Michael Caballero

DATA H195A

Data Source Documentation

While I am open to using more sources of data (from other social media platforms) if I have the time to do so in my thesis, I currently am just going to be using data from Twitter. I will be using the Twitter API to collect Twitter data. This data is not just the actual text of the Tweet but also Tweet annotations which allows searching for Tweets about certain topics or entities, and metrics data like the number of retweets and the number of followers of the account. Twitter is currently rebuilding the Twitter API to include a tailoring offering for researchers with elevated access for free. I am not sure if the full version of this option will be available when I begin to collect data but am not reliant on this option being available. So far, Twitter has released an early access version of this new Twitter API they are creating and I am looking into the advanced capabilities that this academic developer account provides. I believe there could be difficulty in gathering historical data with the general version of the unpaid Twitter API, but Twitter suggests that the new version of their API for researchers solves this difficulty. I am working on applying for an account to be considered a researcher and thus qualify for the early access features of the new Twitter API. The data will be gathered using Twitter's recommended code libraries in conjunction with Twitter's API following Twitter's tutorials.

I believe Twitter owns their data but permits developers to use that data following the Developer Agreement and Policy that all developers have to agree to in order to create an account. I do not believe any of the data will be censored or biased in

any way that I do not expect. I am hoping to make sense of the random noise and bias in the data to poll the opinion of the entire population.

I am also looking into other projects revolving around electoral prediction using Twitter data of the 2020 presidential election. If I am able to find another project similar to mine, I plan on reaching out and seeing if I can access their data. While this provides the benefit of having a large dataset that I could use, I would only add this onto my current dataset as it would restrict my freedom in determining my own methodology for the project.