Covid-19 Sentiment Analysis

Nathan Holdom, Mike McVey, Scott Newhard, Andrew Rebits, Ben Robbins
(holdomna, mcveymic, newhards, rebitsan, robbi155)@msu.edu
Michigan State University

ABSTRACT

The goal for this project was to identify trends in sentiment analysis on COVID-19 in Michigan and see how that sentiment would change over time. The model created predicts the sentiment of a tweets relating to COVID-19 based on positive, negative, or neutral values. The data used to train and validate the model, as well as the data used in prediction was collected by finding tweets with specific keywords (see Keywords section below) included in them. Potentially these finding could reveal a pattern between Covid cases and the sentiment of the tweets in the area, which could help notify communities that are likely to see a rise in cases. These findings could also show the general reception of how the government is handling of the pandemic statewide. In the end, the resulting data revealed a completely negative sentiment for Michigan throughout the entire duration of the days recorded. This negative sentiment fluctuated throughout the days while following a general pattern of more negative sentiment towards the start of collection that lead to slightly less negative sentiment towards the end.

Keywords

The keywords used for data collection from twitter were 'covid', 'corona', and ‘quarantine’.

1. INTRODUCTION

Our groups project focused on sentiment analysis. Sentiment analysis refers to intercepting and classifying emotions in data. In our case we used data gathered from twitter in the form of tweets. Tweets being short, formed commentary people post to the social media platform twitter. Specifically, these tweets all contained at least one of our keywords which dealt with the COVID-19 pandemic. Our goal is to create a model that predicts the sentiment of tweets related to COVID-19 and classifies the tweets as being positive, negative, or neutral. Our information deals with Michigan itself, though it could be applied to any geographical location. Our motivation is to be able to detect changes in the sentiment of areas and correlate it to the actual number of cases in that area. Rapid negative feelings are strong sign of worsening cases while gradual positive feelings could mirror a community's improvement[1]. This model could also be applied to understanding a community's response to governmental action dealing with COVID, such as new mandates.

An example of works like this project, can be found in JMIR Public Health and Surveillance in which researchers used Social networking sites such as twitter to track disease outbreaks and provide early warnings accurately[1]. A big correlation in theirs as well as ours is that the data generated from these social media sites is valuable for real time analysis. Both allow for copious amounts of unfiltered data from anywhere a user is online from any device.

A problem that arises in many of the tweets are how reactive and emotional people are when talking about Covide-19 online. This was only exemplified by the fact that 2020 was an election year and this became a very politicized issue. Most of the handling of the pandemic was put on the states backs, so we see a lot of negatively charged tweets dealing with state’s handling weather its supporting or attacking their decisions. Another issue with such a big topic like COVID-19 is people abusing twitters tag and keyword systems to promote tweets that are unrelated to people's sentiment on COVID-19. Something that can affect the sentiment of the tweets is the large number of publications releasing new data multiple times a day. An example of this may be the death toll for the day. This is something is not emotionally charged, and it is a statement of fact though with death and other words the classifier might deem these types of tweets as negative when they are truly neutral in sentiment. The largest portion of people affected by COVID-19 are the elderly which are not frequently users of social media like Twitter. When going through the manual labeling of our tweets one thing we all notice was the sarcastic language used, these tweets for us where very hard to label so they definitely are an issue for the classifier.

From current polling from FiveThirtyEight, 58% of Americans disapprove of the current government's response to COVID-19. In another poll from statewide survey from Michigan the current disapproval of the current government handling of COVID-19 was at 55%. So, the overall public response about the government handling of COVID-19 is negative. This is supported by what our findings in our data. With the issues stated in the challenges above and the cloak of online anonymity we received more flamboyant responses that skew the data. A lot of our data gathering happened through November into December which was the height of the election as well as a huge reflux in new cases inside the united states. Our average sentiment continued to slowly climb throughout the month which would make logical sense and support the overall negative view on the topic. When applying our classifier to test data we were able to correctly classify tweets at 59.6% percent of the time.

2. PRELIMINARIES

The data that we collected consisted of tweets that were specifically about coronavirus. These tweets were put into a JSON format which contained the content of the tweet as well as the date it was tweeted and a user ID attributed to the user that made the tweet. These tweets had a mention of coronavirus related subjects

in order to increase the number of tweets that we can analyze as opposed to limiting our search to tweets only explicitly containing the word "coronavirus". We then added to the data that we collected by manually assigning the tweets a sentiment score, 1 for positive, 0 for neutral and –1 for negative and adding that attribute to the JSON object that contained the tweet data.

Our goal is to train a model to accurately classify the sentiment of future tweets about coronavirus. These results could be used to see which areas are being heavily affected by the virus or virus-related government mandates, as more negative tweets coming from a specific area would indicate that there is a large negative affect in that area and possibly a larger spread of the virus, or a growing distaste for the way that the local or federal government are handling the disease. The idea is that the more an area is impacted by the virus, more people will write tweets regarding the virus and they will be more negative than other less affected areas. This data could potentially be used to predict an outbreak in an area before any CDC (Center for Disease Control) statistics come out regarding positive tests or deaths. The real-time nature of these tweets can provide an insight into how an area is being affected before the CDC or other agencies can aggregate test statistics and present them, which could help provide people in an affected area the information they need to assess the safety of going out to perform tasks outside of their homes. It can also be helpful in assessing the area's sentiment towards the mandates in their area and their effectiveness. This data can be incredibly important when it comes to providing real time updates on the safety of an area which has become increasingly important as the global pandemic continues to wreak havoc on the country.

3. METHODOLOGY

The framework that we produced consisted of data collection, preprocessing, analysis, postprocessing and visualization. Each of these components were a smaller part in a larger data pipeline.

Data collection was done with Tweepy API. Over the course of a week a data collection streamer script was ran continuously. In the end we produced many tweets from all over the world.

Preprocessing was done on this data beforehand and after to ensure that this was the data that we wanted. Tweets that had locations not in Michigan were removed. Tweets that did not contain the keywords were removed. Features were then extracted and formed from raw text after removing stop words, single characters, whitespace, and additional unneeded text. Then, we stemmed the words using NLTK PortStemmer and vectorized the words with Sklearn CountVectorizor. A challenge that came from preprocessing was that we first tried to use positive and negative word vector features to describe the data, however they were not configurable to the forest classifier we used. To solve this we switched to port stemming and vectorization.

After preprocessing was done, we divided a select number of tweets from the data and manually labeled the data as positive (+1), negative (-1), or neutral (0). Generally, a positive label means a positive sentiment, a negative label means a negative sentiment, and a neutral label means a neutral sentiment.

In the training and analysis phase we produced a script that would train a RandomForestClassifier from the sklearn library on the data we had labeled. At first, a challenge of this was that when a model was obtained, we were not able to input new data for predictions because of vector/matrix dimensionality issues. We produced a solution that used all the data for training the model and then got predictions off a subset of the portion of data that was not labeled. Eventually, we produced a prediction of 59.3% using a test size of 30% of the data. One interesting aspect of this project was that our model did not rate many tweets with a positive 1. For the predictions of over 700 tweets, only 2 were positive. An implication of this is that the classifier rated tweets with a 0 even if it had positive sentiment. A solution to this would be to have more strictness in labeling and be clearer about what constituted a 1 or a –1.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

Evaluation measures that we used include the classification report function from sklearn to output accuracy and other information about the predictions.

500 tweets were used for training and validation of our data model. Each tweet was labeled by hand, with each member of our group labeling 100 tweets.

Over the month of November, we collected tweets from all around the world. After these tweets were filtered and preprocessed we had a dataset that had 754 tweets from Michigan. Each of these tweets contained the attributes: created_at, id, text, url, and information on the user.

The project was implemented in python, HTML, and CSS.

4.2 Experimental Results

Code repository*: https://github.com/nathanmlh/Tweet-Sentiment-Analysis

Final model accuracy was 59.3% with a macro average of 71% and a weighted average of 66% with a test size of 30% of the labeling data. Average net sentiment for each day ranged from –0.59 to –0.81. These results coincide with our initial hypothesis that the results would generally be negative. The days we used the program to identify the sentiment of Michigan seem to have a very small deviation, with a standard deviation of only 0.07. The results can be seen on the website displayed in the video.

*please note we did not deploy the website and submitted a video instead. This is if you would like a closer look at our code.

4.3 Discussion

COVID-19 itself is a pandemic, affecting millions of lives across the globe so it is a profoundly serious and somber topic. Knowing this we expected the average sentiment to be more towards negative. Our data confirms this assumption with most days being around -0.7 average net sentiment. The sentiment scale we used meant that –1 would be most negative and 1 would be most positive. Regardless of the number tweets we gathered for a

day this average stayed relatively constant. It is important for this average to be constant so when there is a meaningful change, we can pinpoint a cause of that change. Example this could be when Pfizer gained FDA approval of their vaccine you would expect a larger surge to positive sentiment. A significant impact of the results is a much wider and more frequent way of the immediate unfiltered emotion from people then an example of something like a polling process.

5. CONCLUSIONS

Our model allows for the analysis and quantification of a large volume of tweets. With the broader idea of sentiment analysis using twitter data, we able to get the public reception on any topic or group of keywords. With our case dealing with COVID-19 its application can range from determining a rise or decrease in cases for a geographical area or it could be used for understanding public perception of the handling of COVID-19. For future work it would be extremely helpful to have a system in place that if there was a significant increase or decrease in average sentiment, we could store the top news stories for the day surrounding COVID-19, making it easier in the future to identify the publics change in sentiment.

6. REFERENCES

[1] Alessa, Ali, and Miad Faezipour. “Flu Outbreak Prediction Using Twitter Posts Classification and Linear Regression With Historical Centers for Disease Control and Prevention Reports: Prediction Framework Study.” JMIR public health and surveillance vol. 5,2 e12383. 25 Jun. 2019, doi:10.2196/12383

[2] Bycoffe, Aaron, et al. “How Americans View The Coronavirus Crisis And Trump's Response.” FiveThirtyEight, 12 Dec. 2020, projects.fivethirtyeight.com/coronavirus-polls/.

[3] Chakraborty, Amartya, and Sunanda Bose. “Around the world in 60 days: an exploratory study of impact of COVID-19 on online global news sentiment.” Journal of computational social science, 1-34. 21 Oct. 2020, doi:10.1007/s42001-020-00088-3

[4] Haddad, Ken. “Michigan Voters Back Whitmer's COVID Response, Disapprove of Trump's, Poll Shows.” ClickOnDetroit, WDIV, 28 Oct. 2020, www.clickondetroit.com/news/politics/2020/10/28/michigan-voters-back-whitmers-covid-response-disapprove-of-trumps-poll-shows/.

[5] “Michigan Data.” Coronavirus - Michigan Data, 2020, www.michigan.gov/coronavirus/0,9753,7-406-98163_98173---,00.html.

[6] Polgreen, Philip M et al. “Using internet searches for influenza surveillance.” Clinical infectious diseases : an official publication of the Infectious Diseases Society of America vol. 47,11 (2008): 1443-8. doi:10.1086/593098