Exploring Public Sentiment During the COVID-19 Pandemic Using Machine Learning Methods

Yujia Bai (Grace)

COVID-19 has been the worse pandemic of human history since the Spanish flu in 1918. From the onset of the widely infectious disease, there had been millions of deaths and countless infected individuals suffering from lasting health effects. However, the COVID-19 pandemic has been evolving due to the speedy research conducted on the disease, vaccines that were made widely available in a short amount of time, and also the evolving nature of the virus. Given that the COVID-19 situation has been constantly changing, it would be interesting to investigate sentiments attributed to the pandemic situation in the public sphere. How did the public respond to the pandemic? Were people angry, upset, or happy and optimistic? How did different countries respond to the situation, and how were these responses different or similar from one another? To achieve answers to these questions, I analyzed textual data using sentiment analysis in an attempt to portray a global picture of the public's response to COVID-19. In this paper, I first introduce the dataset that I used for my analysis. Then, I describe my data cleaning procedure and explain the reasoning behind each step. Next, I classify sentiments for the textual data at hand and conduct an unsupervised exploratory analysis of how sentiments differed for groups of individuals in different locations. Finally, I zoom in on countries who show different focuses on topics during COVID-19 and explore the connection of words by developing bigrams with network analysis.

Data Explanation and Exploratory Analysis

The Twitter data came from Kaggle and was scraped by user gabrielpreda on Github. All tweets were scraped using Twitter API and Python. A query was set up for the #covid19 hashtag and ran daily. Below is a brief summary of the data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179108 entries, 0 to 179107
Data columns (total 13 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   user_name         179108 non-null  object
 1   user_location     142337 non-null  object
 2   user_description  168822 non-null  object
 3   user_created      179108 non-null  float64
 4   user_followers    179108 non-null  int64
 5   user_friends      179108 non-null  int64
 6   user_favourites   179108 non-null  int64
 7   user_verified     179108 non-null  bool
 8   date              179108 non-null  object
 9   text              179108 non-null  object
 10  hashtags          127774 non-null  object
 11  source            179031 non-null  object
 12  is_retweet        179108 non-null  bool
dtypes: bool(2), float64(1), int64(3), object(7)
memory usage: 15.4+ MB
```

Variable Explanation

user_name. The user's username, which is quite random, as it could be anything the user chooses. This variable included many symbols and unique characters, making it difficult to process or extract information from.

user_location. The location of the user, which could be set manually or according to a real location. Because Twitter allowed users to set their location names manually and allowed "fake" locations to be set, this variable inevitably also included symbols and special characters that made it difficult to group together. Another challenge was that users do not always show the same location name even if they were in the same country or city. For example, users in New York City might show their location as "New York City" or "USA", and users in the UK may choose to show their location as "London, UK" or simply "UK".

user_description. The profile that users show for themselves on Twitter. Similar to the first two variables, this variable also included numerous symbols and special characters. Hashtags were included in this variable as well.

user_created. IP address at which the user created their account, resembled numerically.

user_followers. The number of followers the user had at the time of data collection.

user_friends. The number of accounts users followed at the time of data collection.

user_favorites. The number of favorite posts the user had at the time of data scraping.

user_verified. Whether the user's identity was verified or not.

Date. The date of the post.

Text. The raw text of the Twitter post.

Hashtags. Hashtags that were included in the post.

Source. Whether the tweet came from an Android phone, iPhone, or the Twitter web application.

is_retweet. Whether the post was retweeted or not.

Data Cleaning

First, I started with dropping duplicates and NAs in the data. In the scraping process, the authors had ensured that duplicates were excluded, but I dropped duplicates once more to ensure that duplicates truly did not exist in the data. When filtering for NA values, I noticed that user_location, user_description, hashtags, and source had the most prominent number of missing values. Though we usually exclude variables with significant numbers of missing data in regular data processing steps, this dataset was a special case: we had over 100,000 rows of data to begin with, and the column with the largest number of missing data (location) only contained about 36,000 missing values. After dropping all NAs, I was left with a data frame with 99,127 rows and 12 columns, which was more than sufficient to complete this project.

After initial cleaning of the data, I realized that certain variables could be deleted. I eventually deleted user_name, user_created, user_description, source, and is_retweet. The reason for deleting user_created and source is similar: these variables do not provide meaningful information for this project, as I do not intend to use IP addresses or device type to perform my

tasks. I deleted user_name because there were too many unique values, and useful information was difficult to extract due to high variety and low patterns. is_retweet was deleted simply because all variables had the same value (False), and no meaningful information could be extracted from this variable. I eventually decided to exclude user_description due to it being relatively arbitrary and included too many special characters, which were difficult to parse. Also, user_description included other languages, such as Russian, and was out of the scope of the current project. I decided to retain the "location" variable although it contained random characters, made-up addresses (e.g. "Stuck in the middle"), and the highest number of missing data. Specific reasons for retaining "location" will be discussed in the exploratory analysis section.

The last step of my general data cleaning process was to remove certain expressions using lambda functions from the 'text' column in my data. Lambda functions are small and simple functions that contain a single expression and allow for applying the same operation on a given subset of data (MACS31300 course notes). One advantage of lambda functions is that there is no need to name the function, thus allowing us to perform operations that are small and use less code. Using regular expressions and lambda functions, I removed mentions ("@"s) from the text column in my twitter data, as most usernames are randomly decided by individuals, and mentions of other users do not necessarily contribute to the meaning or sentiment of the post.

Furthermore, I tested removing all hashtags from the 'text' column, as hashtags were already specified in the data and certainly included 'covid19'. Not removing hashtags might have prevented word clouds from showing other important information, as the number of #covid19 might obscure the number of other frequently appearing words. However, this approach also had its limitations: completely removing hashtags might have taken away information from the

tweets, as hashtags can exist as a part of speech (e.g. “I hate #covid”), and removing a part of speech may create obstacles for using certain algorithms that rely on context or creating bigrams/trigrams that reflect crucial information. Eventually, instead of removing hashtags from the original text column, I created an extra column with hashtags in the texts deleted. Finally, I appended new columns that removed punctuations, tokenized, removed stopwords, and stemmed the ‘text’ column. However, in my analysis, I still chose to use the original text column because my algorithm of choice (VADER) is sensitive to contextual information, capitalization, and negation, and using fully cleaned textual information might not have allowed VADER to be utilized to its full potential.

Exploratory analysis

Frequent words. Upon examining the most frequent words in all tweets, I noticed through observing word clouds (shown in the figure below) that ‘https’ was a frequently occurring “word”. This signaled that twitter users were posting links with the hashtag “#COVID19”, potentially to popularize sites or other media content. Since this did not add useful information to my analysis, I removed links from both the “text” column and the newly created “text” column without hashtags.

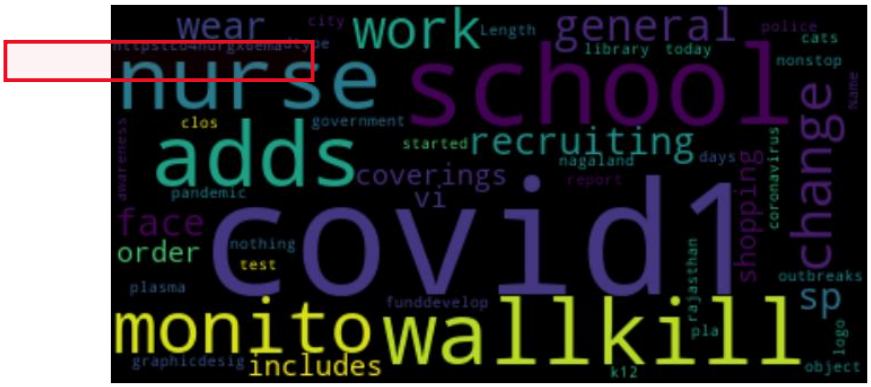


Figure: Text in the red box shows a link

Location. Exploratory analysis showed that the 'location' variable could be useful in extracting insights for different regions. Despite made-up locations, the top 10 locations were all in the United States, India, or the UK (see figures below).

```
{'Pewee Valley, KY': 1,
 'Stuck in the Middle ': 1,
 'Jammu and Kashmir': 1,
 'Новоро́ссия': 1,
 'Gainesville, FL': 1,
 '👇location at link below👇': 1,
 'Dhaka,Bangladesh': 1,
 'Hotel living - various cities!  Who needs a home when hotel
 'Africa': 1,
 'Nagaland, India': 1,
 'Brussels': 1,
 '100+ countries': 1,
 'Graz': 1,
```

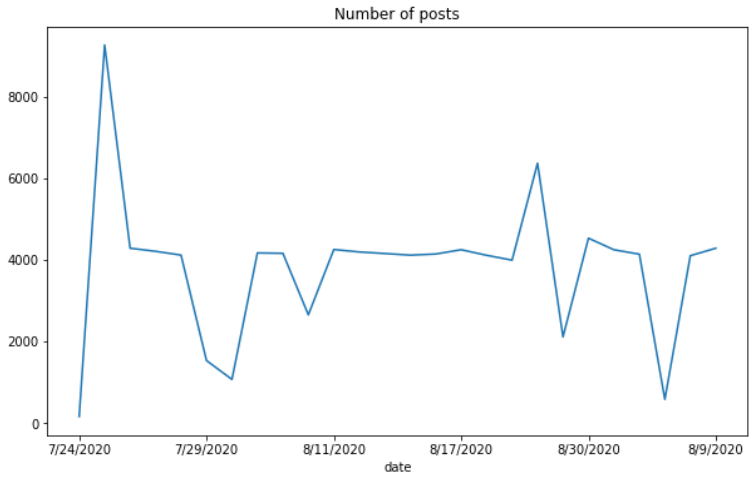
Figure: Sample locations

```
#top 10 locations
twitter_data['user_location'].value_counts()[1:10]
```

```
United States       1728
New Delhi, India    1349
Mumbai, India       1098
Washington, DC       956
London, England      842
New York, NY         765
London               749
United Kingdom       714
New Delhi            676
```

Figure: Top 10 locations by number of occurrences

Date. The dates in this dataset ran from 7/24/2020 to 8/9/2020. Although there were high volumes of tweets over time, the short time span in this dataset does not allow for observations of trends over longer periods of the COVID-19 pandemic. As is shown in the figure below, the number of tweets in the time period does fluctuate quite a bit, and it is possible that there were evident sentiments shifts during the given time period. However, this is not expected, and the short time frame is a drawback of my dataset.



Machine Learning: Sentiment Analysis

The main machine learning task for my project was to conduct sentiment analysis on tweets (the "text" column in my data), classify the sentiment of tweets, and observe trends of sentiment in different groupings.

For this task, I selected the VADER (Valence Aware Dictionary and sEntiment Reasoner) algorithm to conduct sentiment analysis. VADER is a rule-based and lexicon sentiment analysis algorithm that takes into account both the polarity and intensity of texts when classifying sentiment. Based on lexical features labeled by humans as positive or negative, VADER is able to recognize contextual negations, capitalized words (e.g. VADER would most likely recognize that "wow" and "WOW" have different intensities), and even emojis. Given these features, VADER is particularly suitable for text where a mixture of attitudes are expressed, which is often found in social media data. Hence, VADER is suitable for my dataset, which comes from the social media application Twitter. Although pretrained language models such as the transformer-based BERT (Bidirectional Encoder Representations from Transformers) model has been shown to perform well on classifying nuanced sentiment, as different words can have different vectorized representations based on context, implementing the model requires

intense computation and may not be the best choice for the large amount of data I have. VADER, on the other hand, provides a relatively subtle and accurate classification of social media text while using less intensive computation.

I used VADER to calculate four sentiment scores for my data: positive, negative, neutral, and compound, and joined these scores to my dataframe. This took advantage of VADER's ability to not only classify sentiment according to intensity and polarity, but also calculate how positive/negative/neutral a given text was. Typically, when classifying a given text as positive, negative, or neutral, VADER uses a compound score, which is calculated by "summing the valence scores of each word in the lexicon, adjusted according to the rules, then normalized to be between -1 (extremely negative) and 1(extremely positive)." When the compound score >=0.05, the text is classified as positive. When the compound score <=-0.05, the text is classified as negative. When the compound score is larger than -0.05 but smaller than 0.05, the text is classified as neutral. The classification results based on the "text" column is shown below:

	Total	Percentage
neutral	36840	37.16
positive	36755	37.08
negative	25534	25.76

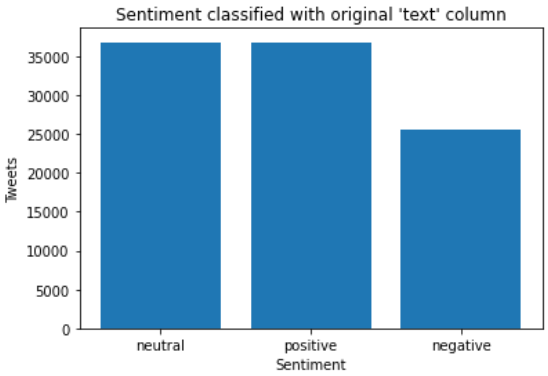


Figure: Proportion and counts of tweets by sentiment

Contrary to common expectations, there were more tweets classified as positive and neutral as compared to negative. While it is possible that individuals during the later July to August period of 2020 were reflecting on their experiences of COVID-19 as positive, these results still required scrutiny. The top 10 tweets of each sentiment category are shown below.

	text	pos
35283	Haha, funny #COVID19	0.855
75424	Gorgeous, funny, harrowing....and true. #COVID19	0.829
1506	Worth a share #COVID19	0.804
139111	Lovely! #COVID19	0.804
45754	Brilliant. #COVID19	0.792
144498	Brilliant #COVID19	0.792
81936	Celebrating #Covid19?	0.787
8572	#Covid19 positive	0.783
92426	survived #COVID´19. bless up! Very thankful.	0.776

Figure: Top 10 positive tweets

	text	neg
41817	Tragic. Scandalous. #covid19	0.865
152928	Hell no !!! Fuck #COVID19	0.848
77549	sighhhhhhhhh :(:(:(#covid19	0.813
12232	I hate #COVID19.	0.787
32025	#COVID19 scam	0.787
95468	I hate #COVID19	0.787
65119	#COVID19 scam	0.787
91000	#COVID19 negative	0.787
104429	Fuck fuck fuck..\n#COVID19	0.778

Figure: Top 10 negative tweets

	text	neu
54246	In #Japan JP, #geishas are grappling with #Cov...	1.0
54198	A rundown of financial programs and guidance i...	1.0
54200	Will Children Spread #COVID19 if They Go Back ...	1.0
54201	SMALL BUSINESSES! Take this short, 15-minute s...	1.0
122436	Is your company concerned about returning to #...	1.0
122432	Facts about #statistics, #treatment and #FaceM...	1.0
54209	CAT panelists discuss possible COVID-19 impact...	1.0
54210	Coronavirus Is Not Over Yet, 4/25/2020, In the...	1.0
122431	Weldone #Pakistan 💪 💪 💪 #COVID19\n ...	1.0

Figure: Top 10 neutral tweets

Observing the top 10 tweets of each category, the way that neutral and negative tweets were classified seemed reasonable. In the top 10 negative tweets, most words in the tweets are negative and even vulgar words humans use to express frustration or negative sentiments. Top 10

neutral tweets also seemed to match the way humans express neutral feelings or opinions, as the tweets are mostly expressing facts and less opinionated.

However, issues arise when observing the top 10 positive tweets: judging with knowledge of how humans express sarcasm, some of the tweets seemed sarcastic (e.g. “Brilliant. #COVID19”) but were classified by VADER as positive. This may have been because the phrase “COVID19” is not in VADER’s dictionary and thus did not receive a sentiment score, and sentiments classified for sentences with “COVID19” rely on other words in the sentence. Since VADER only scores sentiments by calculating word sentiments as a function of their context and does not take into account broader social contexts, sentences involving only extremely positive words (brilliant; great) with “COVID19” could easily be classified as positive. This causes potential sarcasm expressed in tweets to be neglected and is a potential drawback of the VADER algorithm.

Another issue in tweets classified as positive is that, judging from a personal viewpoint, certain tweets did not seem fully relevant to the COVID-19 pandemic. Rather, certain tweets may have been using the #COVID19 hashtag to popularize media content (e.g. “Worth a share #COVID19”). This is because in the beginning of my data cleaning process, all links were removed from the ‘text’ column of my data. Although this saved computational power and removed irrelevant information, it wasn’t able to parse out irrelevant content that added in noise for classifying sentiments of tweets with the #COVID19 hashtag. This is also a potential drawback of VADER.

Given the above issues with tweets classified as positive, individuals during the late July to August time period of the COVID-19 pandemic were not necessarily more “positive” than “negative” about the situation. Instead, the positive sentiments expressed in tweets may have

been due to unsuccessfully classified sarcasm and content irrelevant to COVID-19 but wanted publicity. Therefore, I returned to my data cleaning steps and removed all tweets with links. Eventually, the dataframe had 8,279 rows, which was significantly less than the original dataset, yet still sufficient for unsupervised sentiment analysis. Information about the final dataframe is shown in the figure below, and updated top 10 locations are also shown.

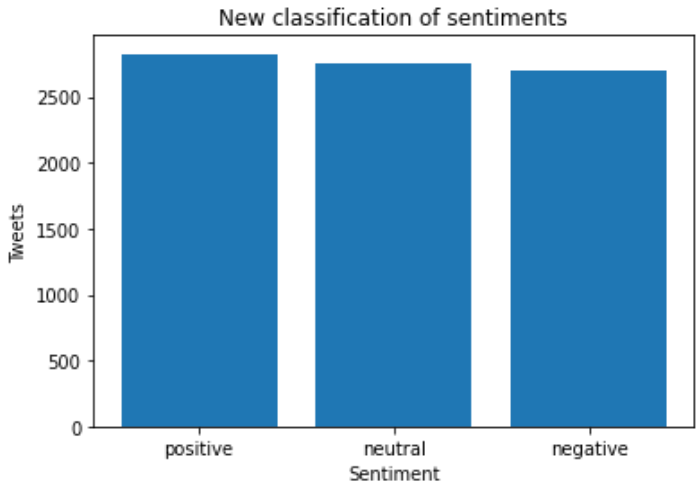
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8279 entries, 57 to 179073
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   user_name         8279 non-null   object
 1   user_location     8279 non-null   object
 2   user_description  8279 non-null   object
 3   user_created      8279 non-null   float64
 4   user_followers    8279 non-null   int64
 5   user_friends      8279 non-null   int64
 6   user_favourites   8279 non-null   int64
 7   user_verified     8279 non-null   bool
 8   date              8279 non-null   object
 9   text              8279 non-null   object
 10  hashtags          8279 non-null   object
 11  source            8279 non-null   object
 12  is_retweet        8279 non-null   bool
dtypes: bool(2), float64(1), int64(3), object(7)
```

```
#top 10 locations
twitter_data['user_location'].value_counts()[1:10]
```

```
India               107
New Delhi, India    105
Hong Kong           100
Canada               76
USA                  70
Ibadan, Nigeria      60
Florida, USA         56
London, England      52
Los Angeles, CA      48
```

Figure: Updated top 10 locations

After cleaning the dataset a second time by removing all tweets with links, sentiment analysis using VADER showed that the three classes of sentiments were quite even:



The only lingering issue was when individuals posted about testing "positive" for COVID, VADER classified their tweets with "positive" sentiment, which hints that in terms of human-level understanding, there were likely more "negative" expressions surrounding COVID-19 than positive ones. However, merely changing the polarity and intensity score of "positive" in the VADER dictionary would not have benefitted my analysis, as the word "positive" more often associates with positive statements. This again inferred drawbacks of the VADER algorithm, such that it relies on existing mappings of vocabularies and does not consider broader social contexts. However, if VADER lexicons were updated with the polarity and intensity of "COVID19", the algorithm may improve significantly on classifying tweets with the hashtag "#COVID19". In the following sections, I moved on to observe word clouds of all tweets and of different countries as well as conduct network analysis with the aim to paint a global picture of responses to the pandemic.

Word clouds

Word clouds of all tweets and top 10 locations were created by taking the 400 most frequent words. Frequent words for different locations were derived by filtering the 'text' column by the top 10 locations identified in the exploratory analysis and subsequently grouped

together if locations were from the same country. Although word clouds of single words have the drawback of breaking down important phrases into separate words, thus stripping away context, we could nevertheless piece together topics and information regarding each sentiment.

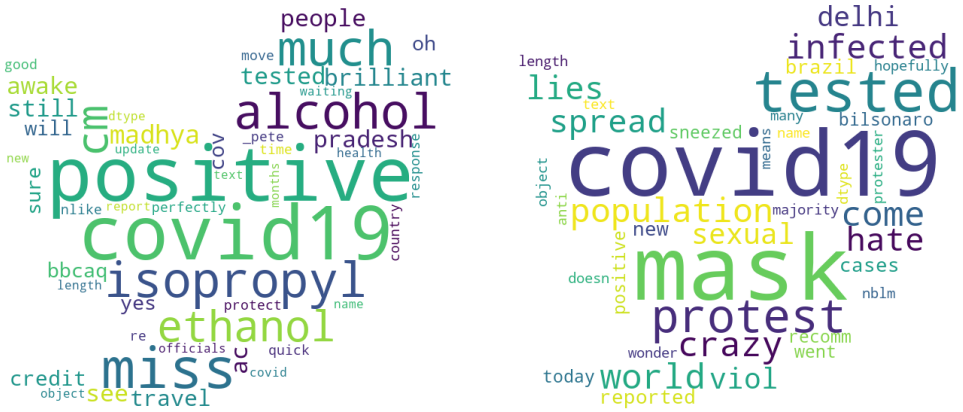


Figure: Positive (left) vs. Negative (right) word clouds

Comparing the positive and negative word clouds, we see that different topics are discussed. For tweets expressing positive sentiment, some frequent topics were isopropyl, alcohol, ethanol, Madhya, Pradesh, and “positive”. Although “positive” was evidently a frequent word in positive tweets, users may have been expressing that they “tested positive”, and I did not fully consider the word to be fully positive in terms of sentiment. Isopropyl, alcohol, and ethanol were discussed frequently likely due to people around the world trying to find disinfectants to protect themselves from the COVID-19 virus. These words could have also been discussed frequently because former president Donald Trump suggested injecting disinfectants was a safe way to protect oneself against COVID-19, although I would have expected more negative responses to his suggestion. Another frequently discussed topic was likely Madhya Pradesh, a city in India. Madhya Pradesh appearing in the positive word cloud was reasonable: at the beginning stages of COVID-19, India experienced tremendous hardship with medical resources and treating the infected, leading to horrific numbers of deaths. However, by June 2020, Madhya

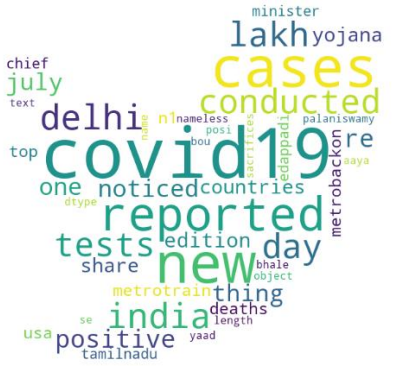
Pradesh had planned to lockdown restrictions in most areas, and citizens in Madhya Pradesh were likely celebrating or feeling hopeful about their situation.

As for words in negative tweets, mask, protest, population, spread, world, lies, crazy, even hate, were frequent topics. Tweets expressing negative sentiment seemed to hint towards prevailing negative topics during the pandemic: protests in cities where citizens opposed lockdowns and mask orders, individuals thinking that COVID-19 was a hoax ("lie"), the quick spread of the virus to countries around the world, and generally judging the COVID-19 situation as "crazy", as humans had not experienced such a pandemic ever since the Spanish flu in 1918.

Next, I grouped tweets by the top 10 locations and combined tweets from the same countries. The figures are shown below. In the top 10 locations, six different countries/cities were identified: India, Hong Kong, Ibadan (Nigeria), Canada, the United States, and London (UK). In the six word clouds, "covid19" appeared to be the most frequently mentioned word for most locations. Although the word cloud of Hong Kong did not show "covid19", "new cases" likely referred to the COVID-19 virus.

Noticeably, tweets from Ibadan, Nigeria differed from all other word clouds. Instead of widely discussing "covid19", individuals (assumed to be) tweeting in this location were discussing religious topics such as "church", "god", "father", and "worship" more frequently than other words. Also, the most frequently mentioned word "gracepoint" appeared to be the name of a Christian church in Ibadan. Upon searching for facts surrounding religion in Nigeria, this became non-surprising. Nigeria is known to be a religious country, with almost half of its population Christian, rendering it Africa's most populated Christian body. This finding reflected that not all countries responded to the pandemic with identical focuses, and that the global response to COVID-19 was likely diverse.

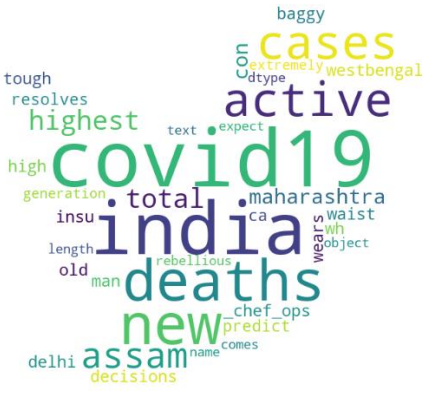
Another special finding in the word clouds was in the word cloud of the United States. This word cloud combined all tweets in the locations “United States”, “Florida, USA”, and “Los Angeles, CA” and identified the top 400 frequently mentioned words. In the word cloud, words related to American politics appeared, such as “trump”, “bidenharris2020” and “pre-inauguration”. Although it was difficult to judge the sentiments related to these political indices from the word cloud, given the context of Trump’s response to COVID-19, I predicted that individuals in the US were expressing frustration against Trump and hoping for Biden and Harris to win the 2020 election. Also, the appearance of these phrases signaled that COVID-19 in the US was not only a pandemic, but also relevant to politics.



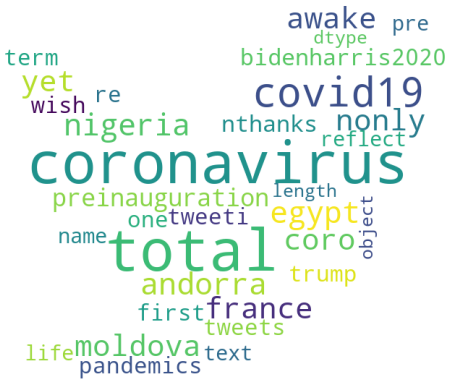
1 India

2 Hong Kong

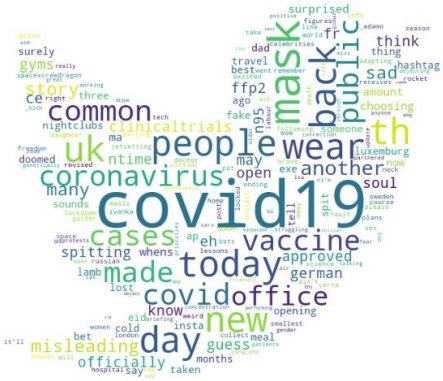
3 Ibadan, Nigeria



4 Canada



5 USA

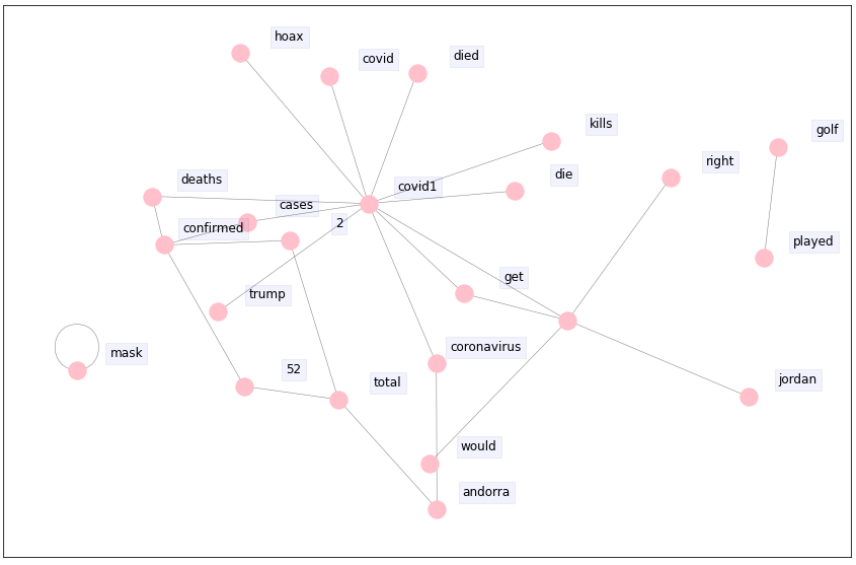


6 London, UK

Figures: Word clouds by top 10 locations (grouped by country/city)

Network Analysis

Given the above results, I conducted network analysis on tweets from Ibadan, Nigeria and in the US. In a network of bigrams, nodes show unique words, and edges show bigrams. By counting the number of edges, we could calculate how often a bigram appears in our data. After performing basic network analysis, I noticed that certain bigrams were clustered together, while others seemed to have a larger distance. An example is shown in the bigram network below of tweets posted with locations in the US:



Therefore, I decided to further conduct community detection using the Girvan-Newman algorithm. Briefly stated, the Girvan-Newman algorithm conducts community detection by removing edges of the given graph. Typically, the algorithm first removes the edge with the highest level of betweenness centrality, defined by the number of shortest paths between two nodes an edge governs, recalculates edge betweenness, and repeats this process until the network has no edges left. This exposes the community structure and allows for detection of communities. In the figures shown below, the top 50 bigrams in both locations were selected, and all bigrams appeared more than once.

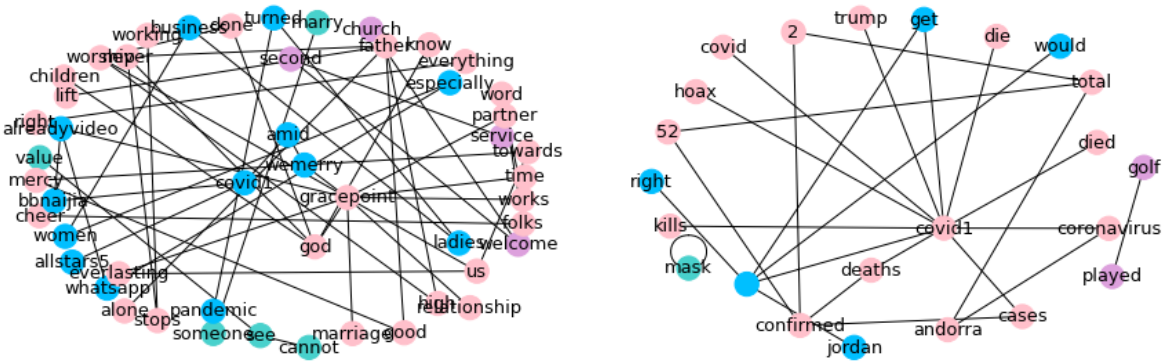


Figure: Network of bigrams with communities, Ibadan, Nigeria (left) and USA (right)

Comparing the two networks of bigrams with communities detected, we see that there were significantly less bigrams in tweets located in the US than Ibadan, Nigeria. Interestingly, for tweets in the US, “Trump” was detected to be in the same community as multiple words with negative connotations, such as “die”, “deaths”, and “kills”. This confirmed my hypothesis for the word cloud earlier: Trump may have been associated with negative events during the beginning of COVID-19. On the other hand, tweets located in Ibadan involved more bigrams, with the majority relating to religion, as can be seen with the pink nodes in the bigram network on the left. Also, bigrams related to religion seemed to involve family, as “partner”, “marriage”, and “children” are in the same community of bigrams as religious concepts. Seeing from these tweets

and analyses, it can be inferred that while family and religion are a crucial part of Ibadan citizens' lives, even during special times like COVID-19, individuals in the US tended to focus more on the COVID-19 situation itself and political implications more often. Although the Girvan-Newman approach is relatively computationally expensive when networks are large, as the algorithm computes the edge betweenness for all nodes in a graph, it nevertheless extracted useful insights for my purposes.

Future Directions

If I was given unlimited time and resources, I would scrape data from both Twitter and news media (headlines) during longer time intervals and analyze whether sentiments of these two media entities coincided or were disparate from one another. I would also try to conduct causal analysis on whether news headlines were driving public sentiment on Twitter. Secondly, I would join my dataset to data from the World City Database to investigate how events related to COVID-19 might have unfolded in different latitudes and longitudes. I would also try to find a way to detect fake locations, remove these locations, and conduct a more thorough analysis of tweets from different countries by combining tweets from different countries more exhaustively. This would assist me in producing a more thorough global view of responses to the COVID-19 pandemic. A third task I would do if given unlimited time and resources is to compare the performance of different sentiment classification models. For example, I might compare sentiment classification results using VADER and BERT in depth and discuss theoretical differences and issues for the two algorithms. Finally, though sarcasm is a sentiment classification topic that even state-of-the-art research struggles with, with unlimited resources, I would thoroughly explore my data, observe words or information that appear when Twitter posts show sarcasm, and try to update the VADER dictionary to potentially detect sarcasm.