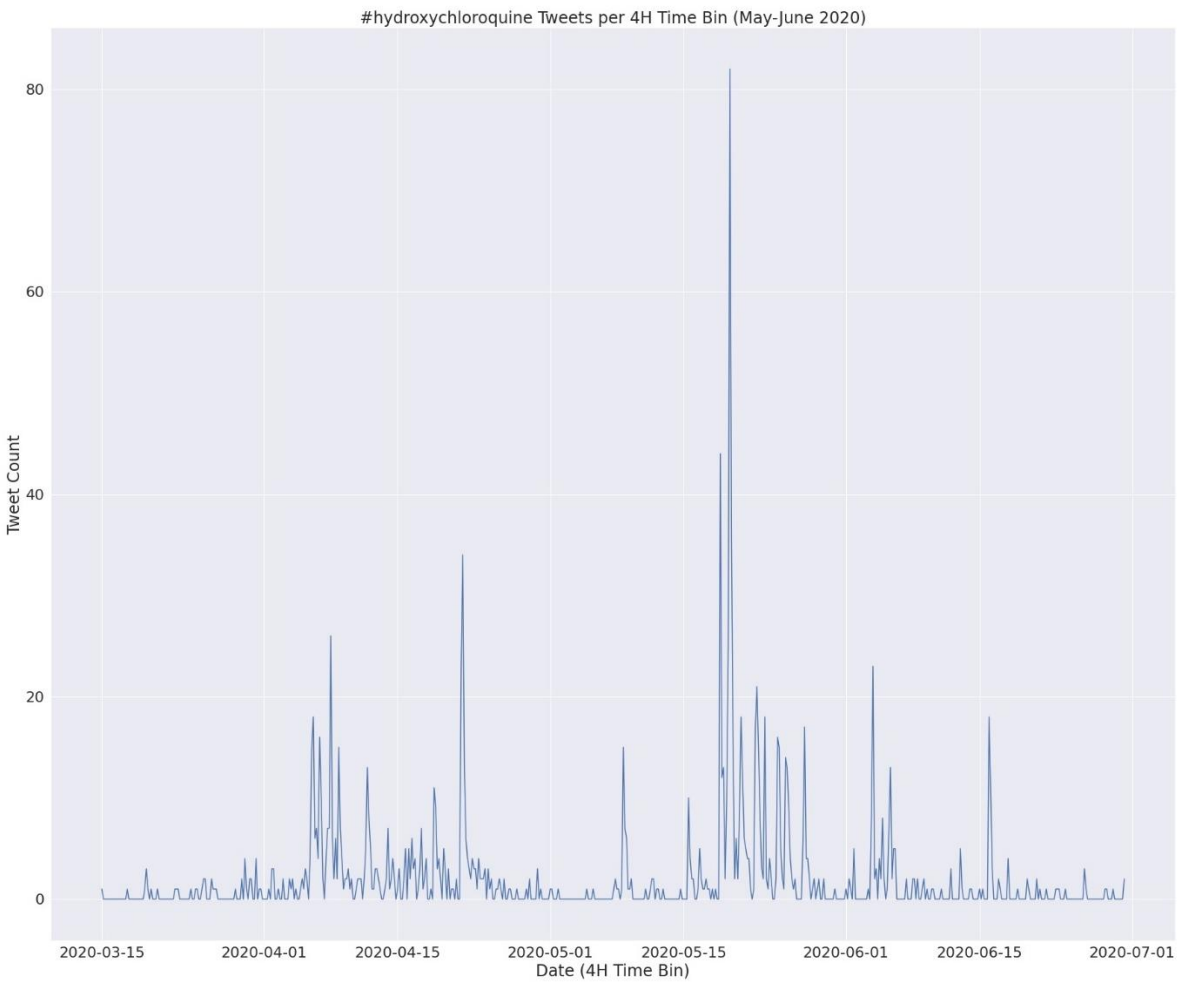

Figure 09. Tweet Count for English statuses originating in the United States that contain the hashtag #hydroxychloroquine. We see a notable burst of activity around 18 May 2020.

This plot gives a temporal perspective of the number of English Tweets over time originating in the U.S. that contain the #hydroxychloroquine hashtag. The very first mention of the drug occurred in mid-March, with some notable activity in early- to mid-April of 2020. There is one prominent spike in activity with the #hydroxychloroquine hashtag, and that occurs on 18 May 2020, when then-President Donald Trump announced that he was taking hydroxychloroquine for an active COVID-19 infection.[28]

Now that we have a bit of a temporal understanding of this data, we can graph the interactions spatially to gain an understanding of the social networks underlying this data.

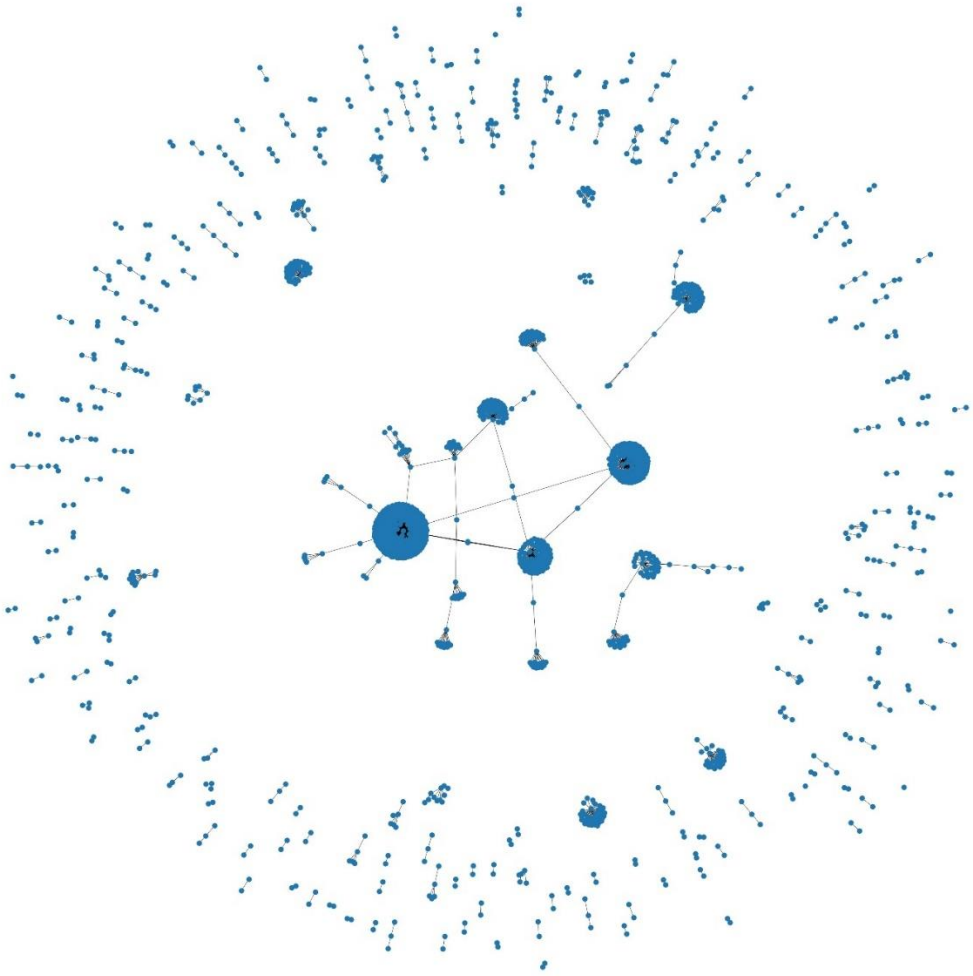

Figure 10. #Hydroxychloroquine network displaying Twitter users (nodes) and their interactions (edges).

This network was created from the #hydroxychloroquine hashtag dataset using NetworkX.[29] Nodes represent Twitter users and edges represent their interactions (replies, mentions, and retweets). These interactions do not capture manual retweets (usually indicated by the 'RT' indicator at the beginning of the Tweet text). However, we note that these is an interconnected group of users in the center of the network, and surrounding that cluster are all of the individual non-interacting or minimally-interacting users.

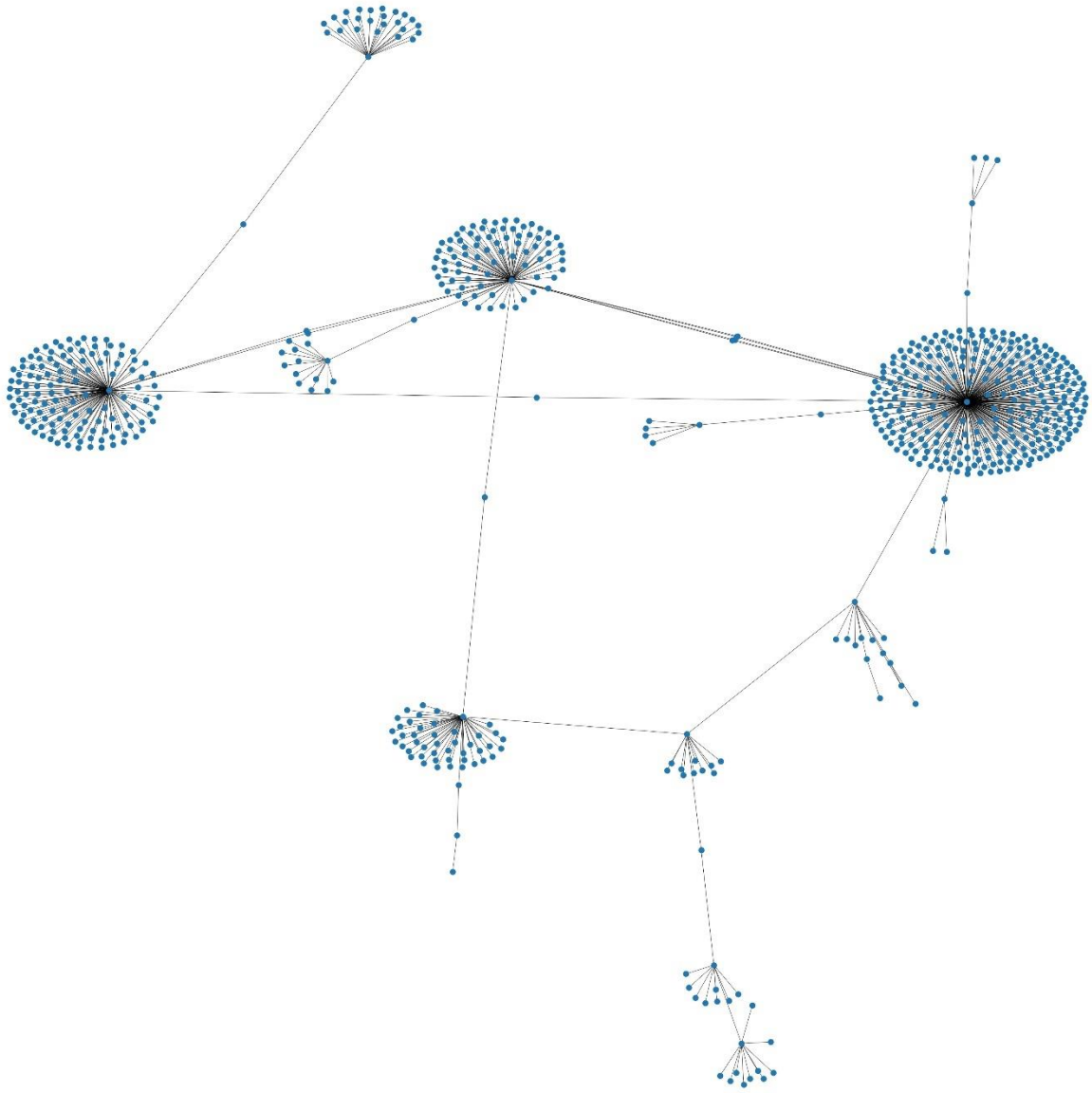

Figure 11. #Hydroxychloroquine Subnetwork

This depicts the largest subnetwork of Figure 10; however, both Figure 10 and Figure 11 were much less dense that I was expecting. I decided after seeing the visualizations to figure out how to represent these interactions more accurately.

The #hydroxychloroquine data did not contain as many tweet objects as I had expected, especially in March 2020, when the hashtag was introduced. It is possible that the hashtag itself is prone to misspelling or abbreviation, so the hashtag filter did not capture these relevant Tweets. It is also possible that the original data collection methods failed to obtain or select for Tweets containing the hydroxychloroquine hashtag. Upon reviewing the keywords.txt file outlining the search terms used to build the original dataset, I noticed that all references to hydroxychloroquine, chloroquine, or HCQ were notably absent.[30] This was somewhat surprising, given that the keywords did include terms like ‘Sinophobia’ and ‘Kungflu,’ which suggest that the keyword selection was typically synchronized with social topics and biases of interest as they related to COVID-19.[30]

To expand my network analysis to include a greater proportion of relevant tweets, I isolated all Tweet objects from the final dataset that had the hashtags ‘hydroxychloroquine,’ ‘chloroquine,’ or ‘HCQ’ and/or those words in the main text of their Tweet.

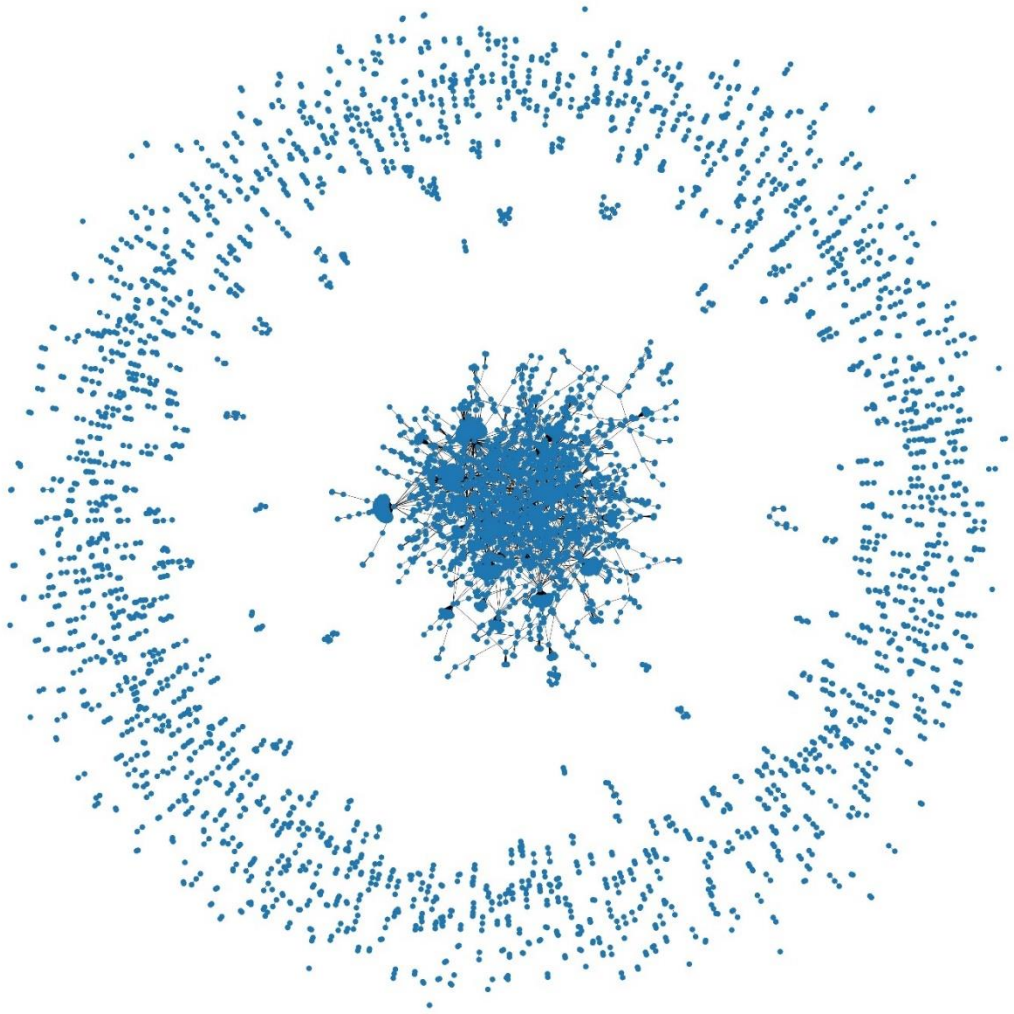

Figure 12. Hydroxychloroquine network, expanded to include both related hashtags and keywords. The density of this network is much greater, indicating that there is more activity surrounding these topics than the original analysis would suggest.

Overall, Figure 12 looks much more like what I expected to see with this analysis. This network consists of 10,352 nodes connected by 9880 edges. The max degree (number of connections from a single node) is 717, but the average degree is approximately 1.909 (mode = 1). In this graph, there is 1136 connected components, which is much greater than we saw in the previous network.

Interestingly, the number of non-interacting or minimally-interacting users on the periphery of this network graph is also much greater than in our previous analysis.

Figure 13. Largest subgraph of Figure 12

The density and interconnectedness of this network is interesting, especially the highly concentrated clusters in blue. This subgraph contains 7087 nodes connected by 7682 edges. The max degree is the same as for Figure 12 because that user is part of the subgraph. As expected, the average degree is slightly higher (~2.1679), but the mode is the same.

Next Steps

Ideally, I would like to apply sentiment analysis to each node produced in the COVID-19 misinformation networks. This would produce a network graph similar to those found in the literature; however, it is prohibitively expensive and the free trial is also rate limited. I do, however, have code available that adds the sentiment score, sentiment magnitude, an interpretation of the sentiment (strongly positive, positive, neutral, negative, strongly negative, mixed), and a mapping to a color so the node associated with the Tweet in the network will reflect the sentiment.

I would also like to delve into more nuanced frequency analysis of the Twitter object data. Specifically, I would like to compare the lexical diversity of the Tweets and hashtags between those with favorable/positive Tweet sentiment on topics of misinformation compared to those with neutral or negative sentiment regarding those topics. This could provide insight into different communication strategies among groups. Do these metrics tend to cluster in our interaction network graphs?

In addition to continuing statistical analyses of the data, I would also like to verify the effect that removing the default NLTK stopwords from the Tweet text has on sentiment analysis results. This effect is likely dataset dependent, and we should investigate whether removing default NLTK stopwords is appropriate or not for our downstream analysis.

I was intrigued by the ways in which location data from the Twitter objects can be processed and elucidated. I have preliminary analysis exploring the relationship between profile locations that map to the same geographic location and the Levenshtein distance between them. If this metric is added as an additional location preprocessing step, I am curious if the mapping of similar locations from profile location data will be more accurate.

References

[1] N. Zhu *et al.*, “A novel coronavirus from patients with pneumonia in China, 2019,” *N. Engl. J. Med.*, vol. 382, no. 8, pp. 727–733, Feb. 2020.

[2] J. V. Lazarus *et al.*, “COVID-SCORE: A global survey to assess public perceptions of government responses to COVID-19 (COVID-SCORE-10),” *PLoS One*, vol. 15, no. 10 October, Oct. 2020.

[3] K. M. C. Malecki, J. A. Keating, and N. Safdar, “Crisis Communication and Public Perception of COVID-19 Risk in the Era of Social Media,” *Clin. Infect. Dis.*, vol. 72, no. 4, pp. 697–702, Feb. 2021.

[4] J. Roozenbeek *et al.*, “Susceptibility to misinformation about COVID-19 around the world: Susceptibility to COVID misinformation,” *R. Soc. Open Sci.*, vol. 7, no. 10, Oct. 2020.

[5] P. Bansal *et al.*, “Hydroxychloroquine: a comprehensive review and its controversial role in coronavirus disease 2019,” *Ann. Med.*, vol. 53, no. 1, pp. 117–134, 2020.

[6] E. Chen, K. Lerman, and E. Ferrara, “Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set,” *JMIR Public Heal. Surveill.*, vol. 6, no. 2, p. e19273, Apr. 2020.

[7] E. Chen, E. Lerman, and K. Ferrara, “COVID-19-TweetIDs,” 2020. [Online]. Available: https://github.com/echen102/COVID-19-TweetIDs. [Accessed: 05-May-2021].

[8] “Naming the coronavirus disease (COVID-19) and the virus that causes it.” [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. [Accessed: 05-May-2021].

[9] “FileZilla - Client Features.” [Online]. Available: https://filezilla-project.org/client_features.php. [Accessed: 06-May-2021].

[10] “DocNow/twarc: A command line tool (and Python library) for archiving Twitter JSON.” [Online]. Available: https://github.com/DocNow/twarc. [Accessed: 07-May-2021].

[11] L. Sloan and J. Morgan, “Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter,” *PLoS One*, vol. 10, no. 11, Nov. 2015.

[12] “Twitter Developer Tutorial: Tweet geospatial metadata,” 2021. [Online]. Available: https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata. [Accessed: 08-May-2021].

[13] “Welcome to GeoPy’s documentation! — GeoPy 2.1.0 documentation.” [Online]. Available: https://geopy.readthedocs.io/en/stable/. [Accessed: 09-May-2021].

[14] “Folium — Folium 0.12.1 documentation.” [Online]. Available: https://python-visualization.github.io/folium/. [Accessed: 09-May-2021].

[15] “Pricing Table | Google Maps Platform | Google Cloud.” [Online]. Available: https://cloud.google.com/maps-platform/pricing/sheet. [Accessed: 05-May-2021].

[16] “mdredze/carmen-python: Geolocation for Twitter.” [Online]. Available: https://github.com/mdredze/carmen-python. [Accessed: 09-May-2021].

[17] M. Dredze, M. J. Paul, S. Bergsma, and H. Tran, “Carmen: A Twitter Geolocation System with Applications to Public Health.”

[18] B. Sen-Crowe, M. Sutherland, M. McKenney, and A. Elkbuli, “The Florida COVID-19 mystery: Lessons to be learned,” *American Journal of Emergency Medicine*, vol. 0, no. 0. W.B. Saunders, 2020.
[19] S. Burke and T. Solc, “Unidecode · PyPI,” 2021. [Online]. Available: https://pypi.org/project/Unidecode/. [Accessed: 09-May-2021].
[20] L. Espinosa *et al.*, “Epitweetr: Early Warning of Public Health Threats Using Twitter Data,” *SSRN Electron. J.*, Apr. 2021.
[21] “re — Regular expression operations — Python 3.9.4 documentation.” [Online]. Available: https://docs.python.org/3/library/re.html. [Accessed: 03-May-2021].
[22] T. Kim and K. Wurster, “carpedm20/emoji: emoji terminal output for Python.” [Online]. Available: https://github.com/carpedm20/emoji/. [Accessed: 09-May-2021].
[23] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media Inc., 2009.
[24] “nltk/nltk: NLTK Source.” [Online]. Available: https://github.com/nltk/nltk. [Accessed: 07-May-2021].
[25] C. D. Manning, P. Raghavan, and H. Schutze, “Introduction to Information Retrieval,” *Cambridge University Press*, 2009. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/properties-of-naive-bayes-1.html. [Accessed: 18-Nov-2020].
[26] “Lit | Definition of Lit by Merriam-Webster.” [Online]. Available: https://www.merriam-webster.com/dictionary/lit. [Accessed: 03-May-2021].
[27] “What Does Lit Mean | Slang Definition of Lit | Merriam-Webster.” [Online]. Available: https://www.merriam-webster.com/words-at-play/lit-meaning-origin. [Accessed: 03-May-2021].
[28] N. Carvajal and K. Liptak, “Donald Trump says he is taking hydroxychloroquine though health experts question its effectiveness - CNNPolitics,” *CNN*, 19-May-2020. [Online]. Available: https://www.cnn.com/2020/05/18/politics/donald-trump-hydroxychloroquine-coronavirus/index.html. [Accessed: 09-May-2021].
[29] “NetworkX — NetworkX documentation.” [Online]. Available: https://networkx.org/. [Accessed: 05-Mar-2021].
[30] E. Chen, “COVID-19-TweetIDs/keywords.txt at master · echen102/COVID-19-TweetIDs.” [Online]. Available: https://github.com/echen102/COVID-19-TweetIDs/blob/master/keywords.txt. [Accessed: 09-May-2021].

Appendix A: Keywords and Accounts

These are the keywords and accounts used to generate the original COVID-19 Twitter dataset. The dates listed indicate when each keyword/account was added to the search or stream Twitter API.

Keywords:

Coronavirus	1/28/2020
Koronavirus	1/28/2020
Corona	1/28/2020
CDC	1/28/2020
Wuhancoronavirus	1/28/2020
Wuhanlockdown	1/28/2020
Ncov	1/28/2020
Wuhan	1/28/2020
N95	1/28/2020
Kungflu	1/28/2020
Epidemic	1/28/2020
outbreak	1/28/2020
Sinophobia	1/28/2020
China	1/28/2020
covid-19	2/16/2020
corona virus	3/2/2020
covid	3/6/2020
covid19	3/6/2020
sars-cov-2	3/6/2020
COVID−19	3/8/2020
COVD	3/12/2020
pandemic	3/12/2020
coronapocalypse	3/13/2020
canceleverything	3/13/2020
Coronials	3/13/2020
SocialDistancingNow	3/13/2020
Social Distancing	3/13/2020
SocialDistancing	3/13/2020
panicbuy	3/14/2020
panic buy	3/14/2020
panicbuying	3/14/2020
panic buying	3/14/2020
14DayQuarantine	3/14/2020
DuringMy14DayQuarantine	3/14/2020
panic shop	3/14/2020
panic shopping	3/14/2020
panicshop	3/14/2020
InMyQuarantineSurvivalKit	3/14/2020
panic-buy	3/14/2020
panic-shop	3/14/2020
coronakindness	3/15/2020
quarantinelife	3/16/2020
chinese virus	3/16/2020
chinesevirus	3/16/2020
stayhomechallenge	3/16/2020
stay home challenge	3/16/2020
sflockdown	3/16/2020
DontBeASpreader	3/16/2020
lockdown	3/16/2020
lock down	3/16/2020
shelteringinplace	3/18/2020
sheltering in place	3/18/2020
staysafestayhome	3/18/2020
stay safe stay home	3/18/2020
trumppandemic	3/18/2020
trump pandemic	3/18/2020
flattenthecurve	3/18/2020
flatten the curve	3/18/2020
china virus	3/18/2020
chinavirus	3/18/2020
quarentinelife	3/19/2020
PPEshortage	3/19/2020
saferathome	3/19/2020
stayathome	3/19/2020
stay at home	3/19/2020
stay home	3/19/2020
stayhome	3/19/2020
GetMePPE	3/21/2020
covidiot	3/26/2020
epitwitter	3/28/2020
pandemie	3/31/2020
wear a mask	6/28/2020
wearamask	6/28/2020
kung flu	6/28/2020
covididiot	6/28/2020
COVID__19	7/9/2020

Accounts:

PneumoniaWuhan	1/28/2020
CoronaVirusInfo	1/28/2020
V2019N	1/28/2020
CDCemergency	1/28/2020
CDCgov	1/28/2020
WHO	1/28/2020
HHSGov	1/28/2020
NIAIDNews	1/28/2020
drtedros	3/15/2020

Appendix B: Hydroxychloroquine Dataset Hashtags

The hydroxychloroquine dataset is a comprehensive collection of English language Tweets originating from the United States containing the hashtag #hydroxychloroquine. This is a dictionary of all hashtags from that data with the associated totals.

Hashtag	Count
hydroxychloroquine	1341
COVID19	261
coronavirus	119
azithromycin	37
Covid_19	33
chloroquine	26
covid19	20
Trump	12
Covid19	11
COVID	9
Texas	7
HCQ	7
COVIDãƒ¼19	6
Cornyn	6
NYC	5
SARSCoV2	5
lupus	5
covid	4
UnitedKingdom	4
Coronavirus	4
zinc	3
LazarusEffect	3
RECOVERY	3
Lupus	3
quack	3
FakeNews	3
vaccine	2
pandemic	2
clinicaltrials	2
hydroxychloroquine_zpack	2
Remdesivir	2
arrhythmias	2
remdesivir	2
StayAtHome	1
medicine	1
study	1
WuhanCoronavirus	1
disgusting	1
PunkTrump	1
FishTankCleaner	1
masks	1
rightwing	1
hcq	1
TrumpCrimeFamily	1
Queens	1
FauciFraud	1
grifters	1
COVIDfoam	1
Brazil	1
MAGA	1
MichaelCohen	1
Chloroquine	1
macrolide	1
peramivir	1
DemocratGovernors	1
immunoglobulin	1
COVID_19	1
JoeBiden	1
WithoutMyHCQ	1
BigPharma	1
hydroxychloroquineandazithromycin	1

Appendix B, Continued

Hashtag	Count
[CENSORED]	1
surgisphere	1
WhiteHousedoctor	1
resist	1
NIH	1
CoronavirusLiar	1
TrumpIsAnIdiot	1
TheGreatAwakening	1
NursingHomeSlaughter	1
WWG1WGA	1
Kaletra	1
socialdistancing	1
UseHydroxychloroquine	1
FoxNews	1
WhiteHouse	1
ThesePeopleAreSick	1
FactsMatter	1
liar	1
TrumpLiesPeopleDie	1
IDSAGuidelines	1
MondayVibes	1
RudyGiuliani	1
resistance	1
physician	1
OrangeClown	1
VoteJoe	1
PressConference	1
raoult	1
FDA	1
trump	1
rheumatology	1
falsecure	1
TrumpIsNotADoctor	1
25thAmendmentNow	1
antimalariadrug	1
LADAorg	1
PusherInChief	1
nursinghomes	1
Birx	1
TheMoreYouKnow	1
CORONAVIRUS	1
Hydroxychloroquinekills	1
stock	1
VitaminC	1

Hashtag	Count
FAIL	1
OPENAMERICANOW	1
TrumpIsALiar	1
PressBriefing	1
TrumpVirus	1
TrumpLiesAmericansDie	1
ZelenkoProtocol	1
FollowTheMoney	1
WorldThisWeek	1
Donation	1
Raoult	1
TrumpPills	1
MotivationMonday	1
BernieSanders	1
Fauci	1
plaquenil	1
addzinc	1
COVIDIOTS	1
AIDS	1
DonaldLoser	1
Azithromycine	1
coronavirusnews	1
lopinavir	1
RealEyesRealizeRealLies	1
LupusAwarenessMonth	1
HIV	1
DrugAdministration	1
Novartis	1
AllHealthLive	1
testing	1
Covid	1
quarantine	1
VoteRedToSaveAmerica	1