A Study on US 2020 Election Tweets

# **Kyle Morris Winter 2020 https://github.com/jkylemorris/DSC-Portfolio**

# Which Domain?

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you’re doing with these data.

The data will be coming from Twitter, specifically tweets from October 15th through November 8th. I’ll likely be doing some sentiment analysis - <https://towardsdatascience.com/another-twitter-sentiment-analysis-bb5b01ebad90>

<https://www.toptal.com/python/twitter-data-mining-using-python>

We will of course start with cleaning and importing the data. Fortunately, the Kaggle dataset is in a CSV format which is very easy to work with. I will likely use python - <https://datascienceplus.com/twitter-analysis-with-python/> (this also gives a good visualization – a wordcloud which may be very useful when analyzing these tweets)

I want to look into textblob, because it seems to do sentiment polarity analysis quite well - https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/ . I’m visualizing perhaps a sentiment over time chart – does a certain candidates supporters get angrier as the election gets closer and results aren’t cut and dry? <https://playfairdata.com/how-to-make-a-timeline-in-tableau/>. Some of the dates of certification of votes - <https://abcnews.go.com/Politics/key-post-election-certification-deadlines-dates/story?id=74157237> might be useful to mark milestones.  
 The data was inspired by the Australian Election 2019 Tweets dataset - <https://www.kaggle.com/taniaj/australian-election-2019-tweets>. I found some other articles that involved analysis of tweets - <https://www.japantimes.co.jp/news/2020/12/05/asia-pacific/politics-diplomacy-asia-pacific/china-tweet-australia/>. Some researchers are analyzing tweets to show how a lack of clarity undermines credibility - <https://news.iu.edu/stories/2020/11/iupui/releases/10-covid-19-data-visualization-errors-study.html> Some things to keep in mind https://resources.observepoint.com/blog/10-tips-for-presenting-data.

# Which Data?

What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

We’ll be examining a collection of Tweets gathered from October 15th through November 8th collected by Manchun Hui. The Kaggle dataset can be found at <https://www.kaggle.com/manchunhui/us-election-2020-tweets>.

Columns are as follows:

* created\_at: Date and time of tweet creation
* tweet\_id: Unique ID of the tweet
* tweet: Full tweet text
* likes: Number of likes
* retweet\_count: Number of retweets
* source: Utility used to post tweet
* user\_id: User ID of tweet creator
* user\_name: Username of tweet creator
* user\_screen\_name: Screen name of tweet creator
* user\_description: Description of self by tweet creator
* user\_join\_date: Join date of tweet creator
* user\_followers\_count: Followers count on tweet creator
* user\_location: Location given on tweet creator's profile
* lat: Latitude parsed from user\_location
* long: Longitude parsed from user\_location
* city: City parsed from user\_location
* country: Country parsed from user\_location
* state: State parsed from user\_location
* state\_code: State code parsed from user\_location
* collected\_at: Date and time tweet data was mined from twitter\*

# Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you’ll be proposing your research questions as well as justifications for why you’d offer these data in this way.

I would be interested to compare the Twitter data to the final results – for instance, since we have location data, can we predict the outcome of a state based on the net positivity of tweets originating from that state?

# What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

Sentiment Analysis will be useful, as our initial dataset I believe just pulls #trump and #biden tweets. Sentiment analysis is important in this case because “#Trump is the Antichrist” and “#Trump is our god #MAGA” can be attributed to two different supports of the 45th president.

Some visualization techniques will help with bringing our data to life and telling the story.

# Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

The sheer amount of data could be intimidating. I’ve got about 900 MBs of tweet data, and considering the size of tweets I hope it’ll be enough. As far as sending the project off schedule, there’s so much that can be done with our data that I may have to scale back the scope of the project once patterns emerge lest we become embroiled in figuring out everything.

# Concluding Remarks

Tie it all together. Think of this section as your final report’s abstract.

Technology and real time analysis of data has allowed us to conduct analysis far faster than could be accomplished in the past. A study of polls, statements by voters, and analysis of the zeitgeist previously could have taken months. In this case, our data was ready for us the day after it happened.

Given that the data is in a raw, unrefined form the cleanup step will take significantly more effort than a study designed from the beginning to source required data. However, the tradeoffs will be worth it. We hope to have a sense of different factions of the hivemind that is Twitter reacted in the run up to the election and how that shifted over time. I hope to find a sort of tipping point – can we identify the moment that hope sprang eternal? Will there be an inflection point? These are some of the questions I hope to answer by analyzing this data.