Alex Mead, Wei-Hua Hsu (Wafer)
STAT-696 Natural Language Processing
Dr. Marco Enriquez
27 February 2020

Final Project Proposal

Background

Over the past decade, there has been a surge in social media activism among elected representatives, particularly among members of Congress. These posts range from topics concerning legislative agendas, constituent outreach, party support, and personal messages, among others. This trend has been driven to new heights after the election of Donald Trump, who uses twitter as his primary means of public communication with citizens and other government officials. As such, the polarization of political discussions has been driven to extremes. There is a substantial amount of data available from public social media companies such as Facebook and Twitter which could drive several interesting research questions. Our final project topic would like to focus on the behavior and sentiments of Congressional tweets and facebook posts, with a focus on how these sentiments differ among subjects, and whether these posts could be used to predict bias in the post.

Data Set

The data set we have chosen for our analysis was gathered from Kaggle.com; it is titled, "Political Social Media Posts" and consists of 5000 observations with 21 variables (if this is too small of a dataset, please let us know!). The key variable of interest is the \`text\` variable, which consists of each Congressmember's social media post. This variable is not clean as it has various "hashtag" markers, URL links, punctuation characters, and usernames (such as @user). Other variables of interest include the time of the post, the type of audience, whether there was bias (either neutral or biased), the type of message, the source (such as Facebook or Twitter), and the Congressmember who posted it. The link to the Kaggle source is here: https://www.kaggle.com/crowdflower/political-social-media-posts

Goals/Hypothesis

There are several interesting questions we would like to analyze based on this data set, and each question involves natural language processing of one or more variables. Our ultimate goal is to analyze the sentiment of each tweet and see how it varies by audience, message, and political party; as well as to see if it can predict bias. To this end, we would like to run a logistic regression with bias as the dependent variable and the text, political party, message type, and audience as the dependent variables. This results in the equation: bias ~ text + party + audience + message + ε. Our hypothesis is that the sentiment of the text along with these other predictors can predict whether a tweet will be biased or not.

Additionally, we would like to compare statistics across categories such as how do sentiments among social media posts differ across audience types and message types? For example, do sentiment scores differ across messages that are meant to convey policy information vs. messages that are meant to attack a political opponent vs. messages that are meant to mobilize supporters, etc. We hypothesize that there are key differences in social media post sentiments across these categories.

Methods

We plan to tackle the above project in the following order: tidy and clean the data, create new variables where needed, run our statistical methods, and create visualizations to communicate our results. Cleaning the data will be most time intensive of the project. As stated above, the `text` variable in this dataset is unclean as it has various numeric and other unneeded characters that are not fit for analysis. To clean this variable, we will normalize the text in each post and transfer the hashtags and usernames into new variables so that they can be analyzed later. This will create a clean `text` variable along with two new variables (hashtag and usernames mentioned in the post) that can be used for further analysis. Second, we would like to clean the `label` variable which consists of the Congressmember who posted the message. We would like to create two new variables that denote the political party of the Congressmember and the state that they represent. These new variables could be used to gain more insight on differences in social media posts across regions and political parties, and potentially be added to our statistical models. Once this is complete, we will formulate and run our statistical models. Likewise, we plan to plot the distribution of the social media sentiments to visually show any differences.

If there is anything unclear in the above proposal, please do not hesitate to ask! Also, we are looking forward to any suggestions and feedback you may have. We both like the dataset and are excited about the analysis; however, we are a little concerned about the overall size of the dataset (with it only being 5000 observations); please let us know your thoughts. We may be able to get more data through https://phantombuster.com/signup if needed. Thanks!

References

Eight, F. (2016, November 20). Political Social Media Posts. Retrieved February 27, 2020, from https://www.kaggle.com/crowdflower/political-social-media-posts