# Diachronic Semantic Change in Tweets About COVID-19: An NLP Based Frequency and Sentiment Analysis

Melis Gökalp
Brown University
Providence, Rhode Island
melis_gokalp@brown.edu

**Abstract**

The COVID-19 crisis has influenced our daily discourse and influenced communications drastically in a short period of time. As the world is social distancing, online communication became even more prevalent. Twitter provides real-time information streams that can be used for natural language analysis due to its large number of users. To demonstrate the semantic change surrounding Coronavirus discourse over time, I propose a diachronic analysis of sentiment, word frequencies and word embeddings. It is shown in this paper that the words related to Coronavirus have emerged, changed meaning and sentiment.

This research asks the questions of: “How does language change over time”, “Do some words and concepts gain importance?” and “How do some words change meaning drastically and shift sentiments?”

## 1. Introduction

Our lifestyles changed suddenly between December 2019 and May 2020. The world has not seen such a global crisis since the Second World War. It seems like the effects of this pandemic will stay with us longer. It is then expected that there will be changes in the way we speak and interact with each other.

Twitter has been an extensively used channel of communication, especially during the time of Coronavirus pandemic. The words we associate with death, illness and health started appearing more on the news and the everyday discourse has been revolving around Coronavirus. The pandemic experiences of people around the world have impacted on the language used as well as changing the politics of Coronavirus and its meaning.

My research aims to diachronically analyze semantic patterns in frequency, sentiments and word vector embeddings. To achieve this, I used different methods of Natural Language Processing (NLP) such as word vectors, sentiment analysis using a Naïve Bayesian Classifier and word frequency analysis.

This article will refer to the novel coronavirus disease (COVID-19 or 2019-nCoV) as Coronavirus and keeping in mind that this crisis situation is still ongoing, everything mentioned in this paper is valid as of the date of writing, although this information might change later.

## 2. Previous Research

Twitter sentiment analysis is a popular topic in linguistics and natural language processing. My research topic in particular is related to how people use and change language during crisis times. Khatua et al.'s paper about the accuracy of Twitter information during 2014 Zika and 2016 Ebola pandemics was inspiring for me. They classified Tweets from these pandemics using Word2Vec and Stanford's Glove[1]. They created a Twitter corpus and trained on both models with different pretrained weights on Google News and PubMed. I found their choice of sources interesting because News also tend to spread false information and PubMed is too scientific for

[1]Khatua, Aparup, Apalak Khatua, and Erik Cambria. "A Tale of Two Epidemics: Contextual Word2Vec for Classifying Twitter Streams during Outbreaks." *Information Processing & Management* 56, no. 1 (2019): 247–57. https://doi.org/10.1016/j.ipm.2018.10.010.

an average Internet user. Nevertheless, their results are impressive, and they prove how word embeddings can be useful for linguistic tasks such as truth labeling Tweets.

Another paper that I found similar to my research is the Muhammad Imran's paper about classifying crisis-related Tweets. They have used a Naïve Bayes classifier, Support Vector Machine (SVM) to classify around 52 million crisis-related tweets collected during 19 crisis events. Their classification results were as high as 90% for some categories. Both papers mentioned here used cumulative data for every crisis situation. Distinctively, my research will be analyzing change over six months to provide information about how the sentiments and meaning changed through the course of pandemic. I will be combining word vector, frequency and sentiment analysis to find semantic relationships between words in Coronavirus Tweets.

## 3. Data Collection

Since Coronavirus emerged in December 2019, I decided to start collecting data starting from 1st of December 2019, until 15th of May 2020. This time frame is sufficient enough to observe changes and trends in tweets related to COVID-19.

The naming of Coronavirus has been a controversial and sensitive topic and its names in the scientific community and public were different. Having this in mind, I chose the Coronavirus keywords as "coronavirus", "corona", "COVID", "COVID-19". This variety of expressions would help us extract data from official sources as well as personal Twitter accounts. The goal of this research is to analyze the language used in the Twitter accounts belonging to not only news organizations but also individuals who are using Twitter daily. Our perception and sentiments are shaped by the information we receive from news sources and our interactions with people. Official sources may refer to coronavirus as "the novel coronavirus", "COVID-19" or "COVID" while an average Twitter user may use "corona" or "coronavirus". Therefore, this choice of keywords allows us to capture data from a variety sources and have a more accurate representation of the online discourse about Coronavirus.

### 3.1. Potential problems in Twitter data

Although online data mining might sound convenient and objective, there are certain drawbacks and disadvantages of using Twitter data for linguistic analysis purposes. The quality and content of the data is mostly dependent on the intentional choices made by the researcher. Since there are over 500 million Tweets[2] tweeted daily, choosing the most relevant and influential ones is necessary. I chose to scrape 5000 "Top Tweets" per month about Coronavirus, which is decided by Twitter's "Top Tweets" search algorithm. According to Twitter, this is determined by the popularity of the Tweet, based on the number of people interacting with and sharing it based on the relevance of the query. This seems reasonable for the scope of this research since it provides data based on interaction. Nevertheless, Twitter keeps this algorithm confidential and it is free to censor and manipulate their data. Another problem is the abundance of misinformation on Twitter. Although Twitter has taken action to prevent this and started warning users about false claims, the independent fact checkers found that around 60 percent of the false claims about Coronavirus remain online on Twitter.[3] The possible effects of misinformation and censorship should be kept in mind for the objectivity of the results. In fact, after my data collection for this research was completed, the Twitter search queries about "coronavirus" started being blocked and users are shown links to the websites of relevant health authorities. This means that further data collection for this research is no longer possible until

[2] Internet Live Stats. Twitter Statistics. https://www.internetlivestats.com/twitter-statistics/
[3] Timberg, Craig. Washington Post, "On Twitter, almost 60 percent of false claims about coronavirus remain online"https://www.washingtonpost.com/technology/2020/04/07/twitter-almost-60-percent-false-claims-about-coronavirus-remain-online-without-warning-label/

Twitter decides to unblock the search functionality for Coronavirus. As much as Twitter has taken these steps to prevent the spread of misinformation, this is a form of information control and might have further implications such as preventing extensive data collection for research purposes.

Scraping Twitter data is a challenging task because of the limitations of the software and the nature of the online environment. Since raw Twitter data needs to be pre-processed and cleaned up for model training and analysis, the search query only included only English Tweets. Twitter's official API only allows the use of tweets in the last week, which makes it useless for diachronic analysis. As an alternative, I used a Python library called GetOldTweets3, which saves Twitter data by making search query requests to Twitter's search engine. I have decided to use 15-day intervals to split the data so that the diachronic change could be portrayed clearly. A summary of the Twitter data collected is given below:

**Table 1.** Number of Top Tweets containing "coronavirus", "corona", "COVID", "COVID-19", extracted in 15-day intervals

| Date interval | Tweets | Words | Tokens |
|---|---|---|---|
| 12/01/2019-12/15/2019 | 167 | 2180 | 1400 |
| 12/15/2019-01/01/2020 | 167 | 2392 | 1378 |
| 01/01/2020-01/15/2020 | 2500 | 33846 | 6926 |
| 01/15/2020-02/01/2019 | 2500 | 35988 | 6878 |
| 02/01/2020-02/15/2020 | 2500 | 38118 | 7162 |
| 02/15/2020-03/01/2020 | 2500 | 38050 | 7698 |
| 03/01/2020-03/15/2020 | 2500 | 39053 | 8557 |
| 03/15/2020-04/01/2020 | 2500 | 39165 | 8782 |
| 04/01/2020-04/15/2020 | 2500 | 38895 | 9113 |
| 04/15/2020-05/01/2020 | 2500 | 39185 | 8966 |
| 05/01/2020-05/15/2020 | 2500 | 39153 | 9077 |
| Total | 22834 | 346025 | 75937 |

The number of Tweets containing the Coronavirus keywords from December 2019 is only 334 because the Coronavirus pandemic was not known. This gives us an opportunity to use December 2019 data as a control group and observe the linguistic change in the following months. The meaning of the words "corona" and "coronavirus" was different and natural language processing analysis allows us to observe this change.

There are 346,025 words in the corpus I created for this research, which is not very large compared to the standards of computational linguistics research today. Given the constraints of computing power and time, I decided to limit the number of Tweets per month to this number and I believe that it is sufficient enough to make meaningful analysis and address the research question.

### 3.2. Pre-processing Data

The raw Twitter data needs to be cleaned, parsed and tokenized to be ready for analyzing. By removing special characters, numbers and links from the Tweets, we only leave the words that will be added to our corpus. Internet text requires more pre-processing than regular texts because emojis, links and hashtags also need to be filtered out. In order to get rid of useless data, English stop words are removed, the remaining text is made all lower-case and tokenized. I used the Natural Language Toolkit (NLTK) for Python in order to find the list of stop words and filter them out. A sample Tweet before and after pre-processing is given below to illustrate this process:

(1) "#BREAKING - Jamaica sees 13 coronavirus recoveries; 2 additional cases May 13th Jamaica COVID19 by the Numbers Confirmed cases: 509 (+2) Tests completed: 7,552 (+87) Recovered: 113 (+13) Deaths: 9 pic.twitter.com/4Ja8R9SlL0"

(2) [“jamaica”, “sees”, “coronavirus”, “recoveries”, “additional”, “cases”, “may”, “jamaica”, “covid19”, “numbers”, “confirmed”, “cases”, “tests”, “completed”, “recovered”, “deaths”]

As it is seen in (1), the Tweet text sample contains punctuation, hashtag and numbers, which are removed in the pre-processed version in (2). Each word becomes a token and contribute to the corpus. This tokenized Tweet will be increasing the number of words related to testing, recovery and death in Coronavirus. For different kinds of analysis, I have prepared tokenized text for monthly and 15-day intervals.

## 4. Methodology

### 4.1. Research Objectives and Questions

Since my goal is to find out how the Twitter discourse related to Coronavirus has changed over 6 months, using a statistical approach is a reasonable idea. I aim to compare the frequencies of words, sentiments, word similarities by using a wide range of tools and visualize them. I will also try to answer the following questions:

How did the keywords related to Coronavirus “coronavirus”, “corona”, “COVID”, “COVID-19” appear or change meaning?

How did our sentiment towards Coronavirus change online overtime?

As the politics and geography of the pandemic shifted how did we start thinking about it differently?

What does the increase in frequency of appearance of new words tell us?

**Frequency Distribution**

Frequency Distribution is informative about the topic and genre of the text. It can be used as an effective method of showing which words have gained importance over time. After gathering tokens from the data preprocessing phase, I used counting tools in order to find word frequencies for every 15 days and every month, as well as cumulative. I tried both NLTK and another Python NLP library called TextBlob. TextBlob is a more optimizable tool that does not remove non-dictionary words or random utterings. Although it might be viewed as noise in the data, I found it useful to extract more information from the dataset such as abbreviations or neologisms. Therefore, I decided to represent both frequency distributions and include them in my findings.

### 4.2. Word Vectors

I aim to find semantic relationships between words in my Tweet corpus and map them in order to show how words have changed meaning throughout the past 6 months. I created word vectors by training the Tweet text tokens on a word embedding model called Word2Vec. Word2Vec is used for representing words in a vector form. It is useful for finding semantically related words by calculating the nearest neighbor distance of every word within the corpus. The Euclidean distance for points d(a,b) is:

$$d(a,b) = \sum_{1 \le i,j \le 28} (a_{ij} - b_{ij})^2$$

The Euclidean distance is calculated for every neighbor of the word and when the nearest two distances are proportioned to create a similarity score.

These models can be customized for the dataset by adjusting the window size, vector dimensions and minimum vector size. By mapping the words to a vector space, I also visualized all of the corpus and the distance between the words. The closer the words, the more related they are. I also listed the most similar words to a given keyword such as "coronavirus" by listing their similarity coefficients. The change in the most similar words also indicate a change in the meaning of that word and how we semantically associate it with other words. The word "corona" is a dramatic example because it used to commonly refer to a beer brand but now it is related to a pandemic.

### 4.3. Sentiment Analysis

Tweets are used to communicate knowledge, thoughts and feelings. In order to find out how Twitter users' sentiments about Coronavirus have changed, I performed a sentiment analysis using the VADER library based on Hutto and Gilbert's paper with the same title.[4] VADER is a rule-based classification model that uses a Naïve Bayes algorithm to classify texts by scoring them for negativity, positivity or neutral. Then it creates a compound score using these three components. If the compound score is between -0.05 and 0.05, the text is neutral, if it is above 0.05 then it is positive and if it is below -0.05, it is negative. VADER is human-centric and pre-trained on social media text, as it is sensitive to use of slang words, exclamations and capitalization. This made VADER a great tool for performing sentiment analysis. After I generated negative, positive, neutral and compound scores for every Tweet, I took their averages for every 15 days and plotted the sentiment change in texts. I expected to find the sentiment worsen about Coronavirus.

[4] Hutto C, Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: International AAAI Conference on Weblogs and Social Media. AAAI; 2014. p. 216–225.

## 5. Results

I have gathered a large variety of results using the methods described in the methodology section. I visualized and documented these results by creating tables, graphs, frequency distribution charts and word mappings. Some of the data shows correlations and there are trends that indicate a semantic change in meaning of the words related to Coronavirus.

### 5.1. Frequency Analysis

I calculated the word frequencies in the Tweet corpus using both TextBlob and NLTK's frequency distribution functions as explained in my methodology. The Lexical distribution plot of the 20 most frequent words using NLTK are given:

**Figure 1: Lexical Dispersion of 20 Most Frequent Words using NLTK**

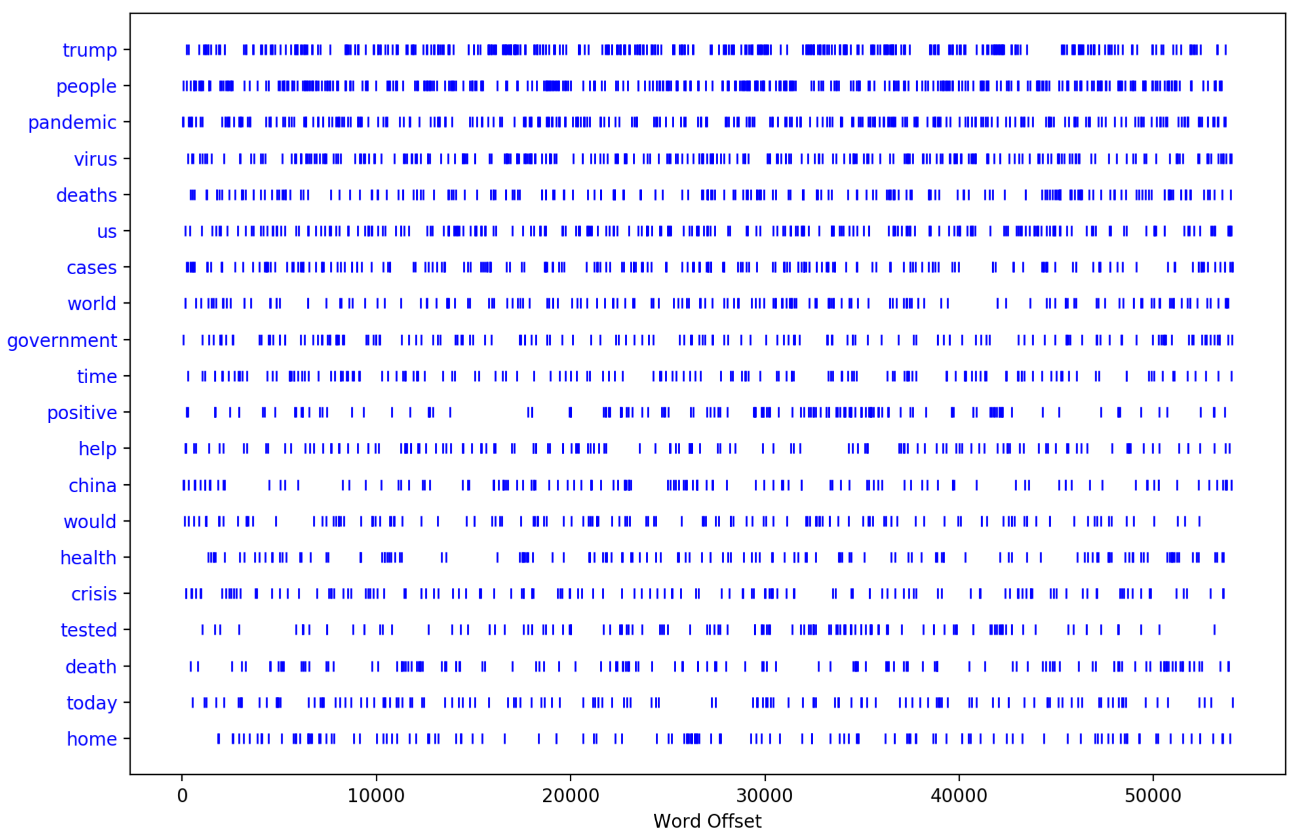
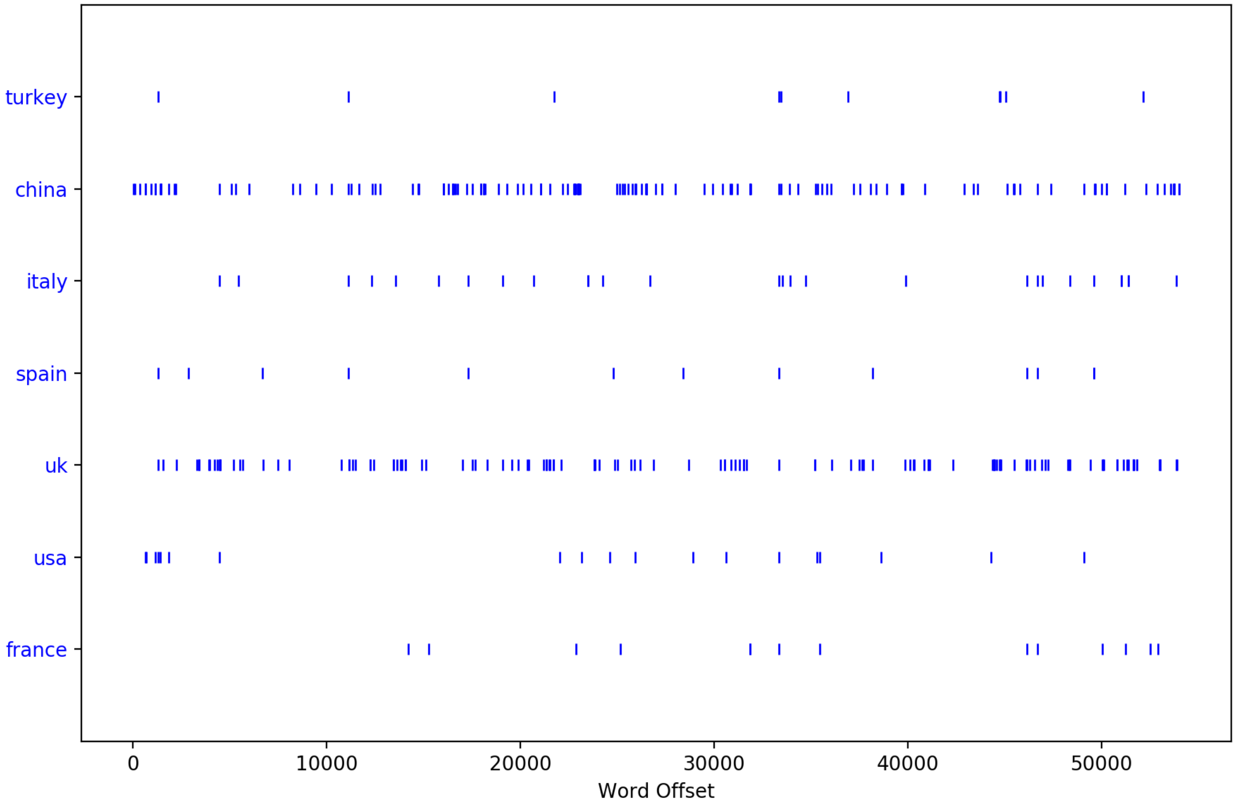


**Figure 2 Lexical Distribution of Country names using NLTK**




Lexical dispersions are also indicative of the change over time of these words. If they become more or less frequent, it indicates that Coronavirus related words occur less with these words. For instance, among country names, China has a very dense lexical dispersion starting from the beginning of the word offset. This indicates that China has been involved in the Coronavirus discourse from the beginning whereas France and Italy have appeared towards the middle and increased in frequency. This can be considered as a result of the geographical shift of the pandemic epicenter from Asia to Europe. The disadvantage of using NLTK for this task is that it removes some of the abbreviations such as the UK and the USA. I chose not to include "us" as an abbreviation for the United States, however even "usa" returned low frequency.

**Figure 4 Frequencies over time of Highest Frequency Words Using NLTK**

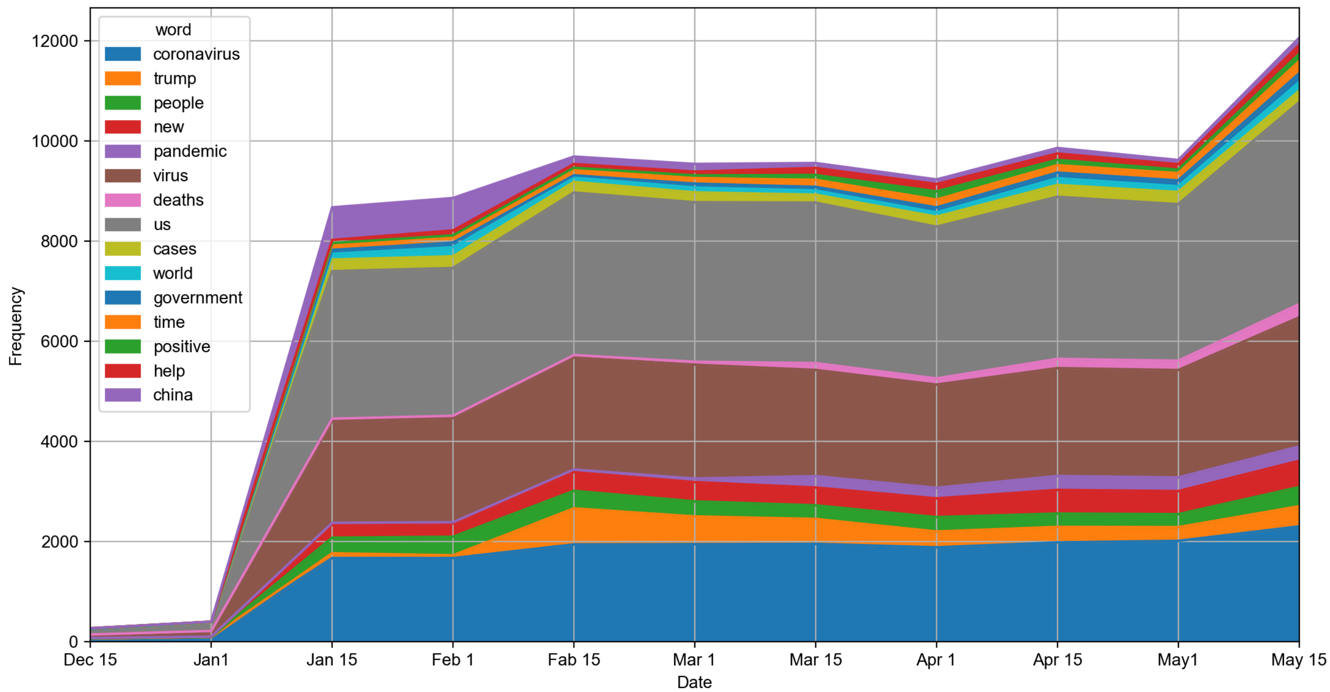
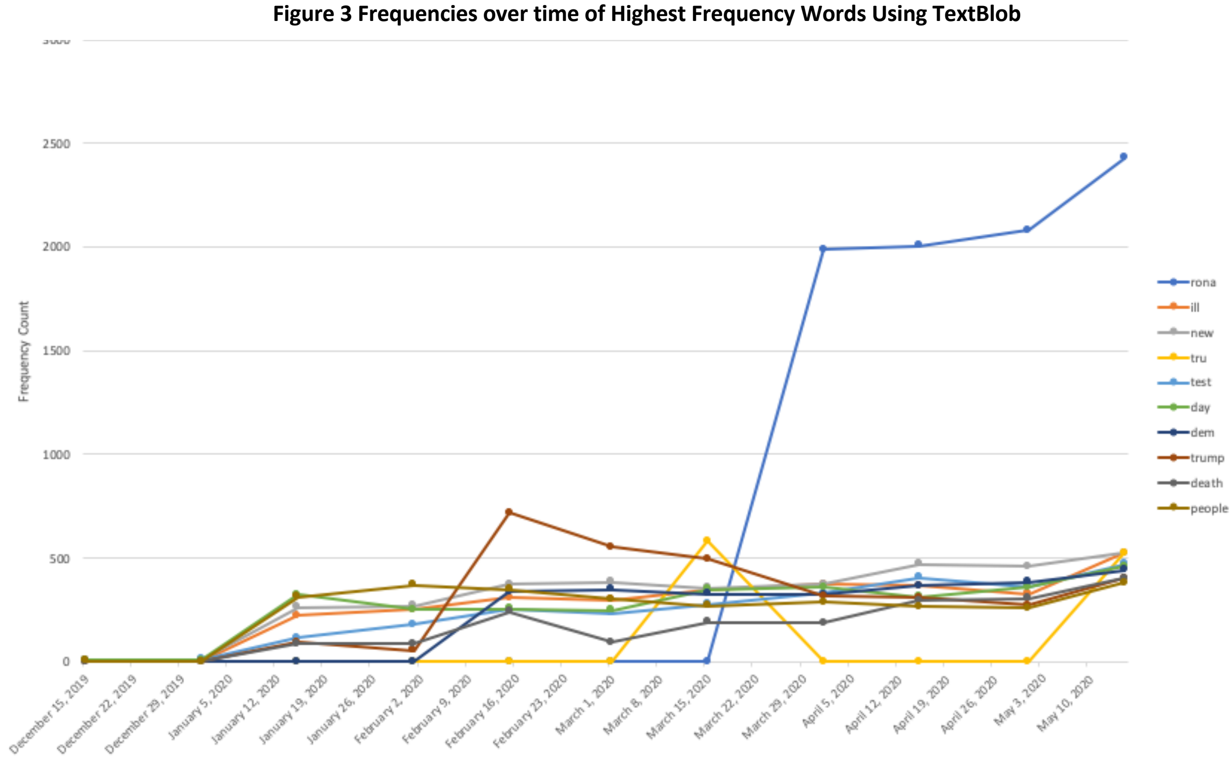



**Figure 3 Frequencies over time of Highest Frequency Words Using TextBlob**



The two graphs using TextBlob and NLTK are quite different although they both represent the most common monthly frequencies of the same dataset. TextBlob includes shorter words such as “tru”, “ill” and “dem”. These are most likely internet abbreviations and slang considering their high frequency within the corpus.

The Table 2 showing the change of most frequent words related to Coronavirus had very interesting findings. The first two columns representing the Tweets from December 2019 had the expected words “drink” and “beer” in them because of the Corona beer. However, “constellation”, “solar” and “borealis” and “Centennial” were unexpected. After doing some research, I learned that corona refers to “something suggesting a crown” or “a usually colored circle often seen around and close to a luminous body”[5]. NASA also completed a Solar Probe, which was an exciting event for the scientific community.[6] Additionally, Corona Borealis is a star constellation that resembles a crown. These associations indicate that the word Corona used to be more exclusive to the scientific community. After January 2020, the word was associated with words such as China, Wuhan, emergency, virus and outbreak. This complete change in meaning and its association with the Coronavirus crisis is a very interesting finding. The word “pandemic” appears in March 2020, which corresponds to the date of March 11, 2020 when the WHO declared the Coronaviruses crisis a global pandemic[7].

[5] *Merriam-Webster.com Dictionary*, s.v. “corona,” accessed May 12, 2020, https://www.merriam-webster.com/dictionary/corona.

[6] *NASA. “First NASA Parker Solar Probe Results Reveal Surprising Details About Our Sun”* accessed May 12, 2020, https://www.nasa.gov/press-release/first-nasa-parker-solar-probe-results-reveal-surprising-details-about-our-sun

[7]Ducharme, Jamie *New York Times,* “World Health Organization Declares COVID-19 a 'Pandemic.' Here's What That Means*”.* https://time.com/5791661/who-coronavirus-pandemic-declaration/

**Table 1 The Most Frequent Words in Coronavirus Tweets over time**

| 15-Dec-2019 | 1-Jan-2020 | 15-Jan-2020 | 1-Feb-2020 | 15-Feb-2020 | 1-Mar-2020 | 15-Mar-2020 | 1-Apr-2020 | 15-Apr-2020 | 1-May-2020 | 15-May-2020 |
|---|---|---|---|---|---|---|---|---|---|---|
| centennial | del | china | china | trump | virus | people | people | coronavirus | pandemic | trump |
| year | ca | virus | virus | people | cases | trump | pandemic | new | rona | rona |
| new | mar | people | health | health | people | health | health | cases | coronavirus | new |
| one | sun | outbreak | people | first | trump | us | cases | pandemic | cases | pandemic |
| get | borealis | health | outbreak | us | health | cases | new | deaths | today | virus |
| christmas | solar | wuhan | wuhan | virus | hoax | pandemic | us | us | health | deaths |
| time | time | cases | cases | cases | new | new | help | health | covid | us |
| del | constellation | us | confirmed | state | case | help | positive | trump | us | cases |
| like | like | chinese | chinese | death | outbreak | today | trump | state | deaths | one |
| today | game | confirmed | novel | hoax | us | positive | today | covid | positive | response |
| mar | constellation | new | chinese | new | one | one | response | today | response | like |
| drink | team | novel | first | washington | flu | deaths | covid | help | help | world |
| today | season | first | emergency | spread | china | please | crisis | workers | trump | government |
| back | football | emergency | spread | outbreak | spread | response | get | patients | patients | time |
| people | centennial | spread | us | died | first | need | due | care | workers | positive |
| way | new | public | new | says | confirmed | crisis | please | time | support | help |
| even | taylor | case | world | flu | public | due | home | world | virus | china |
| paris | christmas | one | global | case | korea | like | support | home | crisis | would |
| dawson | day | like | public | president | disease | spread | spread | crisis | tested | health |
| great | beer | flu | says | said | time | virus | patients | death | may | crisis |

### 5.2. Word Vectors, Semantic shifts and Mapping

In order to find words with similar meanings and show the semantic change for word "corona", the word vector embeddings were trained for the corpus of tweets for each month and the five most similar words were included in the results. The similarity score is between 0 and 1.0, closer to 1 indicates that the words are closely related and situated closely in the vector space.

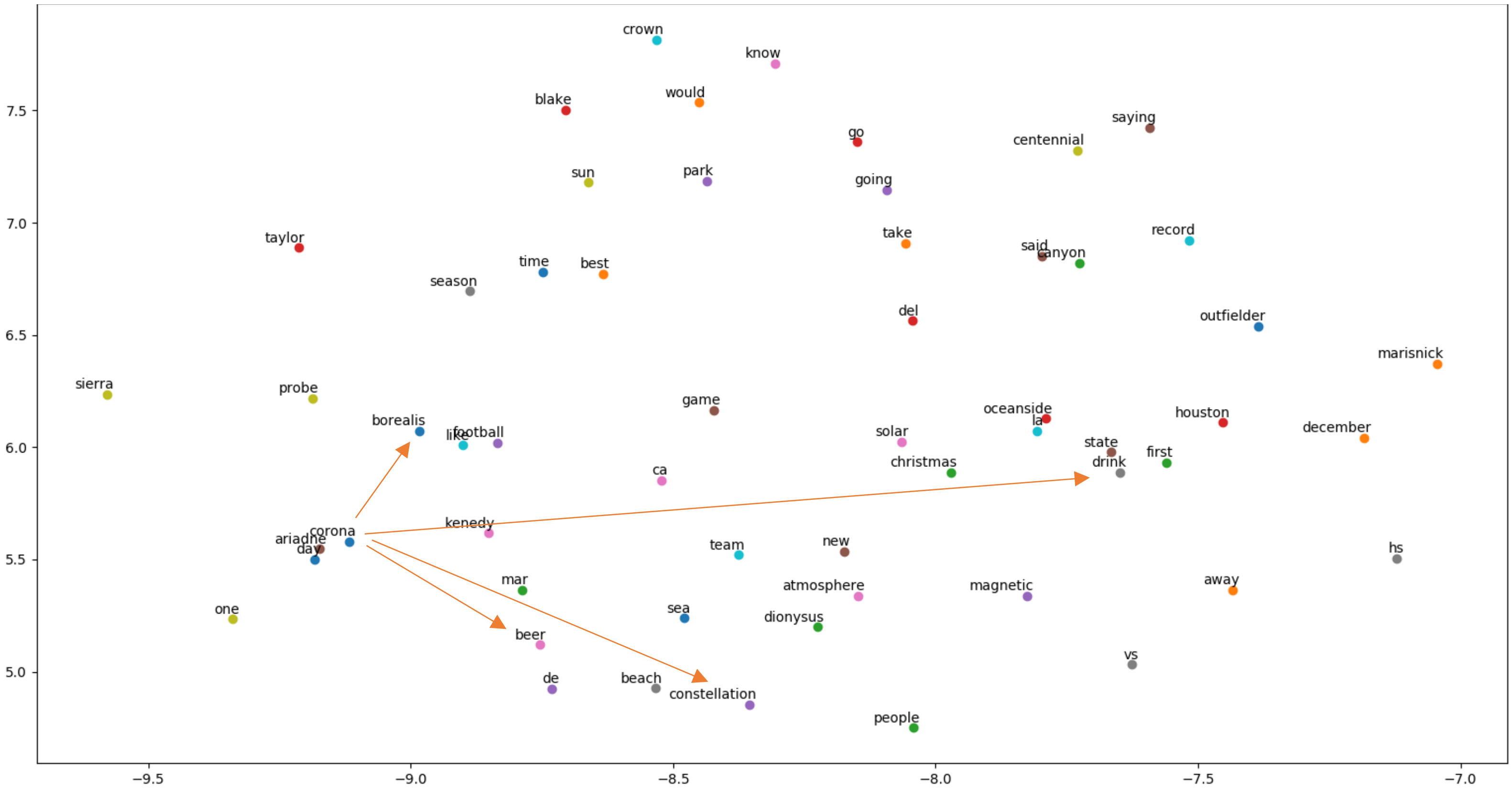
**Table 3.** Top five word-vectors closest to "corona" every month. Given in (word, similarity) pairs

| December 2019 | January 2020 | February 2020 | March 2020 | April 2020 | May 2020 |
|---|---|---|---|---|---|
| new, 0.69 | showed, 0.62 | fears, 0.59 | parties, 0.62 | induced, 0.69 | spend, 0.74 |
| title, 0.49 | fears, 0.59 | short, 0.58 | fears, 0.56 | false, 0.68 | false, 0.68 |
| houston, 0.45 | short, 0.58 | hoaxes, 0.56 | paycheck, 0.66 | air, 0.65 | desperate, 0.68 |
| nasa, 0.41 | crowded, 0.55 | escape, 0.56 | air, 0.65 | track, 0.63 | air, 0.65 |
| drink, 0.38 | real, 0.55 | scares, 0.54 | failed, 0.59 | incarcerated, 0.61 | violence, 0.63 |

The results here are quite striking since the closest related words to "corona" included "Houston", "NASA" and "drink". The relationship with NASA is consistent with the idea that "corona" was a word mostly belonging to the scientific community before January 2020.

The word vector mappings below are done by projecting multidimensional word vectors to a 2D space. The mapping in Figure 6 contains 27601 different words with 20 vectors each. Only a zoomed in versions of these maps were included because they are very clustered and hard to read. The distances between words in these mappings indicate a semantic similarity.

**Figure 5 Word Embedding Map of Corona Tweets from December 2019**




Some of the words that "corona" is related to are annotated with red arrows in Figure 5. There are many scientific words close to "corona" such as "constellation", "borealis", "atmosphere" and "solar", which supports my claim about the semantic meaning of corona being scientific before January 2020. It is also seen that corona is associated with "beer" and "drink" as well as positive words related to a relaxed lifestyle such as "beach", "best", "sun" and "sea".

The mappings in Figure 6 and 7 reflect the dramatic semantic change over time. The word "covid19" is related to "concerns" and "restrictions". It is also interesting that "doctors" and "pray" are clustered together. The words "chinese", "rome", "racism" and "plague" also appear in close proximity, which might indicate the reaction against the rise of racism against Asian people.

**Figure 7 Word Embedding Map of All Coronavirus Tweets**

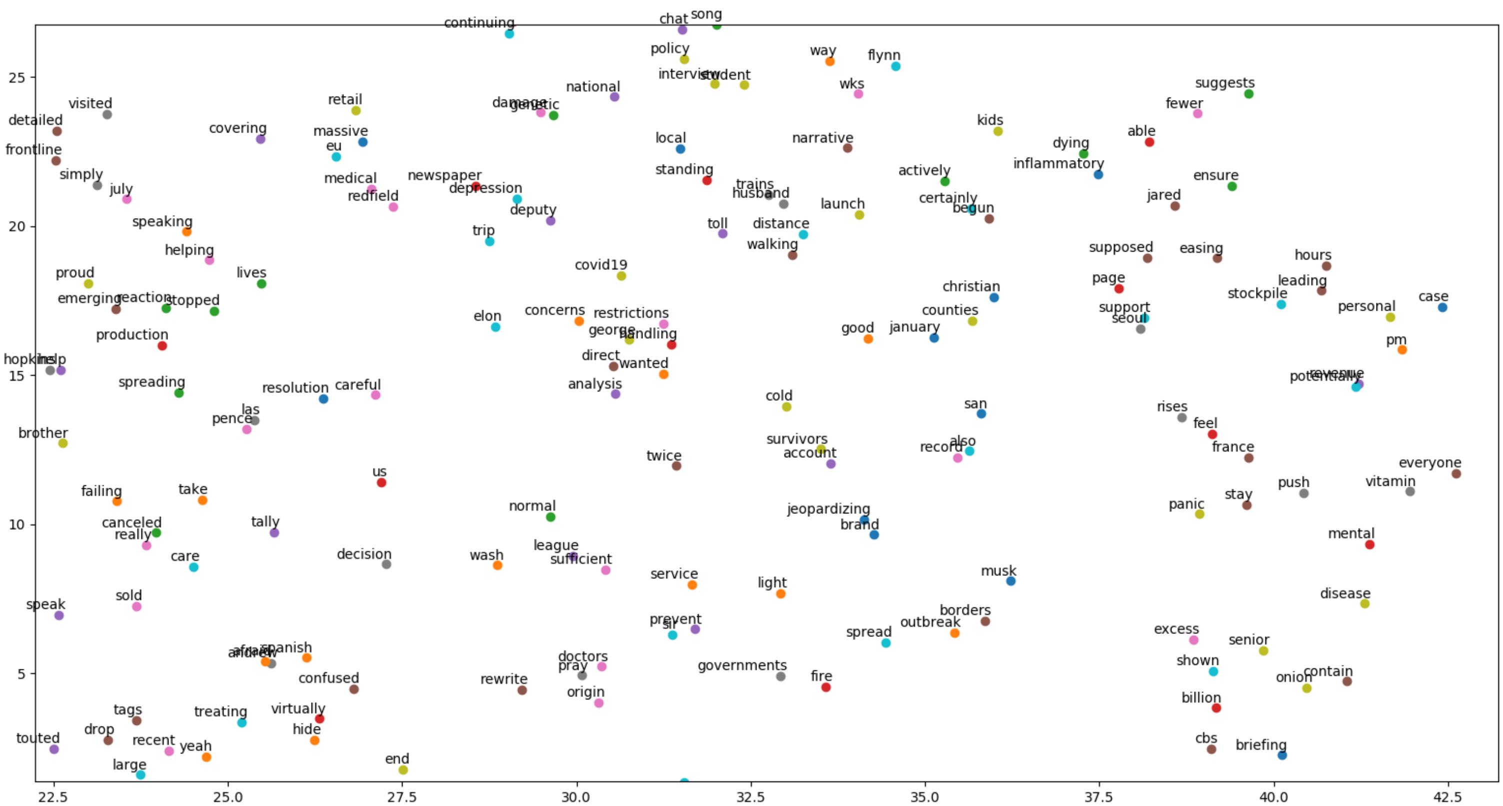
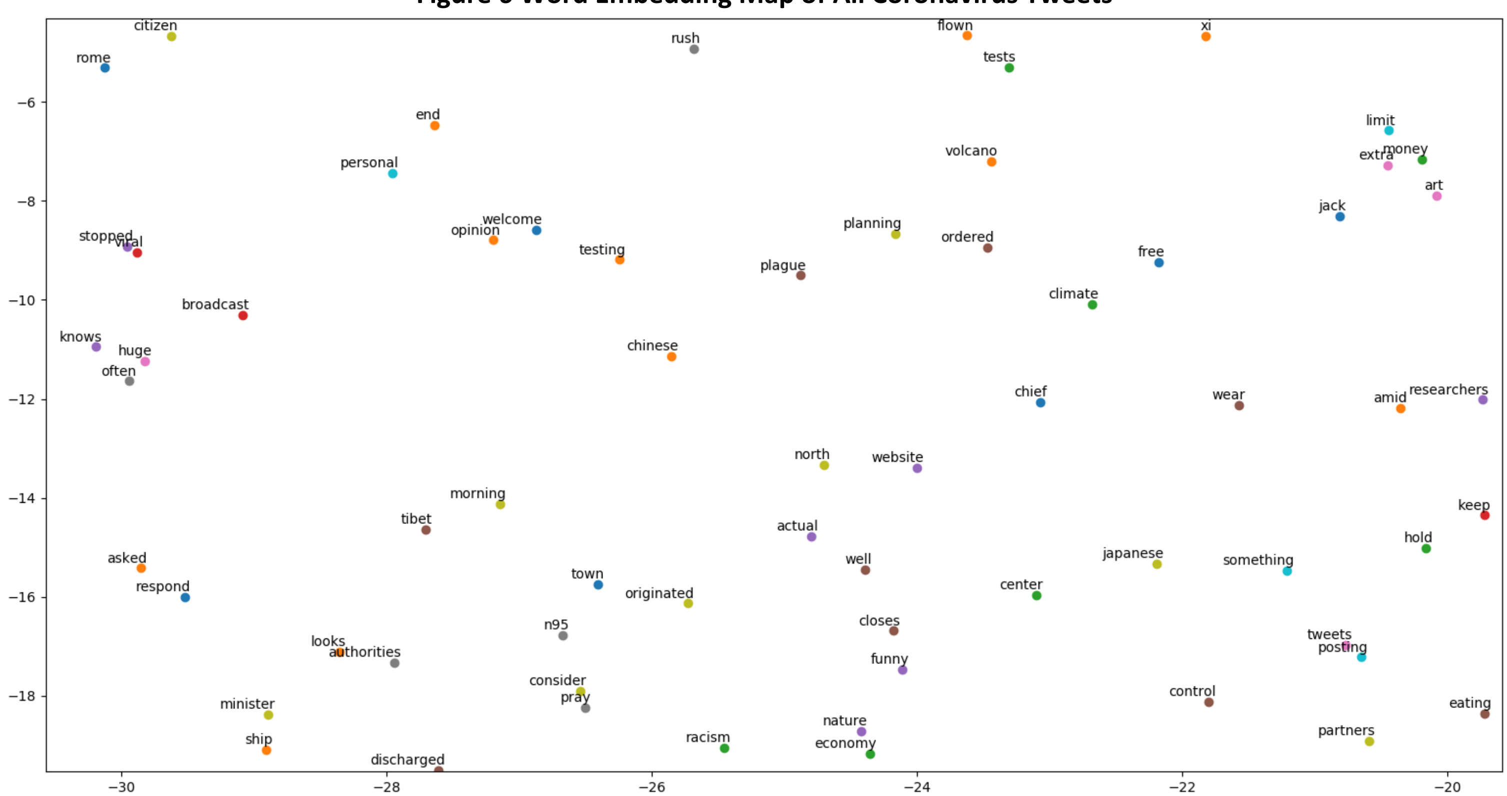
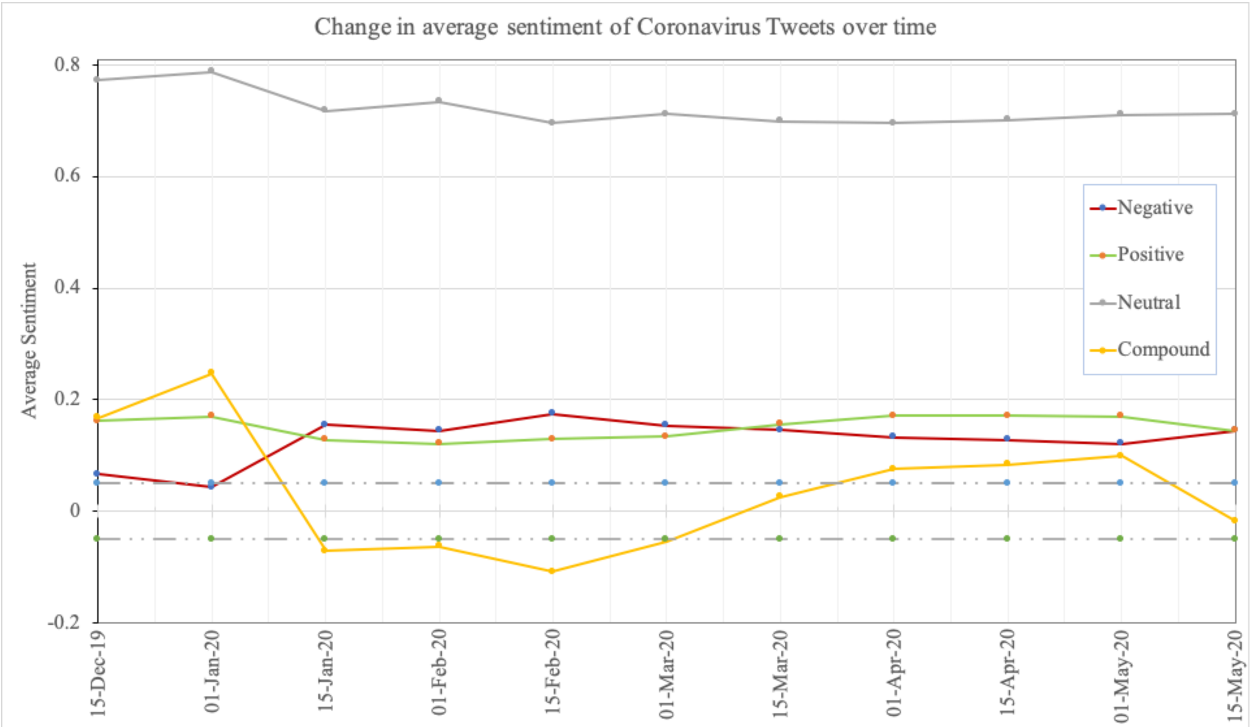



**Figure 6 Word Embedding Map of All Coronavirus Tweets**



### 5.3. Sentiment Analysis




The horizontal dashed lines at y=0.05 and y=-0.05 indicate the neutral zone for the compound sentiment. It is seen that the sentiment about Coronavirus related Tweets was positive until January 15, 2020, then we observe a further decrease in February 2020. However, the sentiment becomes rather neutral in March and briefly positive in April 2020. Then it becomes negative again in May. This might be related to the global outlook and spread of the pandemic. It is important to note that the neutral sentiment is also decreasing, which might indicate the news sources and sources that had neutral tone before are shifting towards a negative narrative about Coronavirus.

## 6. Conclusion and discussion

Throughout the analysis of Tweets about Coronavirus, I was already expecting to see an increase in words with negative connotations and changes in frequency, however the semantic shifts and vector space maps were the most interesting findings. The word "Corona" has gained a completely new meaning in Twitter discourse and perhaps changed ownership. I claim that the word "corona" used to belong to only the scientific community, however it now belongs to a broad range of speakers. The increase in its use frequency distribution table also indicates that its number of users have increased and now is owned by a general audience.

Using both TextBlob and NLTK for frequency distribution allowed me to find an interesting result. In fact, I came across a new slang term "rona", which is a nickname for Coronavirus. The use of "rona" in Coronavirus Tweets has increased suddenly between April and May 2020. This word was not included at all in NLTK frequency distribution, probably because it was not counted as a proper word. Yet, I personally witnessed slang term "rona" and even "miss rona" to be used online while referring to Coronavirus. Figure 3 demonstrates the increase in frequency of "rona" over March, April and May 2020. While it was not used at all in February 2020, "rona" was used over 2,000 times among 39,000 words in Tweets dating to March 2020. This increase cannot be coincidental. Its increasing use shows that Twitter users embrace this word in their Coronavirus discourse. Therefore, I claim that "rona" might be a neologism that is establishing in the Internet slang and even everyday speech.

### 6.1. Politics of Coronavirus

Word embeddings and frequencies tell us meaning that are likely to be ideological. I found some of the results to be political and they are worth discussing. The frequency table in Table 2 shows the constant association of Coronavirus with China even though the pandemic epicenter has shifted from China to Europe in March 2020. The word "trump" has also increased frequency in March and May 2020. This change might be a result of Donald Trump's blaming

China for the spread of Coronavirus and the controversy surrounding his attitude about the pandemic. Another possibility is that that there is increasing racism online against Asians, as the Network Contagion Research Institute reported increase in hate anti-Asian attacks on social media[8]. This indicates that some of these Tweets could be racist attacks as well as defense against them. In addition, the frequency analysis in Figure 2 shows that the only countries associated with Coronavirus words is China and the UK, which is an interesting outcome considering that the pandemic affected many other countries and mainland Europe was considered the epicenter of the pandemic during April 2020.

Another point of concern is that my dataset only consists of English Tweets, which makes it much closer to the Western political ideas and values. I would imagine that Twitter data would be much different if it were only in Chinese, Spanish or Russian. Furthermore, as I have discussed in the data collection section, Twitter data is very likely to be censored by governments and Twitter itself. Although I made efforts to make my dataset close to an objective human interaction, the Tweets were chosen by a biased algorithm.

### 6.2. Further improvements

With a great amount of data and statistics, I found representing the Twitter data and visualizing my results challenging. Data visualization is useful for demonstrating concepts and trends that we cannot otherwise comprehend. Twitter data needs to be analyzed both individually and as a part of the whole dataset. I mapped the 20-dimensional word vectors on a 2-D space, however there can be better ways of representing multidimensional space without losing information.

[8] Asmelash, Leah. CNN, "With the spread of coronavirus came a surge in anti-Asian racism online, new research says" https://edition.cnn.com/2020/04/10/us/sinophobic-racism-rise-coronavirus-research-trnd/index.html

Future researchers could find a more optimal way that will demonstrate all of the data gathered. Since this research was done using a small dataset, this analysis can be performed again on a dataset of millions of Tweets. Semantic change over time is worth exploring and Machine Learning algorithms could study how the meanings of words change.

**References**

Khatua, Aparup, Apalak Khatua, and Erik Cambria. “A Tale of Two Epidemics: Contextual Word2Vec for Classifying Twitter Streams during Outbreaks.” *Information Processing & Management* 56, no. 1 (2019): 247–57. https://doi.org/10.1016/j.ipm.2018.10.010.

Internet Live Stats. *Twitter Statistics*. https://www.internetlivestats.com/twitter-statistics/

Timberg, Craig. *Washington Post,* “On Twitter, almost 60 percent of false claims about coronavirus remain online”https://www.washingtonpost.com/technology/2020/04/07/twitter-almost-60-percent-false-claims-about-coronavirus-remain-online-without-warning-label

Hutto C, Gilbert E. VADER*: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*. In: International AAAI Conference on Weblogs and Social Media. AAAI; 2014.

*Merriam-Webster.com Dictionary*, s.v. “corona”, https://www.merriam-webster.com/dictionary/corona.

NASA. *“First NASA Parker Solar Probe Results Reveal Surprising Details About Our Sun”* accessed May 12, 2020, https://www.nasa.gov/press-release/first-nasa-parker-solar-probe-results-reveal-surprising-details-about-our-sun

Ducharme, Jamie *New York Times,* “World Health Organization Declares COVID-19 a 'Pandemic.' Here's What That Means*”.* https://time.com/5791661/who-coronavirus-pandemic-declaration/