COVID-19

TWITTER AND WHO

BAN 675 Spring 2020

Christina B.

ABSTRACT

This project will analyze the sentiment and topics of Twitter users' response to COVID-19 in comparison with the World Health Organization (WHO) daily highlights. It is important to understand the statements coming from WHO as their messages directly affect the general public. Analyzing the public sentiment will tell us how people are responding to the current situation. The objective of this project is to understand and perform analysis on public language versus official language regarding the coronavirus. The data set from Twitter and the WHO reports were cleaned, and sentiment analysis was performed on the tweets and WHO reports for March 9th, 19th, and 30th. This will allow us to see and compare the changes since the start of the month to the end of the month. March 19th was the day the shelter in place was announced and will let us understand the responses surrounding that event.

I. INTRODUCTION

Covid-19, the pandemic that has infected millions of people and caused thousands of deaths in all over the world, has now become even more devastating to people in the USA. A lot of businesses have been negatively affected. The health care systems are struggling with the rapid increase of cases day by day. The shelter-in-place has been put into effect forcing people to stay home. The novel coronavirus has quickly spread across the country and has swept most parts of the world. Because of this, we've seen businesses and schools close, dramatic changes in commute and travel, increases in unemployment, etc. But we have also seen large support for first responders, communities coming together to support local businesses and provide for the elderly, larger presence of online resources, etc. With all the free time, people are also spending more time than ever on social media. The question posted here is how do people react to this pandemic? The purpose of this paper is to examine how the WHO and Twitter users react to the pandemic. We will be looking at tweets with the hashtag #coronavirus, #coronavirusoutbreak, #coronavirusPandemic, #covid19, and #covid_19. We will identify and divide the tweets into positive and negative categories, determine the most frequently used words overall and in each category, and identify the topics in the documents. This will help us to understand the overall sentiment regarding Coronavirus and learn how people are impacted by the pandemic. The data set from Kaggle.com includes tweets created for the month of March in multiple csv files. There are 22 columns for each csv file with over 500,000 records. We will be analyzing 5,000 records for March 9th, 19th, and 30th of 2020. Attributes in the files are defined as the following:

- Status_id: The ID of the actual Tweet.
- User_id: The ID of the user account that Tweeted.
- Created_at: The date and time of the Tweet.

- Screen_name: The screen name of the account that Tweeted.
- Text: The text of the Tweet.
- Source: The type of app used.
- Reply_to_status_id: The ID of the Tweet to which this was a reply.
- Reply_to_user_id: The ID of the user to whom this Tweet was a reply.
- Reply_to_screen_name: The screen name of the user to whom this Tweet was a reply.
- Is_quote: Whether this Tweet is a quote of another Tweet.
- Is_retweet: Whether this Tweet is a retweet.
- Favourites_count: The number of favourites this Tweet has received.
- Retweet_count: The number of times this Tweet has been retweeted.
- Country_code: The country code of the account that Tweeted.
- Place_full_name: The name of the place of the account that Tweeted.
- Place_type: A description of the type of place corresponding with place_full_name.
- Followers_count: The number of followers of the account that Tweeted.
- Friends_count: The number of friends of the account that Tweeted.
- Account_lang: The language of the account that Tweeted.
- Account_created_at: The date and time that the account that Tweeted was created.
- Verified: Whether the account that Tweeted is verified.
- Lang: The language of the Tweet.

II. RESEARCH QUESTION

With the recent Covid-19 outbreaks, the world has gone into a state of emergency to fight against this pandemic. As we start practicing social distancing, many people start to work from home to lower the infection rate. In this paper, we will investigate in: How did Twitter react to Covid-19? Do the WHO reports influence how the public tweets about COVID-19?

2.1 Data preprocessing

Before any further analysis, it is always important for us to clean the data. The original data is obtained from Kaggle with 20 csv files and over 5k rows in each. To begin with, the dataset is transformed into a corpus. Then, the corpus is pre-processed with making all characters lowercase, removing all punctuation marks, white spaces, and common words (stop words). Besides the default English stopword package, we also create some additional stop words into the list. These words include

‘Covid’, ‘Corona Virus’, ‘Covid-19’, etc. This is because these words have very high frequency but they do not provide us with any important information. Also, in this analysis, we are just interested in Tweets in English, which is defined by column lang==en. After the corpus is pre-processed, the tokenization method is used to grab the word combinations.

III. MARCH 9TH

3.1 Topic Modeling Tweets

March 9th is the first day our resource had a large data set available for tweets. This day is approximately when individual states began to declare states of emergencies due to Covid-19. There was no mass public opinion on the virus at this time, so there is a weak concentration of topics outside of the initial outbreak and the growing number of cases in topic #0. For topic #1, we see President Trump begin to tweet in response to the outbreak, which may be massly retweeted by the public. In topic #2 there are tweets about the virus being spread from China into the U.S.

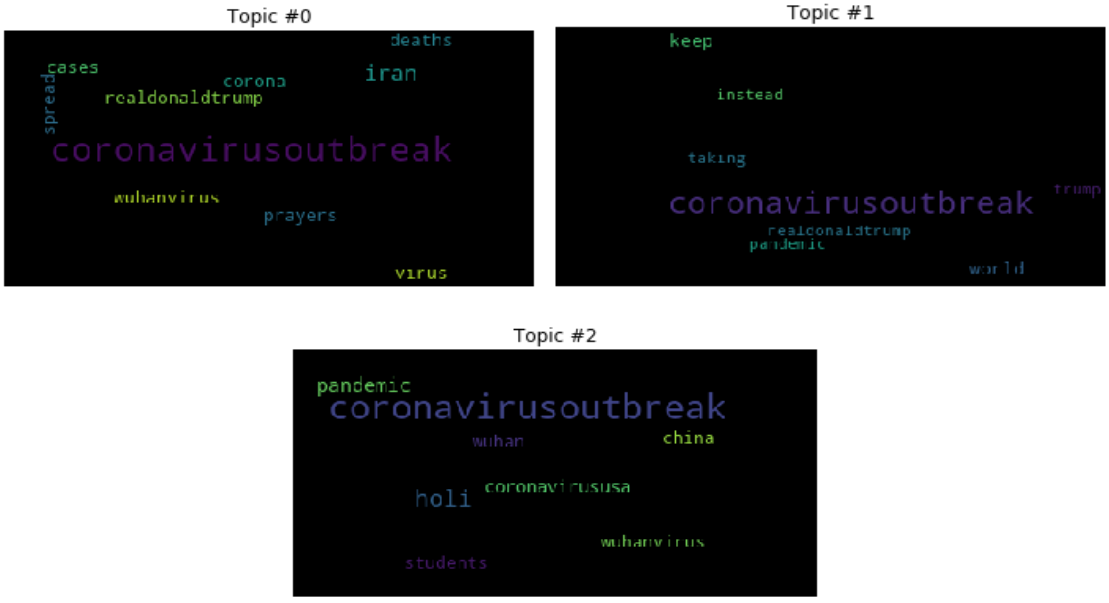

FIGURE 3.1. TOPIC MODELING FOR MARCH 9 TWEETS.

3.2. Word Frequency and Sentiment Analysis on March 9 Tweets

On this date, the most common words included ‘coronavirusoutbreak’, ‘Italy’, and ‘cases’. This suggests a rising awareness of the virus by the public as the number of cases in Italy grows. We also

see that the mentionings and tweets by President Trump are increasing because he is listed twice: once by his twitter handle and again by name.

At this point during the outbreak, there was not a strong use of positive or negative words in tweets. Therefore there are minimal levels of high frequency words available in the data set.

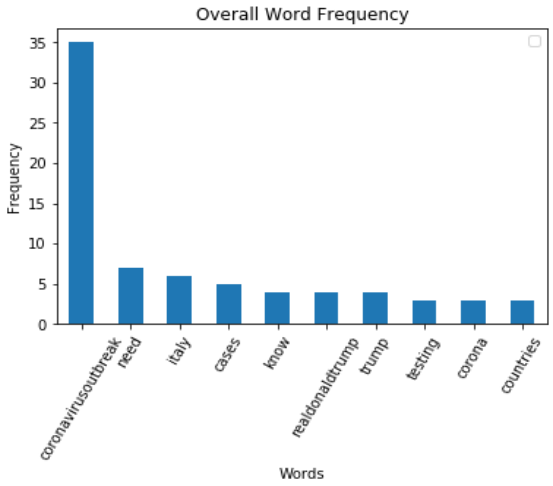

FIGURE 3.2.1 OVERALL WORD FREQUENCY FOR MARCH 9 TWEETS.

3.3. Topic Modeling WHO Report

Topic modeling was used to identify the different topics the WHO is addressing in their daily highlights. We can see from Topic #0 below and as well as Topic #1, that the documents are primarily talking about continuing support and preparation for the need of supplies. While topic #2 is talking about the updated protocol through operation support and logistics (OSL) and the emergence of a global pandemic.

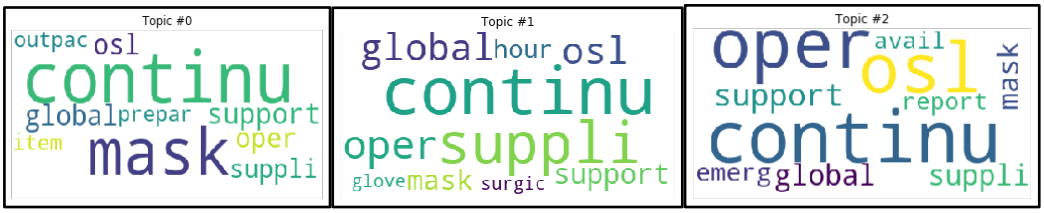

FIGURE 3.3.1. TOPIC MODELING FOR MARCH 9 WHO REPORT.

3.4. Word Frequency and Sentiment Analysis WHO Report

The most common words in the daily WHO reports are similar to the topics. Most frequently used words include ‘continue’, ‘support’, and ‘operations’. At this point, the WHO reports of the need for preparation. OSL includes program support and operation activities to coordinate on health operations. During this time, the assessment of supply needs are still taking place.

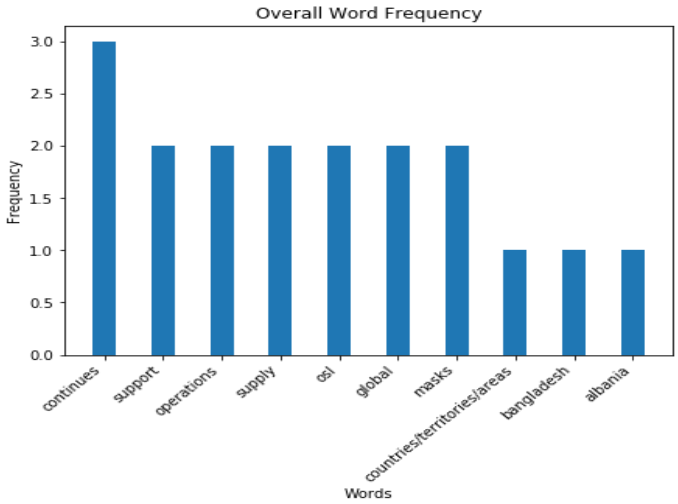

FIGURE 3.4.1 OVERALL WORD FREQUENCY FOR MARCH 9 WHO REPORT

Positive words used in the WHO reports include 'support' and 'availability' while the negative words used include 'outbreak' and 'emergency'. From this we can extrapolate that support is available in countries of outbreak. The word 'emergency' has negative connotations, but stresses the level of importance the virus should be held to.

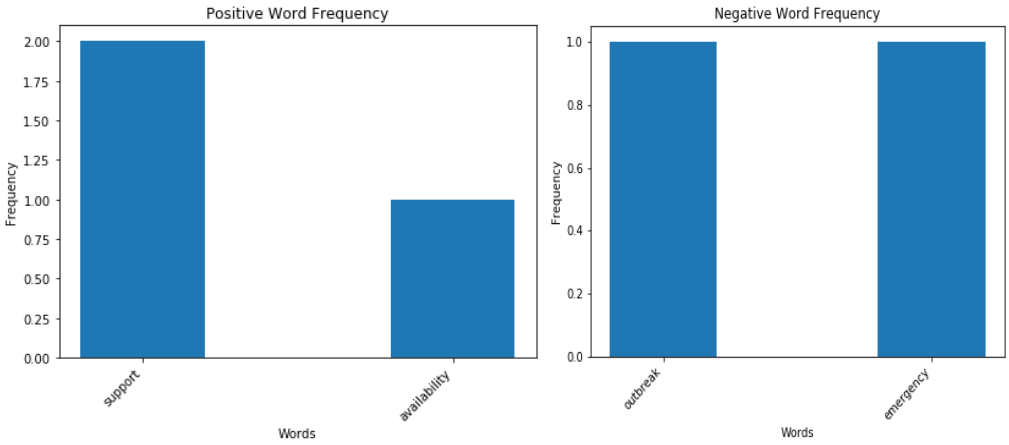

FIGURE 3.4.2. SENTIMENT WORD FREQUENCY FOR MARCH 9 WHO REPORTS

IV. MARCH 19TH

4.1 Topic Modeling Tweets

On March 19th, the Shelter-in-place was announced. Globally there were 209,839 cases and the United States had 7,087 cases. Three topics were selected to identify the different subjects that were tweeted on this day. Topic# 0 tells us that users were tweeting about locations that had the largest impact

with the virus. During this time, China and Italy had the most cases. Topic# 1 tells us that Twitter users were tweeting about Trump and the fear users were experiencing. Lastly, Topic# 2 tells us Twitter users were tweeting about the effects of the coronavirus. People were urged to practice social distancing and to stay at home.

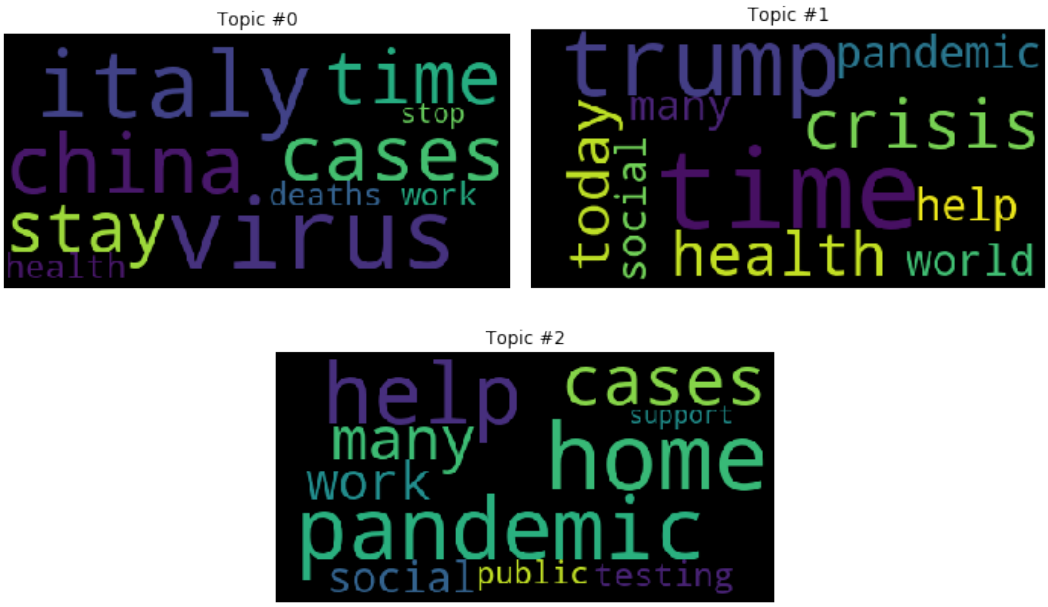

FIGURE 4.1. TOPIC MODELING FOR MARCH 19 TWEETS.

4.2. Word Frequency and Sentiment Analysis on March 19th Tweets

On March 19th, the most common words that were tweeted in relation to coronavirus included ‘time’, ‘stay’, ‘home’, and ‘work’. Users tweeting ‘time’ most frequently tells us that they are trying to understand the current moment and adjust to the new reality of how shelter in place is affecting their lives.

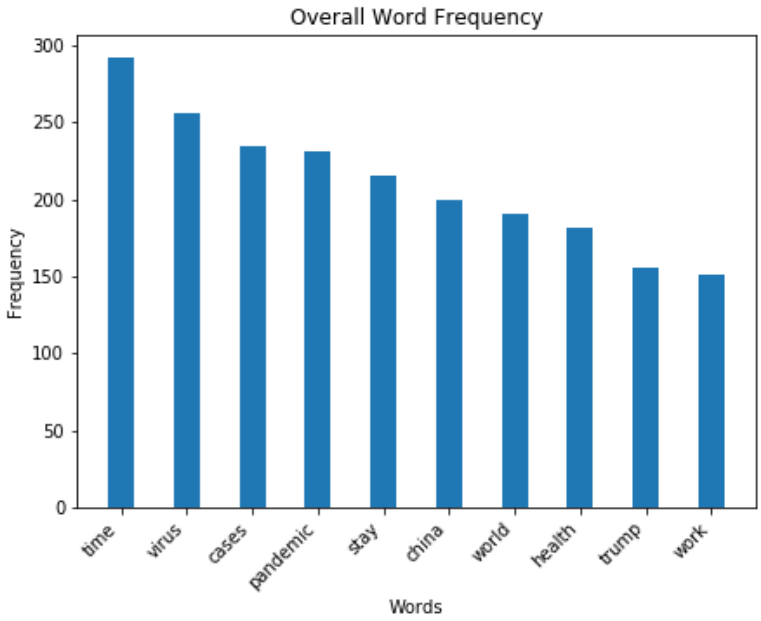

FIGURE 4.2.1 OVERALL WORD FREQUENCY FOR MARCH 19 TWEETS.

For words with positive sentiment, words such as ‘safe’, ‘support’, ‘care’, and ‘thank’ were frequently tweeted. This suggests that users were tweeting about the announcement of the Shelter in Place with a sense of safety and regards to protecting their health, and being appreciative.

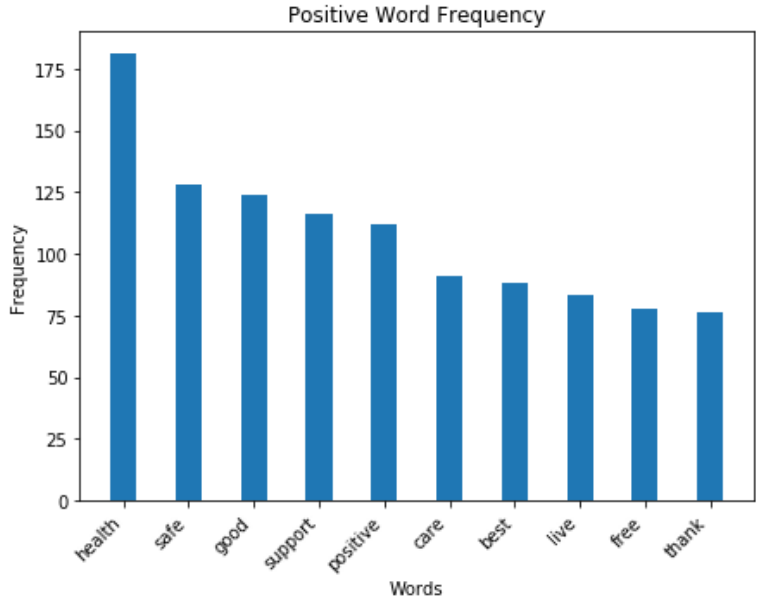

FIGURE 4.2.2. POSITIVE WORD FREQUENCY MARCH 19 TWEETS.

For words with negative sentiment, most commonly tweeted words included ‘crisis’, ‘death’, ‘sick’, and ‘fight’. This tells us that users were tweeting the negative effects of the coronavirus.

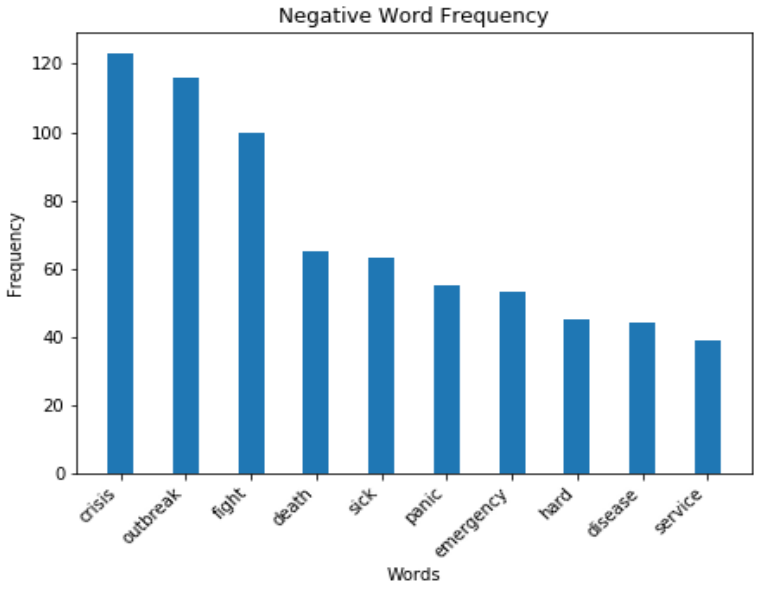

FIGURE 4.2.3. NEGATIVE WORD FREQUENCY MARCH 19 TWEETS.

When comparing the positive word frequency with the negative word frequency for the tweets on March 19th, we see that words such as 'health', 'safe', and 'good' were more frequently tweeted than words such as 'crisis', 'outbreak', and 'fight'. This tells us that users were more concerned for their own safety and health during the shelter in place announcement and tweeted with a more positive sentiment.

4.3. Topic Modeling WHO Report

We can see from Topic #0 below and as well as Topic #1, that the WHO Report for March 19th is primarily talking about the ongoing investigations on the new cases of infections in different areas. While topic# 2 is talking about the updated protocol on the infections and identifying the antibodies. Generally these topics will be similar since the WHO is addressing the issue of coronavirus.

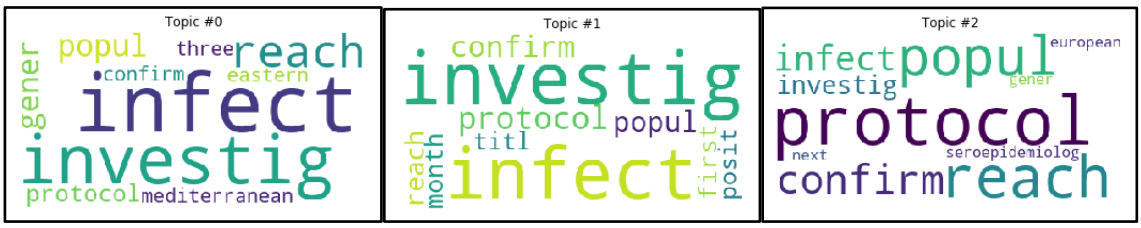

FIGURE 4.3.1. TOPIC MODELING FOR MARCH19 WHO REPORT.

4.4. Word Frequency and Sentiment Analysis WHO Report

The most common words in the daily WHO reports were mainly discussing new Covid-19 cases in different areas. These terms are primarily informative and a collection of numerical data for the general public.

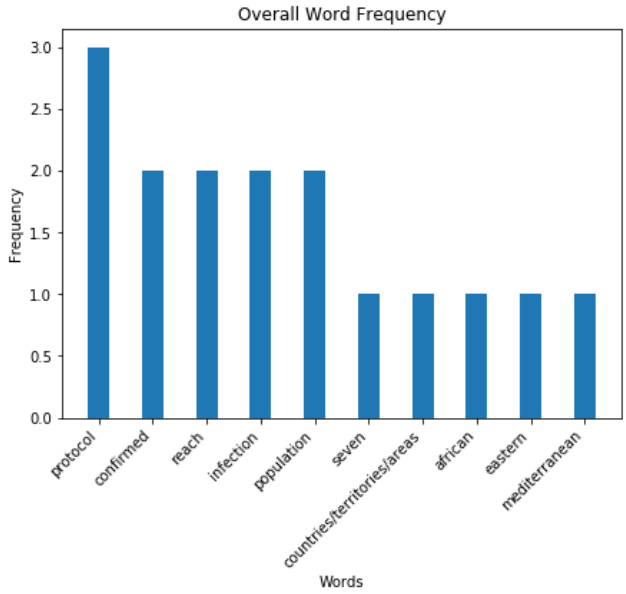

FIGURE 4.4.1 OVERALL WORD FREQUENCY FOR MARCH 19 WHO REPORT

For words with positive sentiment, words such as 'health', 'essential', 'inform', and 'understand' were frequently used suggesting the resources to be provided and the actions to be taken.

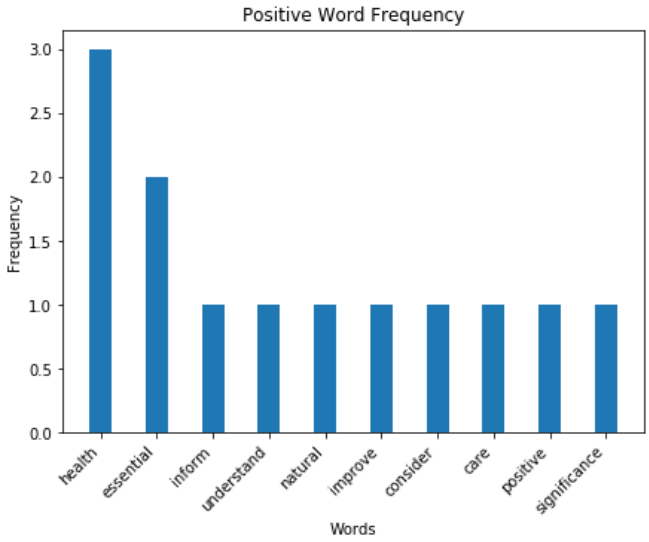

FIGURE 4.4.2. POSITIVE WORD FREQUENCY FOR MARCH 19 WHO REPORTS

While words associated with negative sentiment included 'infection', 'decline', 'outbreak', and 'help' suggest panic and the significance of the situation.

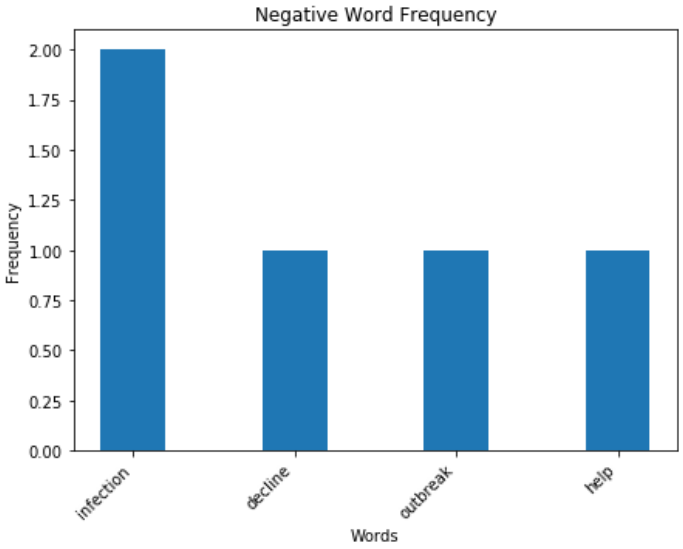

FIGURE 4.4.3. NEGATIVE WORD FREQUENCY FOR MARCH 19 WHO REPORT

When comparing the positive word frequency with the negative word frequency, we can see that words such as 'health' and 'essential' were more frequently used than 'infection' and 'outbreak' which tells us that the WHO is reporting highlights with a more positive sentiment.

V. MARCH 30TH

5.1. Topic modeling Tweets

As of March 30th, the number of people infected by Covid-19 in the United States went up to over 120,000 cases (more than double compared to March 19th) as the economy continued to struggle with the ongoing effects of the worldwide pandemic. We will continue to investigate how people's feelings change toward the novel coronavirus. First of all, we will look at trending tweets regarding feelings toward this situation. A word cloud is presented by topic modeling to demonstrate what people are talking toward Covid-19.

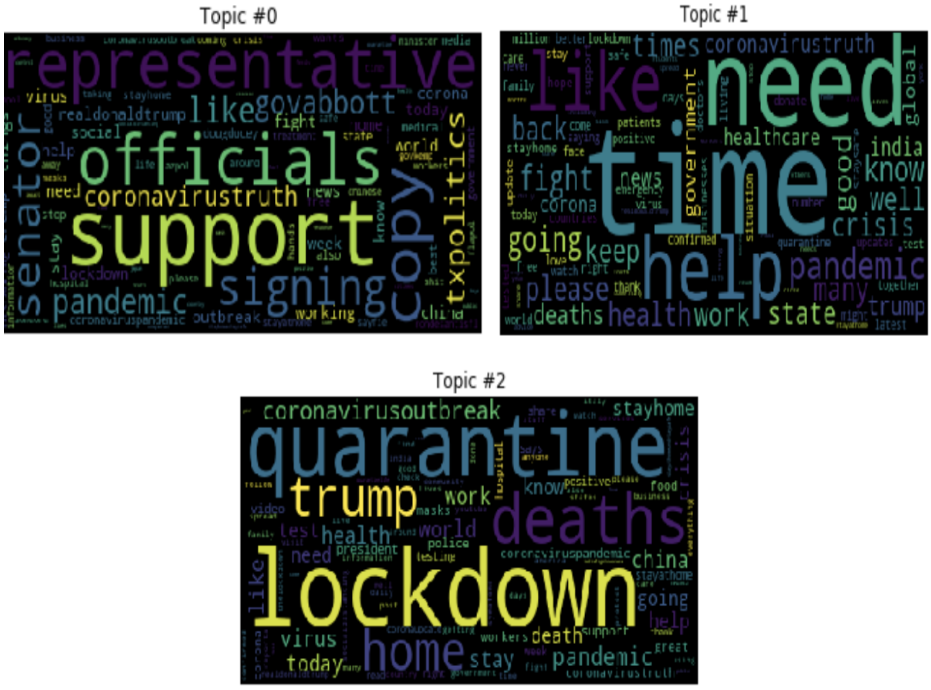

FIGURE 5.1. TOPIC MODELING FOR MARCH 30 TWEETS.

As Covid-19 rapidly spreads, people are paying more attention to supporting the first responders. It was the first sign of positivity that people were united and fighting together. In topic 0, politics seem to be involved in people's minds since a lot of words like 'official', 'representative', 'senator', 'txpolitics' were trending. In addition, in topic 1, people have been seeking for help while the need for health care increases rapidly due to the surging amount of confirmed cases. In the last topic, the shelter-in-place was still in order. People were talking about the quarantine, lockdown, and stay-at-home. Also, the death tweets also increase gradually as an indication of the rise in the number of deaths. In general, besides negative tweets like 'quarantine', 'lockdown', and 'deaths', people also offered a lot of support and help to each other.

5.2. Topic Dominance

After getting 3 most frequent topics, we will continue to evaluate which topic people talk about the most or can be considered the most important topic. Topic distribution is an important point to pay attention to in this analysis. Each Tweet consists of various topics. However, each Tweet will always have one topic that has the highest distribution score. This is called the topic dominance. Out of the 5000 tweets, we will continue to analyze which topic people are talking about the most using topic contribution score.

```
df_dominant_topic.head()
```

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	1.0	0.9540	support, officials, representative, signing, c...	[science, hurting, government, response, repor...
1	1	1.0	0.5696	support, officials, representative, signing, c...	[coronavirusoutbreak, china, unexpectedly, cut...
2	2	1.0	0.9137	support, officials, representative, signing, c...	[citizen, york, commits, suicide, infected, co...
3	3	0.0	0.9589	home, time, pandemic, stay, lockdown, quaranti...	[hmrc, launched, information, service, whatsap...
4	4	2.0	0.9586	health, says, help, today, lockdown, spread, v...	[andekhaasach, infestations, like, specially, ...

FIGURE 5.2.1. TOPIC DOMINANT FOR MARCH 30 TWEETS.

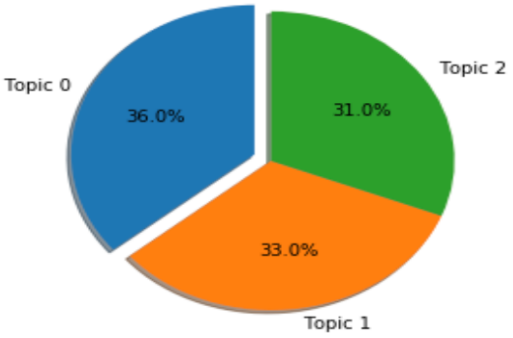

FIGURE 5.2.2. TOPIC DOMINANT FOR MARCH 30 TWEETS.

As shown by the pie chart above, people generally talk proportionately about all three topics. Even through this pandemic, tweets are assumed to be speaking of fear and anger, but the results turned out surprisingly opposite. People seem to pay more attention to supporting others and politics.

5.3. Sentiment analysis Tweets

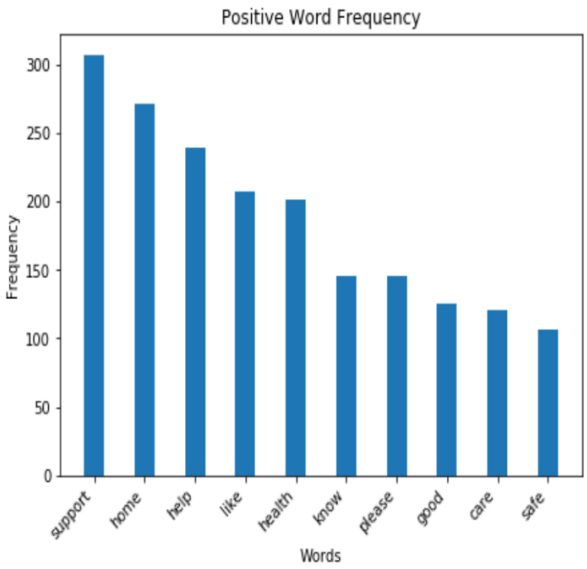

FIGURE 5.3.1. POSITIVE WORD FREQUENCY FOR MARCH 30 TWEETS.

For positive words that people tweet the most frequently are ‘support’, ‘home’, ‘care’, ‘safe’, etc.

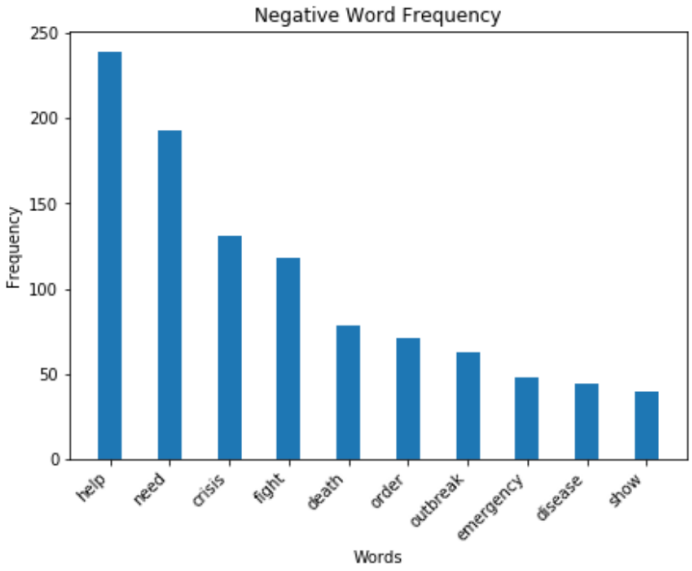

FIGURE 5.3.2. NEGATIVE WORD FREQUENCY FOR MARCH 30 TWEETS.

For negative words, people tweet mostly about getting help. They are worried about a crisis and outbreak. The death rate is also mentioned more often.

5.4. Topic Modeling WHO Report

Topic# 0 is mainly about the essential services and maintaining health in the countries affected. Topic# 1 and Topic #2 are similar to Topic# 0 where it is addressing the people's health and maintaining our current state.

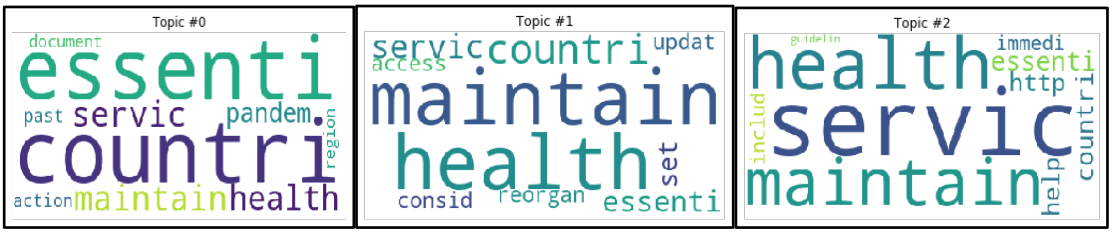

FIGURE 5.4.1. TOPIC MODELING FOR MARCH 30 WHO REPORT.

5.5. Word Frequency and Sentiment Analysis WHO Report

In the overall word frequency for the March 30th WHO Report we see that words such as 'countries', 'maintain', 'health', and 'services' were primarily used. This tells us that the cases around the world are still continuing and our health is what is important to maintain.

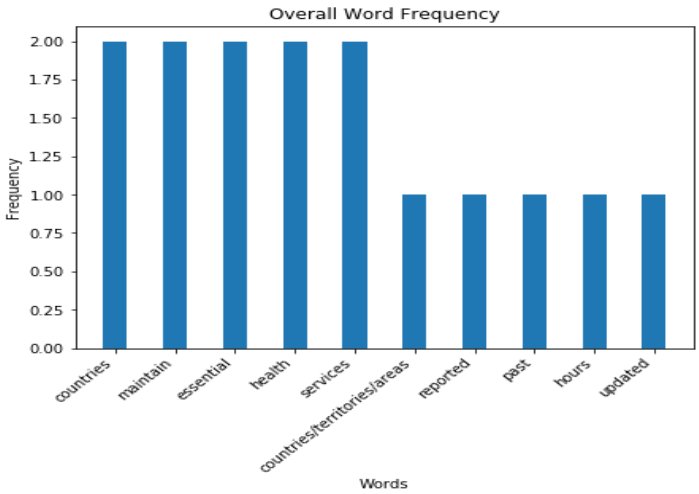

FIGURE 5.5.1. OVERALL WORD FREQUENCY FOR MARCH 30 WHO REPORT

When analyzing the word frequencies for positive and negative sentiment, we see that the WHO Report used words with a positive sentiment with a higher frequency. Although 'help' is categorized as both a negative and positive sentiment, we see that the numbers of words associated with a positive sentiment is greater than the words associated with a negative sentiment.

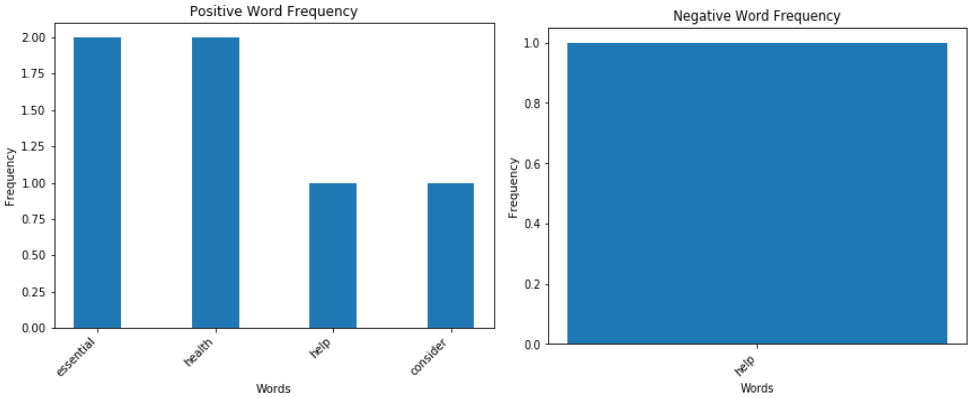

FIGURE 5.5.2. POSITIVE AND NEGATIVE WORD FREQUENCY FOR MARCH 19 WHO REPORTS

VI. RESULTS

Figure 6.1 depicts the polarity and subjectivity of the WHO reports for each day. There is little variation in the sentiment in each report. On average, the polarity was 0.05 meaning the emotion within the report is neutral. The average subjectivity of the reports was 0.37 meaning the reports are more objective. Overall, the reports stay true to the facts of daily activity related to COVID-19. The reports would serve as a neutral baseline for analyzing the sentiment of public opinion via Twitter.

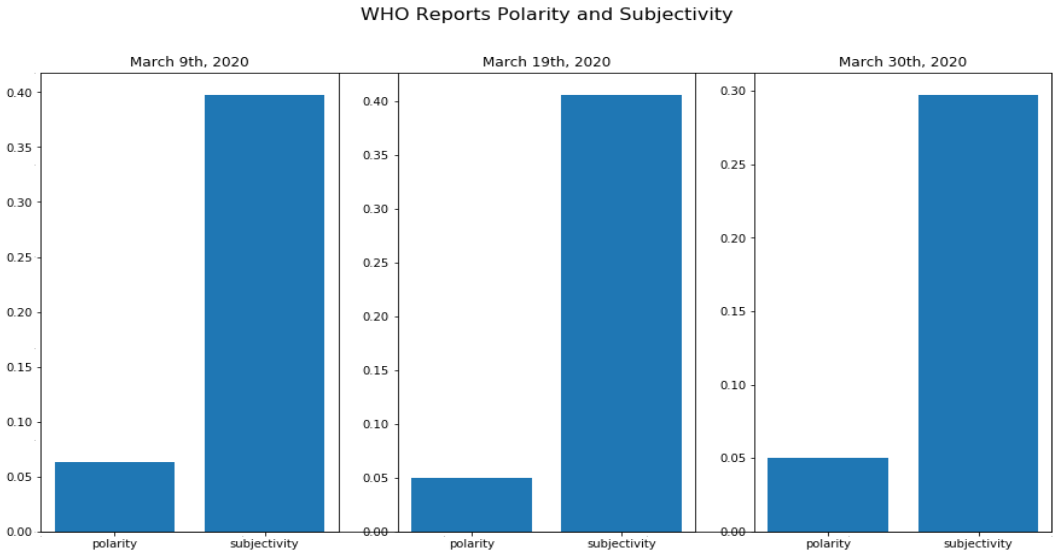

FIGURE 6.1 Sentiment Analysis of WHO Reports by Date

In figure 6.2, we analyze the sentiment of tweets by date. We categorized the sentiment of each tweet and created a histogram of the polarity. We see that extreme negative and positive polarity are consistent for each day. While there is a gradual decrease in the number of slightly negative tweets over time. The number tweets with neutral emotion are lowest when lockdowns were implemented, and are highest when lockdowns are extended. Tweets with slightly positive emotional language are highest when lockdowns are implemented and lowest prior to the declaration of a pandemic. This suggests that fear was greatest prior to lockdown and the public was in favor of extending lockdown in support of flattening the curve and supporting health care and essential workers.

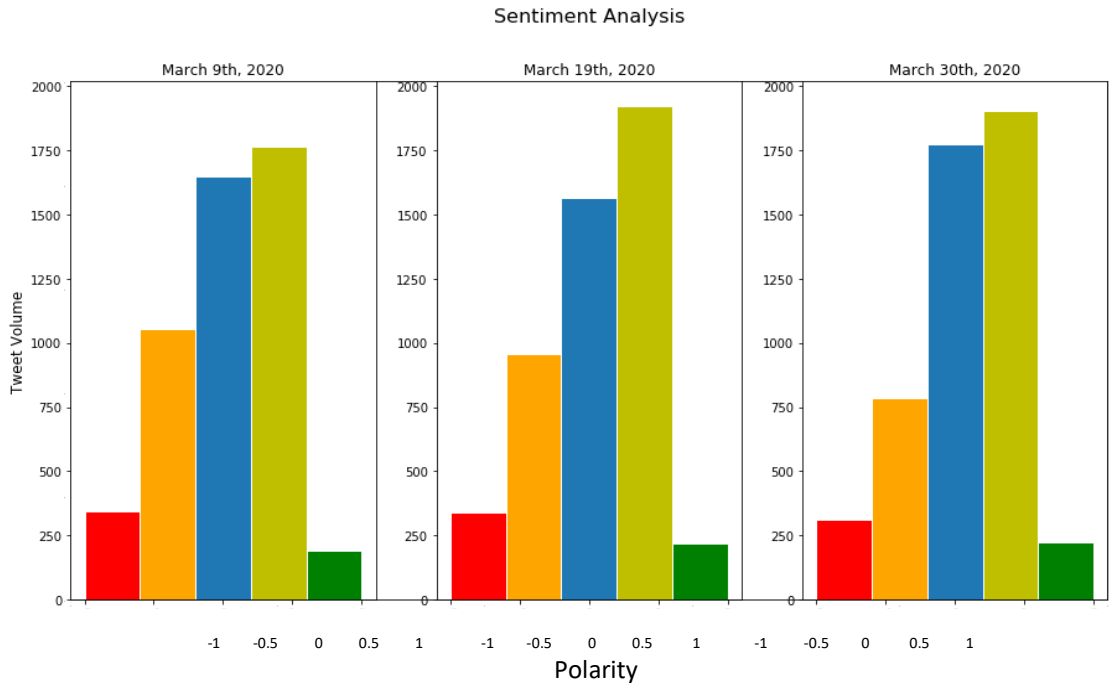

FIGURE 6.2. Sentiment Analysis of Tweets by Date

VII. CONCLUSION

In conclusion, WHO reports have little influence on public opinion on COVID-19 on Twitter. The emotional language used in tweets do not depend on the daily highlights in the WHO reports. The WHO reports are true reports that remain generally neutral and objective according to the sentiment analysis by date. Yet the number of tweets with medium sentiment with slightly positive and negative emotion vary by date. However, tweets with extremely negative or positive polarity are stable for each date. We can infer that initial reports have more influence on opinion, but over time public opinion gradually depends more on government action to the virus rather than reports on COVID-19 daily activity.

However, there are similarities in the topics between the WHO reports and tweets. Key topics throughout all three days were for support and maintenance of health. We see cohesion on relaying messaging on the importance of health and support of governments and other countries. This supports the presence of a relationship between official and public texts. In this sense, there is transparency in information and dialogue between these two platforms.

VIII. REFERENCES

Smith, Shane. "Coronavirus (covid19) Tweets." Kaggle, Mar. 2020, www.kaggle.com/smid80/coronavirus-covid19-tweets/data#2020-03-12%20Coronavirus%2Tweets.CSV. (accessed May 10, 2020)

World Health Organization. "COVID-19 Situation Reports." *World Health Organization*, World Health Organization, Mar. 2020, www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. (accessed May 10, 2020)