Sentiment Analysis of Tweets During the COVID-19 Pandemic

Jordan Chervin
jcc874@nyu.edu

Silas Mann
manns03@nyu.edu

Chenjie Su
cs5998@nyu.edu

Yuan Zhao
yz6452@nyu.edu

Abstract

The public health crisis created by the COVID-19 pandemic has manifested a variety of emotions among the general public. In this assignment, we seek to classify emotions expressed through Twitter data using unsupervised clustering methods. We will evaluate the labels generated by this technique by comparing them with the labels assigned using the VADER technique. We will also investigate how trends in the proportion of positive and negative tweets compare with major milestones of the pandemic.

1 Introduction

As we are entering the ninth month of the Coronavirus pandemic in the United States, we have seen a great variety in the ways that people express their implicit and explicit attitudes towards the virus through their actions. For example, whether or not one wears a mask and social distances conveys two different attitudes. The 7 PM cheers demonstrated gratitude for our healthcare workers, and hoarding toilet paper demonstrated fear. Our primary objective for this project is to characterize sentiments toward the coronavirus pandemic expressed through tweets using an unsupervised learning technique - namely clustering - and our secondary objective is to correlate these sentiments with major milestones of the pandemic in the United States.

2 Related Work

Previous studies (Medford et al., 2020) have analyzed tweets by matching hashtags and keyword frequency to understand public sentiment toward transmission prevention, vaccination, and racial prejudice, and they have concluded that negative sentiment and emotion paralleled the incidence of cases of the COVID-19 outbreak. Kabir et al. (Kabir and Madria) analyzed real-time COVID-19 Twitter data to understand the psychology and behavior during the ongoing pandemic to provide insight into managing the economic and social crisis. Wang et al. (Wang et al., 2020) utilized a BERT model to classify sentiment categories and a TF-IDF model to summarize the topics of posts.

3 Data

Our dataset comes from a repository of COVID-19 tweet IDs actively maintained by the Panacea Lab from Georgia State University (Banda et al., 2020). This lab has accumulated over 200 million tweets daily since March 22. Due to the sheer volume of the dataset and computational limitations, we hydrated tweets on a weekly basis over twenty Fridays from March 27th to August 7th, which captures the first wave of the pandemic in America, as well as the mounting and often conflicting public concern.

For each day, the Panacea Lab saves ‘dailies’ including one file containing tweet IDs for all tweets and retweets and another file for a cleaned version with no retweets. We downloaded the cleaned version for each Friday and randomly sampled 50,000 tweet IDs from each. After gaining access to the Twitter API, we used the Social Media Mining Toolkit (SMMT), a set of Python tools also provided by the Panacea Lab, to request metadata for each of the 1,000,000 tweet IDs in our sampled dataset. The data returned by the API are provided in JSON format, which we then flattened into CSV format for use in analysis. While we initially experimented with pre-processing tools provided in the SMMT collection, we ultimately opted to retain all emoticons (typographic), emoji (pictographic), and URLs for use in analysis. After filtering to tweets in the English language, we had a corpus of over 500 thousand tweets with which to work.

4 Methods

The primary goal is to use unsupervised learning to cluster the tweets into positive and negative sentiment categories. In order to evaluate the labels applied by our unsupervised method, we needed to assign "ground-truth" labels to our tweets. A simplified flow chart of our approach is seen in Figure 1.

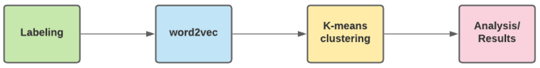

Figure 1: Methodology

We used the VADER Sentiment Analysis Module to label our tweets as positive, negative, or neutral (Gilbert and Hutto, 2014). VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically designed for social media, so it handles emoticons, emojis, and URLs. The sentiment lexicon is sensitive to both polarity and intensity, so in theory this tool could be generalized to other contexts or domains.

VADER generates four scores for each tweet: a positive, negative, neutral, and compound score. The positive, negative, and neutral scores sum to one for each tweet, so each score is a proportion of the extent to which the sentence exhibits sentiment, providing a multidimensional measure of sentiment. The compound score is a normalized value on a scale of -1 (extremely negative) to +1 (extremely positive).

Because the compound scores are continuous values, we used the recommended threshold of 0.05 to classify the labeled tweets into the three discrete categories: positive, negative, neutral. A compound score of 0 is precisely neutral.

We also used a recurrent neural network model to label our tweets with a more nuanced sentiment, one of Ekman's six basic emotions: joy, sadness, anger, fear, disgust, and surprise (Colnerîĉ and Demsar, 2018). A benefit of this labeling schema is that the RNN works at the character level, so an entire tweet can be passed without preprocessing the input. The two aforementioned labeling schemas are independent of each other.

Our unsupervised clustering is based on previous work of unsupervised sentiment analysis of a dataset of Polish reviews (Vashishtha and Susan, 2019). We start by initializing and training a word2vec model in order to generate word embeddings. The benefit of word2vec is that it produces short, dense vectors. A lower-dimensional embedding means that there are fewer weights for a model to learn in downstream processes, and a dense vector may generalize better and prevent overfitting. Finally, dense vectors better represent synonymy. We settled on the following parameters: window = 4 and size = 300.

Our next step is to take the word embedding vectors generated by word2vec and plug them into sklearn's K-means clustering to generate two cluster centroids: one positive and one negative.

Then, each word was assigned a weighted sentiment score based on the cluster to which it belonged, as well as its tfidf score. These scores were used to create two vectors for each sentence: one where each word was replaced with its weighted sentiment score, and one where each word was replaced with its tfidf score. The dot product of these two vectors indicates whether the overall sentiment of the tweet was positive or negative.

In an effort to more directly investigate sentiment towards the coronavirus, we manually constructed a list of discriminative hashtags for both the positive and negative sentiments. Our hypothesis was that if we can generate embeddings for the words in tweets containing these hashtags, we would force our centroids to better align along our dimension of interest. We included the hashtag symbol because a hashtag acts as a stamp on the tweet and can stand alone to convey sentiment. Hashtags representing positive tweets (i.e. "respect" the virus) include #StayHome, #FlattenTheCurve, #MaskUp, #WearADamnMask, and #Covidiot (only someone who is judging a #covidiot for being a #covidiot would use this hashtag, implying that the author of the tweet respects the coronavirus and the necessary interventions). Hashtags representing negative tweets (i.e. "disrespect" the virus) include #NoMask, #Scamdemic, #Plandemic, #ReopenAmerica, #EndTheLockdown, #ChinaVirus, and #FauciFraud.

This resulted in two lists of arrays of individual length=300, one for words from tweets with any of the positive hashtags, and one for words from tweets with any of the negative hashtags. We needed to flatten these lists to be of size 2 x 300, so we attempted two approaches: sum and average. These two methods did not produce significantly different results compared to each other; however, it did have an effect compared to our baseline clustering without seeding the centroids.

5 Results

The unsupervised clustering method described earlier was more effective on the dataset of Polish reviews than it was for our Twitter dataset. We believe this difference in performance is due to the natural tendency for reviews to trend more positive or more negative in a single dimension, whereas there is more variability and neutral sentiment within Twitter data. Additionally, neutral reviews were removed from the Polish dataset to enhance polarity during the training step.

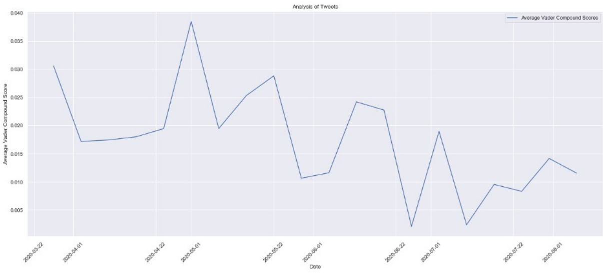

Figure 2: Average VADER Compound Score Over Time

The plot shown in Figure 2 depicts the average VADER compound score sentiment over time. The day where the average compound score was most positive was on May 1. On April 29, the NIH reported positive data from clinical trials of Remdesivir (Beigel et al., 2020). There was another peak on May 22, a day after the United States and AstraZeneca formed a vaccine deal (Koirala et al., 2020). The day where the average compound score was the most negative was after June 22. On this day, a study was published that suggested 80% of cases in March went undetected (Aspelund et al., 2020). Later that week, the White House Coronavirus Task Force held a briefing to address the rising number of cases in the South (Briefing, 2020). Mid-July also showed a low average compound score, and on July 16, a new record number of daily COVID-19 cases were reported (WHO, 2020).

Notably, by visual inspection, the word cloud in Figure 3 for 'disrespect' appears to include more angry sentiment and references to the president, whereas the 'respect' word cloud in Figure 4 appears to contain more news and news organizations. We believe that news articles may be using encouraging hashtags to tag articles indiscriminately, causing our positive sentiment clustering to erroneously skew towards news posts. Figure 5 shows the overall clustering of positive and negative tweets.

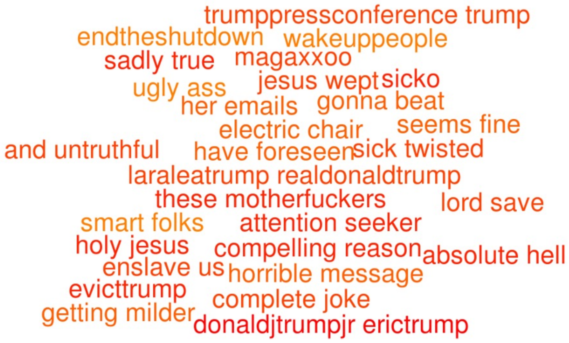

Figure 3: 'disrespect' word cloud generated from tweets labeled 'negative'

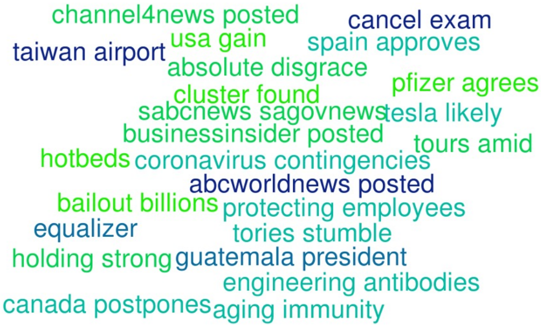

Figure 4: 'respect' word cloud generated from tweets labeled 'positive'

6 Discussion and Further Investigation

There are a number of modifications we would suggest for future work on this topic:

1. Because VADER is built to handle emojis and URLs, we retained them in our corpus. However, we hypothesize that URLs may be oversaturating our dataset. Since URLs do not carry any sentiment in and of themselves, it would be interesting to compare our results without this noise.

2. Due to the volume of tweets and computational limitations, we chose to use tweets from regular intervals: once weekly over twenty weeks. However, we think that this does not capture the granularity of any observable trends or changes in sentiment that may be better captured by using a continuous stream of tweets.

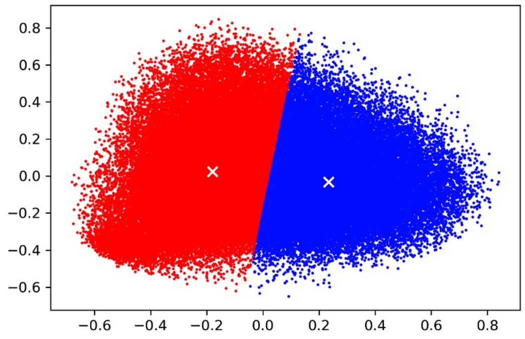

Figure 5: Cluster of Positive and Negative Tweets

3. It is easier to compare different clustering methodologies and approaches when the binary classification is positive or negative. However, we do not think this is the best way to represent the sentiments of tweets from the pandemic, and therefore, using VADER as our ground truth labeling may be inherently flawed. We posit that binary labels such as “respect” and “disrespect” may be more adequate classifications for tweets related to the pandemic. Perhaps this labeling schema would also help mitigate the effect of other global events unrelated to the pandemic which may be saturating the dataset and affecting the sentiment trends.

4. It would be an enlightening investigation to filter tweets in English by city to compare trends in sentiment with different regional milestones and timelines. For example, case numbers were rising in the South while they remained stable in the Northeast, and Seattle was a hotspot just before New York City was. We expect that there would be significant variation in the proportion of positive and negative tweets, especially in regard to certain non-therapeutic interventions like wearing a mask and lockdown. This is because the virus was politicized, and implementing these measures were left up to the discretion of the individual states. We would also be interested to see if there is a correlation between trends in tweet sentiment and regional case count.

5. We would like to reattempt our K-means clustering method using 6 clusters to correspond with each of Ekman’s 6 basic emotions.

6. Instead of using word2vec to generate our initial word embeddings, we could use ELMo or BERT to construct contextual word embeddings. The benefit of contextual word embeddings is that different representations are learned for polysemous words.

7 Appendix

We also applied supervised sentiment analysis over time and compared with unsupervised sentiment analysis to assess if the two correspond. Figure 6 shows the proportion of all VADER-labeled positive, negative, and neutral English tweets over time, and Figure 7 is the percentage of positive and negative English tweets over time using our unsupervised labeling method. Although there are differences, the two have similar trends. We identified a similar pattern when we filtered all tweets originating rom New York City as shown in Figures 8 and 9. The vertical lines in Figure 9 represent the phased re-opening in New York City.

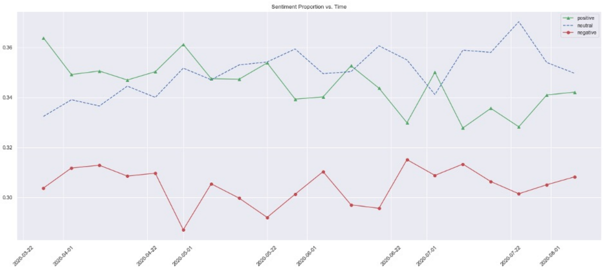

Figure 6: Proportion of VADER-labeled Positive, Negative, and Neutral English Tweets Over Time

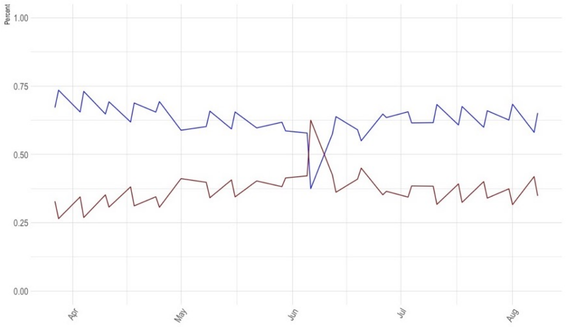

Figure 7: Percentage of Positive and Negative English Tweets Using Unsupervised Labeling

References

Karl Aspelund, Michael Droste, James H Stock, and Christopher D Walker. 2020. Identification and es-

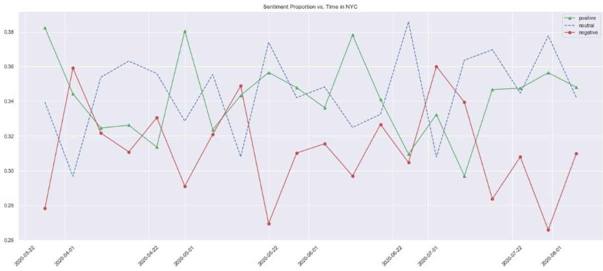

Figure 8: VADER-labeled English Tweets, New York City

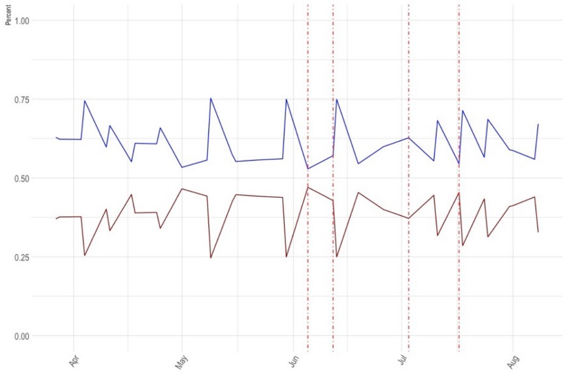

Figure 9: New York City Tweets with Unsupervised Labeling Method

timation of undetected covid-19 cases using testing data from iceland. *NBER Working Paper*, (w27528).

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalin, and Gerardo Chowell. 2020. A large-scale covid-19 twitter chatter dataset for open scientific research – an international collaboration.

John H Beigel, Kay M Tomashek, Lori E Dodd, Aneesh K Mehta, Barry S Zingman, Andre C Kalil, Elizabeth Hohmann, Helen Y Chu, Annie Luetkemeyer, Susan Kline, et al. 2020. Remdesivir for the treatment of covid-19—preliminary report. *The New England journal of medicine*.

James S Brady Press Briefing. 2020. Remarks by president trump, vice president pence, and members of the coronavirus task force in press briefing.

Niko Colneriĉ and Janez Demsar. 2018. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE transactions on affective computing*.

CHE Gilbert and Erric Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, volume 81, page 82.

Md Yasin Kabir and Sanjay Madria. Coronavis: A real-time covid-19 tweets data analyzer.

Archana Koirala, Ye Jin Joo, Ameneh Khatami, Clayton Chiu, and Philip N Britton. 2020. Vaccines for covid-19: The current state of play. *Paediatric respiratory reviews*, 35:43–49.

Richard J Medford, Sameh N Saleh, Andrew Sumarsono, Trish M Perl, and Christoph U Lehmann. 2020. An" infodemic": Leveraging high-volume twitter data to understand public sentiment for the covid-19 outbreak. *medRxiv*.

Srishti Vashishtha and Seba Susan. 2019. Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Systems with Applications*, 138:112834.

Tianyi Wang, Ke Lu, Kam Pui Chow, and Qing Zhu. 2020. Covid-19 sensing: Negative sentiment analysis on social media in china via bert model. *Ieee Access*, 8:138162–138169.

WHO. 2020. Coronavirus disease (covid-19): weekly epidemiological, update 1.