Sentiment Analysis of Covid-19 Tweets Using Naive Bayes Classification

Nico Espinosa Dice

May 14, 2020

Contents

1 Introduction.

For my project, I conducted sentiment analysis on Tweets regarding the Covid-19 pandemic using a Naive Bayes classifier. Through this project, I examined the relationship between the severity of the virus and its impact on public's reaction to it, particularly on social media. In doing so, I hoped to develop an understanding of how the public's response – particularly the hysteria many felt as the pandemic approached their area – is correlated to the actual virus's spread.

1.1 Experimental Question.

Is the public's sentiment when discussing the virus correlated to the virus's actual spread?

1.1.1 Hypothesis.

I believe that negative sentiment in Tweets regarding Covid-19 will be positively correlated with the number of confirmed cases of Covid-19.

2 Datasets.

All datasets used in this project are available publicly and are linked below.

2.1 Dataset: Covid-19 Tweets.

The dataset that included Tweets regarding Covid-19 was provided by Panacea Lab, and it is available publicly on Zenodo and Github. The dataset consists of timestamped Tweets related to the Covid-19 pandemic.

2.2 Dataset: Covid-19 Cases.

The dataset that included statistics related to the development of Covid-19 cases, as well as recoveries and deaths, was compiled by the Johns Hopkins University Center for Systems Science and Engineering. The data is available publicly on the Humanitarian Data Exchange.

2.3 Dataset: Twitter Sentiment Corpus

The dataset that included the Twitter sentiment corpus is available publicly on Github.

2.4 Exploratory Data Analysis.

2.4.1 Data Features.

The dataset of Tweets regarding Covid-19 contained the following features:

- Tweet's Date
- Tweet's Text

The dataset of Covid-19 cases contained the following features:

- Country/Region
- Date
- Number of Confirmed Covid-19 Cases per Date
- Number of Confirmed Covid-19 Recoveries Per Date
- Number of Confirmed Covid-19 Deaths Per Date

2.4.2 Data Shape.

The Twitter dataset contained 7500 Tweets. The Covid-19 dataset contained the cases each day over a three and a half month period, from late January to early May.

3 Mathematics of Naive Bayes Classifiers.

Naive Bayes classifiers are a Bayesian network model, a type of "probabilistic classifier," that uses Bayes' theorem to create classifications.

3.1 Bayes' Theorem.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}, \text{ where} \tag{1}$$

$$P(A \mid B) : \text{The probability that } A \text{ is true given that } B \text{ is true,} \tag{2}$$

$$P(B \mid A) : \text{The probability that } B \text{ is true given that } A \text{ is true,} \tag{3}$$

$$P(A) : \text{The probability that } A \text{ is true,} \tag{4}$$

$$P(B) : \text{The probability that } B \text{ is true.} \tag{5}$$

4 Naive Bayes Classifier Implementation.

First, the Naive Bayes classifier was trained on the Twitter sentiment corpus. The corpus contained thousands of Tweets, each labeled with a sentiment.

Next, the model was applied to the Covid-19 Tweets dataset. Each Covid-19 Tweet was assigned a sentiment – positive, neutral, negative, or irrelevant – based on the predictions of the model.

Next, the data was partitioned into weeks, and the percentages of each sentiment of the Tweets during each week was recorded.

Finally, the correlation between the percentage of positive/negative Tweets and the number of Covid-19 cases confirmed/recovered/deaths were measured.

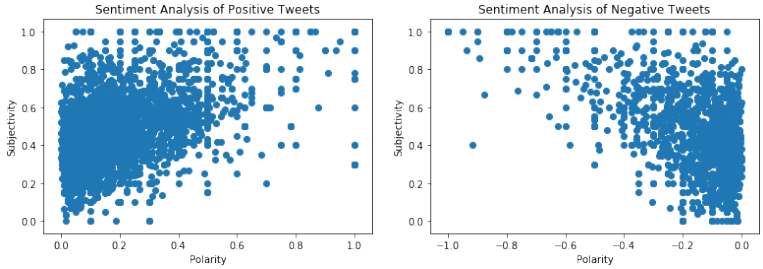
4.1 Code.

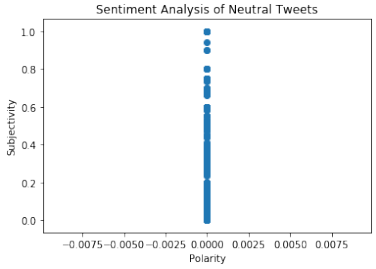
The Naive Bayes classifier was coded in Python using the Natural Language Toolkit. Additionally, the TextBlob library was used to create an ensemble of sentiment analysis models. The datasets were imported and explored using the Pandas library. The code is available on my Github.

5 Results.

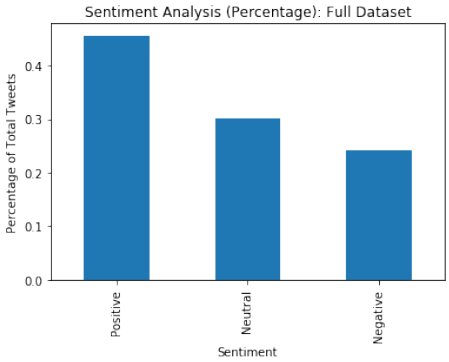
5.1 TextBlob Model.

Using the TextBlob model, I classified the subjectivity and polarity of each Tweet. The relationship between subjectivity and polarity is presented in graphs below:

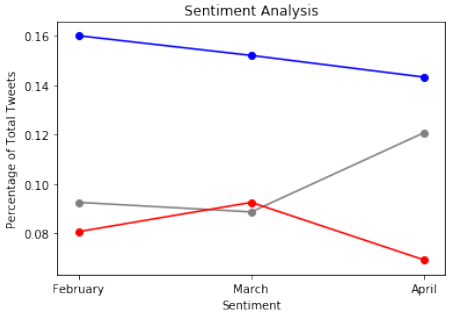




The Tweets were classified in the following percentages:



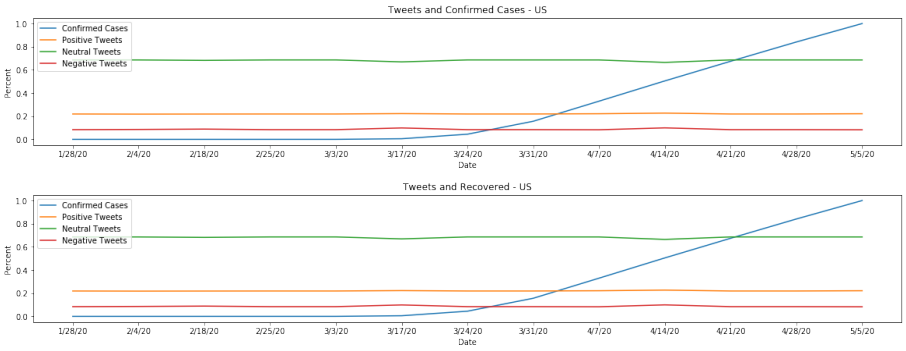
Overall, the TextBlob model classified the following sentiment percentages for each month:

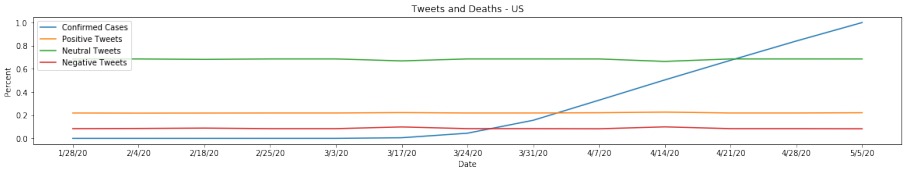


The blue line refers to Tweets classified as "positive." The red line refers to Tweets classified as "negative." The grey line refers to Tweets classified as "neutral."

5.2 Naive Bayes Classifier.

The Naive Bayes classifier produced the following results:





Furthermore, using the Naive Bayes classifier's sentiment predictions of each Tweet, the following correlation table was created:

	Cases	Normalized	Positive Sentiment	Neutral Sentiment	Negative Sentiment
Cases	1.000000	1.000000	0.292486	0.017619	-0.120203
Normalized	1.000000	1.000000	0.292486	0.017619	-0.120203
Positive Sentiment	0.292486	0.292486	1.000000	-0.861067	0.760407
Neutral Sentiment	0.017619	0.017619	-0.861067	1.000000	-0.984273
Negative Sentiment	-0.120203	-0.120203	0.760407	-0.984273	1.000000

The correlation for Covid-19 confirmed cases.

	Cases	Normalized	Positive Sentiment	Neutral Sentiment	Negative Sentiment
Cases	1.000000	1.000000	0.203056	0.087071	-0.173824
Normalized	1.000000	1.000000	0.203056	0.087071	-0.173824
Positive Sentiment	0.203056	0.203056	1.000000	-0.861067	0.760407
Neutral Sentiment	0.087071	0.087071	-0.861067	1.000000	-0.984273
Negative Sentiment	-0.173824	-0.173824	0.760407	-0.984273	1.000000

The correlation for Covid-19 confirmed recoveries.

	Cases	Normalized	Positive Sentiment	Neutral Sentiment	Negative Sentiment
Cases	1.000000	1.000000	0.251551	0.037319	-0.130999
Normalized	1.000000	1.000000	0.251551	0.037319	-0.130999
Positive Sentiment	0.251551	0.251551	1.000000	-0.861067	0.760407
Neutral Sentiment	0.037319	0.037319	-0.861067	1.000000	-0.984273
Negative Sentiment	-0.130999	-0.130999	0.760407	-0.984273	1.000000

The correlation for Covid-19 confirmed deaths.

6 Discussion.

The results show a few noticeable trends. First, there is a significant difference between the Naive Bayes classifier and the TextBlob model. This difference can be partly attributed to the fact that the Naive Bayes classifier was trained on Tweets, whereas the TextBlob model was trained on general text in several

forms of communication. The language of Tweets and our general writing is quite different, so this likely explains a significant portion of the discrepancy.

We see that the TextBlob model classified a significantly greater percentage of the Tweets as positive than the Naive Bayes classifier did.

Next, the TextBlob model shows that the percentage of positive Tweets decreased steadily from February to April. The Naive Bayes classifier also showed a decrease in the percentage of positive tweets, though not as noticeably as the TextBlob model. This suggests that as the pandemic spread in the United States, which began to take place serious in March and April, the public's sentiment – in this case people's Tweets – reflected a negative sentiment. This suggests that we may be able to gain some insight into the spread of the virus by monitoring the sentiment on Twitter.

The TextBlob model also shows a noticeable inverse relationship between the percentage of negative Tweets and the percentage of neutral Tweets. The percentage of negative Tweets dropped significantly from March to April, while the percentage of neutral Tweets increased significantly. The Naive Bayes classifier also showed a decrease in negative Tweets between certain weeks in late March and early April, although the trend is less noticeable. One possible reason for this trend is that many individuals began self-quarantining and sheltering-in -place during the month of March. Furthermore, as people began living with the virus present throughout the United States, they became used to their living adjustments, so their perspectives become less negative and more neutral. Again, I do not claim that this is exactly the reason for the trends; I see this as one possible implication of the trends the models demonstrate.

The correlation table output by the Naive Bayes classifier demonstrates that there are no strong correlations between the sentiment of Tweets and the number of Covid-19 cases. We can see that there is almost no correlation between the percentage of neutral Tweets and the number of confirmed cases, recoveries, or deaths. We see that there is a positive correlation between the positive sentiment of Tweets and recoveries, which seems reasonable. As more people recover from the virus, people's perspectives become more positive. Similarly, the negative sentiment of Tweets is negatively correlated with recoveries, backing the same trend.

There are some trends that are counter-intuitive. For example, positive sentiment of Tweets is positively correlated with confirmed cases. One possible explanation for this trend is that people are, in general, becoming more positive as they are adjusting to changed lives.

In summary, the Naive Bayes classifier and the TextBlob model demonstrate that sentiment analysis applied on Tweets can lead to insights into the perspective of the public as crises occur. However, given the many factors affecting people's sentiments, it is nearly impossible to correlate them with a single problem, even when focusing on Tweets related to the problem itself.

6.1 Future Research.

In future research, I hope to examine the correlation between Tweet sentiment and Covid-19 cases by region, specifically states. I was unable to perform sentiment analysis on the state level because my Twitter developer's application has not been approved. If it is approved, I should be able to use this model to compare the results.

7 References.

7.1 Academic Papers.

- Sentiment Analysis of Twitter Data (1)
- Sentiment Analysis of Twitter Data (2)
- Covid-19 Tweets Dataset and Statistics

7.2 Open-Source Resources.

- Twitter Sentiment Analysis with Explanation (Naive Bayes)
- Creating The Twitter Sentiment Analysis Program in Python with Naive Bayes Classification
- How to Do Sentiment Analysis on a Twitter Account
- Comprehensive Hands on Guide to Twitter Sentiment Analysis with Dataset and Code

7.3 Datasets.

- Covid-19 Tweets
- Covid-19 Cases
- Twitter Sentiment Corpus