Raymond Adams, Scott Sirk

Determining if A State's Sentiment Towards COVID-19 has A Negative Correlation With That States Number of COVID-19 Cases

Introduction:

In March 2020 the World Health Organization declared the Coronavirus a global pandemic. According to the CDC, the virus was first reported on December 31, 2019 in Wuhan, China. Shortly afterwards the virus began spreading very quickly. In fact, the CDC states, "the virus had spread rapidly throughout China and within 1 month to several other countries, including Italy (2), the United States (3), and Germany (4)." [4]

COVID-19 has quickly become one of the most deadly events in history. The Washington Post stated in an article written on November 19, 2020 that nearly 250,000 Americans alone have died due to the virus since February 2020. The deadliest event in US history was the 1918 Spanish flu. This event killed roughly 675,000 Americans alone and amassed an even larger number world wide. With still no vaccine ready to be distributed, COVID-19 is estimated to kill another 250,000 Americans by April 2021. The Washington Post stated, "One of the more conservative disease models currently projects the United States could reach 438,000 deaths" by March 1, 2021 [2]. This number will surpass the number of Americans that died during World War II.

According to the CDC, the virus is spread in multiple ways. The most common way is coming in close contact, within 6 feet, with someone who has the virus. The CDC says, "It spreads through respiratory droplets or small particles, such as those in aerosols, produced when an infected person coughs, sneezes, sings, talks, or breathes." [4] They recommend wearing a mask when in public environments "such as grocery stores, pharmacies, and gas stations" as it can help reduce the spread of the virus [4].

Unfortunately, not everyone believes that the virus can be spread in the ways that the CDC discusses and some do not believe the virus is real or a threat at all. Public attitudes, behaviors, and beliefs towards COVID-19 can help public health officials determine what practices and restrictions should be implemented to mitigate the rate of infections and mortality. Attitudes, behaviors, and beliefs all lead to actions. The actions by the public determine how fast and to what extent the virus will spread. For example, if someone believes that it is important to wear a mask, maintain social distancing, and abide by rules and regulations they are less likely to spread the virus. Whereas someone who has opposing beliefs are more likely to spread the virus, at least this is what we assume.

This paper will attempt to determine whether or not beliefs towards COVID-19 have a direct correlation with the number of confirmed cases. To determine this twitter data will be collected from 4 states Illinois, Indiana, New Mexico, and Texas. Sentiment analysis will be conducted to determine each state's sentiment, or beliefs, towards the coronavirus using a support vector machines(SVM) classifier. Next machine learning will be used to determine if a linear regression model can predict how many cases a state has. These predictions will be matched up with statistics from the Centers for Disease Control and Prevention. If the numbers are close then it is possible to predict the number of cases by state from twitter sentiment. However, if the numbers are not similar it will be concluded that it is not possible to predict the number of cases by state from twitter sentiment.

Research Question:

Does a state's public sentiment towards COVID-19 have a negative correlation with that state's number of COVID-19 cases?

Data:

Data was gathered from Twitter using it's api through the Python Tweepy library. Tweets were collected from Illinois, Indiana, Texas, and New Mexico by using the geo parameter in the Cursor method for Tweepy. Tweets were gathered based on both popularity and how recent they were. The states were randomly chosen using Google's online generator. One thousand tweets were gathered from each state. The data was stored inside of a separate csv file for each state labeled illinois.csv, indiana.csv, texas.csv, and new_mexico.csv. For each state a unique id, the created date, state, and text of tweet was stored in columns. This information was then used to conduct sentiment analysis and predict the number of COVID-19 cases in that state.

Methods:

For this project sentiment analysis was conducted on the text of all tweets related to COVID-19. This task was completed separately for each state that the tweet was collected from. Therefore, sentiment analysis was conducted on the state of Illinois, Indiana, Texas, and New Mexico. To complete this the data was imported into Jupyter notebook using a data manipulation software named pandas. The data frame of each state was analyzed.

Head of data frame containing tweets related to COVID-19

Illinois Sentiment

```
illinois_twitter_data = pd.read_csv('data/states/illinois.csv')
illinois_twitter_data.head()
```

	id	create_date	state	sentiment	text
0	bc337c99-878d-49ed-8b1d-a1bc02e7aa24	2020-12-05 14:56:38	illinois	UNKNOWN	Diagnosing why the US has so blatantly failed ...
1	37636924-8276-45a5-b98f-a209e67b0d75	2020-12-05 00:28:37	illinois	UNKNOWN	BILL NYE is still the science show king. And h...
2	61863ba9-4214-498d-b4fc-8eb0d9354245	2020-12-05 06:06:44	illinois	UNKNOWN	Stellar detailed reporting from Florida on Gov...
3	fd8643a9-b9a7-4024-b25d-8aaf57dc0e52	2020-12-06 19:39:32	illinois	UNKNOWN	RT @Dr2NisreenAlwan: Some facts: ▪Children ca...
4	8a1e1ce8-8a1a-4ccb-96f5-94c654bfdd58	2020-12-06 19:39:32	illinois	UNKNOWN	RT @AbraarKaran: A quarter of all detected #co...

The text was then cleaned by removing the url and special characters. The cleaned text was then stored in a python dictionary. The key of the dictionary was an integer that started from 0 and increased by 1. The text was then stored as a value in the dictionary. TextBlob was then used to gain each tweet's polarity which is the numerical sentiment ranging from -1 to 1. A text is most negative at -1 and most positive at 1. Thus a tweet is considered neutral if it's polarity score is 0.

These values were stored in a new data frame named "*state_sentiment_df". This data frame contained two columns named "polarity" and "tweet". The polarity score will be useful later for the machine learning task. However, for the sentiment analysis we only want to know whether the tweet is negative, neutral, or positive. So a new column was created named "sentiment". A list of conditions were then created and used to label a tweet as negative, neutral, or positive. If the polarity of the tweet was below 0 then the tweet was labeled negative. If the polarity of the tweet was equal to 0 then the tweet was labeled neutral. If the polarity of the tweet was above 0 then the tweet was labeled positive. Seaborn was then used to analyze the sentiment values.

Lastly, linear regression was used to predict the number of COVID-19 cases. The python library scikit-learn was used to complete the linear regression. To conduct this a new data frame composed of 3 columns was created. The average polarity for each state was collected and stored inside of a column named "total_avg_polarity". A second column named "total covid-19 cases" was created and stored the total number of COVID-19 cases by state listed by the Centers for Disease and Control Prevention. The last column was named "state" and stored the name of the state that the features related to.

The x value for the equation was set as "total_avg_polarity". The y value was set as "total covid-19 cases". Scikit-learn's LinearRegression() method was used to fit the model. The intercept and coefficient were then found and used to build a linear regression equation. From this equation a states' average polarity could be set as the x value and used to predict that states' total number of COVID-19 cases.

Results:

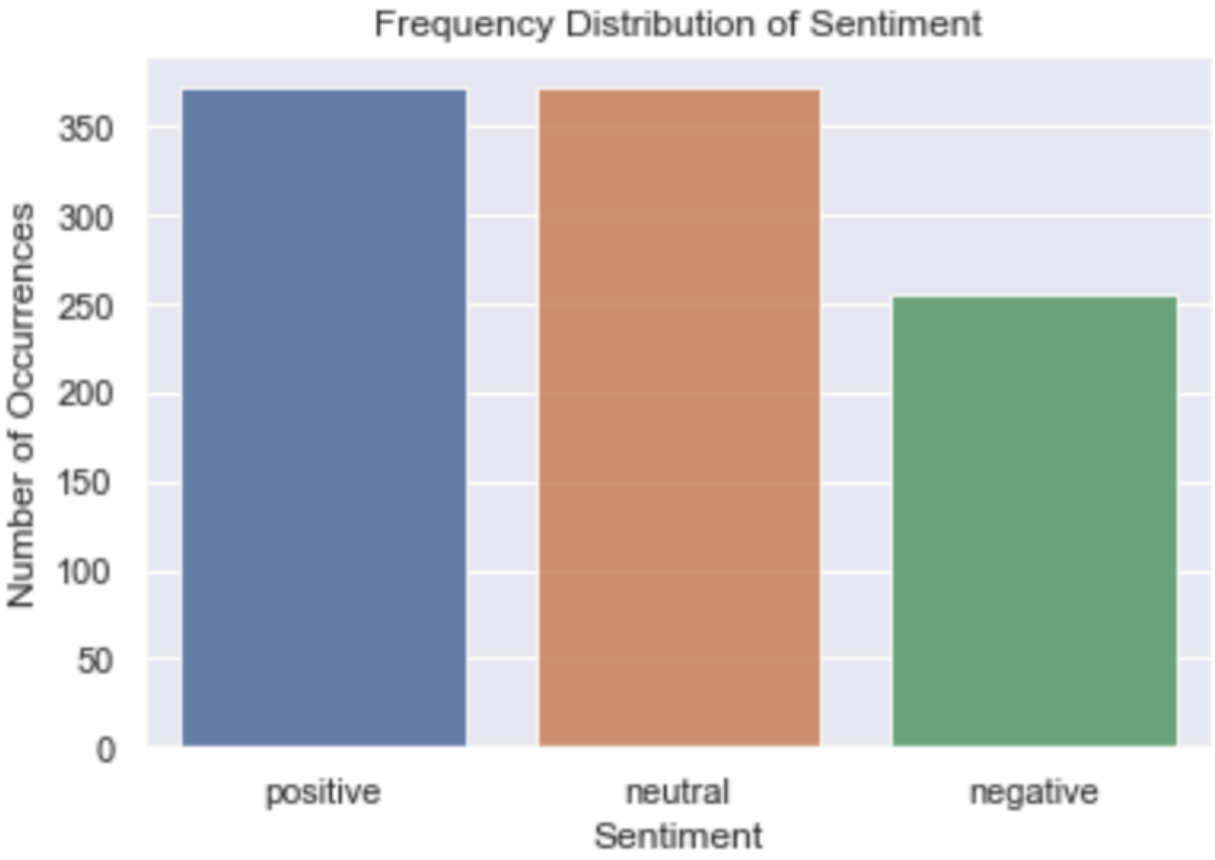
First the raw data was loaded from each of the state csv files into a pandas DataFrame. Using regular expressions URLs were removed from the text and all text was transformed into lowercase. Next each tweet was transformed into a TextBlob() object using the textblob module. This text was then fed into Tweepy to access Twitter's sentiment/polarity detection functionality. Initially a home grown Support Vector Machine(SVM) classifier was used to determine polarity, but since it performed only as well or worse than the baseline Twitter's polarity determination was used instead. This calculated polarity was then saved into a new DataFrame for analysis and visualization.

Data frame containing the polarity score of each tweet and the categorical sentiment of each tweet

	polarity	tweet	sentiment
0	-0.500000	diagnosing why the us has so blatantly failed ...	negative
1	0.000000	bill nye is still the science show king and he...	neutral
2	0.333333	stellar detailed reporting from florida on gov...	positive
3	0.000000	rt dr2nisreenalwan some facts children can get...	neutral
4	-0.250000	rt abraarkaran a quarter of all detected covid...	negative

This process was then repeated for each state so that all data was transformed and ready for analysis. Due to some failures in the data collection there were repeating tweets within each state's data set, and there were also shared tweets between different state datasets. This likely caused the separate data sets to be more similar than they actually are, or it is possible that no matter the region Twitter sentiment is relatively consistent. With that being said the overall sentiment seemed to show that roughly 40% of tweets were positive, 40% of tweets were neutral and the remaining 20% of tweets were negative.

Barplot of the frequency distribution of COVID-19 related tweets from illinois



As stated above, because of the duplication from data collection these values remained largely consistent for each state as well as when the states' sentiments were aggregated. It is unclear if this relatively even split is due to the specifics of the data set being analyzed or if it is reflective of Twitter's sentiment on Covid19. All state data was then combined into a single DataFrame to model a linear regression. Calculated sentiment was passed as the features and the total Covid19 cases in the state were treated as the labels. Using sklearn's LinearRegression module a line was fit to the data. Due to the low number of data points it is not clear if the model was predictive or not.

Conclusion:

There were some failures in data collection that likely caused the analysis to be incorrect. First, much of the data that was collected repeated. This appears to be largely caused by some early decisions made in the data collection process. First tweets were collected using a mix of popular tweets and recent tweets. It is possible that popular tweets could be popular in different regions. Second retweets were not filtered out. Because of that the same tweet could keep showing up in our data set. The last major flaw in the data collection process comes from a possible bug in our collection code. Using Tweepy's search api Twitter was queried for tweets related to Covid19 and 1000 tweets were requested for each run. However, a value was not passed to the count parameter of the search api. It appears that this may have caused Twitter to return the same tweets over and over until it reached 1000 tweets, successfully running but returning poor results.

Because very few states were used in the analysis the answer to the research question appears to be largely inconclusive. It is difficult to tell if sentiment is predictive of Covid19 cases because there are effectively only four data points. While the line drawn by the linear regression fits the data provided due to the sparseness of the data set it is difficult to tell how well the model would predict cases in other states or regions. This could be tested by adding more data points, states, to the model and seeing if it remains predictive or if it's accuracy decreases. Instead of looking at the state level by zooming into counties or cities the model would have access to more data points and possibly generate more conclusive findings.

Preliminary results appear promising, but more research would be required to answer the research question more definitively. Due to some issues with data collection duplicate tweets were present potentially causing prediction to be less accurate. Also there were relatively few data points present for the linear regression to work with so either adding more states or changing the scope to predict on a more detailed scale, like by city, could give the model more data to work with potentially making it more predictive. Due to these issues the research shows that polarity on Twitter is not predictive of Covid case in a state, but further research is required for a definitive result.