CIS 600 Team Project Report

Study of COVID-19 and Election

Instructor: Professor Edmund Yu

Team Members: Leah Luo, Minxuan Qin, Yuxin Li, Bo Li

Date: 12/05/2020

Abstract

2020 must be a memorable year to many Americans. During this year, the United States suffered a severe COVID-19 epidemic. The COVID-19 epidemic has severely affected the lives of many people in the United States. One of our goals is to study the degree of concern and position of American people about the epidemic over time.

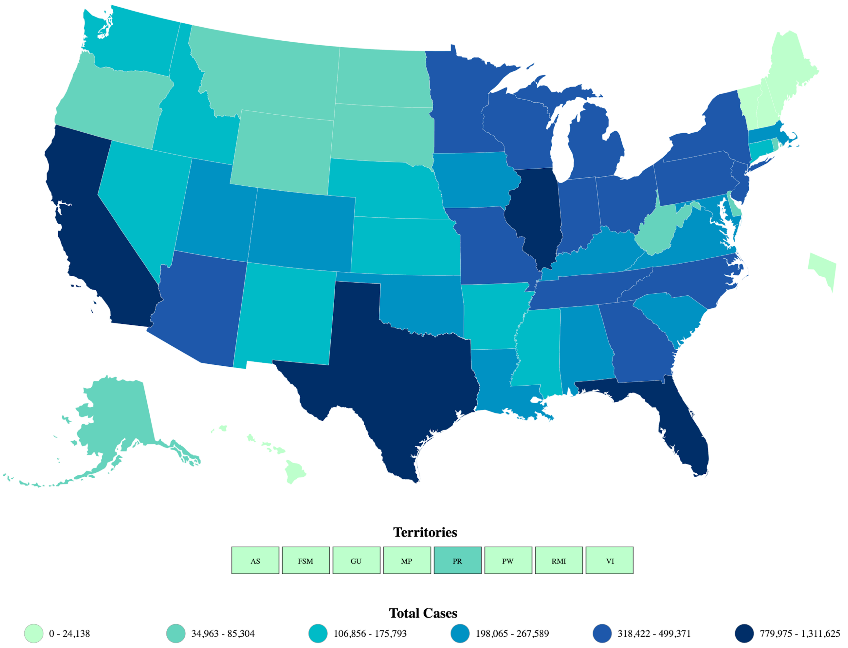
As the 2020 presidential election approaches the final poll, discussions on the topics of the election become more and more popular. During the two presidential debates, COVID-19 is one of the most important debate segments. As the debate continues, the discussion among COVID-19 continuously increases on Twitter. Many users post tweets to express different standpoints to the coronavirus pandemic. Now the winter comes, and the coronavirus comes back again. Our group members are interested in how the election affects the opinions of COVID-19 among individuals whose standpoints might be swayed by the election campaigns.

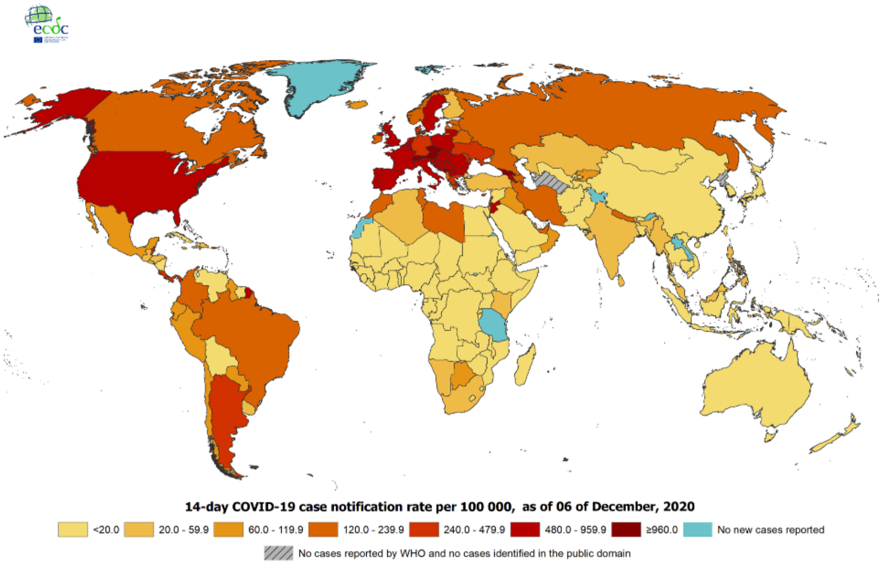
Table of Contents

1 Introduction

COVID-19

COVID-19 broke out in China in January. Unfortunately, the United States has also been affected. COVID-19 began to break out in the United States in March. So far, about 15 million people have been infected, and the death rate is as high as 280,000. In this project, we mainly focus on analyzing the views and attitudes of residents in some states on the COVID-19, and whether everyone's concern and attention to COVID-19 has changed significantly through the passage of time.




ecdc
14-day COVID-19 case notification rate per 100 000, as of 06 of December, 2020
<20.0
20.0 - 59.9
60.0 - 119.9
120.0 - 239.9
240.0 - 479.9
480.0 - 959.9
≥960.0
No new cases reported
No cases reported by WHO and no cases identified in the public domain

Election

The quadrennial presidential election just ended in November 2020. The COVID-19 has been one of the major topics of discussion, both in speeches and debates. Since last year's outbreak of the COVID-19, attitudes toward them have undergone many changes. Beginning in October, the two candidates traveled to various states to give speeches and hold debates as the election approached. We are curious as to how attitudes toward the COVID-19 have changed during this time for those who care about the election and the virus, whether they have changed as a result of politicians' speeches, and what the trends are.

Twitter

We decided to use Twitter as our data. Twitter was chosen for a few main reasons. First, Twitter is one of the most active social media platforms in the U.S. Currently, people use Twitter anytime to express their opinions and interact with others by liking, retweeting, commenting, etc., so we considered the Twitter data to be representative. Secondly, Twitter's development platform is very mature, and it provides rich powerful tools and apis for developers, thus easing the workload of crawling data.

2 Data Collection

Tweepy

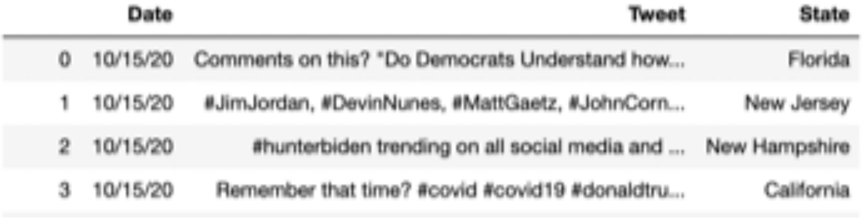
The Tweets dataset is obtained and generated from Twitter via its API. A developer account should apply for keys (consumer_key and consumer_secret) to invoke Twitter API.

The meaningless information should be cleared up when obtaining data from Twitter. For example, emojis take up a lot of space in the text, and they are meaningless. Therefore, we simply replace all emojis with empty strings.

We want to collect all tweets that are related to two presidential candidates Donald Trump and Joe Biden as hashtags. Then we'll save all tweets into CSV file in order that we could easily further process data by cleaning and analyzing.

All tweets are collected from 10/15/2020 to 11/30/2020 in the United States.

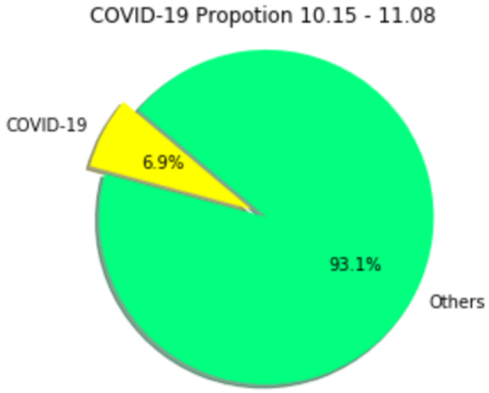
Our analysis is divided into two main parts: analysis before election and analysis after election. We will perform cleaning and sentiment analysis through all tweets in this time period.

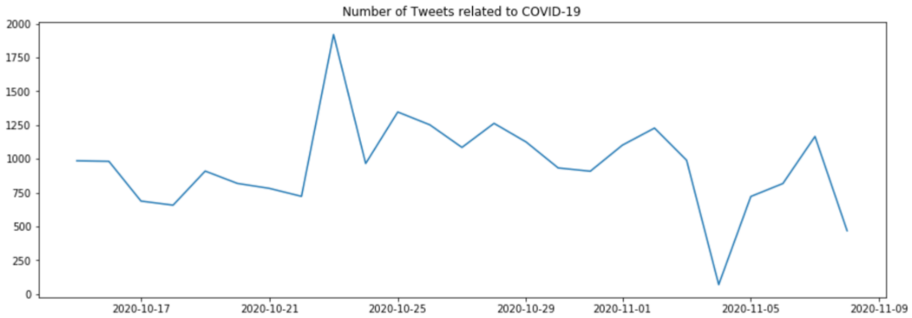


	Date	Tweet	State
0	10/15/20	Comments on this? "Do Democrats Understand how...	Florida
1	10/15/20	#JimJordan, #DevinNunes, #MattGaetz, #JohnCorn...	New Jersey
2	10/15/20	#hunterbiden trending on all social media and ...	New Hampshire
3	10/15/20	Remember that time? #covid #covid19 #donaldtru...	California

Extraction

Because we wanted to do sentiment analysis related to the COVID-19, we first screened the crawled tweets. We created a list of keywords that covered words we thought were related to the COVID-19, such as epidemic, coronavirus, and so on. From this list, we filtered out tweets that were only related to COVID-19 and did a simple analysis. As shown below, among the people who followed the election, COVID-19 was one of the topics of great interest, with tweets about COVID-19 accounting for twelve percent of the total.





3 Model for Sentiment Analysis

Sentiment analysis means the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.
We are going to conduct the sentiment analysis with the data collected.

NLTK

We used NLTK to conduct the analysis.
NLTK refers to The Natural Language Toolkit, it is a leading platform for building python programs to work with the language data. It also includes graphical demonstrations and sample data for developers to train models. The reason we chose NLTK is that it has a suite of text processing libraries for various kinds of analysis, for example, classification, tokenization, stemming, tagging, semantic reasoning, and so on.

Model Training

To train the model for sentiment analysis, we used the training set named "twitter samples" from NLTK package. It is a dataset of sample tweets. There are 5000 tweets with positive sentiments, 5000 with negative, and 20000 with no sentiment. A model will be created and trained on this dataset, and we will evaluate this new model before applying it on our Twitter dataset.
negative_tweets.json: 5000 tweets with negative sentiments
positive_tweets.json: 5000 tweets with positive sentiments tweets.20150430-223406.json: 20000 tweets with no sentiments

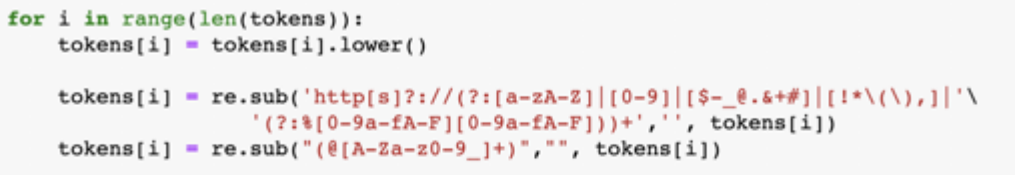
Data Cleaning

Tokenization

Tokenization is a common and necessary process in Natural Language Processing. Generally, tokens can be either words, characters, or subwords. Work Tokenization is the most common one, which separates a piece of text into individual words. In this project, we applied the word tokenization on the training dataset.

Noise and Stopwords

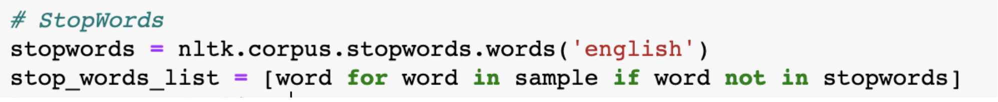
Noise refers to hyperlinks, twitter handles in replies, punctuation, and special characters. We used the regular expression to remove them, as shown below.



```
for i in range(len(tokens)):
    tokens[i] = tokens[i].lower()

    tokens[i] = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+#]|[!*\(\),]|'\
                       '(?:%[0-9a-fA-F][0-9a-fA-F]))+','', tokens[i])
    tokens[i] = re.sub("(@[A-Za-z0-9_]+)","", tokens[i])
```

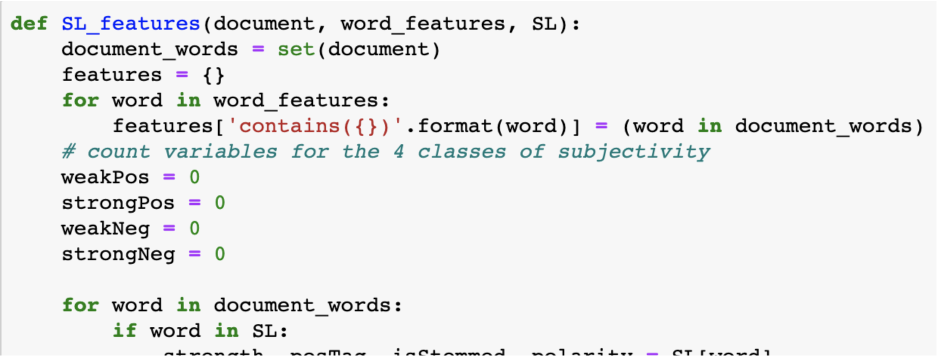
Stopwords refers to the commonly used words, for example, "the", "a", "is", etc.. They are not necessary for the analysis, and therefore we wanted to remove them. NLTK has a list of stopwords.



```
# StopWords
stopwords = nltk.corpus.stopwords.words('english')
stop_words_list = [word for word in sample if word not in stopwords]
```

Subjectivity Feature

We created a Subjectivity feature and function that applies the feature.



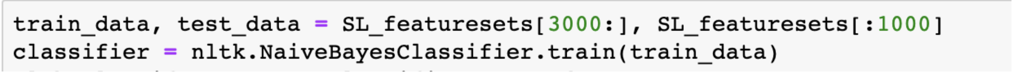
```
def SL_features(document, word_features, SL):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains({})'.format(word)] = (word in document_words)
    # count variables for the 4 classes of subjectivity
    weakPos = 0
    strongPos = 0
    weakNeg = 0
    strongNeg = 0

    for word in document_words:
        if word in SL:
            strength, posTag, isStemmed, polarity = SL[word]
```

NavieBayesClassifier

To build the model, we created the training and test set, and a dataset that has all the positive and negative tweets.

The classifier we used is the NavieBayesClassifier. In statistics, Navie Bayes classifiers are a family of simple “probabilistic classifiers” based on applying Bayes’ theorem with strong independence assumptions between the features. We chose it because it is simple but powerful, and can achieve a high accuracy level.

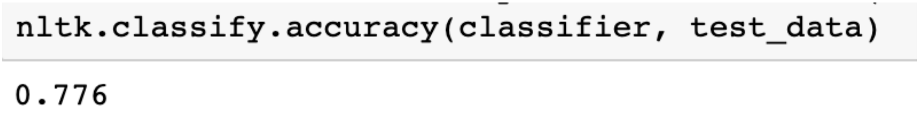


```
train_data, test_data = SL_featuresets[3000:], SL_featuresets[:1000]
classifier = nltk.NaiveBayesClassifier.train(train_data)
```

Model Evaluation

Accuracy

We called the accuracy function in classify, and the accuracy of the new model is 0.776, which is pretty high. This accuracy is calculated with a simple algorithm.

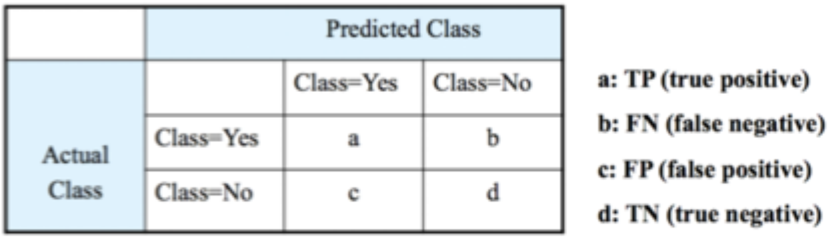


```
nltk.classify.accuracy(classifier, test_data)

0.776
```

F-measure

The matrix below is the result of a test for the number of actual class labels (“Yes” or “No”) that match with the predicted class. For example, TP (true positive) means that the predicted class matches the actual, FP (false positive) maans that the predicted class doesn’t match the actual one, that the predicted class is “Yes” which should be “No”.



		Predicted Class	
Actual Class		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

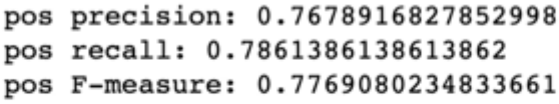
The measure idea is proposed in the Information Retrieval field. We used 2 common measures from this field, which are Precision and Recall. The formulas are shown below.

Precision and Recall can be combined into another measure which is named “F-measure”.

Precision = TP / (TP+FP); Recall = TP/(TP+FN)

*F-measure = 2 * (Recall * Precision) / (Recall + Precision)*

The evaluation results are shown below. The average score is about 0.77, we believe that the new model is accurate enough to conduct the sentiment analysis.

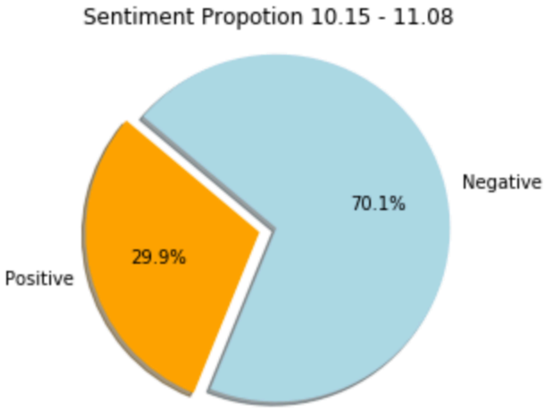


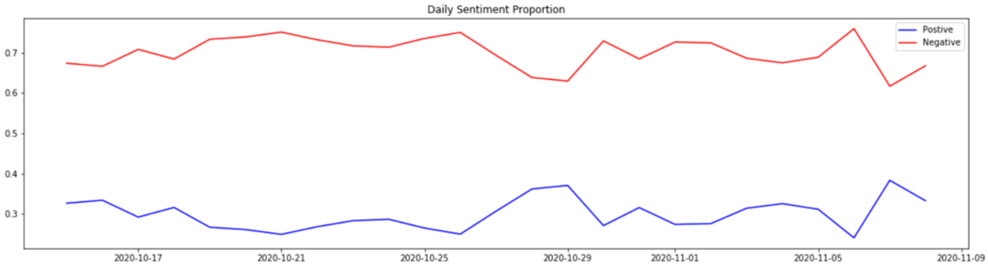
```
pos precision: 0.7678916827852998
pos recall: 0.7861386138613862
pos F-measure: 0.7769080234833661
```

4 Twitter Sentiment Analysis

Sentiment Proportion

We applied the trained model on our dataset. The analysis result indicates that there are 7244 tweets tagged “positive” while 17149 tweets tagged “negative”. The daily proportion analysis is shown below.

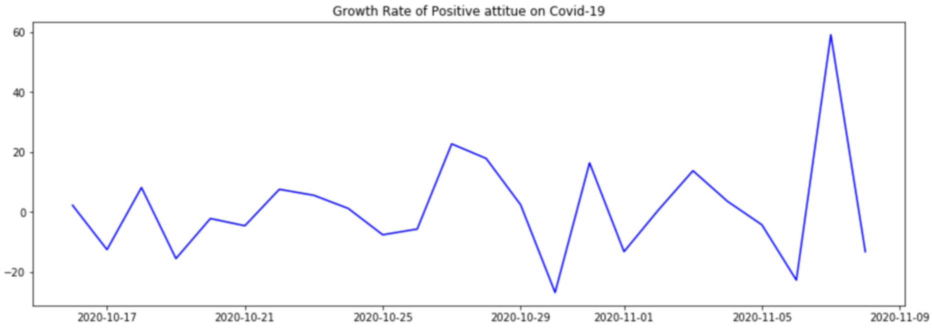


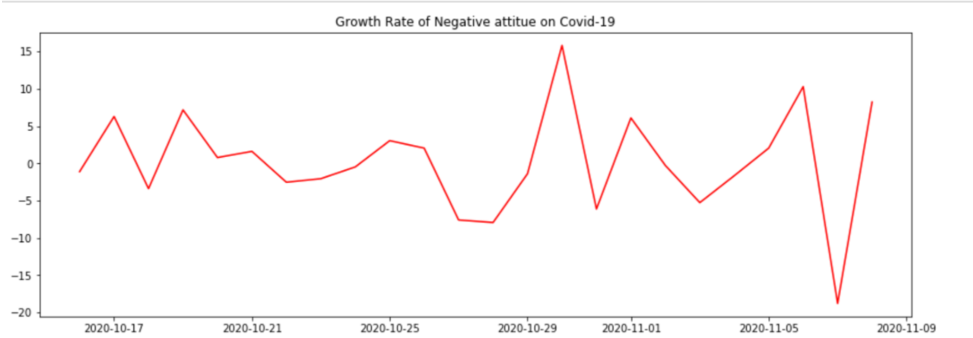


Growth Rate

The growth rate of negative tweets and positive tweets are shown below. Notice that there was a sharp rate around November 7.

Our guess is that this has something to do with voting day. People's moods change considerably around voting day, and there may be more people tweeting their opinions than usual. And the positive and negative changes could be a reflection of people's attitudes toward the future after learning the voting results.

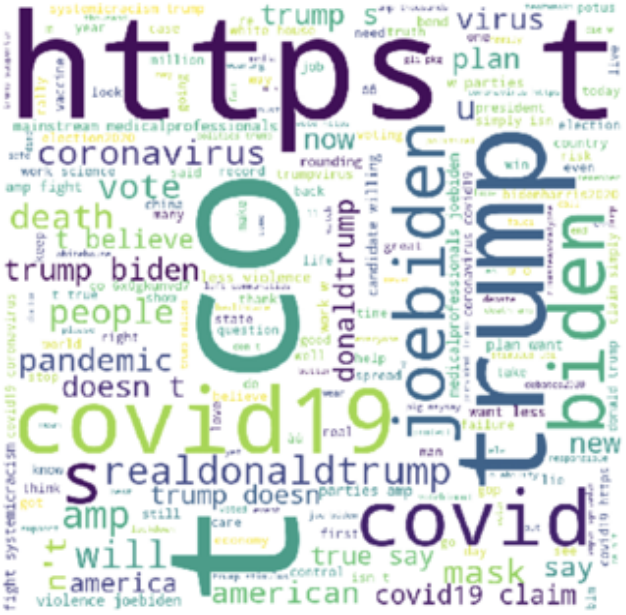




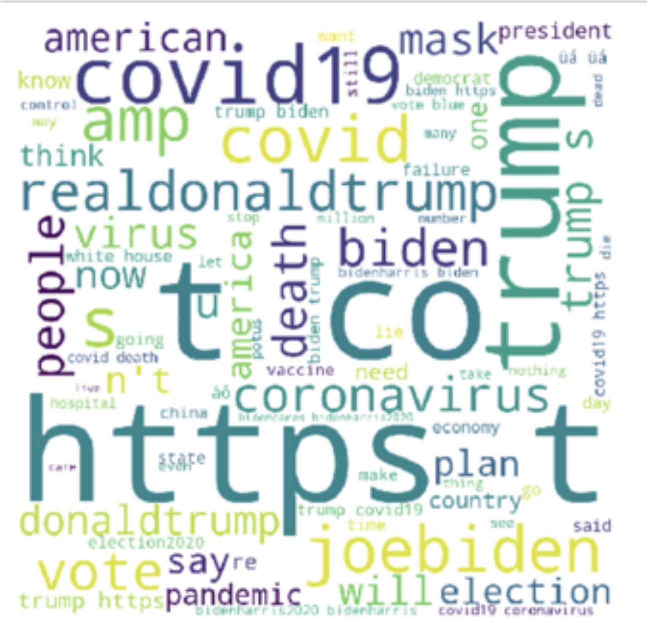
WordCloud

Word Cloud is a technique that represents the text with different size based on the frequency. We wanted to take a research on the word people used most frequently besides stop words. To do so, we downloaded and installed the three necessary modules, which are matplotlib, pandas, and word cloud.

Positive



Negative



Summary

Observing these two graphs, we could see that there is no clear sentimental trend in the use of high-frequency words, most of which are related to COVID-19 such as viruses and epidemics. However, in the word cloud of positive tweets, we can see that two of the high-frequency words are believe and professional. It can be inferred that people who have a positive attitude towards the COVID-19 are more positive about medical care and more positive about the future.

Future Plan

As mentioned above, we are curious about the reason of the sharp change around November 7.

Therefore, we continued to crawl tweets related to the election and the COVID-19 after November 9 for further analysis.

5 Twitter Sentiment Analysis Case Study

Election Speech and Sentiment Analysis

We are curious if it is possible to conduct an analysis on if politicians' speeches have influence on people's opinions towards COVID-19. Our guess is that if their speeches had a greater impact on the public's attitudes about COVID-19, then the public's attitudes about the COVID-19 should have changed more significantly around the time of the speeches.

Timeline

We compiled the timeline of two candidates' speeches in each state through Google search.

Both

On October 15, both Biden and Trump held separate town hall speeches, replacing the cancelled second debate.(Philadelphia, PA)
On October 22, Biden and Trump participated in a second and final debate in Nashville, Tennessee.

Joe Biden

On October 6, Biden made a campaign speech in Gettysburg, Pennsylvania, called "the best of his campaign" by CNN's John Avlon.

Date	City	State
Sat, October 10, 2020	Erie	PA
Mon, October 12, 2020	Toledo, Cincinnati	OH
Tue, October 13, 2020	Pembroke Pines, Miramar	FL
Fri, October, 2020	Detroit	MI
Sun, October 18, 2020	Durham	NC
Wed, October 21, 2020	Philadelphia	PA
Fri, October 23, 2020	Wilmington	DE
Sat, October 24, 2020	Bucks County, Luzerne County	PA
Tue, October 27, 2020	Warm Springs, Atlanta	GA
Thu, October 29, 2020	Broward County, Tampa	FL
Fri, October 30, 2020	Des Moines	IA
Fri, October 30, 2020	Saint Paul	MN
Sat, October 31, 2020	Flint, Detroit	MI
Sun, November 1, 2020	Philadelphia	PA
Mon, November 2, 2020	Cleveland	OH
Mon, November 2, 2020	Beaver County, Pittsburgh, philadelphia	PA
Tue, November 3, 2020	Philadelphia	PA

Donald Trump

Date	City	State

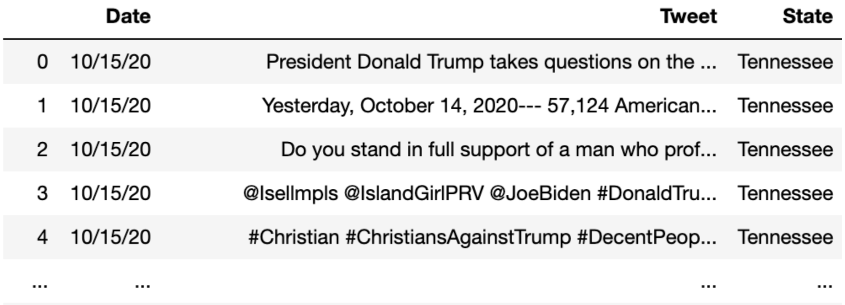
Mon, October 12, 2020	Sanford	FL
Tue, October 13, 2020	Johnstown	PA
Wed, October 14, 2020	Des Moines	IA
Thu, October 15, 2020	Greenville	NC
Fri, October 16, 2020	Ocala	FL
Sat, October 17, 2020	Muskegon	MI
Sun, October 18, 2020	Carson City	NV
Mon, October 19, 2020	Prescott, Tucson	AZ
Tue, October 20, 2020	Erie	PA
Wed, October 21, 2020	Gastonia	NC
Fri, October 23, 2020	The Villages, Pensacola	FL
Sat, October 24, 2020	Lumberton	NC
Sun, October 25, 2020	Manchester	NH
Mon, October 26, 2020	Allentown, Lititz, Martinsbury	PA
Tue, October 27, 2020	Lansing	MI
Wed, October 28, 2020	Bullhead City, Goodyear	AZ
Thu, October 29, 2020	Tampa	FL
Fri, October 30, 2020	Waterford Township	MI
Sat, October 31, 2020	Newtown, Reading, Butler, Montoursville	PA
Sun, November 1, 2020	Washington	MI
Mon, November 2, 2020	Fayetteville	NC
Mon, November 2, 2020	Scranton	PA
Mon, November 2, 2020	Traverse City	MI

Tennessee

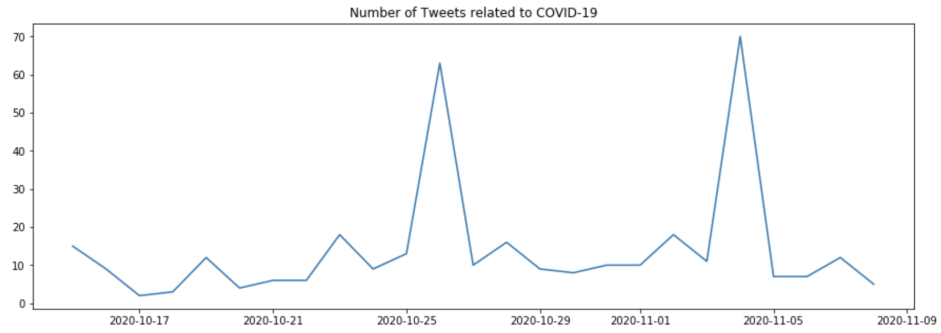
We noticed that they both participated in the second and final debate in Nashville, Tennessee on October 22 2020.

Extraction

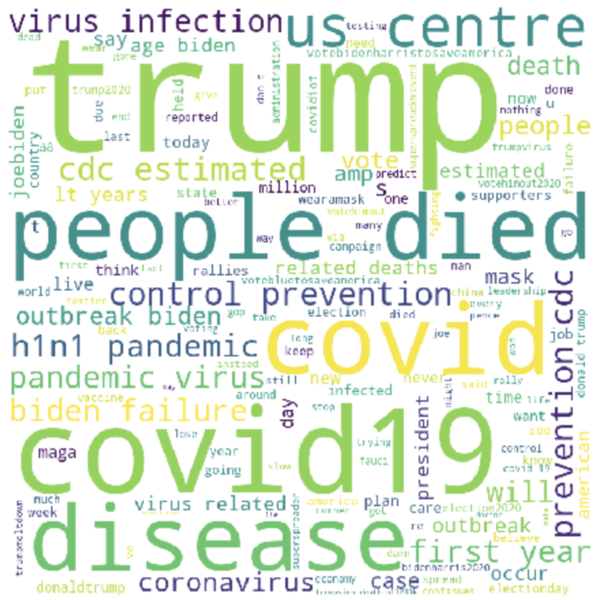
We extracted all the tweets that were posted in Tennessee and conducted the analysis on this new dataset.



	Date	Tweet	State
0	10/15/20	President Donald Trump takes questions on the ...	Tennessee
1	10/15/20	Yesterday, October 14, 2020--- 57,124 American...	Tennessee
2	10/15/20	Do you stand in full support of a man who prof...	Tennessee
3	10/15/20	@Isellmpls @IslandGirlPRV @JoeBiden #DonaldTru...	Tennessee
4	10/15/20	#Christian #ChristiansAgainstTrump #DecentPeop...	Tennessee
...	...	...	...

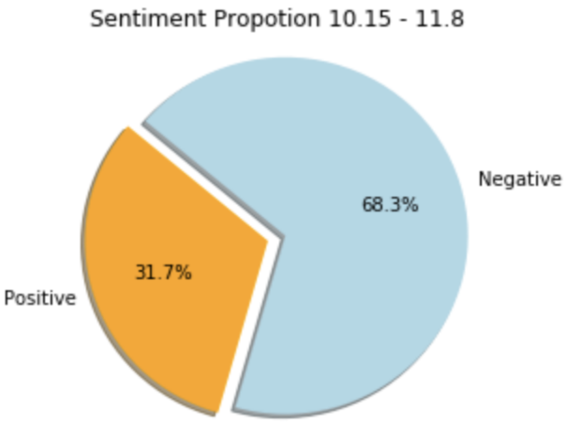


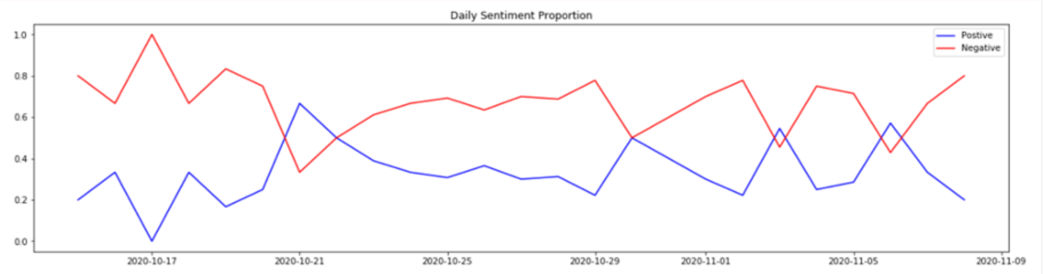
WordCloud



Sentiment Proportion

We conducted the sentiment analysis, utilizing the model we built. The result indicates 31.7% positive tweets and 68.3% negative tweets, as shown below.





Summary

The number of tweets that tagged “positive” decreased after 10/22, while the number of tweets tagged “negative” increased after 10/22. However, the changes are not that sharp that we could conclude that the speeches had a great impact on people in Tennessee.

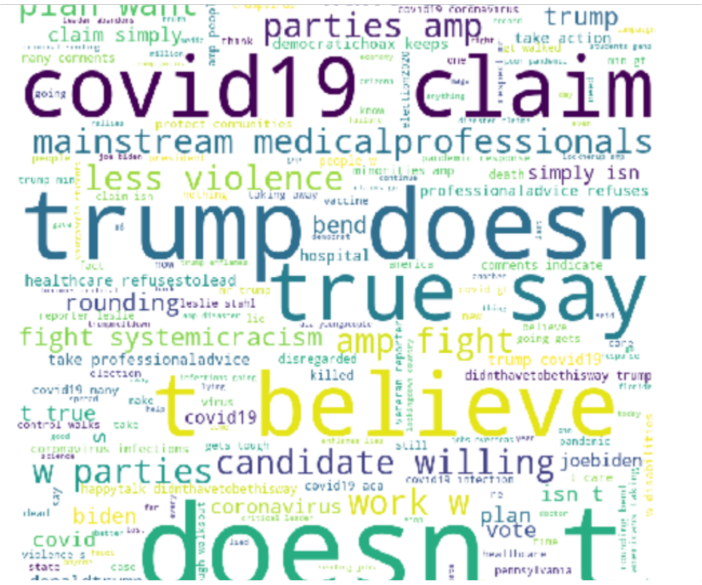
Pennsylvania

We noticed that there are several speeches in Pennsylvania.

Extraction

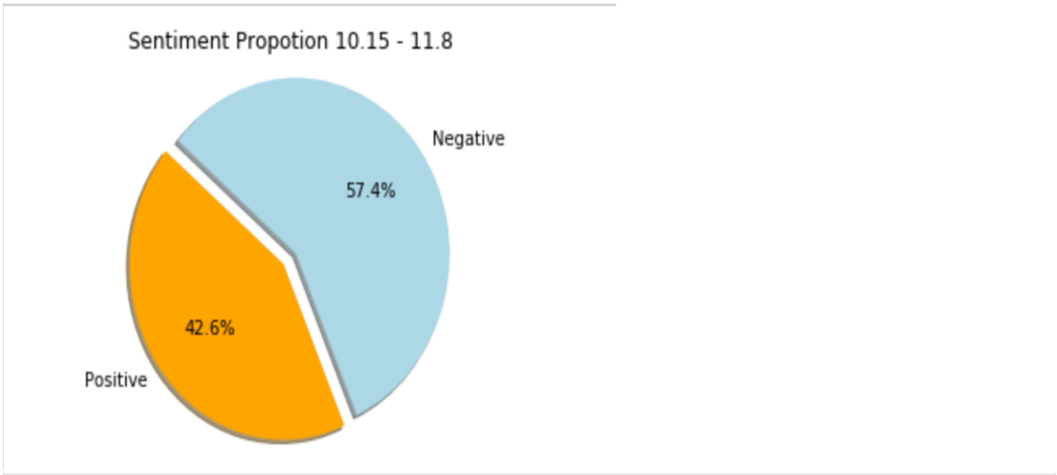
Similar to what we did for Tennessee, we extracted all the tweets posted on Pennsylvania.

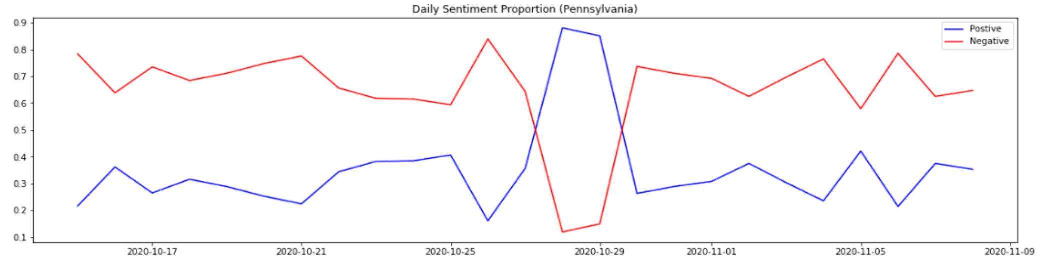
Word Cloud



Sentiment Analysis

We conducted the sentiment analysis with our model. The result shows that 42.6% tweets are tagged positive while 57.4% are tagged negative, as shown below.





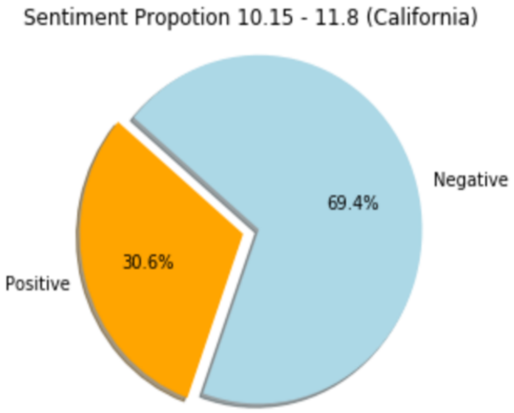
Summary

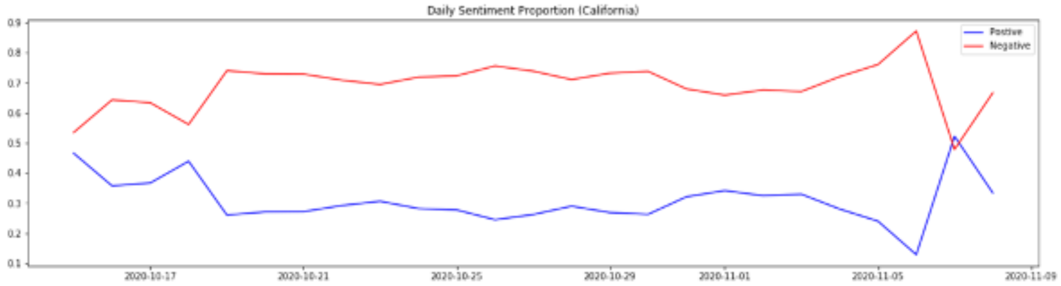
10/21: The number of tweets that tagged "positive" increased after 10/22, while the number of tweets tagged "negative" decreased.
11/1: The number of tweets that tagged "positive" increased slightly after 10/22, while the number of tweets tagged "negative" decreased.

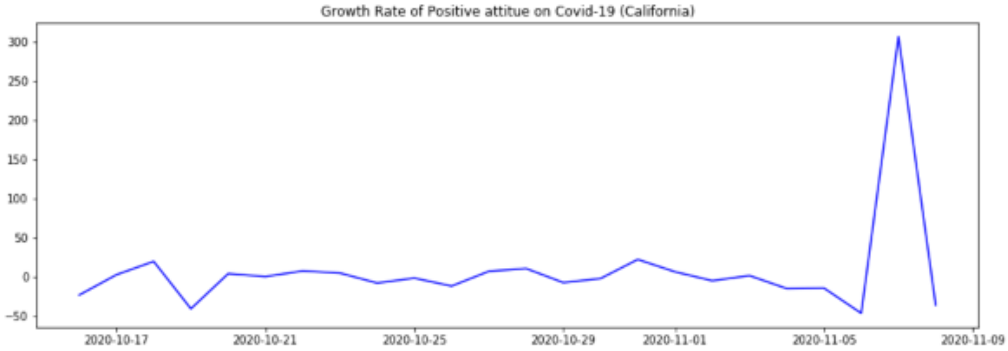
11/3: The number of tweets that tagged “positive” decreased after 10/22, while the number of tweets tagged “negative” increased.
The changes we mentioned above are not that obvious, as shown in the figure. However, from these two cases, we noticed that every state has very different patterns.

California

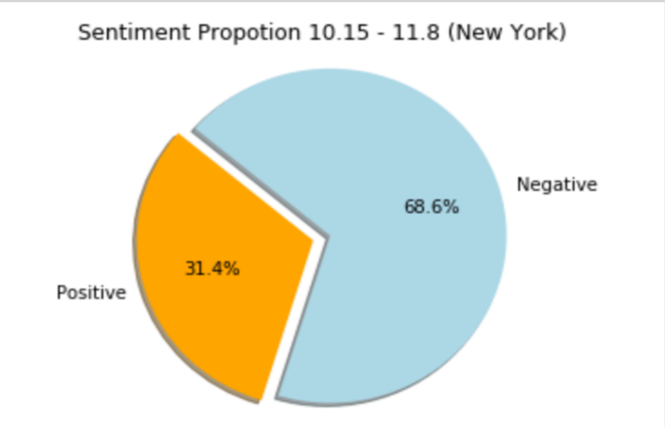
We conduct the sentiment analysis on California.

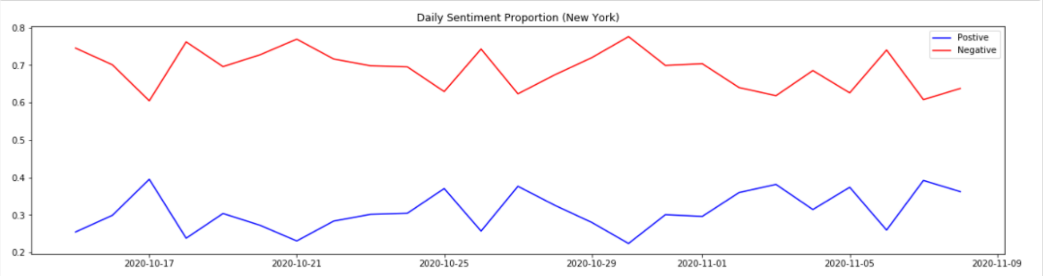


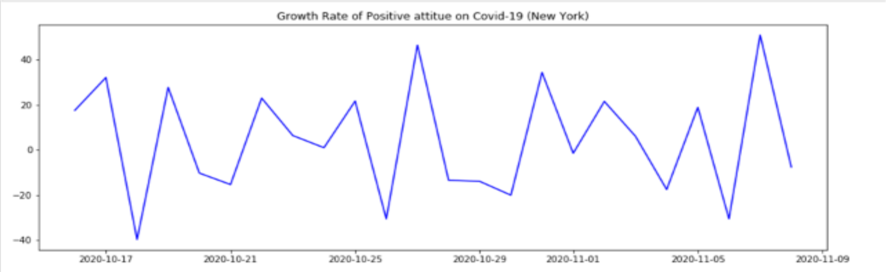


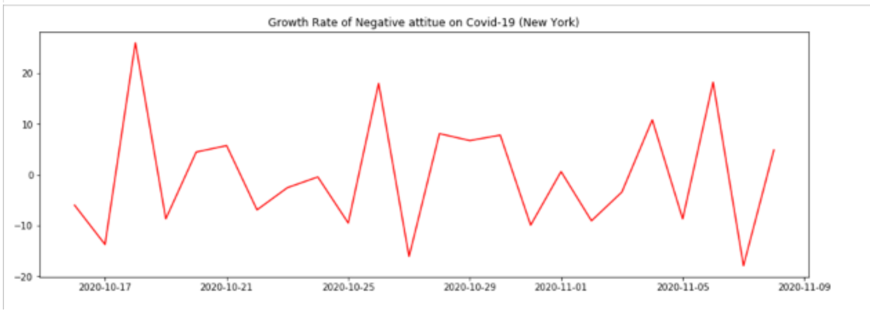


New York


Sentiment Propotion 10.15 - 11.8 (New York)
Negative
68.6%
31.4%
Positive


Daily Sentiment Proportion (New York)
Postive
Negative
0.8
0.7
0.6
0.5
0.4
0.3
0.2
2020-10-17
2020-10-21
2020-10-25
2020-10-29
2020-11-01
2020-11-05
2020-11-09


Growth Rate of Positive attitue on Covid-19 (New York)
40
20
0
−20
−40
2020-10-17
2020-10-21
2020-10-25
2020-10-29
2020-11-01
2020-11-05
2020-11-09

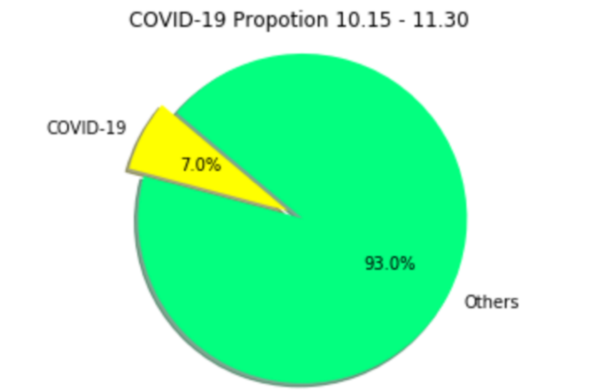

Growth Rate of Negative attitue on Covid-19 (New York)
20
10
0
−10
−20
2020-10-17
2020-10-21
2020-10-25
2020-10-29
2020-11-01
2020-11-05
2020-11-09

6 Data Collection and Sentiment Analysis after Election

We continued to crawl the data to further analyze whether people's attitudes toward the epidemic have changed significantly as a result of the election results.

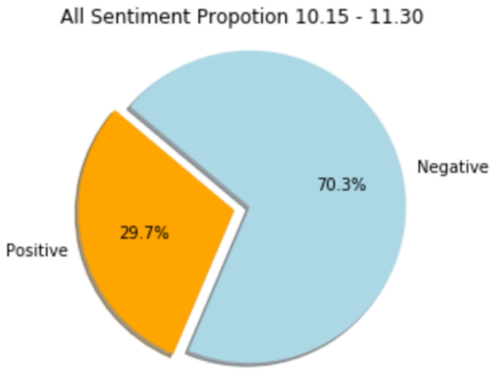
Proportion

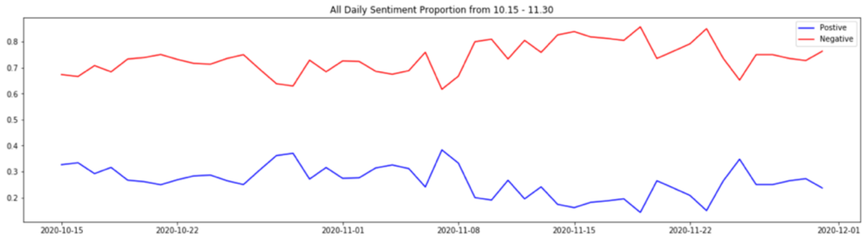
The figure below shows that among all the tweets we crawled, there are 7% related to COVID-19.



Sentiment Analysis

The figure below shows that among all the tweets that related to COVID-19, 29.7% are tagged positive while 70.3% are tagged negative. We could not see much changes from this figure so another analysis will be conducted.





The results of the analysis are somewhat surprising: after November 9, we can see that the number of negative tweets increased, while the number of positive tweets decreased. Despite the small change, it can be concluded that people continue to be mostly pessimistic about the COVID-19, and the number of tweets tends to increase.

7 Conclusion

The speeches do not have a significant impact on people's attitudes towards the COVID-19. However, attitudes have changed very significantly at certain times, such as around November 7.

In addition, the word cloud analysis shows that each state has a different focus on the epidemic. The trend of change in people's attitudes is also different in each state. Surprisingly, people's attitudes toward the COVID- 19 did not seem to change much as a result of the politician's speech. In Pennsylvania, for example, we would have expected some change in attitudes or an increase in tweets after a few major speeches, but there was not. We concluded that there were several reasons for this. First, when both candidates are traveling to different states before and after the election, sometimes several times a day, people may not pay attention to each speech, even if the candidate is speaking in their own state. Secondly, as the election approaches, people are already politically entrenched and are not swayed by speeches or debates, so our analysis shows little change. Last but not least, people's attitudes toward the COVID- 19 are more dependent on reality, i.e., current epidemic control efforts, epidemic trends, and progress of vaccine research. It can be seen that overall most people are pessimistic regardless of the date. It has been almost a year since the outbreak of the COVID-19 at the beginning of this year, so attitudes are not likely to change much as a result of a few presentations.

The epidemic has been going on for almost a year now, and our overall analysis shows that attitudes are still generally negative about the epidemic. We hope that the epidemic can be alleviated as soon as possible and people's attitude will be more optimistic.

8 Reference

1. Precision and recall. (2020, November 10). Retrieved December 08, 2020, from https://en.wikipedia.org/wiki/Precision_and_recall
2. Naive Bayes classifier. (2020, November 11). Retrieved December 08, 2020, from https://en.wikipedia.org/wiki/Naive_Bayes_classifier
3. F-score. (2020, November 06). Retrieved December 08, 2020, from https://en.wikipedia.org/wiki/F-score

4. Joe Biden 2020 presidential campaign. (2020, October 22). Retrieved October 26, 2020, from https://en.wikipedia.org/wiki/Joe_Biden_2020_presidential_campaign
5. Donald Trump 2020 presidential campaign. (2020, October 25). Retrieved October 26, 2020, from https://en.wikipedia.org/wiki/Donald_Trump_2020_presidential_campaign
6. John Hopkins Coronavirus Resource Center Home. (n.d.). Retrieved October 26, 2020, from https://coronavirus.jhu.edu/
7. Donald J. Trump for President. (n.d.). Retrieved October 26, 2020, from https://www.donaldjtrump.com/events/
8. Coronavirus Disease 2019 (COVID-19). (n.d.). Retrieved December 08, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/index.html