




2020

SAC Deloitte NLP COVID-19 Consortium Report

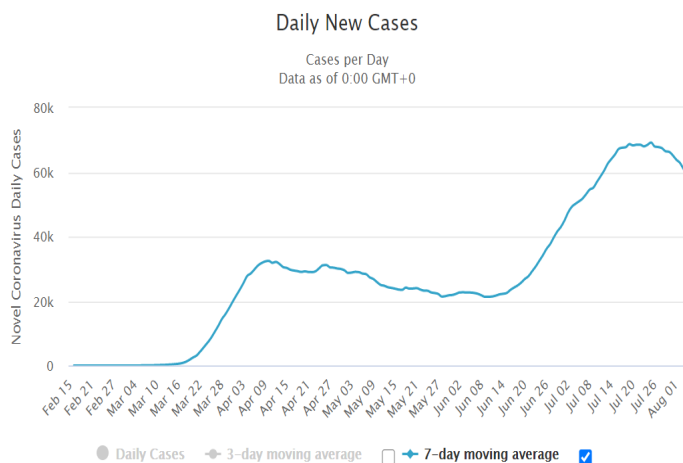
Aishwarya Bhangale
Jiahong Yu
Zexi Luo
Xinhua Tang

“Nothing in life is to be feared; it is only to be understood. Now is the time to understand more, so that we may fear less.” — Marie Curie



1. Introduction

4.9 million. With 1.4% of the United States population infected with the deadliest disease of our times, COVID - 19, people are eager to learn from this tragic pandemic. This report aims to highlight insights into what the average twitter user thinks about the United States' response to this pandemic.



With cases on the rise and White House top disease experts, Dr. Anthony Fauci and Dr. Deborah Birx expressed the epidemic as extraordinarily widespread across the rural and urban areas of the United States. The symptoms of the disease are as common as an upset stomach or mild fatigue, two weeks for a person to know they have a highly contagious, airborne and mutable disease is simply too long a period on a planet housing

over 7.8 billion humans.

As student participants of the Smith Deloitte SAC COVID-19 Consortium Workshop, we focused our research on the response phase of the U.S. Government from March-May 2020. Before we begin, we would like to thank Tulio Tablada(Data Scientist @Deloitte), Connor Welch(Data Scientist @ Deloitte), Maxwell Cutler(TA) and Xichen Wei(T.A.) for this humble opportunity of performing Natural Language Processing and Analytics on Twitter data collected for insight gathering purposes.

This dataset includes variables such as I.D. (the unique identification series for each user), Location (from where the tweet was tweeted), and Processed Tweet (text data with limited cleaning). With 20,620,442 records, our first task was to clean the data for a smooth and scientific analysis project.

2. Data Cleaning

Upon analysis, we discovered that there were no stop words in the tweets provided. However, it was required to deal with swear terms, weblinks and replace emoticons with their textual form. While replacing emojis with their original text form, we removed the "_" symbol between the text words that make up the emoji.

Our approach to cleaning 20 million tweets was to use a combination of the GPU/TPU ability of Google Colabs and divide the primary dataset into five smaller datasets of 3 G.B. in size each, so that each of us could use Colab in parallel, speeding the analysis up. These cleaned tweets were saved in a new column, "Further_Cleaned_Text".

Furthermore, we cleaned the "Location" column by subsetting and mapping each Location to its full name. For a few of these, we had multiple duplicate values ("ny", "new york" for the state of New York). We made sure we got all the unique values for that particular state and mapped these to only one common state name. Now that the state information was standardized, we could utilize this information and make statewide analysis.

3. Text Processing & Initial Data Analysis

3.1 Tokenizing

After cleaning the dataset, we tokenized the documents with CountVectorizer from Sci-kit learn, and, at the same time, threw away certain characters, such as punctuation and residual stop words. After that, we conducted both stemming and lemmatization to prepare the data for a better analysis and modeling result.

3.2 Word Cloud

Our team imported the word cloud function from the 'wordcloud' package. We then managed to create a basic word cloud for each subset of the data(5 million records each) to get a brief idea of what most of these tweets are about. Unsurprisingly, we found that the most important words (Fig 1) are the common pandemic-related

keywords such as 'COVID', 'corona', 'virus', etc. These words were expected to be universal, yet not valuable, in terms of an in-depth analysis.



(Fig 1: word cloud from subset 1)

To excavate more valuable keywords from the tweets, we excluded a list of common words that were believed to be mundane. Then, we remade the word cloud and uncovered some unexpected keywords from the tweet word clouds.

A fascinating insight here is that "loudly crying" , "nursing home" , " skin tone" , "chinese" , and "black" are words that frequently occur in tweets.

After referring to the actual text data that contains these terms, we came up with the following interpretations of why these words may have been used so frequently :

- (1) "Skin stone", "loudly crying", and "nurse home" are three emojis conquering all five word clouds. On the one hand, we can roughly draw a user graph for twitter that most users are young people, who. On the other hand, the high frequency of "skin stone" illustrates that when people apply emoji, they care about each emoji's skin tones.
- (2) Words like "call chinese", "call black", "black people" indicate that there might be a wave of racial separation from March to May 2020. Many tweets of these months were mentioning races in a surprisingly high frequency. In particular, Black and Asian people were referred to the most during the COVID-19 outbreak(Horowitz & Tamir, 2020). Racism has always been a tragic issue in human history. During this historical event of a global pandemic, such a problem repeatedly appeared, testing people around the world of their moral values and

beliefs in equality. Coronavirus is not specifically for one or several races, but how it began spreading, how it transferred, and how it became a worldwide threat might build boundaries between different races. The developed nationalism can be easily seen during a pandemic period among tweets. For one thing, people express their gratitude for medical personnel. For another thing, people refuse to accept they might be the original cause of large spread.

4. Machine Learning Application

4.1. TF-IDF

One of the simple methods to capture features is the term frequency-inverse document frequency, also known as TF-IDF. To evaluate how important a word is to a tweet, we adjusted each term in each document using the TF-IDF method. Given that the TF-IDF adjustment can be seen as a standard procedure that improves modeling and analysis processes, we calculated TF-IDF based on each of the five subsets divided from the 20 million tweets.

4.2. Clustering

4.2.1 K-means Clustering

Given the assumption that there must be several mainstream opinions among the tweets, we believed that several clusters appear in our dataset, and we just don't know how many and what they are. Therefore, we chose K-means, a classic and one of the most popular clustering algorithms, to apply our first try as the machine learning algorithm.

First and foremost, we had to figure out the number of clusters, K , in our dataset. The methods of determining the optimal K were to try different values onto the dataset, set a performance measure, and compare the performances of different K s. This "try and true" methodology requires enormous time and computing power since the clustering would run once for each K , mainly when the entire dataset contains over 20 million records of text data. Given the situation, we decided to try to find the optimal K on a

small portion of sample data, using a less computationally power-consuming method than the elbow method mentioned in the workshops.

We ran a Silhouette Analysis on 1 million records, 5% of all data. Silhouette analysis decides the optimal K quantitatively, outputting a Silhouette score, between -1 and 1, for each value of K. It's believed to be a bit faster than the elbow method because there is no need to make a graph, and it's more scientific since the elbow method involves subjective judgment of the line graph's shape while Silhouette doesn't. The decision of K is made by comparing different Silhouette scores and picking the highest one, among the ones scored over 0.5.

After trying out several possible Ks ranging from 2 to 7, we found that all of the Ks scored over 0, yet far less than 0.5. Either there were no apparent clusters in our dataset, or that K-means doesn't fit on our dataset well enough.

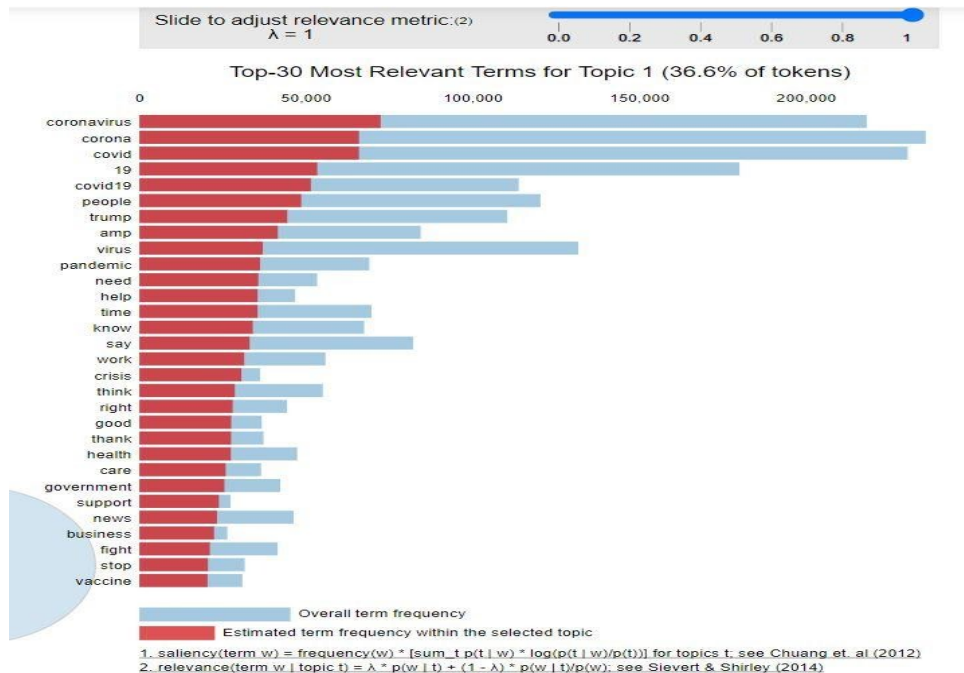
With the results from Silhouette analysis, we decided to move on to a second method, Latent Dirichlet Allocation(LDA).

4.2.2 Latent Dirichlet Allocation(LDA)

Similar to using K-means clustering, we still need to determine the number of topics using LDA. By repeating the strategy as before, we utilized a relatively small portion of data to determine the most critical parameter in our final model, the number of topics. The evaluation score this time is the log-likelihood. Same as before, the higher the log-likelihood, the better. With sklearn's grid search function, the calculation of log-likelihoods and their comparison happens behind the scene. The output shows that the best `n_components`, number of topics, appear to be 3. This was conducted on a unigram term-document matrix adjusted by the TF-IDF method.

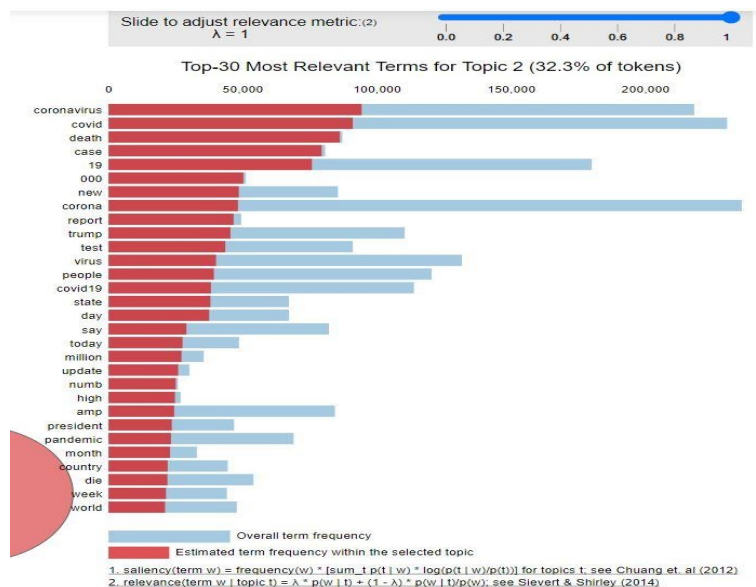
We used the `LatentDirichletAllocation` class from `sklearn.decomposition` library to carry out this part of our analysis. With the optimal value of `k=3`, we used this to segregate the tweets into three topics, 0,1 and 2. We added this topic number allocation as a new column "LDA_Topic" to the original data frame upon having done this.

We managed to visualize the most critical terms in each LDA topics with `pyLDAvis`:



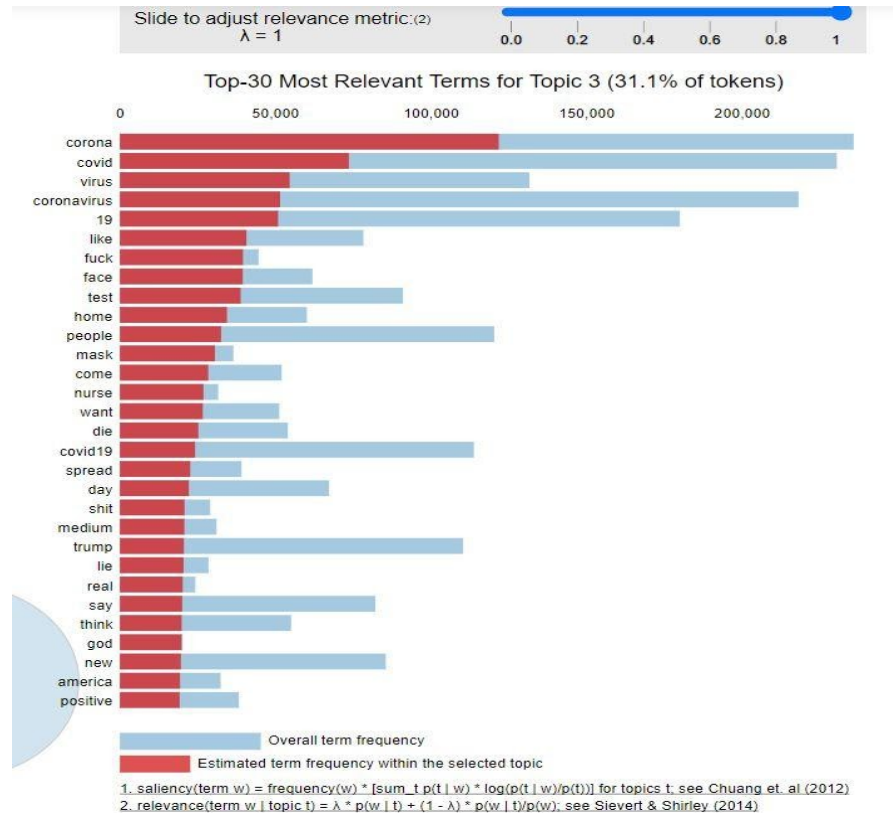
(Fig 2: Top 30 most relevant terms for topic 1)

Cluster 1 terms (Fig 2) are more neutral, just like news reports, regarding how many patients and death each week, etc.



(Fig 3:Top 30 most relevant terms for topic 2)

Cluster 2 (Fig 3) could be defined as positive encouragement due to the importance of 'help' 'mask', 'love', etc.



(Fig 4:Top 30 most relevant terms for topic 3)

Cluster 3 (Fig 4) could be defined as negative complaints, people complaining 'trump', 'wrong', 'china', 'lie' etc.

4.3 Sentiment Analysis

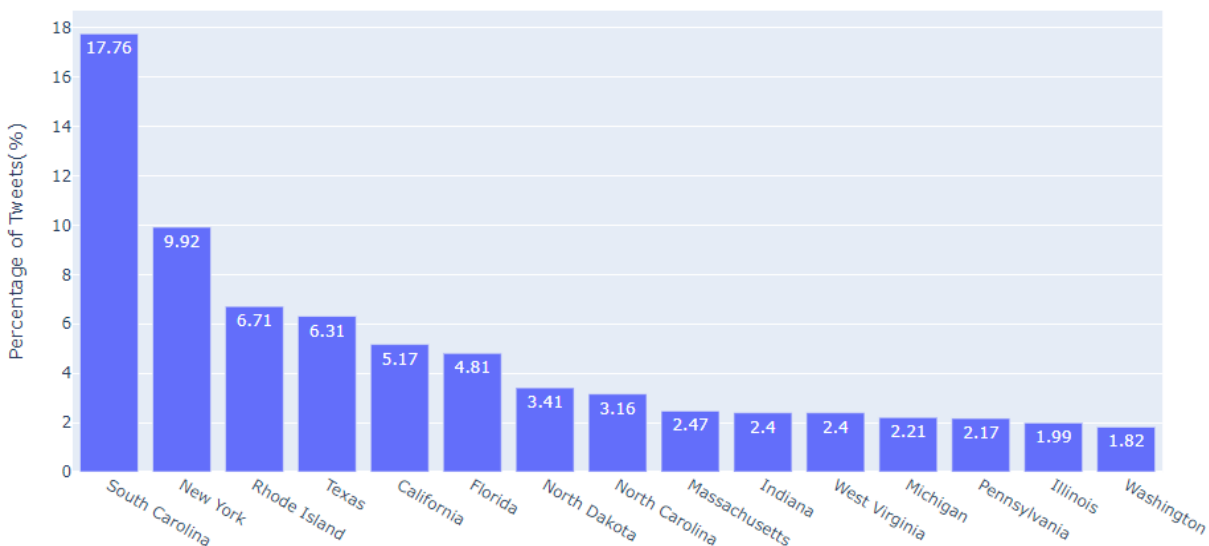
Our group then moved to sentiment analysis, hoping to find more about people's attitudes towards the pandemic.

We chose the "vaderSentiment" package and utilized "SentimentIntensityAnalyzer". Since our documents were stored as float objects in the data frame, we had to convert each value from the 'further_processed_tweet' column to strings before doing polarity score analysis. We then saved the polarity sentiment score in a new column for each corresponding tweet. Tweets that scored over or at 0.05 were classified as positive, neutral if they are between -0.05 and 0.05, and negative for ones below -0.05.

Our group then proceeded to unstack the sentiment scoring table by putting each sentiment category ('pos', 'neg', 'neu') into a separate column. Other than these counts, we also calculated the percentage of each sentiment category within the state(using the tweet counts of that sentiment category divided by the total number of tweets in that state). Considering the corona pandemic is a national crisis, we also calculated the percentage of one state's pos/neg/neu tweets in the federal pos/neg/new overall counts. Surprisingly, we found some unusual patterns in the derived percentages.

One of the patterns that caught our attention was that South Carolina had a significantly high percentage of positive tweets between March and May (Fig 5). 17.76% of overall positive tweets came from South Carolina, though it only had 5 million people. According to worldometers, South Carolina had a constant number of cases during Late March - Mid April(wordometer, 2020). It is reasonable to believe that South Carolina residents had faith in the S.C. government's response and chose not to panic. Till the day of the report, South Carolina's total number of COVID cases is 103,909, much lower than the average incidence rate of the U.S.

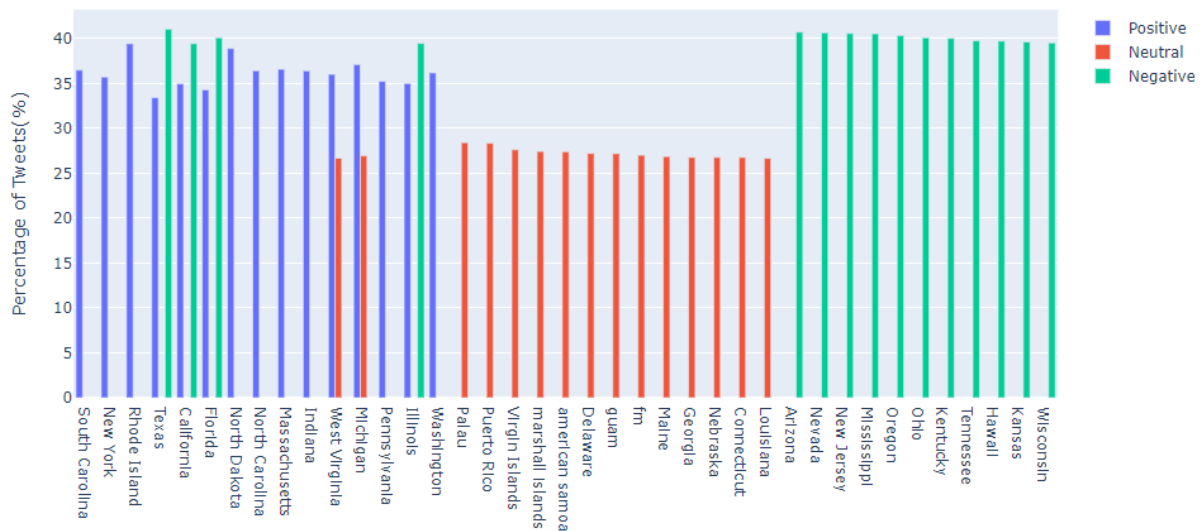
Top 15 states within highest % of Positive Tweets Overall



South Carolina had constant number of cases during this time(Late March - Mid April)

(Fig 5 : percentage of tweets for each state)

Our team also created a chart to show the top 15 states (Fig 6) that have the highest percentage in each sentiment category within that state:



(Fig 6: sentiment category for each state)

From the chart, we could see that the states with a great economy (New York, California, Texas, etc.) tended to be more positive than those with a small/medium economy. We believed this phenomenon was due to those large states having the most excellent medical services, infrastructures, and prosperity in the U.S. Thus, the residents of the large states, in general, were more confident than those of the small/medium states.

However, having an excellent economy might not be a good thing in this pandemic. Many rich states had a large population and were also the transits of the U.S. aviation/water transportation network, which means they were even more attractive to asymptomatic carriers. According to [augustahealth](#), the critical factor that made this coronavirus so influential was because an infected patient might not show any symptoms as long as fourteen days ([augustahealth,2020](#)). In some large economy states like Illinois, Texas, California, and Florida, the percentage of negative tweets were also top 15 and even higher than the percentage of positive tweets within these states.

5. Conclusion

The emoji usage of people showed that people cared about their emojis' skin tone and expressed their sympathy to the victims and medical personnel of the pandemic most of the time. However, there is also an observable wave of racism towards minorities in the tweets between March and May 2020.

For the large economy states, residents tended to have more extreme views on this pandemic than the others. They were proud of their states' medical services while worrying about many asymptomatic carriers at the same time. Also, states with a relatively constant number of COVID cases between March and May 2020 were more positive than those who did not.

According to our Latent Dirichlet Allocation, all the tweets could be classified into three topics: one for neutral news reporting, one for positive encouragement, and the other for complaint and blame.

References

1. Ruiz, Neil, G., Horowitz, Julianna, M., Tamir, Christine. July 1st, 2020, "Many Black and Asian Americans Say They Have Experienced Discrimination Amid the COVID-19 Outbreak", <https://www.pewsocialtrends.org/2020/07/01/many-black-and-asian-americans-say-they-have-experienced-discrimination-amid-the-covid-19-outbreak/>
2. Wordometers, 2020, <https://www.worldometers.info/coronavirus/usa/south-carolina/>
3. AugustaHealth, 2020, <https://www.augustahealth.com/health-focused/covid-19-asymptomatic-carriers-and-antibody-tests#:~:text=%22An%20asymptomatic%20carrier%20is%20someone,Dr.>
4. Weiß, Martin, et al. "Age-Related Differences in Emoji Evaluation." Experimental Aging Research (2020): 1-17.