Political Tweets
Version 1.0 (October 1, 2019)

Pavlos Protopapas, Kevin Rader, and Chris Tanner

Harvard University
CS109A

1 Problem Statement

We live in a digital world that is ever-changing, from the sources of our information, to the credibility we place in various online mediums, and the ease and patterns of disseminating information. Online text is incredibly powerful. Further social media is ubiquitous, and Twitter alone produces, on average, 500 million tweets a day. Individual tweets can have profound effect.

Another global trend is that the political landscape is becoming increasingly polarizing. Within the United States, we are experiencing a unique period in history. To this end, tweets, especially political ones, are influential and ripe for analysis.

This past week marked the beginning of an inquiry to impeach Donald Trump from the position of U.S. President. Consequently, on Saturday and Sunday, Donald Trump tweeted 87 times. Many of his tweets are direct responses to news stories. Thus, there is often a direct relation between the content of news and the content of his tweets – the nature and sentiments of his tweets vary, but it is often in reaction to current events. On the other hand, his tweets also cause the creation of news stories.

For this project, you are asked to analyze tweets from Donald Trump. If you prefer, you could alternatively focus on any other political entity or organization. This project is deliberately highly open-ended in its direction, for the sake of allowing you to investigate interesting problems during this incredibly rare time in politics.

For example, perhaps you may be interested in seeing if there is a correlation between the economy and Trump's tweets. Although the economy is based on an infinite number of latent variables, there is allegedly a correlation between Trump's tweets and the market sentiment of US Treasury Bonds[1]. One could study this to validate any correlation. If you wish to use data that concerns the market of raw goods such as soybeans, wheat, and corn, we can provide such for you.

Alternatively, one could craft a project involving news articles from the New York Times (or any other online news outlet), in attempt to find correlations and make predictions as to what and when Trump will tweet. Further, one could create a predictive model to estimate the nature of his tweets, per your own

[1] Wikipedia.org: Volfefe Index

defined categories. For example, you could create your own dataset of his tweets and label each tweet as being one of N categories (e.g., defensive, offensive, both, possibly with another independent axis for the sentiments of positive, negative, neutral).

It could be interesting to investigate his tweets over a long time period, in attempt to notice any changes on a larger scale.

2 Data Resources

We will not provide any data for this project. Donald Trump's tweets are available online at a few locations, such as:

– Trump Twitter Archive (click here)
– CNN's archive (click here)

You will likely need to scrape web pages using Beautiful Soup. We hope for you to be creative in your project scope and usage of datasets.

3 High-level Project Goals

1. Obtain tweets either from simply downloading a pre-existing archive or by scraping them the web.
2. Optionally use an additional non-Twitter dataset to supplement your tweets.
3. Per the milestones, define a specific task, explore it with the aforementioned data, perform appropriate EDA and refinement, build a predictive model toward your task.
4. Evaluate your models on your designated development set, then evaluate your best performing model on the test set. Perform an error analysis, noting the best features, worst features, and any trends in misclassification.
5. Discuss model weaknesses and future ideas for improvements.

Note, this project should not be interpreted as being any reflection as to the CS109A staff's or Harvard's political views. This is purely an academic project whereby we want students to explore data and make statistical predictions in an area that is vastly affected by many confounding, real-world complex factors. The goal is for students to illustrate learned skills that align with our course goals.