Covid-19 Twitter Analysis Final Project

By Noah Skole, Steven Caione, and Nathan Minkwitz

Background

- Initially we wanted to analyze what twitter users have the most influence on climate change by promoting awareness on their twitter account.
    - Politicians had enough Climate Change data
    - Athletes, Celebrities, CEOs did not
- Instead, we decided to analyze what twitter users have the most impact on spreading awareness around the Covid-19 Pandemic

UN Sustainable Development Goal

- With our original project goals to analyze the influence of twitter users around climate change, we planned on connecting our analysis to the UN Sustainable Development Goal of Climate Action.
- However with our new objectives to analyze the influence of twitter users around the Covid-19 Pandemic, our analysis and findings are relating to the UN Sustainable Development Goal of Good Health and Well-Being

Project Questions

1. Which group has the most influence over promoting good health and well being throughout the Pandemic?
2. Who are the top performers or influencers of the 4 groups?
3. What kind of language are these people using to get their points across?
4. As the pandemic progresses how does the sentiment of tweets change?
5. Is there a correlation between the sentiment and covid cases and deaths over the duration of the pandemic?

Capture

- We defined four different groups of twitter users to collect data from: Politicians, Celebrities, CEO’s, and Athletes
- We manually crawled each twitter users profile and the loaded each user into a json file individually.
- Raw Data consisted of just over 60,000 Tweets
- Cleaned Data consisted of just over 45,500 tweets in the time frame of COVID-19 pandemic (further cleaning was done when data was processed)
    - Further categorizing into Covid-19 related tweets made a dataset of 5,173
- Second Dataset was COVID-19 data from Johns Hopkins University, obtained from Kaggle
- Within the dataset we analyzed United States confirmed cases and United States deaths

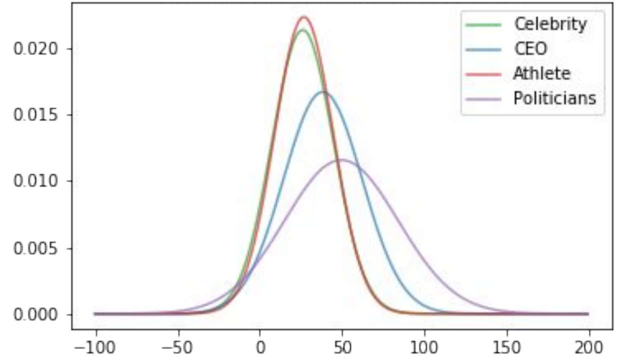
Process

- We removed outliers pertaining to tweets that occurred prior to the Pandemic
- We defined a series of keywords used to categorize each tweet as either covid or non-covid related
- We used grouping functionality to first visualize and compare the tweet, retweet, and favorite counts of covid related tweets among the 4 groups
- Then for each group we visualized the the top 20 tweet counts to identify the most active users
- In addition, for each group we visualized the top 5 most favorited and retweeted tweets of users to identify which users might influence the most people.
- Utilized Textblob to assist in creating Polarity, Subjectivity, and finally a Sentiment column within the Data Frame

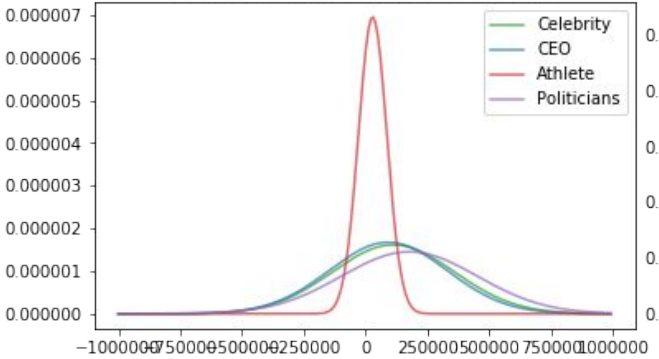
Process Continued

- Means and Standard Deviations were calculated for each group’s tweet count, retweet count, and favorite count
- Normal Distributions were formed of each group’s tweet count, retweet count, and favorite count

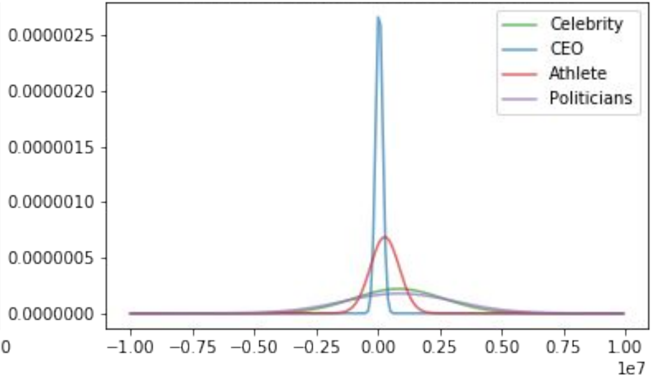
Tweet Count Normal Distribution



Retweet Count Normal Distribution



Favorite Count Normal Distribution



Means and Standard Deviations

Celebrity Means : Tweet Count, Retweet Count, Favorite Count: **(26.093023255813954, 111698.23255813954, 826876.0465116279)**

Celebrity Stds : Tweet Count, Retweet Count, Favorite Count: **(18.695318937266432, 246781.3208954377, 1817521.1964478532)**

CEO Means : Tweet Count, Retweet Count, Favorite Count: **(38.31111111111111, 90797.46666666666, 44066.2)**

CEO Stds : Tweet Count, Retweet Count, Favorite Count: **(23.91635254782633, 238245.66235501922, 142991.59428944194)**

Athlete Means : Tweet Count, Retweet Count, Favorite Count: **(26.86842105263158, 27876.815789473683, 268278.1842105263)**

Athlete Stds : Tweet Count, Retweet Count, Favorite Count: **(17.873687519581765, 57304.06670152405, 581898.7865591568)**

Politician Means : Tweet Count, Retweet Count, Favorite Count: **(49.59016393442623, 176370.80327868852, 846543.475409836)**

Politician Stds : Tweet Count, Retweet Count, Favorite Count: **(34.50670323728436, 274530.5576178372, 2242398.0009241863)**

Analyze - Probability

- We calculated the probability of a covid related tweet among all of our twitter users gathered by dividing the total number of covid related tweets under the defined keywords by the total number of tweets crawled
- We also performed a similar calculation on each of the 4 groups and divided them by the total number of tweets crawled

probability that a tweet is covid related or not covid related of all groups/users: 0.11496498584066911

probability that a tweet is covid related for CEO group: 0.024762364718020766

probability that a tweet is covid related for Celebrity group: 0.017122911773099467

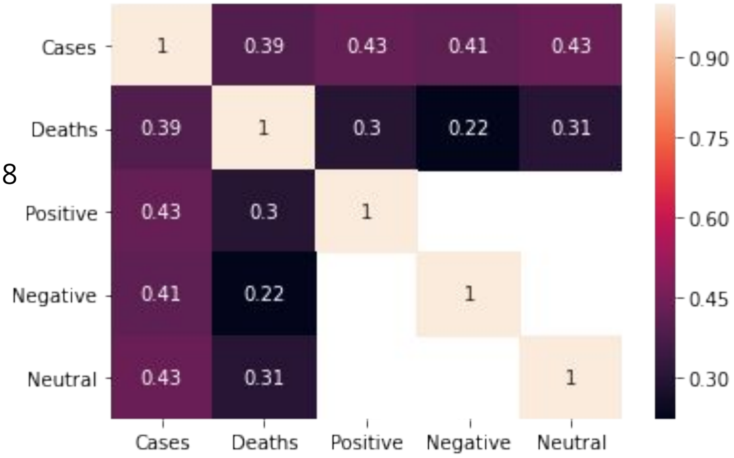
probability that a tweet is covid related for Athlete group: 0.013325137751629969

probability that a tweet is covid related for Politician group: 0.05975457159791891

Analyze - Correlation

- We computed the correlations for COVID-19 cases with all of the sentiments positive, negative, and neutral
- We also computed COVID-19 deaths with the positive, negative, and neutral sentiments
- Correlation Calculations:
    - **Positive** Sentiment & COVID-19 **Cases**: 0.4285877834628595
    - **Negative** Sentiment & COVID-19 **Cases:** 0.410780527886592
    - **Neutral** Sentiment & COVID-19 **Cases**: 0.4332843631006731
    - **Positive** Sentiment & COVID-19 **Deaths**: 0.3019833140206021
    - **Negative** Sentiment & COVID-19 **Deaths**: 0.22232472667053338
    - **Neutral** Sentiment & COVID-19 **Deaths**: 0.30690298361029017

Correlation Heat Map:
COVID-19 vs Positive, Negative, Neutral Sentiment



Analyze - Hypothesis

- **H0**: politicians group tweet, retweet, and favorite count means are all less than the tweet, retweet, and favorite count means of the other three groups
- **H1:** politicians group tweet, retweet, and favorite count means are all greater than the tweet, retweet, and favorite count means of the other three groups

Analyze - Hypothesis

Tweet Count			
Group 1	**Group 2**	**T-Statistic**	**P-Value**
Politicians	CEO	1.8832	0.0625
Politicians	Celebrity	4.0611	**9.60206141790e-05**
Politicians	Athlete	3.7527	**0.000298032**

Politicians have larger mean than Celebrities → **Reject Null Hypothesis**

Politicians have larger mean than Athletes → **Reject Null Hypothesis**

Politicians do not have statistically significant larger mean than CEO → Accept Null Hypothesis

Analyze - Hypothesis

Retweet Count			
Group 1	**Group 2**	**T-Statistic**	**P-Value**
Politicians	CEO	1.6762	0.0967
Politicians	Celebrity	1.2328	0.2205
Politicians	Athlete	3.2841	**0.0014**

Politicians have larger mean than Athletes → **Reject Null Hypothesis**

Politicians do not have statistically significant larger mean than CEOs → Accept Null Hypothesis

Politicians do not have statistically significant larger mean than Celebrity → Accept Null Hypothesis

Analyze - Hypothesis

Favorite Count

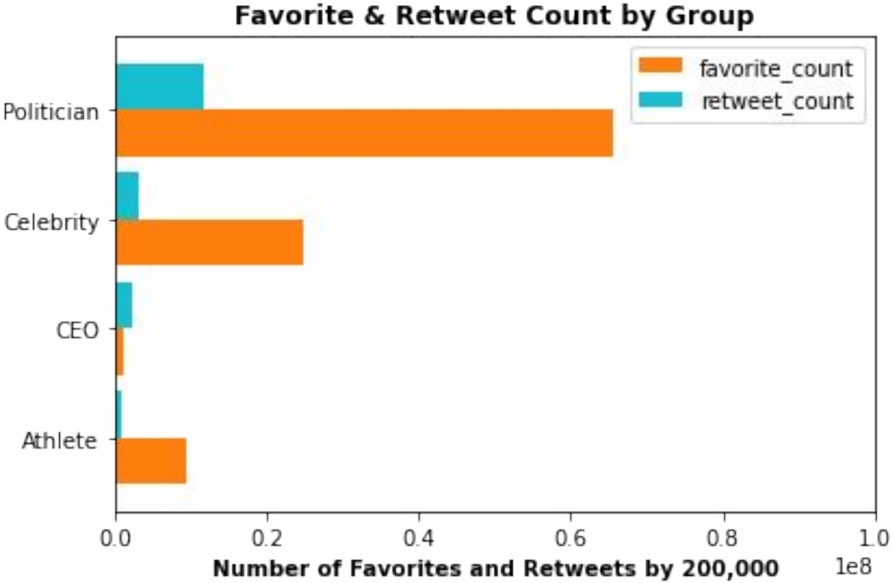
Group 1	**Group 2**	**T-Statistic**	**P-Value**
Politicians	CEO	2.39404	0.01845
Politicians	Celebrity	0.04753	0.96218
Politicians	Athlete	1.55464	0.12329

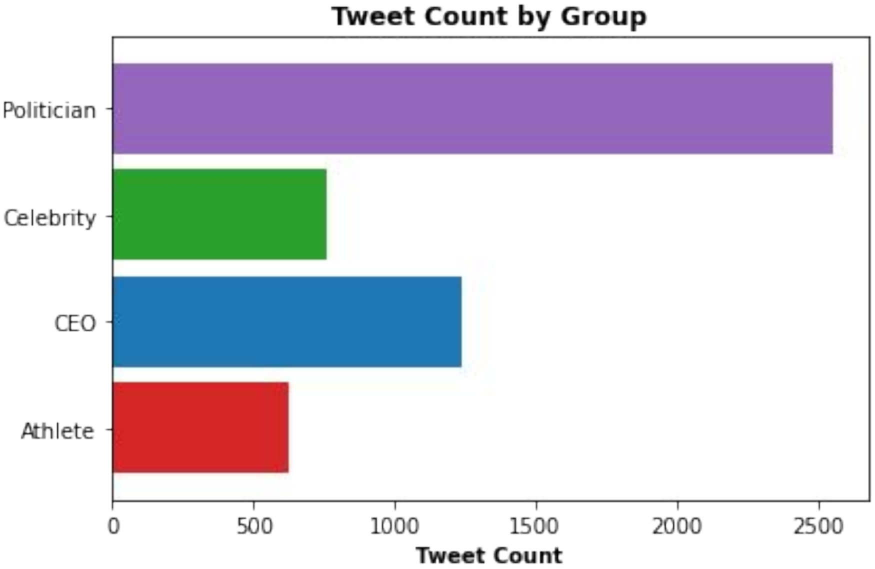
Politicians do not have statistically significant larger mean than CEO → Accept Null Hypothesis

Politicians do not have statistically significant larger mean than Celebrity → Accept Null Hypothesis

Politicians do not have statistically significant larger mean than CEO → Accept Null Hypothesis

Communicate

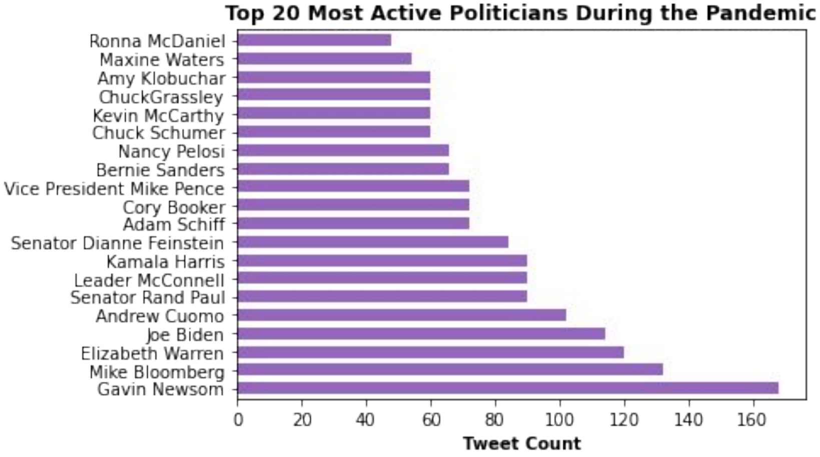


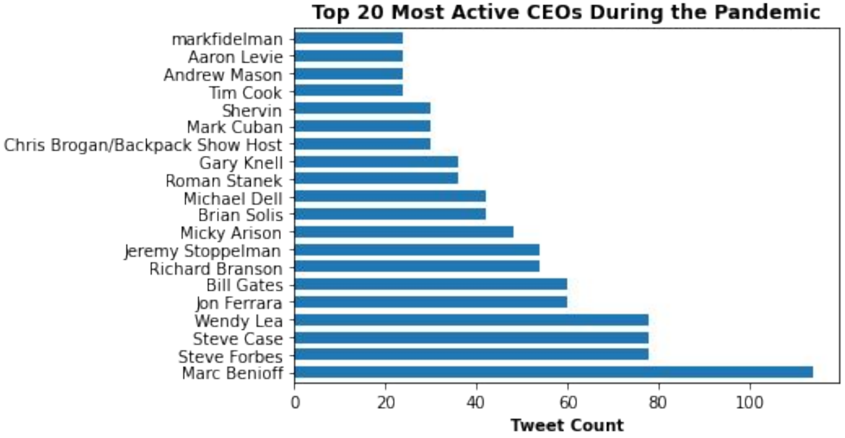


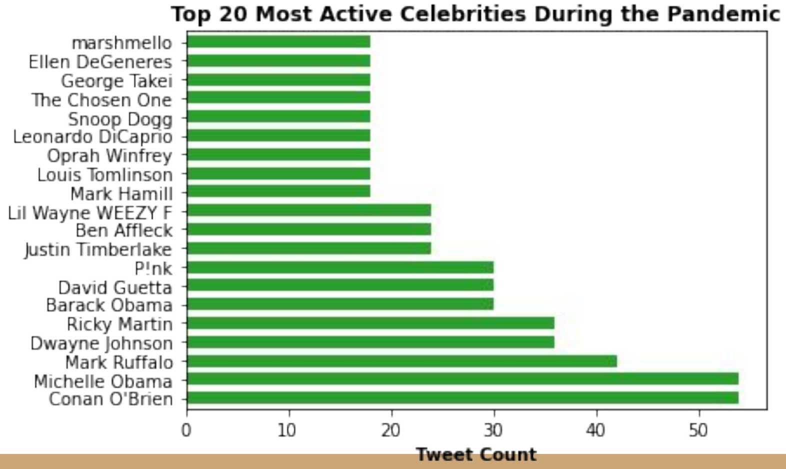
Visualization to show the distribution of COVID related favorite and retweet counts by group

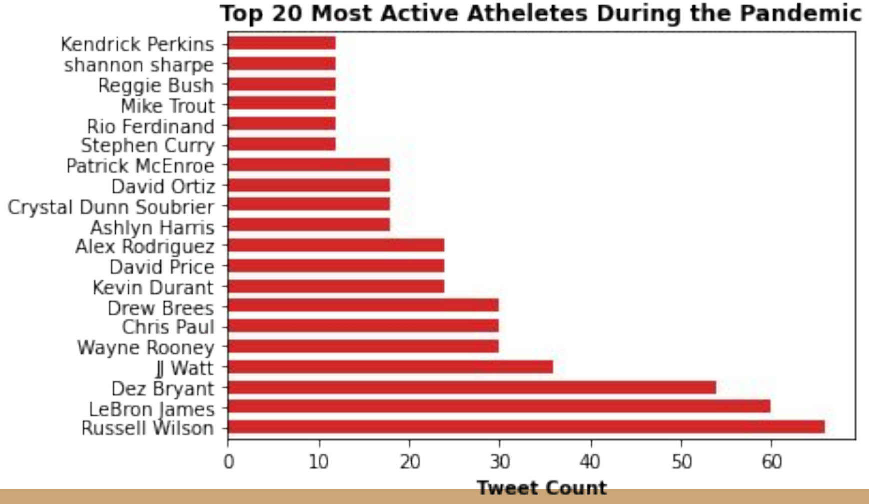
Visualization of COVID related tweets by tweet count to show which groups tweet tweet most about the pandemic

Communicate

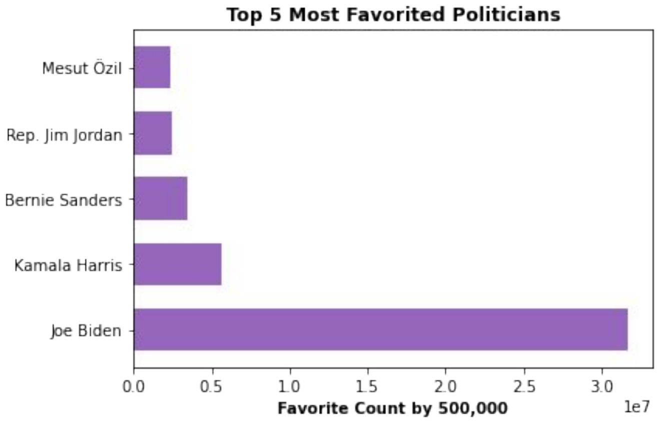

Top 20 Most Active Politicians During the Pandemic

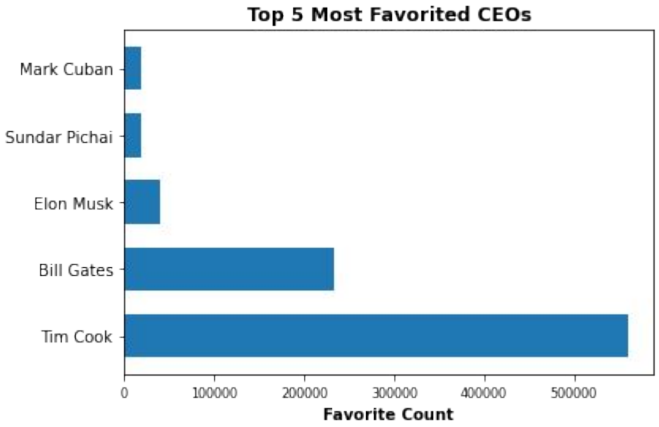

Top 20 Most Active CEOs During the Pandemic

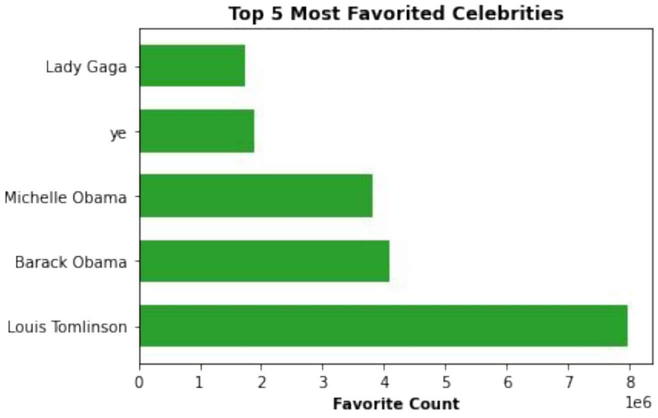

Top 20 Most Active Celebrities During the Pandemic

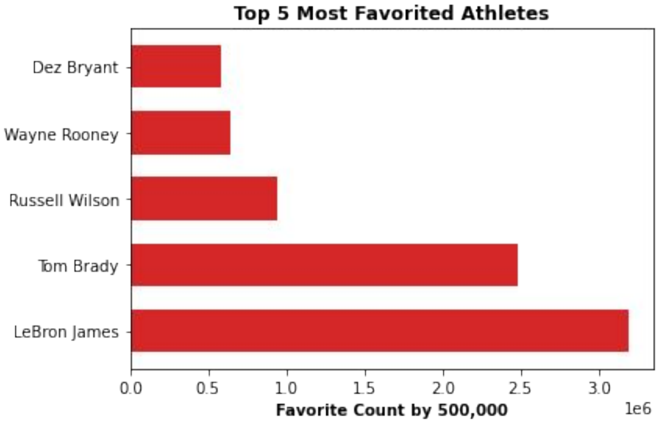

Top 20 Most Active Atheletes During the Pandemic

Communicate


Top 5 Most Favorited Politicians

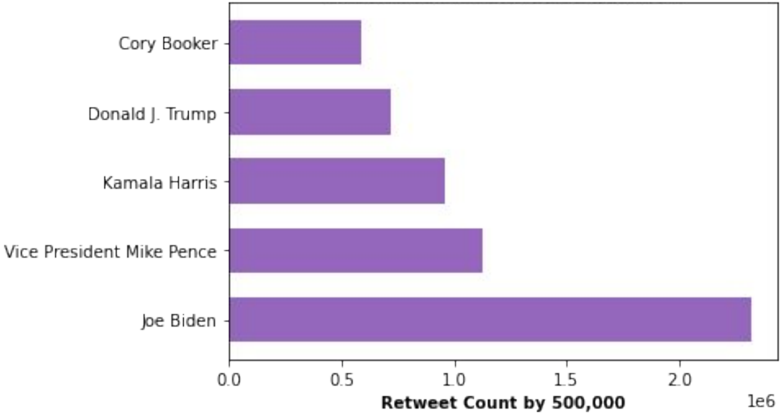

Top 5 Most Favorited CEOs


Top 5 Most Favorited Celebrities

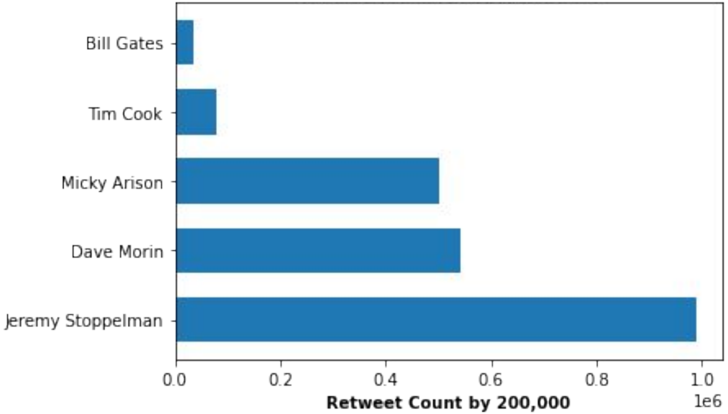

Top 5 Most Favorited Athletes

Communicate

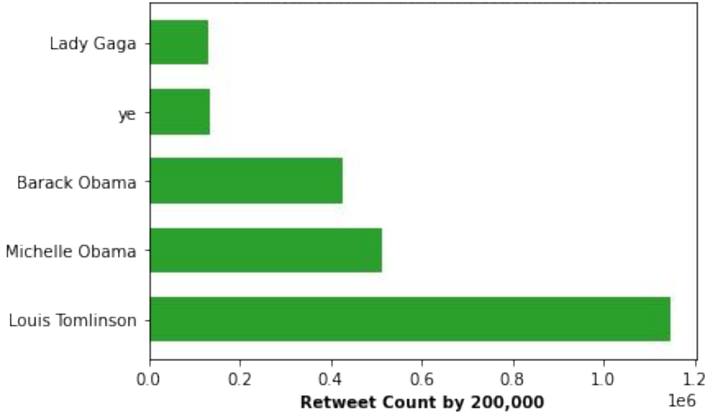
Top 5 Most Retweeted Politicians



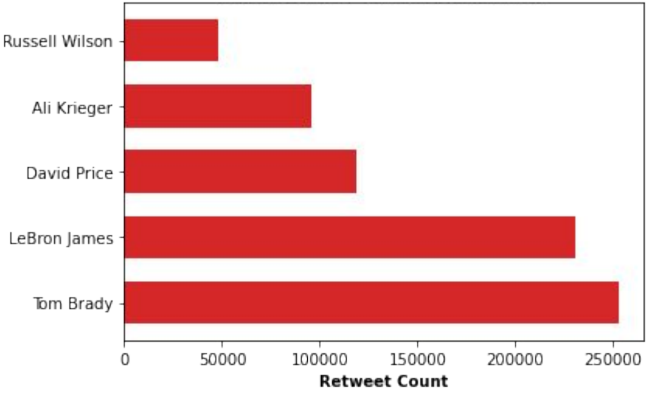
Top 5 Most Retweeted CEOs



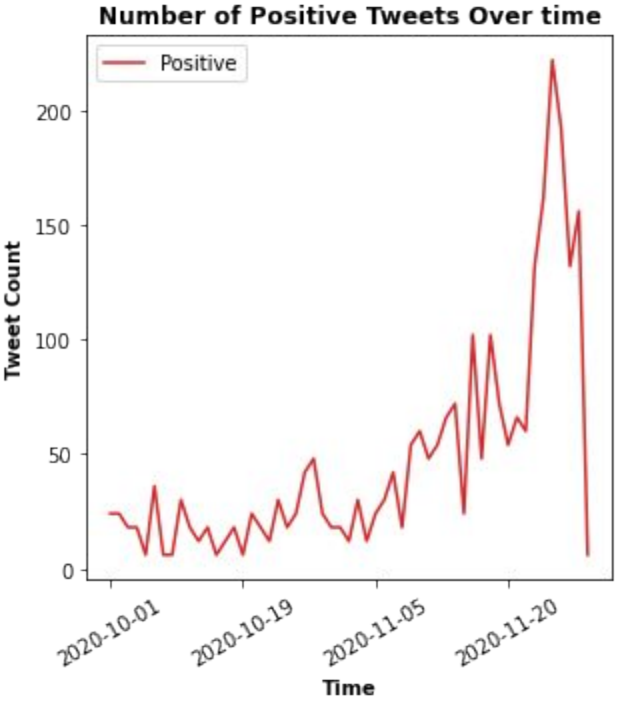
Top 5 Most Retweeted Celebrities

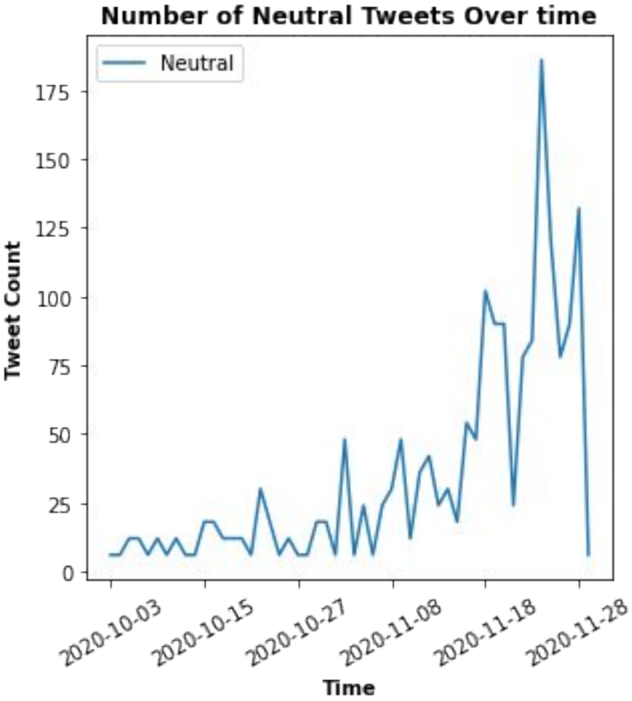


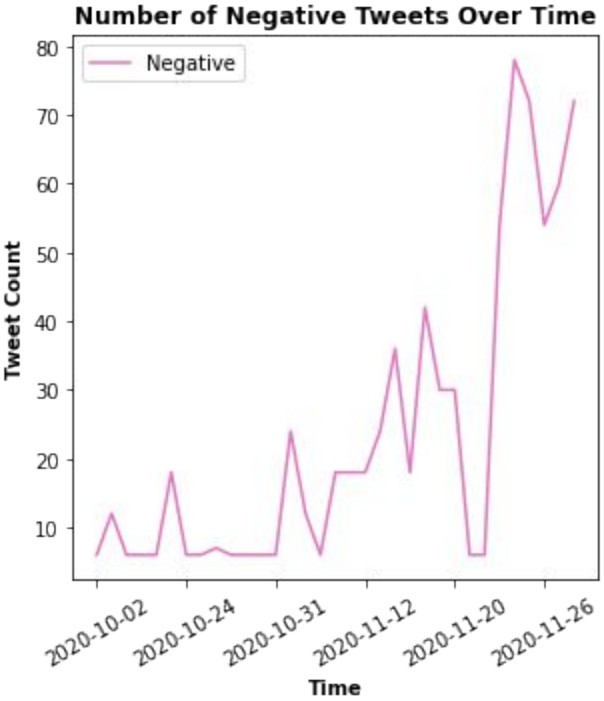
Top 5 Most Retweeted Athletes



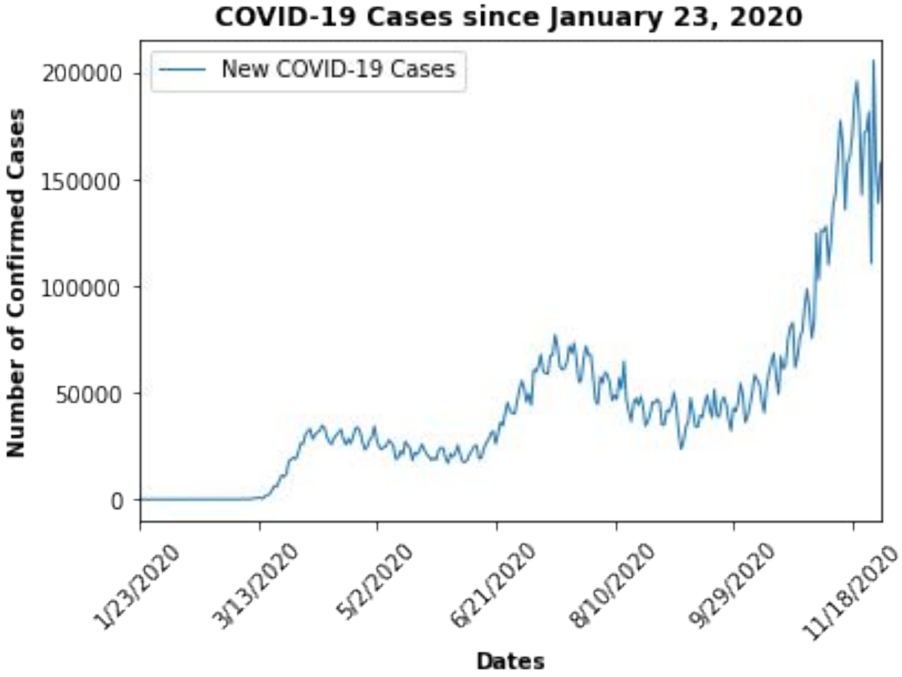
Communicate

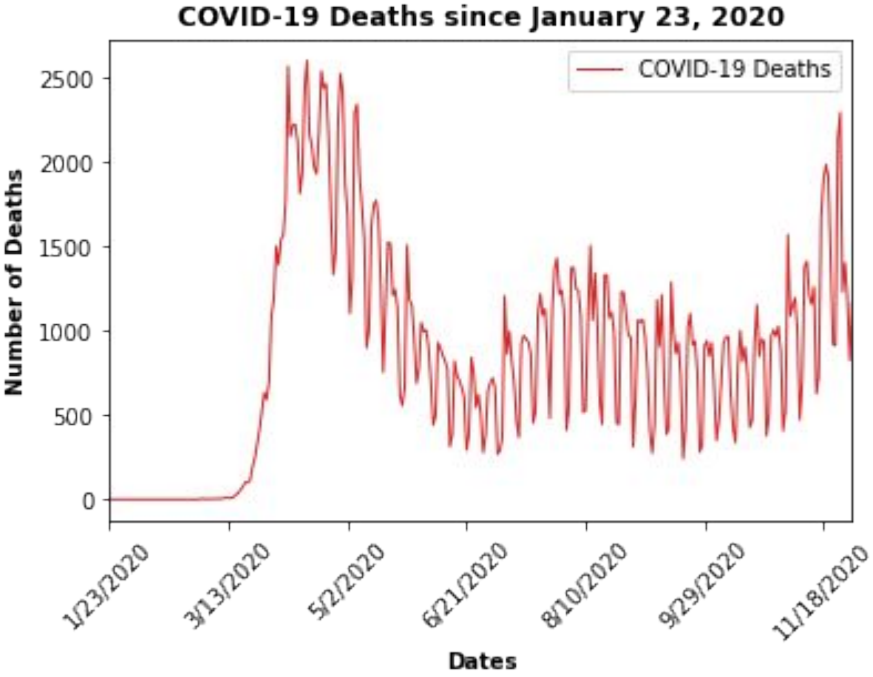






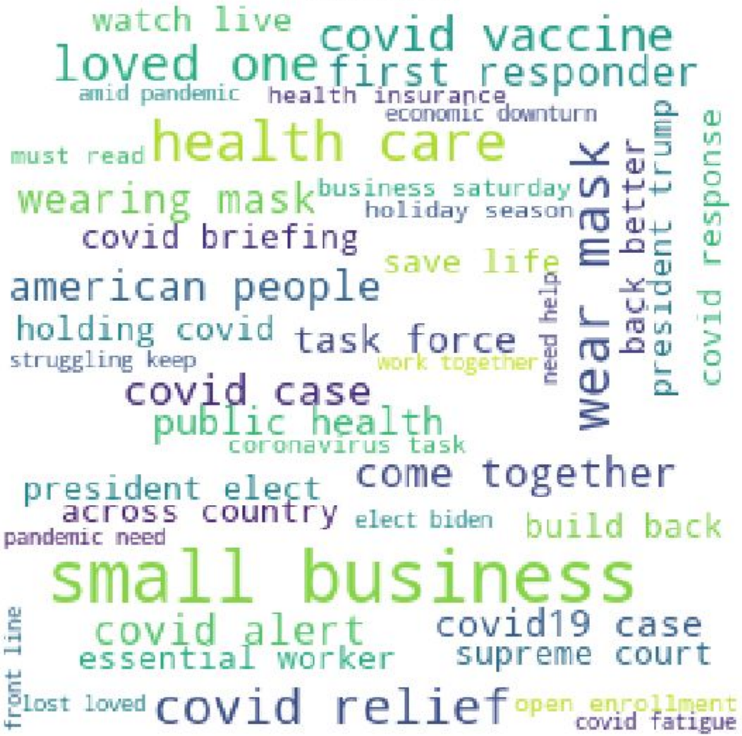
Communicate



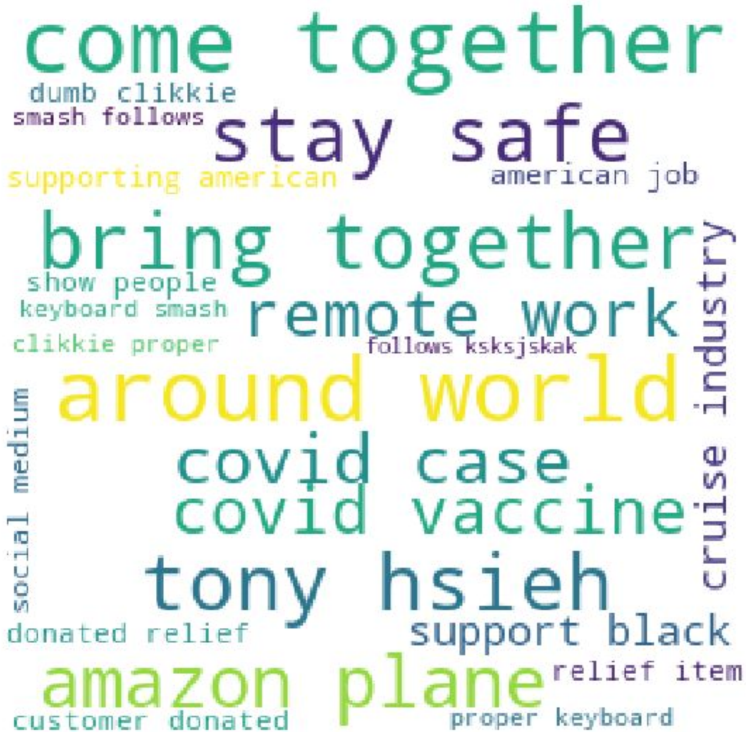


Communicate

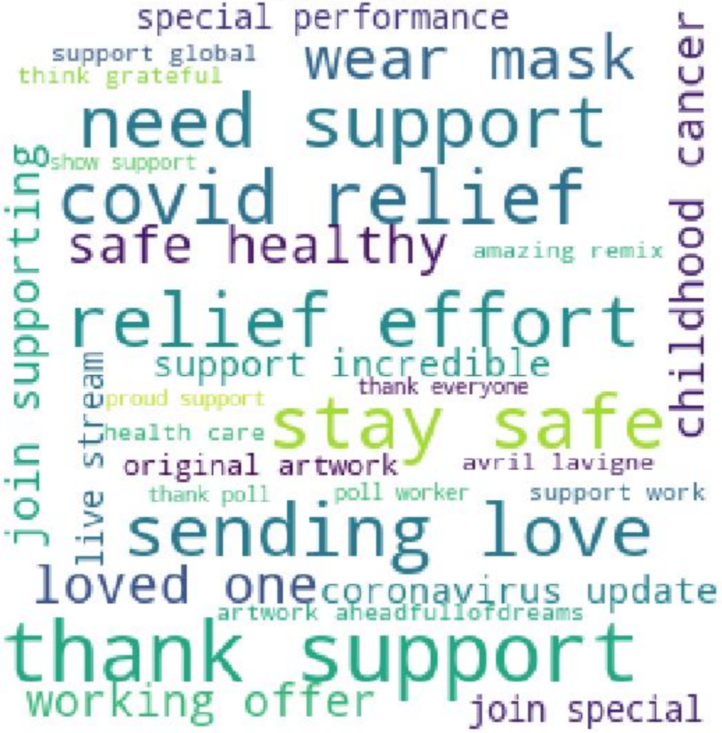
Politicians

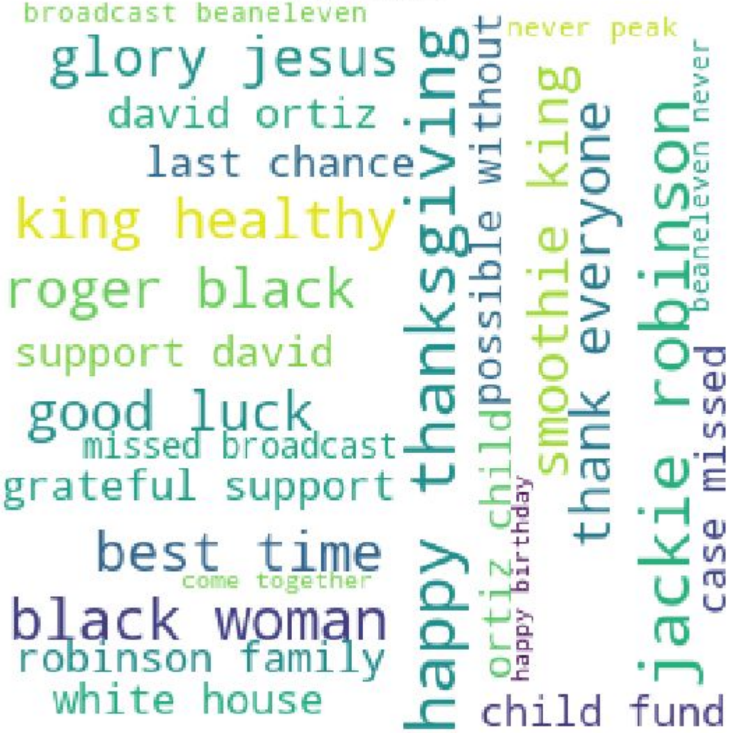


CEOs



Communicate


Celebrities


Athletes

Machine Learning Failure

- We tried implementing a machine learning technique to predict the sentiment of the tweets we analyzed
  - We tried using KNN and Bayes models similar to the ones learned in class
  - We tried alternative methods of training and testing to predict sentiment
  - Moral of the story, we failed, and did not complete the machine learning step

The End