## Introduction

The political conflict between Donald Trump, representing the Republican Party, and former Vice President Joe Biden, representing the Democratic Party was reflected in the discussions among users on online social networks like Twitter, Instagram, Facebook etc.

When you think of politicians and Twitter, chances are President Donald J. Trump comes to mind. Ever since he launched his run for office in 2015, Trump has become infamous for what many have described as derogatory, negative, and somewhat inflammatory tweets. Turns out that with even as little as 280 characters, he can communicate a whole spectrum of emotions, “facts”, and, believe it or not, opinions.

Over the past few years, the use of political Twitter accounts has skyrocketed. Nowadays it seems like every political leader and their entire family has a Twitter account, and they all feel the need to share their opinions with all of us on the platform. In fact, many leaders use Twitter as their primary mode of communication to the public rather than addresses or newsletters. However, this has caused some interesting problems. Often tweets aren’t vetted, and when coming from “personal” accounts rather than “official government-run communications” they can be quite polarizing and/or problematic.

## Research Questions

We wanted to investigate how different leaders around the world used Twitter. Specifically, we oriented our analysis around the following questions:

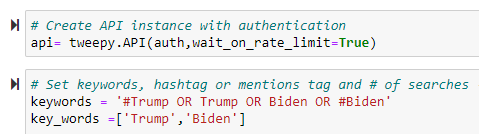
1. How people are reacting to the daily news and relate that to presidential election?
2. What are the most frequently used keywords used in political tweets? How do these differ by candidate?
3. Can we predict the tweet sentiment using ML model and compare that with industry products like Google and TextBlob?

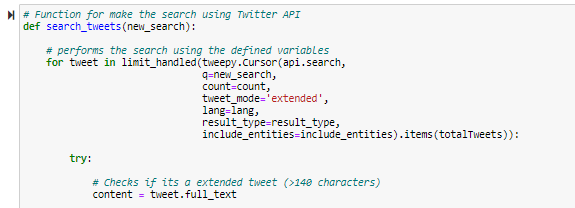
## Data



Our first step was to gather data. We knew we wanted to gather tweets that had key words “Trump” or “#Trump” or “Biden” or “#Biden” over 11 days at random time of a day. Due to infrastructure limitation, we took a sample of 10K tweets per day.

First, we tried Twitter’s native API but soon realized that they had quite strict limits on how much data a single person could download. We used twitter API to download the data.

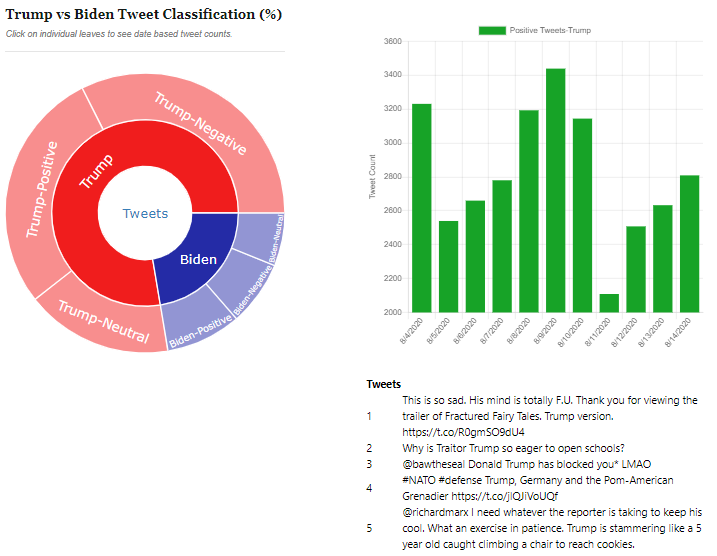




After downloading the data, we focused on cleaning up the tweets and translating them. We removed URLs, @ tags, punctuations regular expressions and normalized all of the tweets to be lower case. We used tweets that had 5 or more words.

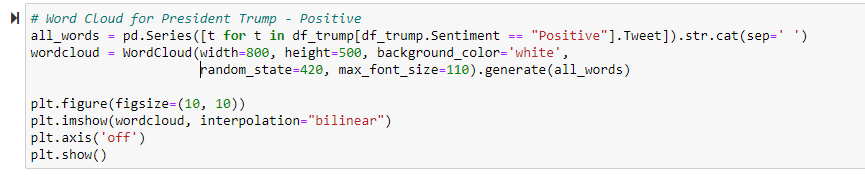
## Analysis

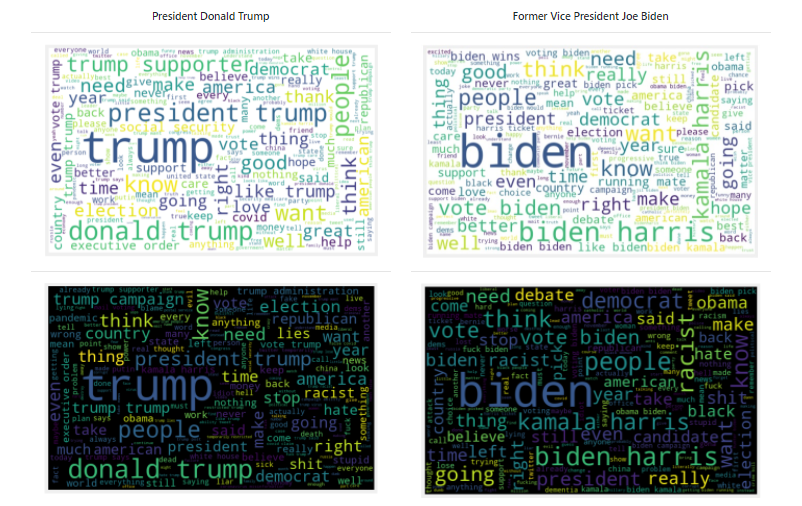
First, we wanted to analyze basic information. We were surprised to see tweets linked with President Donald Trump (77.64%) were significantly higher than Joe Biden (22.36%).



Next, we wanted to look at some of the keywords being used by the users. We limited this analysis to English accounts because, during translation, the text data can be changed enough to not accurately reflect word usage. Word clouds are a great way to easily visualize how certain keywords stand out in a large collection of text.

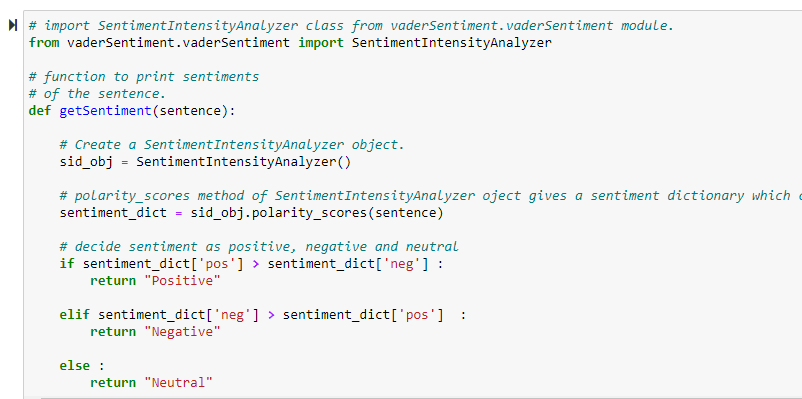
We used the [WordCloud](https://pypi.org/project/wordcloud/) Python library to generate the images by feeding in the raw tweet data.



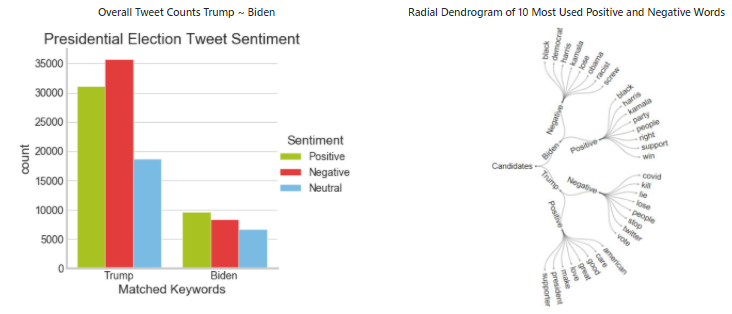


After gathering our baseline data, we shifted our focus to sentiment analysis and decided to use VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. This provided us with four data points for each tweet – Positive, Negative, Neutral and Compound score.

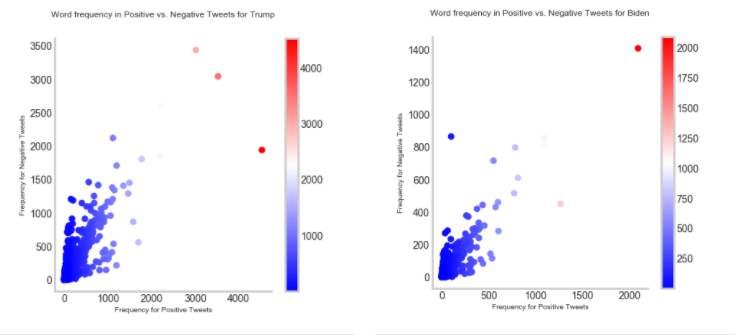
We compared positive and negative scores to determine final sentiment.



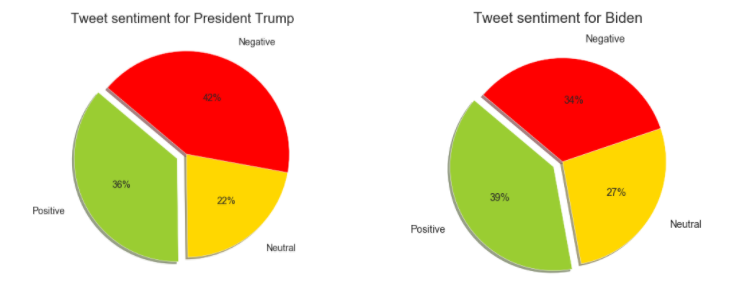
We compared the average sentiment scores and looked at the top 10 words in each sentiment category.



Also, we looked positive and negative word frequency for each candidate.

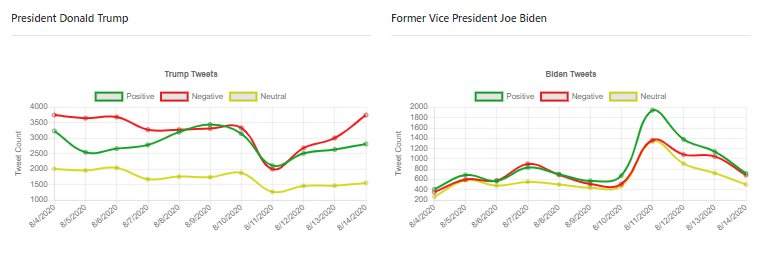


As tweet volume for President trump is much higher than former Vice President Joe Biden, we looked at % sentiment using pie chart.

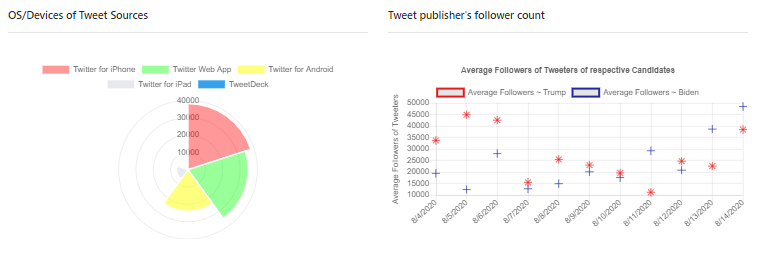


A huge part of Twitter conversation revolves around news and politics. That makes it an excellent place to measure public opinion, especially during election campaigns.

First, we were able to count the number of positive and negative mentions for each candidate during a period of time. Following graph shows Trump and Biden’s tweets based on sentiment. Joe Biden announced his VP choice on August 11 and one can see user sentiment to the news.



Interesting to understand the twitter user, followers and devices used to tweet the message.



## Machine Learning Model (AAHA Model)

Machine learning models can do pretty amazing things, even when it comes to something as subjective as the human language. With a machine learning, it’s simple to get started with sentiment analysis of the Twitter data.

We compared Logistic Regression, Naive Bayes, and Support Vector Machines learning algorithms to build the review predictions. Final approach that does a combined "vote" of all three models.

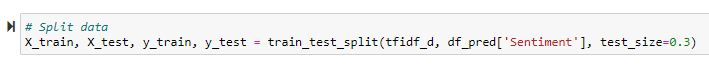
Since we have limited data, we used cross-validation to split the data 10 ways and measure accuracy in an unbiased way.

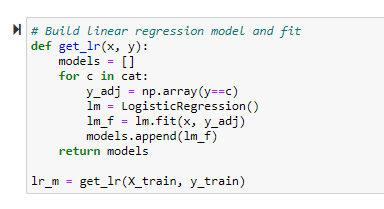


### After evaluating all the models, we selected Logistic Regression model. There are three models that are build for each category: Positive, Negative and Neutral.

### Training the model of choice:

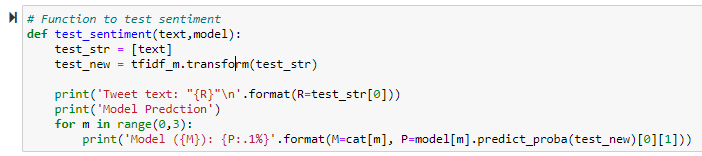
* For training models, data was split 70%/30% where 30% will be used for prediction to gauge final accuracy of the model.
* For each review, the model that scores the highest will tell us which kind of review it likely is.

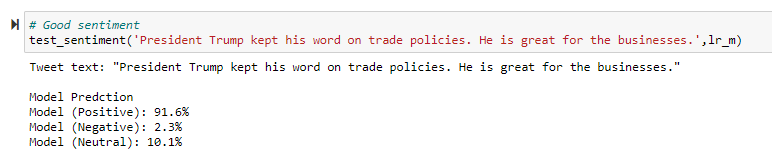




### Test output:

* Built a test function which allows to supply review to see how well the model will predict it's rating.
* For simplicity of testing, function allow only one review at a time.
* The program uses the stored TFIDF matrix to tokenize and transform our new review which is then fed to all three of logistic regression models.
* Each model has an independent assessment of how likely it is that our review is a positive hit.

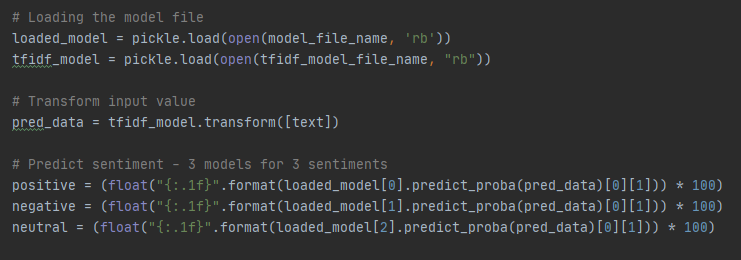


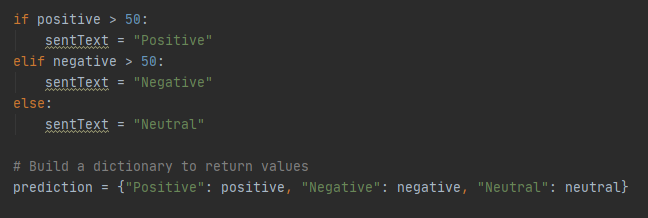


### Save Model:



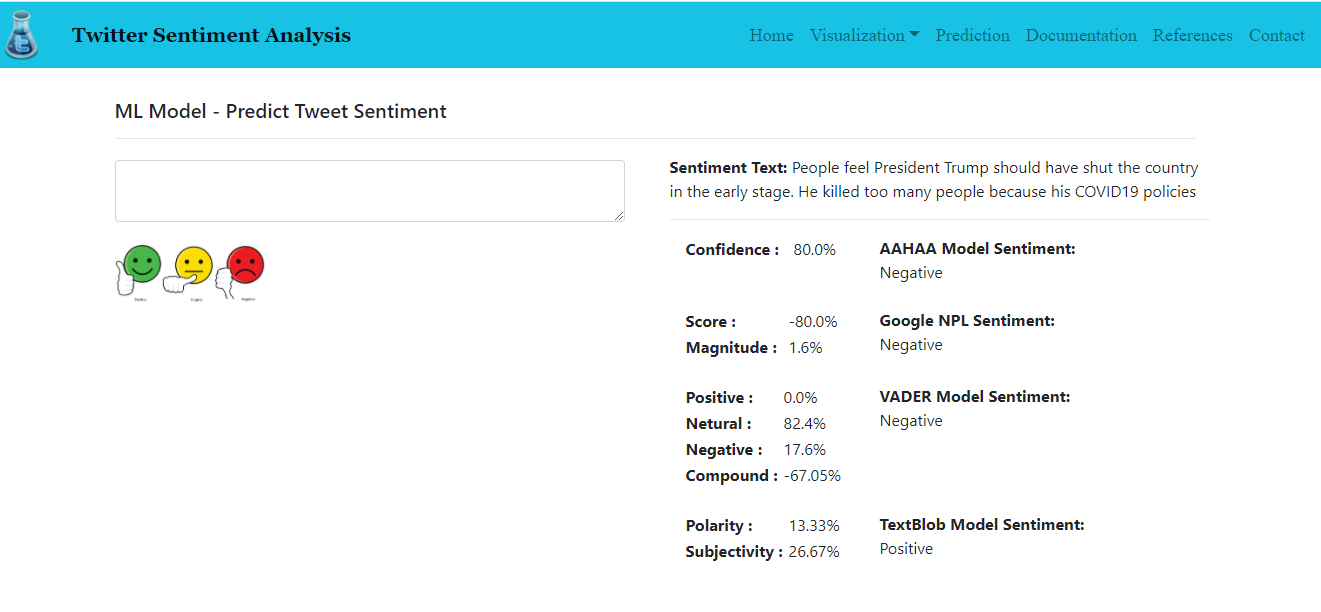
Loading the model:





Model Testing:

We build user interface screen to test AAHA Twitter Sentiment model that predicts user entered tweet sentiment. Results are compared with Google NLP model, Vader Model and TextBlob model.



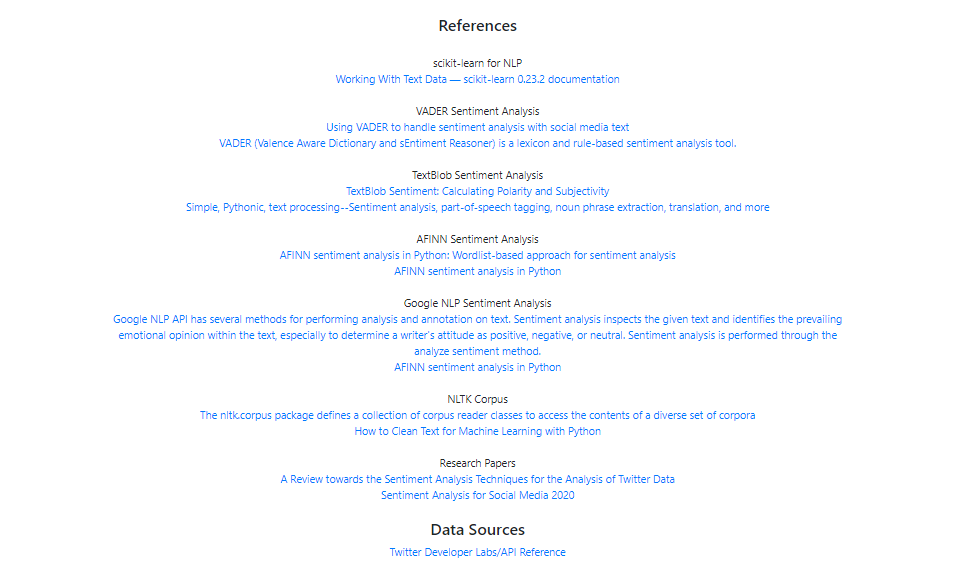
## Conclusion

While our data set was smaller, we were surprised at just how much insight could be drawn from this data. There are clear, explainable differences between the sentiments for two political leaders.

If anything, this shows that more and more, Twitter is reflective of a politician’s *true* beliefs, and sometimes is more revealing than any other source.

AAHAA Model gave fairly same results across industry sentiment analysis API’s.

## References



## Credits

* Medium.com articles
* Towardsdatascience.com articles
* Dimitrios Effrosynidis (Text Cleaning Techniques)
* Twitter Developer Page
* Paul Arias (RU Bootcamp Instructor)

## Team Members

* Ajay Patil
* Akshita Parasrampuria
* Hasti Patel
* Archna Ashish
* Aslam Momin

## Declaration

The authors declare and solemnly affirm that this research has neither been funded by any political or religious groups nor are the authors in any way affiliated to any institutions with direct or indirect access to groups with biased interests. This research work has been carried out in the interests of technology and academics.