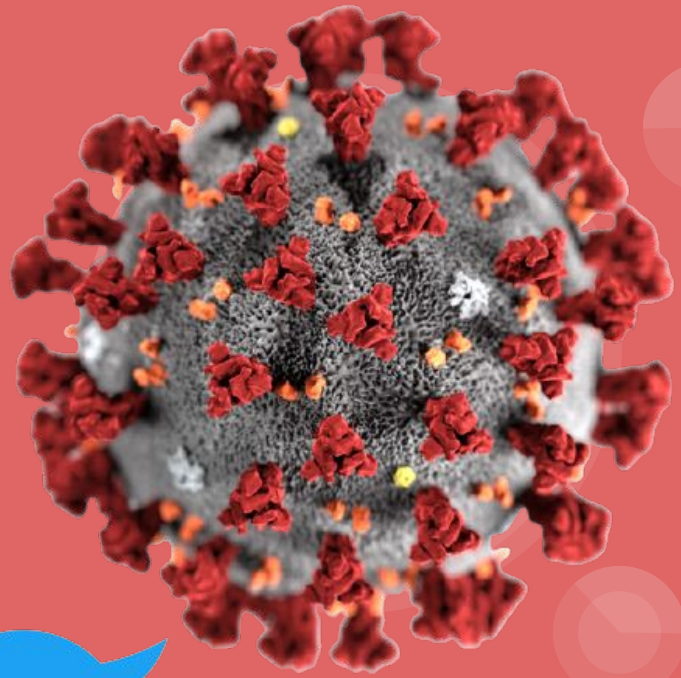
COVID-19 Twitter Sentiment Analysis



Christopher Bach
Khalid Hamid Fallous
Jay Hirpara
Jing Tang
Graham Thomas
David Wetherhold

Current State

Covid-19 Sentiment Analysis: Current State

- Twitter Covid-19 Sentiment Analysis
- Reason topic was selected
- Description of the source of data
- Questions the team hopes to answer with the data
- Description of the data exploration phase of the project
- Description of the analysis phase of the project

Twitter COVID-19 Sentiment Analysis

This project aims to collect data from Twitter and perform sentiment analysis surrounding COVID-19 vaccines. We will perform the ETL using python and a SQL database, and also come up with some machine learning algorithms to possibly predict trends related to the virus and it's vaccines. The project will bring some meaningful discussions whether getting vaccinated stands in the way of individual personal liberty, including, but not limited to, the topic of employers requirement to have staff vaccinated. This project ultimately will be able to identify any correlation between changes in daily inoculation rates and changes in twitter sentiment surrounding COVID-19.

Why did we choose Twitter COVID-19 Sentiment Analysis?

- We chose this topic because it is quite topical and has immediate importance to our lives and directly connects to current events.
- We also believed that the topic would provide answers to actual questions in regards to how sentiment can be captured in regards to COVID-19.
- Our topic also involves technology that is directly relevant to data analysis in the modern world

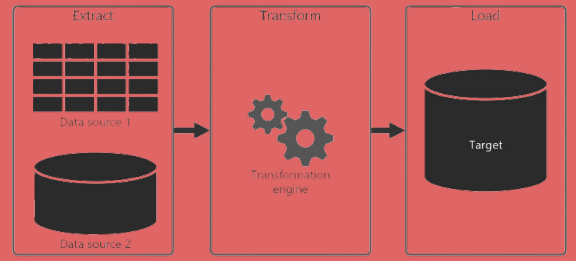
Our Source Data

Twitter.com - We used the Twitter API to gain access to tweets (limited to the past 7 days)

Kaggle.com - We supplemented Twitter API tweets with aggregated tweets from Kaggle (dating back to Dec. 2020)

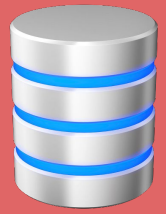
CDC.gov - CDC data allowed us to access inoculation numbers and rates as well as Covid-19 cases and deaths

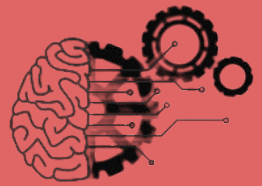
Python



ETL

Database





Machine Learning Algorithms

Questions Upon Further Analysis

Is there any correlation between tweet sentiment and vaccination rates?

If there is a correlation, can we use tweets to predict how vaccines can be distributed/marketed?

Sentiment Analysis: VADER vs. Textblob?

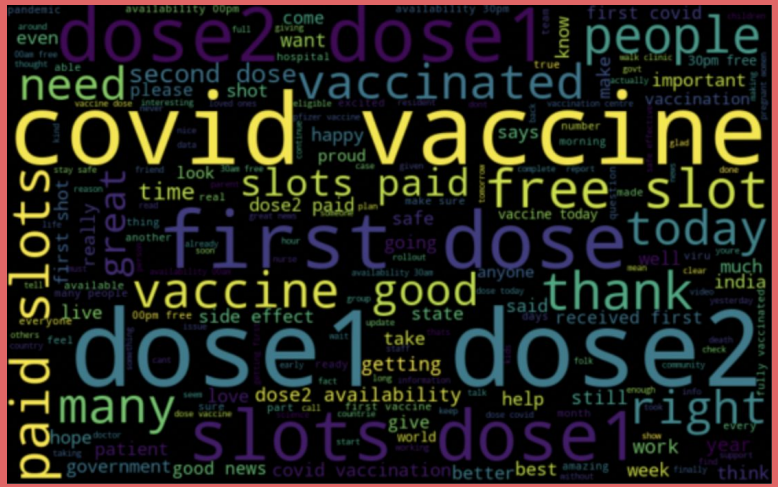
- Vader is optimized for social media data and can yield good results when used with data from Twitter, Facebook, etc. As the above result shows the polarity of the word and their probabilities of being pos, neg, neu, and compound.
- Textblob sentiment analyzer returns two properties for a given input sentence:
- Polarity is a float that lies between [-1,1], -1 indicates negative sentiment and +1 indicates positive sentiment.
- Subjectivity is also a float that lies in the range of [0,1]. Subjectivity sentences generally refer to opinion, emotion, or judgment.

Data Exploration/Analysis/Visualization

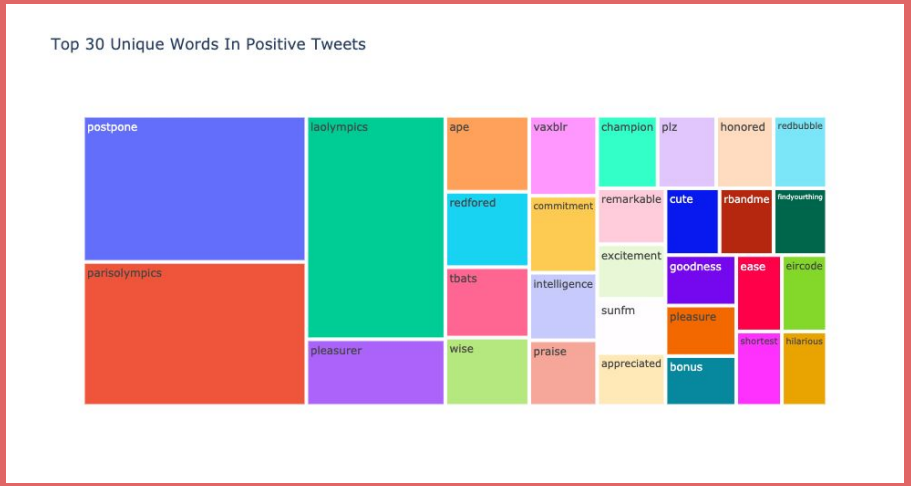
What common words appear in our sentiment analysis?

- Positive:

Textblob



VADER

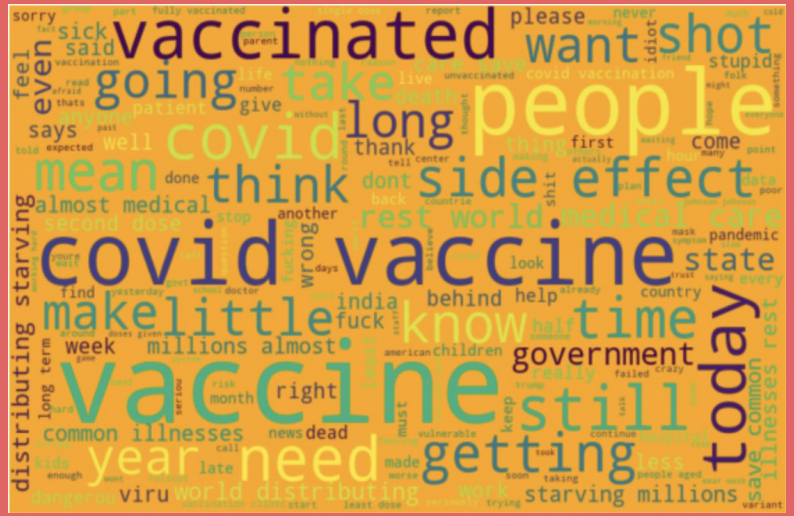


Data Exploration/Analysis/Visualization

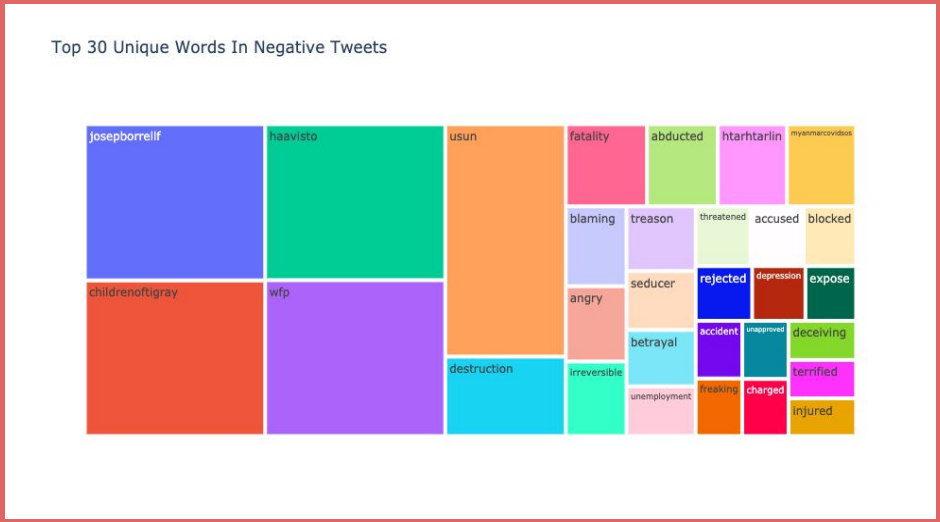
What common words appear in our sentiment analysis?

- Negative:

Textblob



VADER

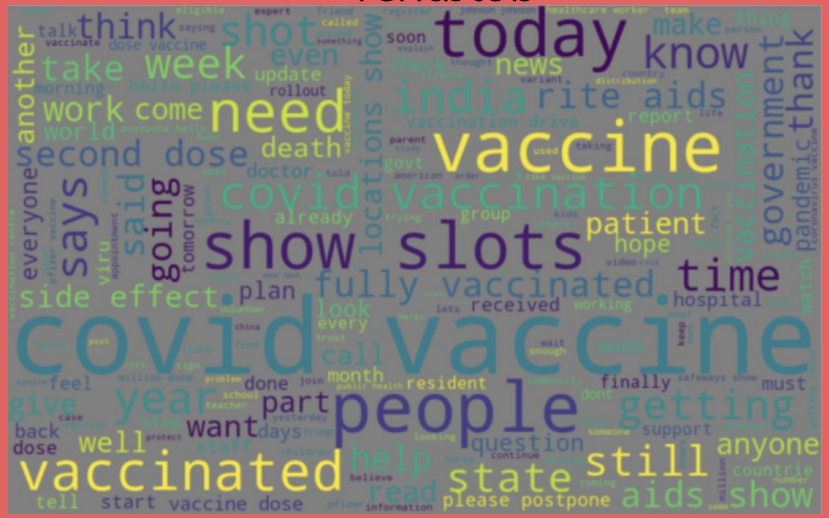


Data Exploration/Analysis/Visualization

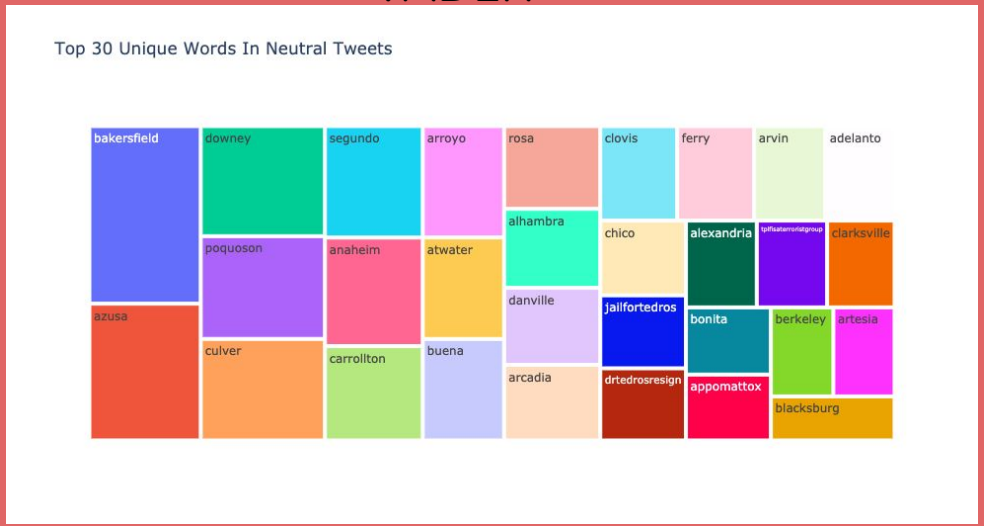
What common words appear in our sentiment analysis?

- Neutral:

Textblob

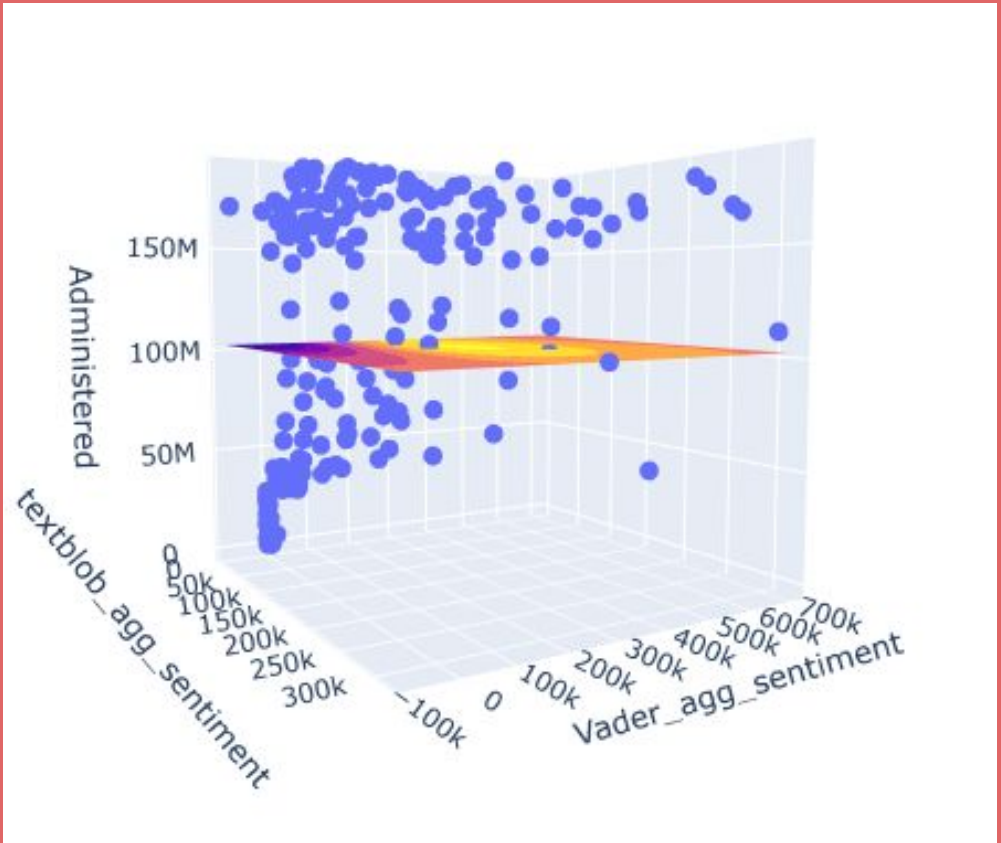


VADER

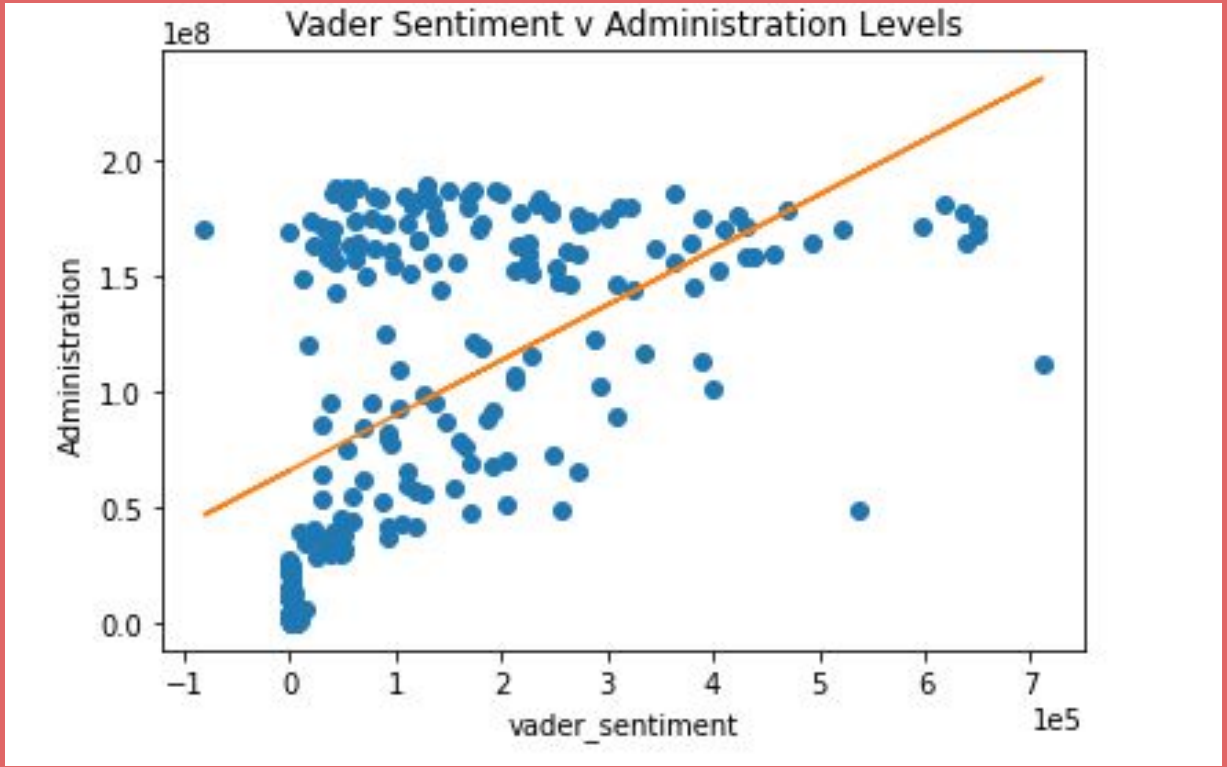


Data Exploration/Tableau 

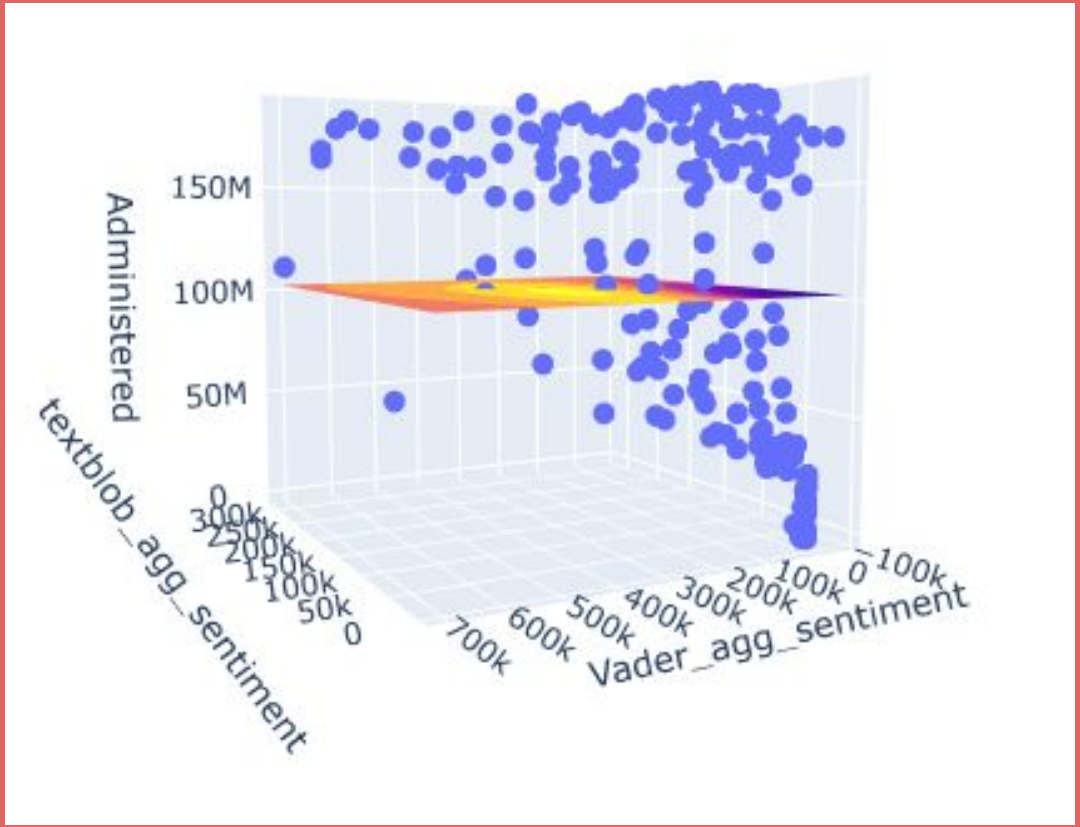
Regression Analysis (a)



Regression Analysis (b)



Regression Analysis (d)



Challenges and Limitations

- Facebook, Instagram and TikTok were all considered initially, but did not have the necessary data readily available.
- Some members ran into issues with gaining Academic Twitter accounts to be able to access the Twitter API.
- After gaining access to tweets our original goal of using the location of tweets was not possible due to most tweets not having geotag data
- The Twitter API was very limited to the amount of data we could pull
- Using academic accounts only allows access back to 7 days of tweets. We could not get twitter's full archive search without having a twitter scholar account.
- The group decided to use a Kaggle Dataset, which provided us with thousands of tweets from December 20, 2020 when the first vaccine was announced.
- Group ran into a machine learning natural language paradox, where we noticed an issue within our sentiment analysis. When analyzing tweets for Covid-19 Vaccination sentiment (pro/anti-vaccine) when running into a tweet such as “I hate anti-vaxxers”, this would return a negative sentiment when this person is actually pro-vaccine.


Questions?