UNIVERSITY OF MARYLAND
18 56

ROBERT H. SMITH
SCHOOL OF BUSINESS

SMITH ANALYTICS CONSORTIUM
Data Series Workshop

Twitter Sentiment Analysis on COVID-19 Pandemic

Group 6

Weibo Chen, Xiaohui (Sophie) Li, Yifan He and Madhavi Mundada

Date: 08/27/20

Table of Contents

Abstract 2

Introduction 2

Dataset and Exploratory Analysis 3

Methodologies 4

Results and Discussions 7

Conclusions 12

Appendices 14

References 19

Abstract

In the face of the unprecedented Coronavirus disease 2019 (COVID-19) health crisis, nationwide lockdowns and uncertain times, people experienced stress, worry, fear, disgust and sadness as the pandemic changed lives in a dramatic way. Social media has become a significant interface for the common public to share information and articulate feelings. As the increased sophistication of the natural language text interpretation and processing technologies has created possibilities of analyzing public sentiments toward the pandemic, we used natural language processing (NLP) to analyse public response and sentiments in different four states (California, Texas, Kansas and New York) of the US on Twitter from late-March to mid-April. We applied word cloud and N-grams to understand critical trends of public reaction and analyzed sentiment people expressed in tweets. We found people were interested in the government policies regarding social distancing, testing of the COVID-19, as well as the new cases and deaths of the diseases in the period studied. And we learned people's sentiments are more negative rather than positive toward the pandemic, and there is no significant discrepancy between states in the sentiment distributions.

1 Introduction

COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, Hubei, China, and was classified as a pandemic by the World Health Organization (WHO) in March 2020. As of 13 August 2020, more than 20.6 million cases have been reported across 188 countries and territories, resulting in more than 749,000 deaths. The health impact of the outbreak has been so huge that it has been compared to dreaded epidemics and pandemics of the past like 'The Great Influenza' (Spanish flu of 1918), or the Black Death (form of bubonic plague) [1]. Moreover, the swift and massive shock of the pandemic and shutdown measures to contain it have plunged the global economy into a severe contraction. According to World Bank forecasts, the global economy will shrink by 5.2% this year. That would represent the deepest recession since the Second World War. [2]

The rise in emphasis on artificial intelligence (AI) methods for textual analytics and natural language processing (NLP) followed the tremendous increase in public reliance on social media. Researchers and practitioners mine massive textual and unstructured datasets to generate insights about mass behavior, thoughts and emotions on a wide variety of issues such as product reviews, political opinions and trends, motivational principles and stock market sentiment. [3] In the face of the unprecedented health crisis, uncomfortable nationwide lockdowns and uncertain economic conditions, people experienced more stress, worry, fear, disgust and sadness as COVID-19 changed lives in a dramatic way. As a result, social media has become a significant interface for the common public to share information and articulate feelings of the unpredicted situations. As the public rely on social media to exchange facts and opinions, the increased sophistication of the natural language text interpretation and processing technologies has created possibilities of analyzing public sentiments toward the pandemic and related social distancing measures [4,5].

In this work, we analyzed public response and sentiments expressed in different states of the US over one of the most popular social media interfaces, Twitter, from late-March to mid-April when majority of states had declared emergency in different types.[6] In this study, due to immense database size and limited computing power, we focused our analysis mainly on four States across the US. These four states are California, Texas, New York and Kansas. The analysis of this report mainly focuses on four areas:

- First, textual data visualization is used to identify critical trends of public reaction and sentiment for COVID-19 outbreak. We created word clouds for different states to search for insights.
- Second, after visualization of the most prominent aspects in common from the tweets, we identified not only most frequently used words but also word pairs using N-gram identifications in a text corpus.
- Third, sentiment analysis is used to understand people's sentiment in different states and comparison analysis is done to study people's sentiment differences related to the number of COVID-19 cases in different states.
- Last but not least, machine learning K-Nearest Neighbor is used to find clusters in tweets in different states.

2 Dataset and Exploratory Analysis

2.1 The COVID-19 dataset

The Natural Language Processing allows a machine to process a natural human language and translates it to a format that the machine understands. In our analysis, we leveraged NLP to gain information from unstructured text on Twitter through various analytic tools. The dataset was extracted in the period of COVID-19 between late-March and mid-April 2020. For each tweet related to the COVID-19, its user ID, location, and full tweet was extracted.

2.2 Exploratory Analysis

The data contains 20 million rows. The location is restricted to the United States. We have the user twitter ID instead of the tweet ID, hence extracting the timestamp or other more relevant information related to the tweet is not possible.

With the current information, we decided to conduct analysis based on the state distribution. Hence the tweets were grouped by state. Only the top ten states are displayed in Figure 1 below for better view of the graph. These include - South Carolina, New York, Texas, etc. This exploratory data analysis later helped us to target states for further analysis.

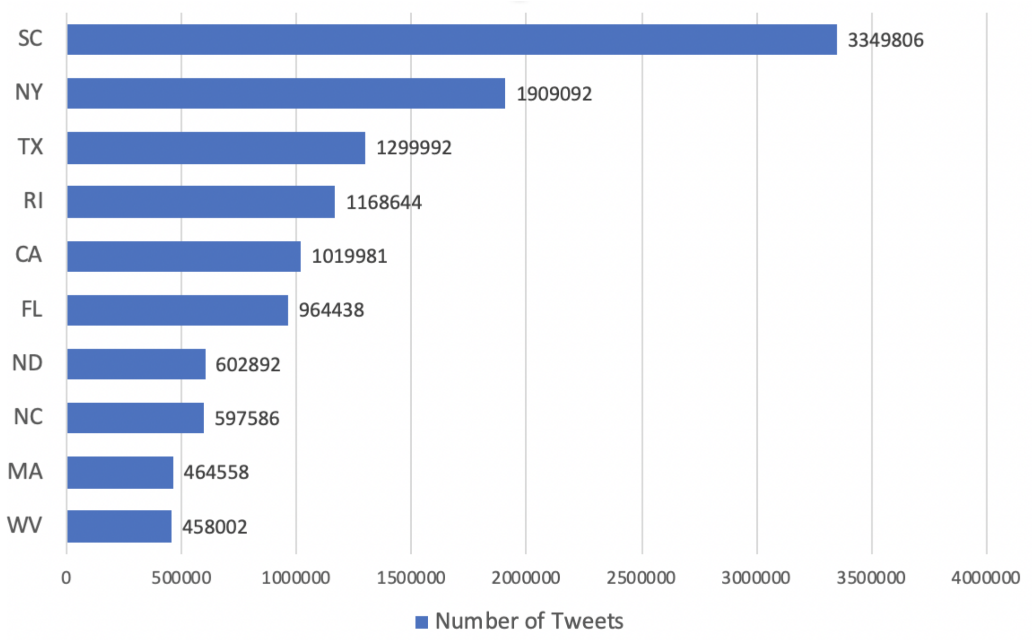


Figure 1 - Top Ten States with the Highest Number of Tweets

3 Methodologies

3.1 Preprocessing data

After loading the dataset, first we preprocessed it. In the cleaning phase, for the ‘location’ column, we uniformatted the name of states. For the text, we removed the specified punctuations(#$&@ etc.) but not ‘!’ to better understand their sentiment. Then we removed the URL and ‘RT’, and converted the emoji to unicode. We removed stop words from text and print POS tagging and placed them into an iterable SpaCy pipeline. We used 'lemma_' since it standardized accounting for grammar and parts-of-speech.

3.2 Choosing the target states

After preprocessing the dataset, we divided the dataset according to states to conduct sampling due to the large dataset volume. Among all the states in the United States, we chose the representative ones based on their cases confirmed, the number of tweets extracted, and the geographical location. As shown in Table 1, we chose New York, California, Texas and Kansas for the following reasons: first of all, we hope to gain a broad understanding of the public reaction to COVID-19 nationwide by picking states with different geographical locations in the US; secondly, we want to gain insight into the difference in public opinion where confirmed COVID-19 cases are at different levels (i.e. New York had more than 13,887 accumulated cases by April 15, whereas Kansas only had 1,504 cases. [7]); lastly, we also considered the tweet count of each state to make sure the significance of analysis. (As shown in Figure 1, California, Texas and New York are among the most tweet counts states, Kansas has 146,769 tweets for analysis).

Table 1 - The Basic Information of Four States

States	Confirmed Cases	Tweets Extracted	Geographical Location
CA	26,686	1,019,981	West Coast
NY	214,454	1,909,092	East Coast
TX	15,907	1,299,992	South
KS	1,504	146,769	Midwestern

3.3 Vectorization

After lemmatization, we built a list of unique words used by the users in the tweets and counted the occurrences of each word in the corpus. This is also known as Bag-of-Words. This resulted in knowing the most frequently used words like - Immunity, mental, guidelines, target, worldwide, etc. Depending upon the context of analysis, n-grams is chosen for further analysis. We chose uni-grams, bi-grams and tri-grams for the analysis to understand which words are used the most separately and in combination. Figure 2 shows ten most common used words in tweets originated from California.

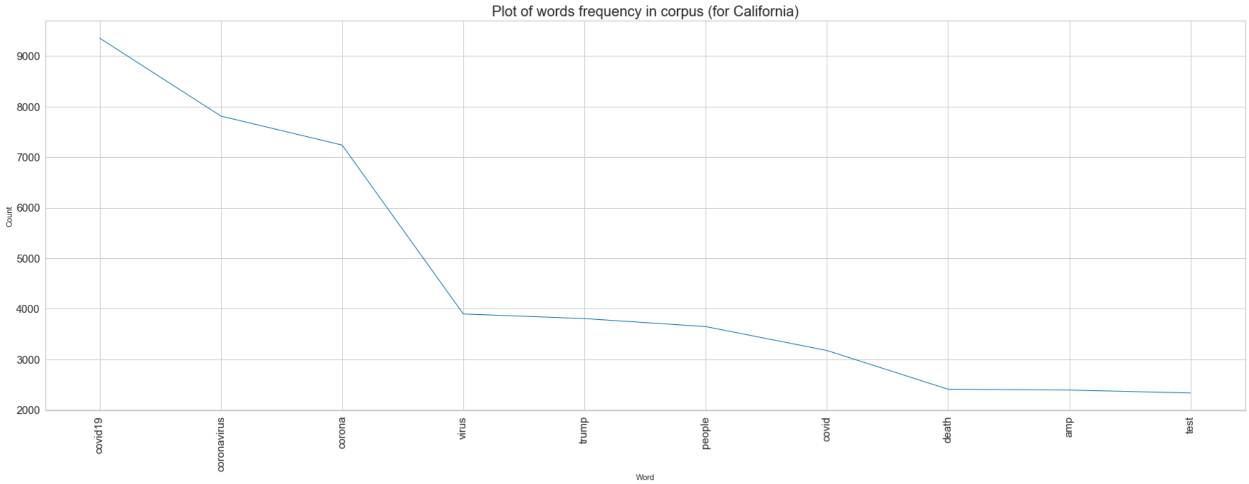


Figure 2 - Ten Most Common Words Used in Tweets Originated from California

Later we created a term document matrix, which determines occurrences of each word in the bag of words in terms of each document (which is a tweet in our case). Using the Term Document Matrix (TDM), the term frequency–inverse document frequency (TFIDF) for each word in the corpus is calculated, which gives the importance of every word with respect to the other words in the corpus as well as the document. The TDM and TFIDF will be different for each state, as the corpus changes according to the state.

3.4 Sentiment analysis

In the step of conducting sentiment analysis, which allows us to glean how the subject at hand feels. We used Valence Aware Dictionary and sentiment Reasoner (VADER)[8] , in which the compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). We specifically pulled out the 'compound score' and set the sentiment as positive one if the score is greater than 0.05, as neutral one if it is between 0.05 and -0.05, as negative one if else. Accordingly, we got the percentage of each type of attitude in each state to make a comparison.

3.5 K-Means analysis

The K-means analysis was used to group similar tweets together depending on the underlying pattern, in this case the context of the tweet or what the tweet refers to.

Post the TFIDF transformation, the document vectors are put through a K-Means clustering algorithm which computes the Euclidean distances amongst these documents and clusters nearby documents together. Using the elbow method, appropriate numbers of clusters are decided for a given corpus. As shown in Figure 3, the best number of clusters are determined to be seven (elbowatk=7), with the lowest distortion score. This data is further used to split the tweets into seven clusters, each focusing on a particular topic. After dividing the tweets of each state into clusters, the Singular Value Decomposition (SVD) is applied to reduce the multiple dimensions into two and to better visualize the result.

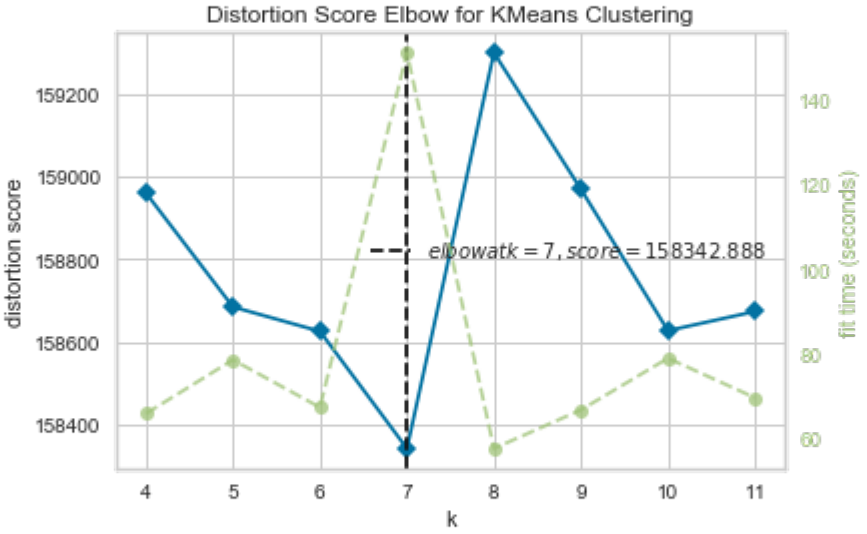
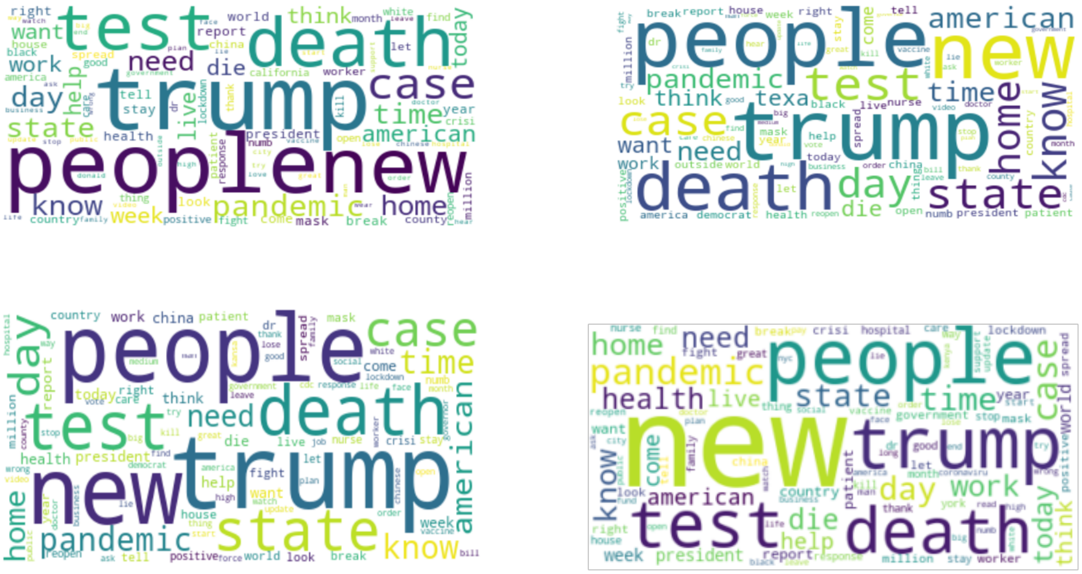


Figure 3 - Deciding Optimal Number of Clusters Using Elbow-method

4. Results and Discussions

4.1 Word Cloud Representation

(a)California **(b) Texas**



(c) Kansas **(d) New York**

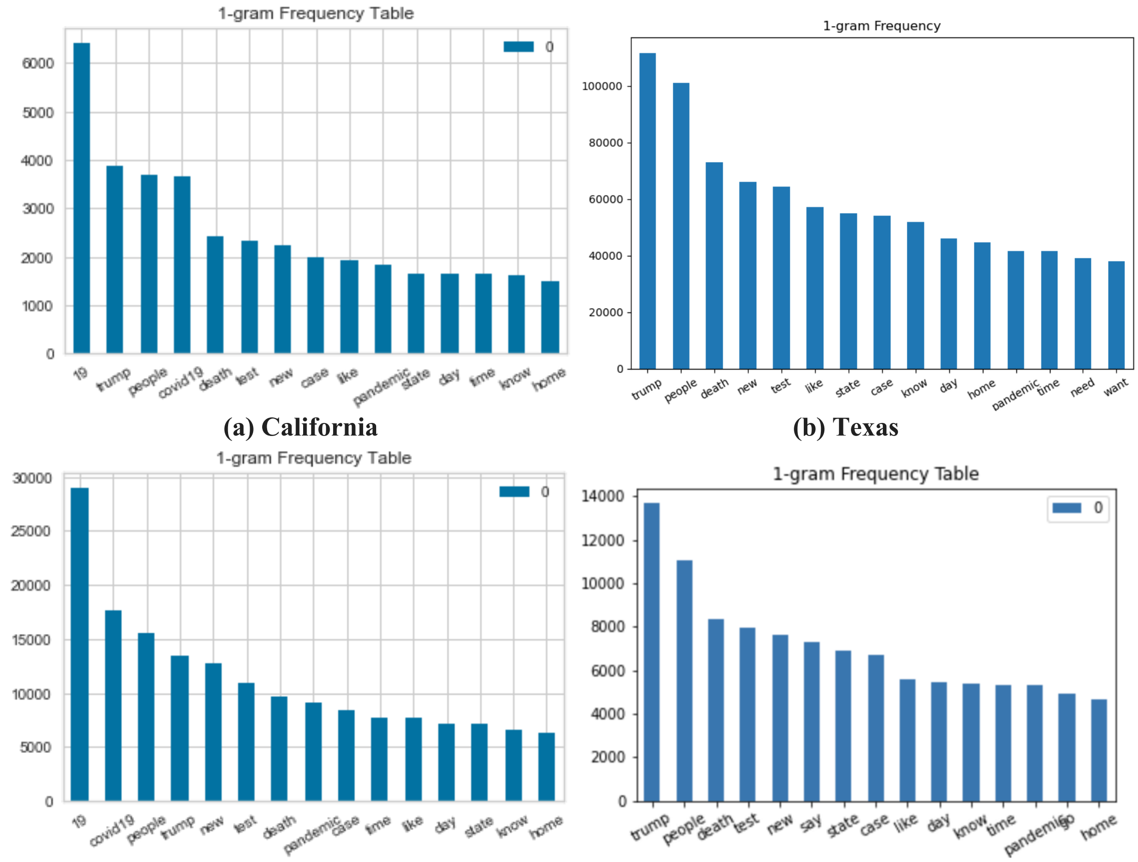
Figure 4 - Word Cloud Representations

In order to generate more informative word clouds, additional words were added to the stop words list such as ‘covid’, ‘corona’, ‘virus’ since these tweets are centered with this topic.

As shown in Figure 4, within all four states we focused, we can infer that “President Trump” is the most frequent topic that Twitter users tend to talk about. Other high frequency words are ‘test’, ‘death’ and ‘new(cases)’, indicating that users also took great care of how the pandemic was going on around the US.

4.2 N-gram Representation

Further to the generation of word clouds, we also utilized n-gram to dig deeper on short phrases that people were concerned about during the hard time. For each state, we used n-gram (where n ∈ [1, 2, 3]) to extract contiguous words, sort them into descending order and plot the bar charts. Though the entire analysis is based on Uni-grams, using bi-grams and tri-grams gave further information on how the combination of words impacted. Here is the result of barchart about the frequency in each state when n=1. In the appendix, the graphs for bi-grams and tri-grams are displayed. Which suggested words like stay home, social distance, donald trump mexican etc. often appeared together.



(a) California **(b) Texas**

(c) New York **(d) Kansas**

Figure 5 - 1-gram Representation for Tweets from Four States

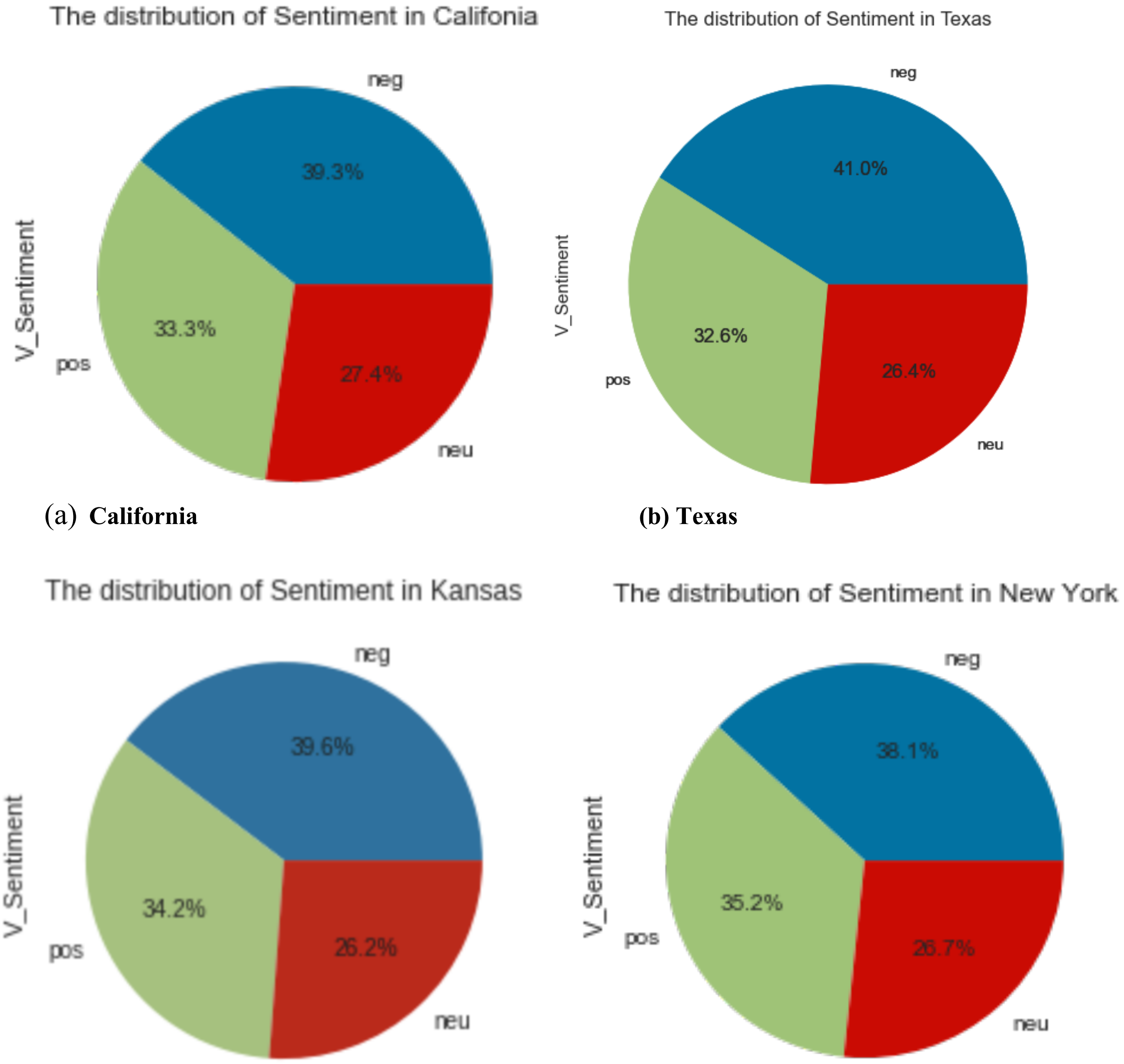
Figure 5 shows that among all four states, the hottest single words are similar , such as ‘people’, ‘trump’, ’death’, and ‘test’, which indicates peoples’ consideration and raised awareness about the latest news of COVID-19.

In 2-gram graphs in California (Figure A1) , the ‘ stay home’ ranks as the second frequent word; also, other safety measures like ‘wear mask’ and ‘social distance’ that may slow down the spread of viruses could also be seen. This implies the posters’ efforts on raising people's awareness about prevention. The words ‘test positive’, ‘nurse home’ shows the concern about measures to newly admitted and readmitted residents with COVID-19. Similarly, those words related to policy and measures were also mentioned by people in New York, Texas and Kansas.

In 3-gram graphs (Figure A1), the key word ‘stay home order ’ shows in all four states. In Texas, both the ‘ban outside hang’ and ‘ hang outside come’ appear, indicating the controversy of the policy. In New York, ‘test positive 19’ and ‘new (covid) 19 cases’ mainly show the concern of the public about the latest trend of COVID-19, most other words much more show people’s feelings toward the virus in March and April (like ‘sucks’) . Kansas is the only state where the protection of wearing masks (‘mask help protect’) shows up in 3-gram, consistent with the time when the local department of health released their suggestion on mask wearing [9].

4.3 Sentiment Analysis

After using VADER in Python for sentiment analysis as described in section 3.4, we plot a pie chart (Figure 6) to demonstrate the sentiment distributions for each of the four states. From the charts, there is no significant difference among states in each category of sentiment (negative, neutral and positive). And all four states showed the highest percentage of negative sentiment, followed by positive sentiment and lowest percentage of neutral sentiment. As expected, people tend to express more negative attitudes toward the pandemic, since the spread of diseases causes casualties and changes people's living styles. From late-March to mid-April was the national pandemic countermeasure beginning to act, it is reasonable that more than half of the people positively, or neutrally believed the U.S. would recover from the pandemic soon. The vision of the current day probably tells a different story as the government's efforts on controlling the pandemic and their outcomes were under people's expectations.



(a) **California**

(b) Texas

(c) Kansas

(d) New York

Figure 6 - Sentiment of Tweets from four States

In the next step, we compared the sentiment across all four states. More than that, some economic indicators and COVID-19 related data (Table 2) were added to facilitate our understanding of the sentiments people expressed in their tweets. From Figure 7, we did not find any positive relationship between people's sentiments and the economic conditions (Referring to Gross Domestic Product and Unemployment Rate). And surprisingly, we found states with higher hospitalization rates share more positive tweets. This may support our preliminary surmise that at the beginning of the pandemic, the negative sentiments on twitter are more focused on social distancing policies rather than the death toll of the pandemic. As people started acknowledging their states having increasing hospitalization rates, they began to recognize the threat of the disease and became accepting the isolation policies.

Table 2: COVID-19 Related Data

States	**Positive**	**Neutual**	**Negative**	**Cases** [7] **(4-15-2020)**	**GDP Q1 Change** [10]	**Unemployment filed (4-18-2020)**[11]	**Average Sentiment**	**Mortality Rate** [7]	**Hospitalization Rate** [7]
TX	32.6%	26.4%	41.0%	15907	-2.5%	7.41%	-8.4%	2.36%	9.67%
CA	33.3%	27.4%	39.3%	26686	-4.7%	13.98%	-6.0%	3.22%	19.35%
KS	34.2%	26.2%	39.6%	1504	-3.1%	7.14%	-5.4%	4.72%	22.74%
NY	35.2%	26.7%	38.1%	214454	-8.2%	14.40%	-2.9%	5.42%	21.54%

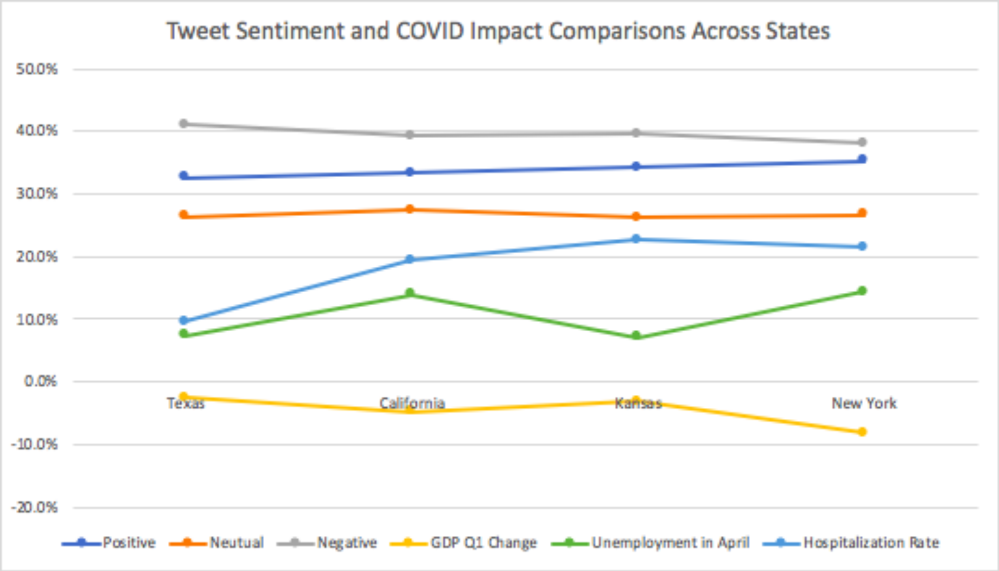


Figure 7 - Tweet Sentiment and COVID-19 impact Comparison Across States

After obtaining the percentage of each sentiment category, numerical values were assigned to each of them: 1 for positive, 0 for neutral, and -1 for negative. The average sentiment scores across all four states were then calculated. As shown in Figure 8, this plot captured the average sentiment against mortality rate of COVID-19 reported by each state by April 15, 2020 (Table 2). A generally positive correlation between the two variables can be inferred. As mentioned above, we found that the higher the mortality rate of a state, the higher average sentiment score in that state. We also took the confirmed case number on April 15 (Table 2) for determining the size of the dots, and we found even though New York had much more cases than other states, people there showed least negative sentiments in their tweets. Whereas Texas had less than 10% of New York's cases, people's sentiments were the most negative. Again, we believe the time series analysis is very important to consider, as the COVID-19 situation changes rapidly, people's sentiments would change dramatically from mainly frustration of isolation and lack of testing to fear of the long-term effects of the health and economic crisis.

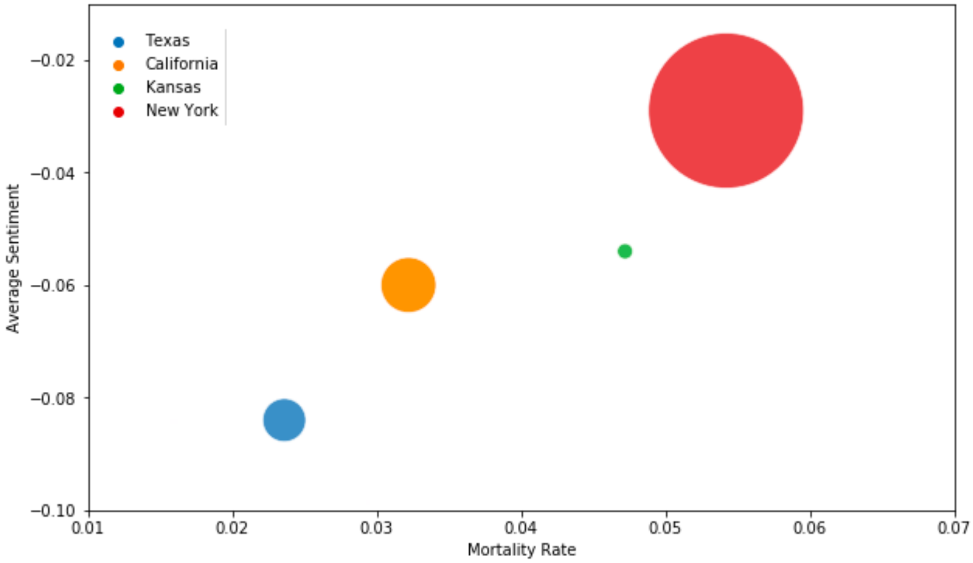


Figure 8 - Average Sentiment Score and COVID-19 Mortality Rate Relations across states

4.4 K-Means Cluster Analysis

From the result obtained from the Elbow method, the K value is used to cluster tweets into different clusters. The Figure 9 illustrates the distribution of clusters for New York, clusters for California and Kansas can be found in Figure A2. Each point belongs to a tweet in the cluster. From the figure we understand that tweets belonging to the same cluster are plotted closely which would be an expected output.

The tweets were divided into seven clusters. After observing the tweets in depth, the topic of the cluster could be easily determined. Like one cluster focused on flights whereas on the other had tweets related to President Donald Trump's actions towards COVID-19 formed another cluster.

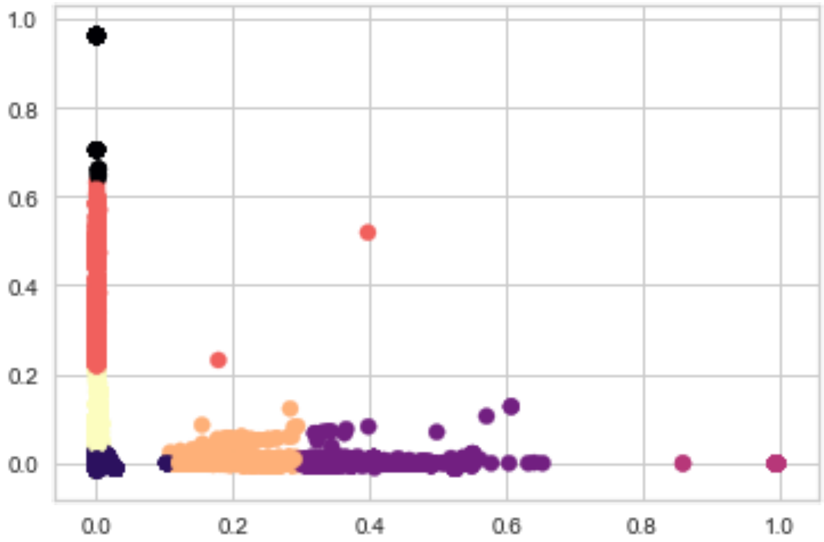


Figure 9 - Representation of Tweets into Different Clusters (New York)

4.5 Limitation and Future Study

In this research, aside from our findings, there remain challenges and holes to the credibility of our study, and there are some suggestions to further researchers in this area.

Limited Data: In the research, the time period of our dataset is only between late-March to early-Apirl and with no timestamp on each tweet, which constrains the exploration about the change along time in each state.

Representativeness of States Chosen: Considering the limited time, we chose four representative states based on the geographical location, confirmed cases and number of tweets, while there may be other states that have populations expressing different attitudes and sentiments toward COVID-19.

Noisy Data: The dataset extracted from Twitter is noisy and unstructured. The text inevitably has misspellings, slang, unknown words, unknown-abbreviations as well as other languages but written using the English alphabet, which all can negatively impact the analysis. Advanced cleaning tools may be required.

Combine Other Machine Learning Algorithm on Clustering: The clustering metric is adopted to divide the tweets into several clusters, while some further step about analyzing the different clusters could be applied. Our suggestion is to use 'scattertext' to compare each cluster virtually,or use the Latent Dirichlet Allocation (LDA) to build the topic model.

5 Conclusions

In this work we used NLP to process the tweets from four states: California, Texas, Kansas, and New York in the time period of late-March to mid-April. We wanted to gain some sight into people's opinion and attitudes towards the unprecedented pandemic and its impact to lives. We found the most popular topics are government policies for social distancing like stay at home orders, face masks and state lockdowns, as well as the threat of the pandemic to people's lives like daily added new cases and death. We also did sentiment analysis on the public's attitudes toward the pandemic. And we found that even though the negative sentiment has the highest percentage across all four states, the gap between the negative and positive sentiment is not large. This indicates that in the beginning of the pandemic, a lot of people still hold high hope for a rapid nationwide recovery from the damage caused by COVID-19. Surprisingly, we also found New York had the highest average sentiment among the four states we analyzed despite it having the highest confirmed cases and deaths in the period. We think this might be due to the negative sentiment expressed mostly because of the frustration caused by social distancing. And when people realized the real threat of the disease, they became more corporate to the isolation measures.

Appendix

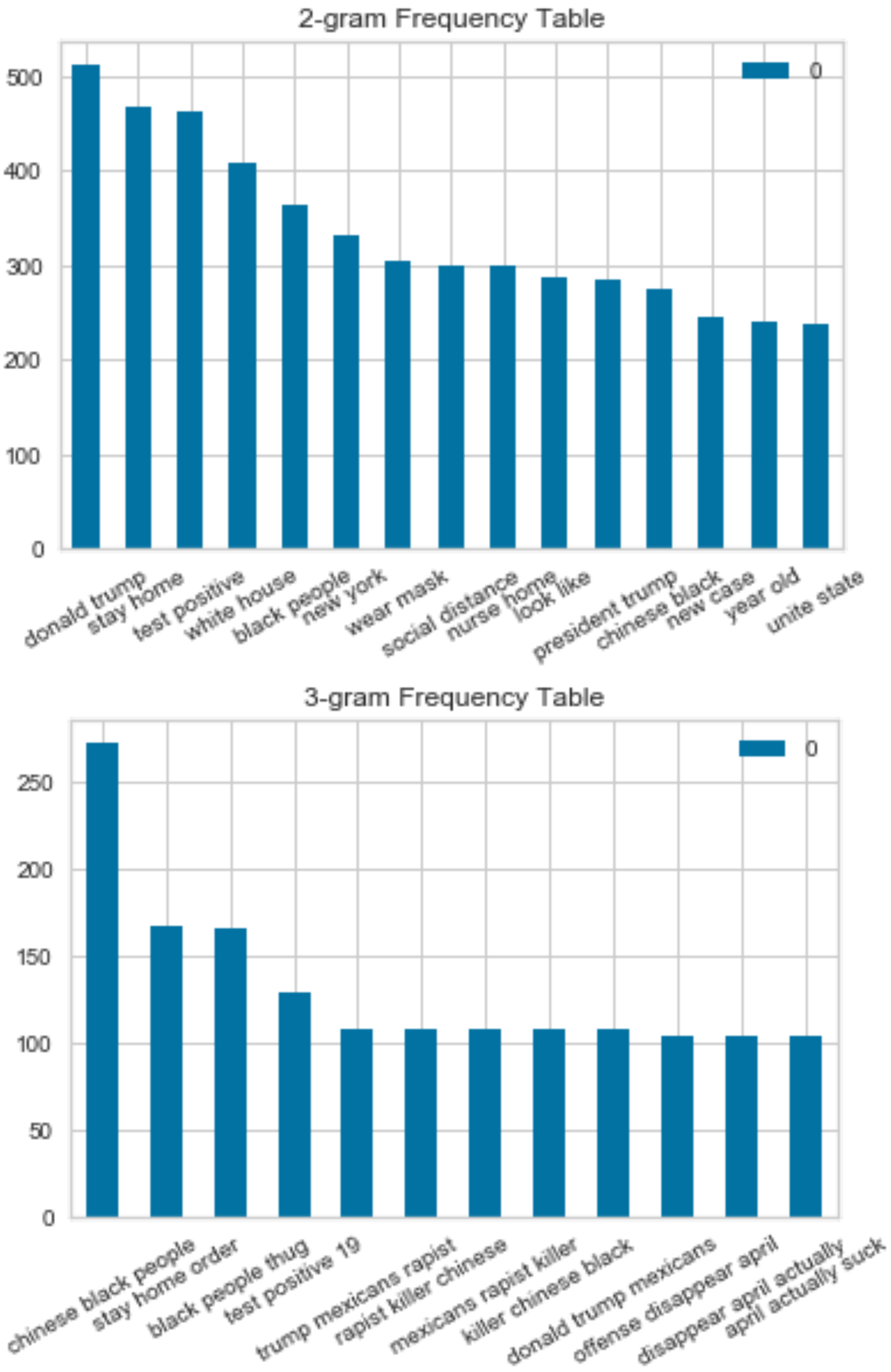


Figure A1-a N-gram Representation for California's Tweets

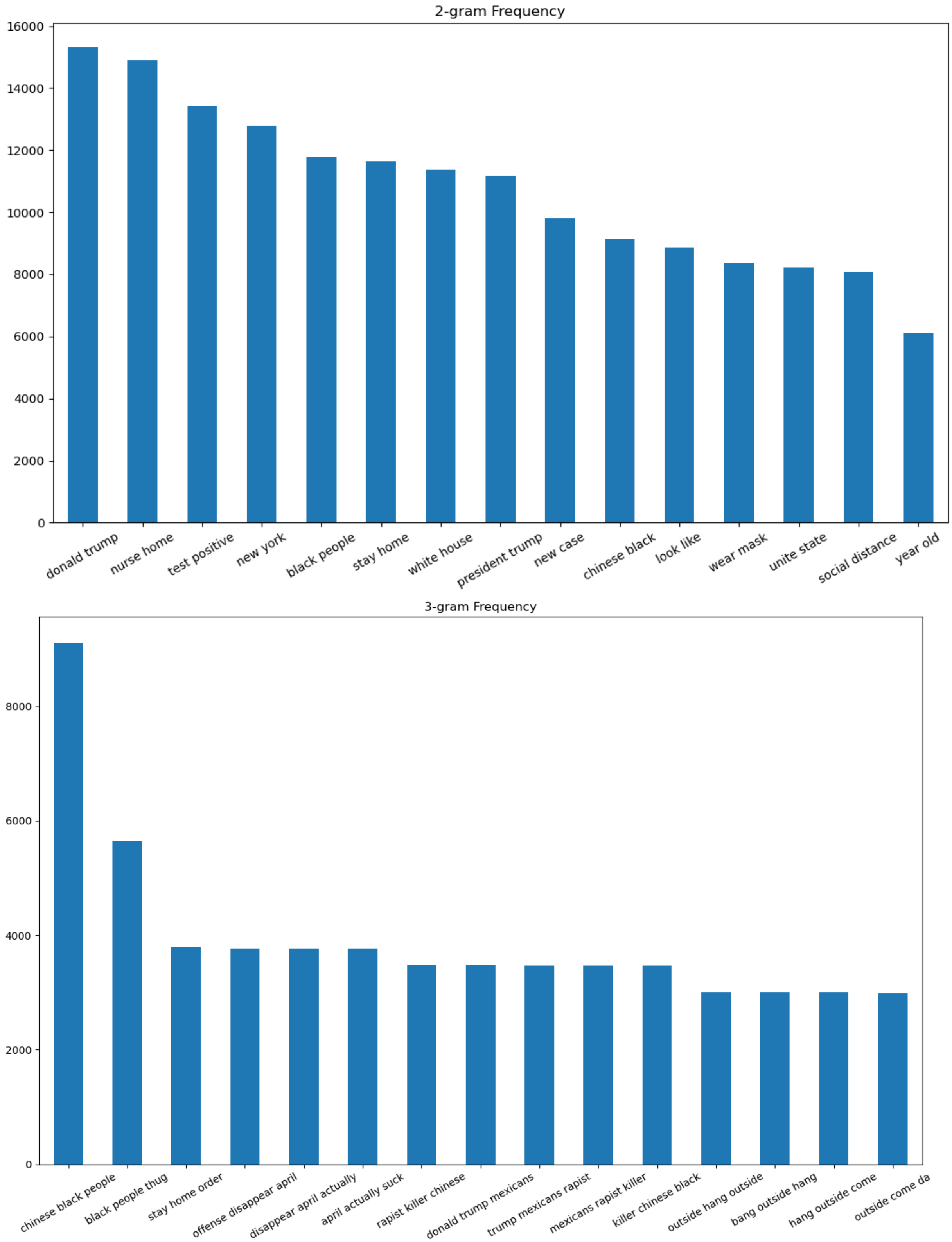


Figure A2-b N-gram Representation for Texas' Tweets

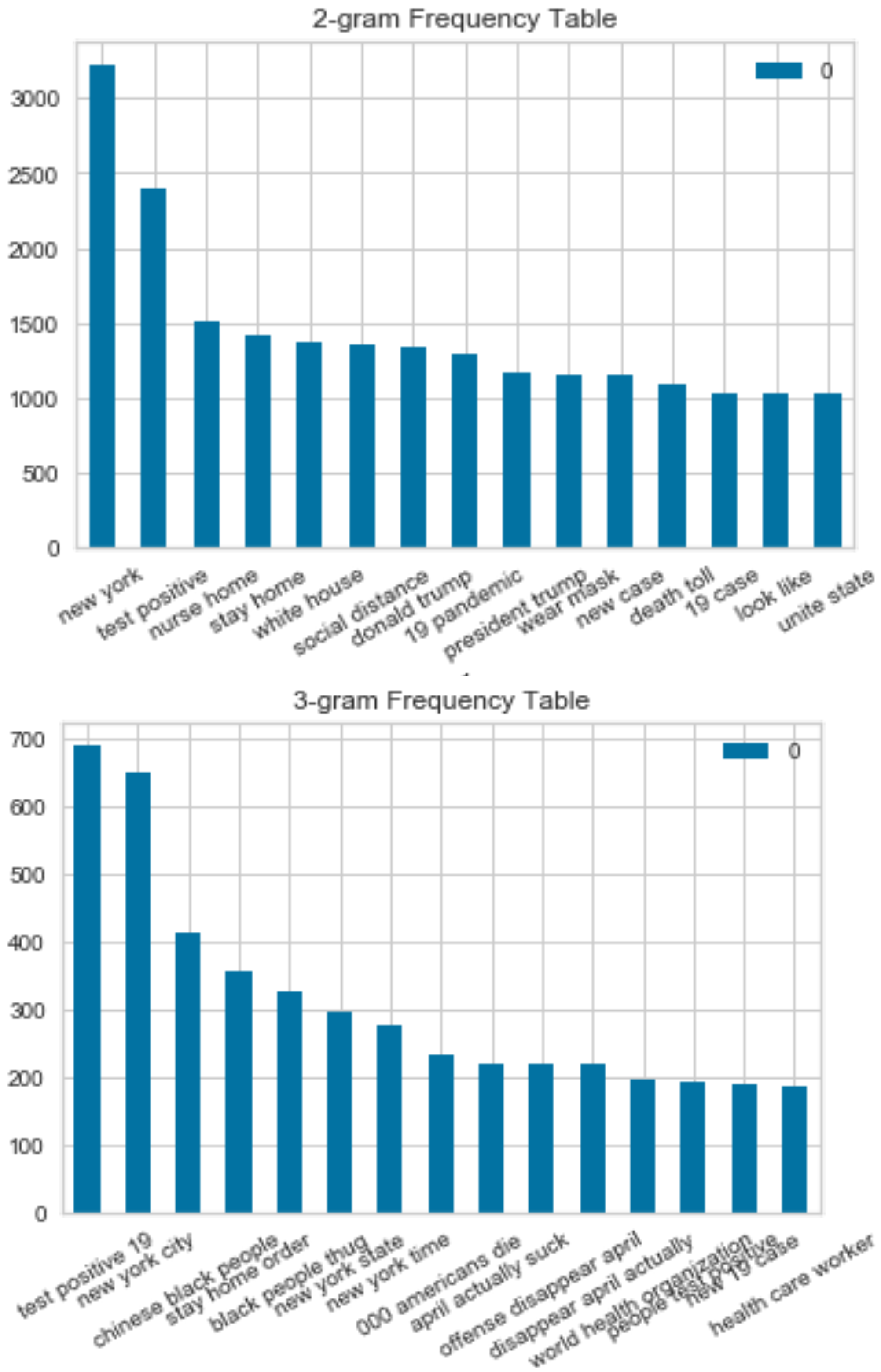


Figure A1-c N-gram Representation for New Yorks' Tweets

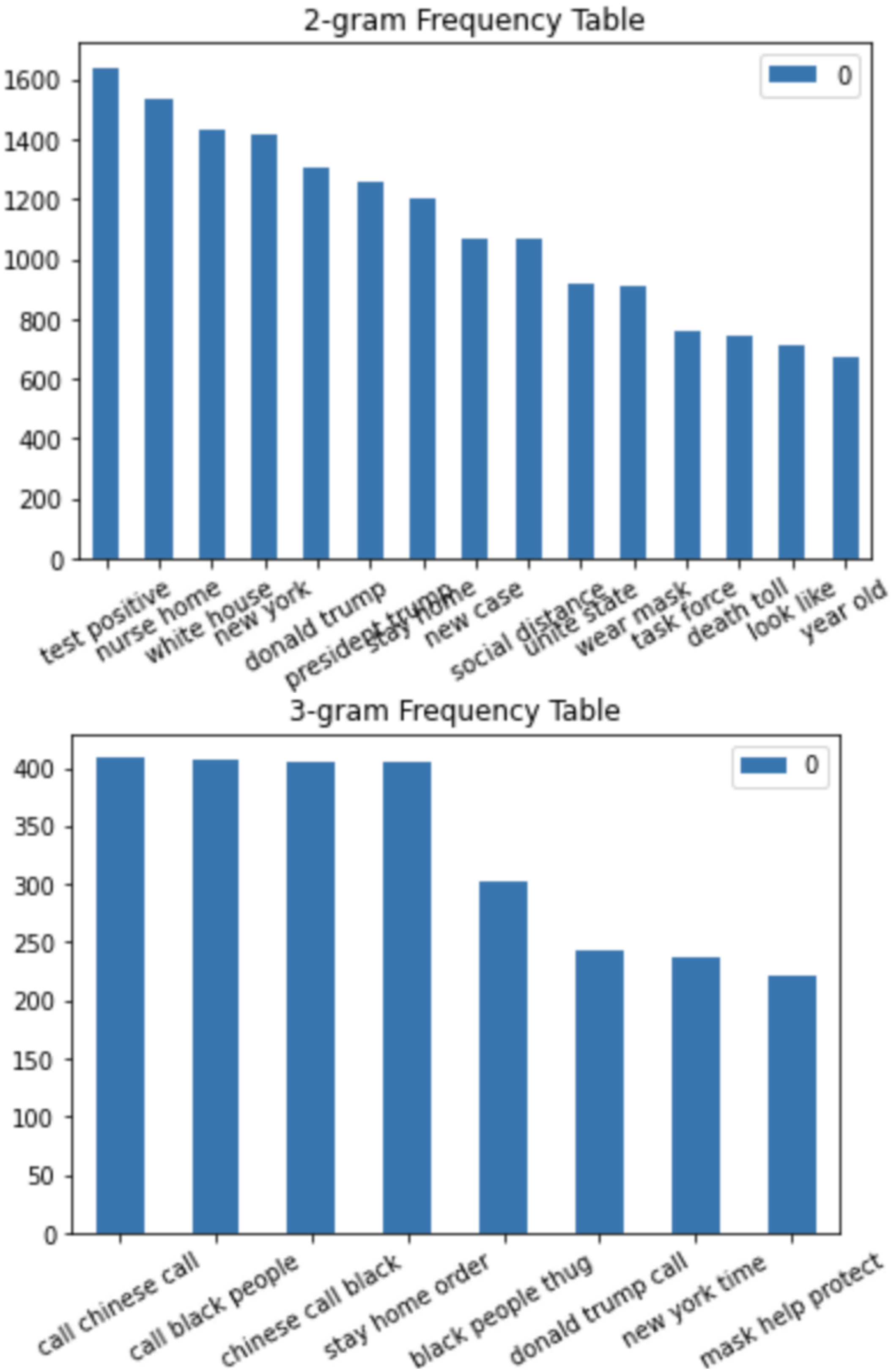
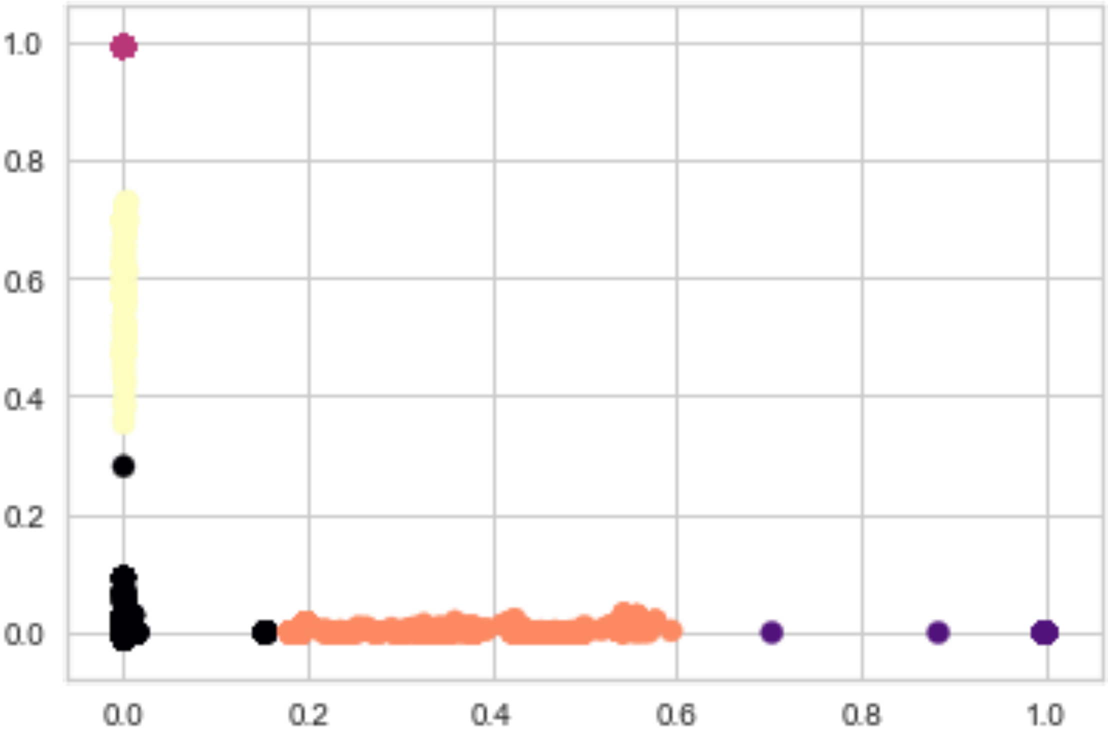


Figure A1-d N-gram Representation for Kansas' Tweets

(1) **California**



(2) **Kansas**

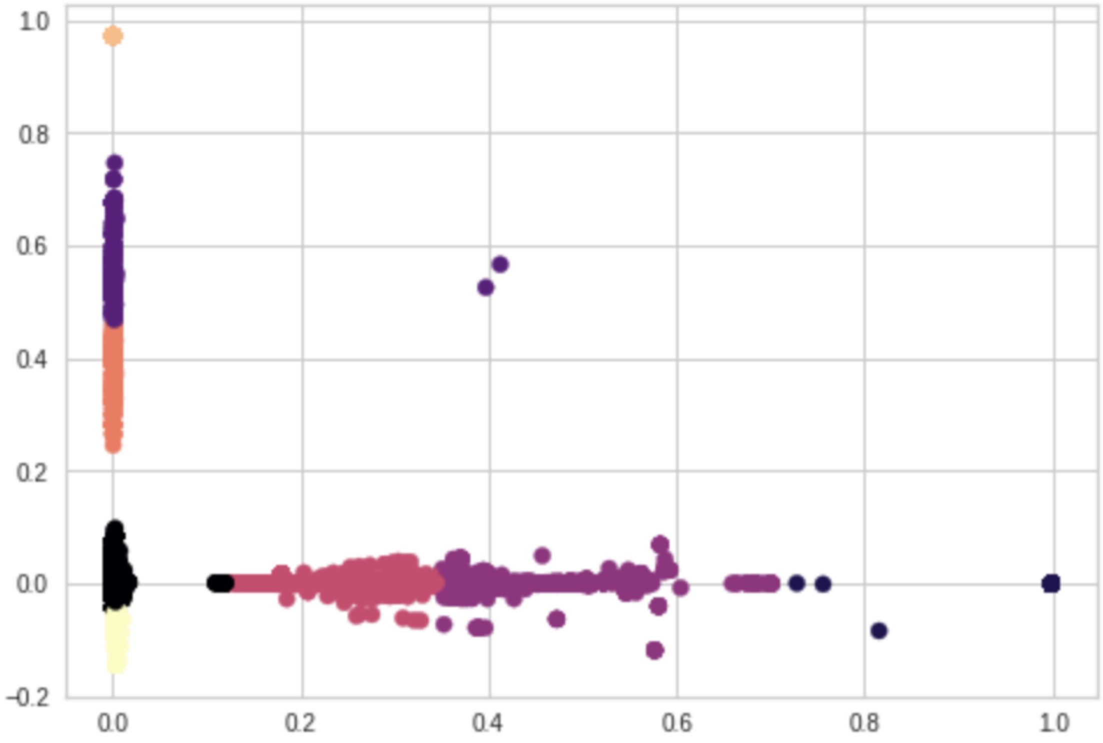


Figure A2 - Representation of Tweets into Different Clusters (California and Kansas)

References

1. Wikipedia (2020), Coronavirus disease 2019. Retrieved from https://en.wikipedia.org/wiki/Coronavirus_disease_2019
2. World Bank Group (2020), Global Economic Prospects Report
3. Samuel, J., Ali, G., Rahman, M., Esawi, E., and Samuel, Y. (2020), COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification, *Information*, *11*, 314, www.mdpi.com/journal/information
4. Bhat, M., Qadri, M., Geg, N., Kundroo, M., Ahanger, N., and Agarwal, B. (2020) Sentiment analysis of social media response on the Covid19 outbreak. *Brain, Behavior, and Immunity*, *87*, 136-137
5. Emtiaz Ahmed, M., Rafiqul Islam Rabin, M., and Naz Chowdhury, F. (2020) COVID-19: Social Media Sentiment Analysis on Reopening. Retrieved from https://ui.adsabs.harvard.edu/abs/2020arXiv200600804E/abstract
6. N.D. (2020), Status of State COVID-19 Emergency Orders. Retrieved from https://www.nga.org/state-covid-19-emergency-orders/
7. Dong, E., Du, H., and Gardner, L. (2020) An interactive web-based dashboard to track COVID-19 in real time. *Correspondence, 20*(5), 533-534
8. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
9. KMBC News (2020), Governor recommends all Kansans wear face masks while out in public. Retrieved from: https://www.kmbc.com/article/governor-recommends-all-kansans-wear-face-masks-while-out-in-public-covid-19-coronavirus/32143168
10. Bureau of Economic Analysis (2020), Percent Change in Real Gross Domestic Product (GDP) by State and Region, 2019:Q1–2020:Q1. Retrieved from: https://www.bea.gov/news/2020/gross-domestic-product-state-1st-quarter-2020
11. United States Department of Labor, Employment & Training Administration (2020), Unemployment Insurance Weekly Claims Data (run date 8/5/2020). Retrieved from: https://oui.doleta.gov/unemploy/claims.asp