COGS9: Introduction to Data Science

Final Project

[redacted]

Grading: 10% of overall course grade. 40 points total.

Completed as a group. One submission per group on Gradescope.

Group Member Information:

[redacted]	[redacted]	[redacted]	
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

Question

Is there a relationship between the sentiment of adjectives in Donald Trump's tweets and the political ideology of the politicians he is tweeting about in his tweets since he announced his run for president in June 2015?

Hypothesis

Donald Trump tweets more positively about alt-right politicians compared to other politicians because they reinforce ideologies similar to his own.

Justification

Archives of Donald Trump's (further abbreviated as DJT) tweets found on http://www.trumptwitterarchive.com/ provide a variety of pre-filtered tweets. Section headings such as "What's the worst" contain databases of trump tweets where the string "the worst" is included in the tweet, or "tweets with 'loser'" showcasing every tweet where DJT called someone a "loser." In these filtered databases where tweets contained adjectives with negative connotations, many of the tweets contained left-leaning and more centric politician names in them. We were curious if, on average, there was a difference in the types of adjectives associated with politicians in DJT's tweets. We developed our hypothesis based on the supposition that DJT favors alt-right political theory.

Background Information

Michael Tauberg published an article on Medium detailing text analyses done by Michael himself on DJT's tweets. These analyses revealed DJT relied heavily on simple adjectives to articulate his feelings. It was surprising to see that the vast majority of DJT's adjectives were positive words such as "great" and "good." Another interesting analysis of DJT tweeting about politicians by Tauberg revealed that he tweets much more about politicians with different ideologies than politicians with the same ideologies. For example, tweets containing "Obama/BarackObama" were three times greater than the second most frequent person he tweeted about (Hillary Clinton). Coming away from this article, we learned that our own biases

may be determining what we think DJT's tweets are like which may be significantly different from reality. For example, James Riddell found it surprising that the president tweeted more about 'golf' than 'border" given what he knew about DJT's tweets before starting the project. This article also informed us that there is a lot of data out there. With the types of analyses we have seen so far, our hypothesis seems very testable: easy-to-use databases exist, word-sentiment algorithms are on GitHub, and data visualizations will not be too difficult for laypeople to understand.

In an article published on their website, National Public Radio(NPR) performed an analysis on Trump's Twitter feed showing how he tweets differently about lawmakers of color than white lawmakers. This article reflected one of the key reasons why we chose our question and why it is an interesting topic. The ultimate goal of our study is to gain a better understanding of how bias towards right-wing political ideologies affects the way Donald Trump tweets about politicians across the political spectrum. This article is related because the analysis used a politician's race as the independent variable whereas our independent variable will be political ideology. We ultimately want to understand if our assumption that Donald Trump is biased towards the political right is supported by the way he tweets about politicians, and this article proved that this type of sentiment analysis exists already. While the study reported by NPR focused on specific law-makers such as "The Squad" and Committee Chair members, our study will be more comprehensive and look at a much greater number of politicians and the adjectives DJT has been associating with them in his tweets.

Data

Database Links:
http://www.trumptwitterarchive.com/

An inferential text analysis would be performed. The perfect dataset would have two parts: the first is a dataset with all the tweets attached to each chosen politician's name. Sentiment analysis would be performed on each tweet to get sentiment score for each tweet. The score would be adjusted to match if it was positive, neutral, or negative. Lastly, a second dataset would be made containing the overall percentages of positive, neutral, and negative tweets for each politician that will be moved down the data pipeline to the visualization portion. This ideal dataset differs from our previous one because we realized that it would be too difficult to gauge "political ideology" on a numerical scale. Instead, we will analyze the position of the individual separately in the discussion rather than in the dataset or visualization.
Values are all arbitrary

Name	Tweet Text	Word sentiment score	Adjusted score
Barack Obama	"@BarackObama is corrupt!"	-0.7	-1
Barack Obama	"Barack Obama is worse than me"	-0.6	-1
Jeff Sessions	"Jeff Sessions	0.6	1

	is a great friend and a good guy."		
Jeff Sessions	"Jeff is way better than the corrupt squad"	0.0	0
. . .			

Name	**% of tweets positive**	**% of tweets neutral**	**% of tweets negative**
B. Obama	20	50	30
J. Sessions	60	30	10
...	...	...	...

We will likely have to wrangle a lot of data to get it to look like the one above, and it will be important to take advantage of join operations. For example, finding a dataset on party affiliation of politicians would be great to join with a CSV from the Trump Twitter Archive containing politician names and number of tweets by DJT they are included in. Also, evaluation of left-right political spectrum score will be difficult to find and calculate. Our group has currently decided to make it a relative scale, but more planning will need to be done on how to decide that scale. That section may be switched to categorical if it becomes too complex. We will also need to find a text analysis package that is simple enough to use for first-time users.

The database contains 42059 trump tweets (observations). It is updated regularly! There are a few variables collected: Tweet date, time, contents, and the web app it was posted from. The interface has great text filtering capabilities and allows for exporting the filtered dataset in CSV or JSON format. The main limitation of the dataset is that, while it does have a section that sorts the tweets based on words like "loser" or "dummy" occurring, we would still have some additional sorting/cleaning to do to make it look like our ideal dataset. For example, this dataset does not include the left-right spectrum which is very important to determine how biased Donal Trump is. The dataset is also limited because Trump does not tweet names consistently. There are casing issues and nicknames that DJT may use that are missed by the filter we put in place. We need to be smart in deciding how to group names that may be similar so that we don't accidentally cut out relevant data.

Ethical Considerations

Data scientists must always be aware of possible ethical violations when conducting research. These violations can happen at any stage of the data science process--even after the research has been published. For our specific data science question, we must be careful to avoid collection bias in the data

collection stage. News outlets providing information on DJT are often biased, so it is important to evaluate the source before using data from that source. Since the data we are dealing with are Donald Trump's tweets, the bias in the sources could be a major issue. To navigate this, a raw database (The Trump Twitter Archive) was decided upon. The bias was evaluated by reading the Frequently Asked Questions section of the website. The website creator reported missing about 4000 out of ~42000 tweets in addition to any tweets DJT deleted prior to September 2016. Timestamps were collected using Twitter's official API. Location data is unreliable. Tweets checked every minute and the website is automatically updated.

Certain ethical considerations also arise in data storage. For example, data retention includes a plan to delete data if it is no longer needed. However, The Trump Twitter Archive's Frequently Asked Questions reveal that if Trump deletes a tweet that has been online for more than 30 minutes, it will not get deleted off of The Trump Twitter Archive and will be stored forever. This complicates our analysis because if Trump chooses to delete a tweet, we would not know if we solely relied on The Trump Twitter Archive for our analysis. It is an important ethical consideration our group will have to keep in mind for analysis. One possibility would be to find a dataset of all deleted tweets and then filter those out from our working dataset.

Most of the possible ethical issues in our study come up in the analysis stage. In this portion of the research, we must be careful to avoid dataset bias. We need to be able to identify possible sources of bias in the data and we need steps to address these biases if they arise. In our specific question, this sort of bias would likely come from our sentiment analysis. Sentiment analysis can be a very useful tool, especially in a study like ours, but it comes with limitations. Our research group should be aware of these limitations and have steps in place to look through the results and make any fixes that may come up. One example is the use of "good" versus "not good." If we evaluate those to both be a positive sentiment, then our results may not reflect the true sentiment of the dataset. Another important thing to consider in the analysis stage is making sure our representation of the data is honest. This means all of our visualizations, summary statistics, and reports must be accurate representations of the underlying data. This can be accomplished simply by taking the time to carefully review each of these elements at each step of the project to make sure they are accurate. For example, we need to make sure our data doesn't have any fake tweets. Our group must also make sure that our data analysis has auditability; that is, that our analysis is well documented and that another group would be able to reproduce it and draw the same conclusions. We will do this by using good version control, commenting code thoroughly, and using a knit R markdown document for the final product.

Our potential ethical problems do not stop once the analysis is done. In the modeling section of the research, we must address communication bias. This means that all the possible limitations of the model have been communicated to any parties that have a stake in the research. For our study, the limitations we would communicate would be, again, from the word sentiment analysis. Since we don't necessarily have any stakeholders in our research to communicate this with, we would likely put some kind of disclaimer on our study, describing the limitations of the sentiment analysis and where it may fail.

Even after the study has been deployed, our team needs to be ready to possibly redress or rollback. Redress refers to an organization's plan for a response if a user is harmed by our results. In our particular study, this seems unlikely, but it's important to prepare for the unexpected. Similarly, rollback refers to a method of turning off the model in production if it is needed. It is very important for our research team to have a way to revert changes to our word sentiment analysis or to turn off the model at a moment's notice if it becomes necessary. One of the final things to consider is the all-important unintended use factor. We

must try to identify and prevent possible abuse of the model or use that was not intended when it was made. For our specific research, it is possible that people would draw the wrong conclusions from our results or that people would use the model in ways that are unethical or not what was intended. To avoid this, we must carefully monitor what people are using our model for and consider making the model exclusive without permission.

Analysis Proposal (15 pts)

Name	Keyword	Name (cont.)	Keyword (cont.)
republican	republican	m_rubio	Rubio
democrat	democrat	a_schiff	schiff
r_mueller	Mueller	j_biden	Biden ~~ Hunter
r_guiliani	Rudy	h_clinton	Hillary
s_scalise	Scalise	n_pelosi	Pelosi
s_hannity	Hannity	m_romney	Romney
m_mcconnell	McConnell	m_trump	Melania || FLOTUS

Table 1. Keyword used for initial data collection from trump twitter archive. Data was downloaded as csv files and joined together post-transformation.

Analysis would start by **collecting our data** in the form of Trump's tweets by politician name. Tweets would be sorted using a filter built into the Trump Twitter Archive (the website has already scraped all trump tweets of twitter and attached a number of features to them). Tweets about politicians would be located by searching with a keyword, which may change depending on the politician and how Trump refers to them in tweets. We would likely hand pick the most relevant political figures based on current political trends. We also used search terms to maximize the number of tweets containing a person's name name. In other words, there was no standardization in terms of how tweets were sorted for name. For example, DJT used "Crooked Hillary" to most frequently refer to Hillary Clinton and this would have been lost if we searched for just "Clinton." Also if we searched using "Clinton" we would have contaminated our data with tweets about "Clinton Foundation" and "Bill Clinton." On the other hand, Mitt Romney was most often referred to by his last name Romney. By searching with "Romney" we caught both Mitt Romney and just Romney. If we had used "Mitt," then other words such as "Committee" were coming up in our filtered search.

Next, **data wrangling** will take place. In the Trump Twitter Archive filter interface, features can be eliminated using a feature selection checkbox at the top of the page. Features such as "time of tweet" and "favorite" are not relevant to the study, so they will not be selected. A csv file will be made for each politician. A script will then get the sentiment score for each tweet by politician using sentiment analysis and come up with a % value for the number of positive, negative, and neutral tweets that Trump made of each politician. From this we can calculate the overall percentage of tweets that contain positive, negative and neutral sentiments for each politician.

Exploratory Data Analysis will be performed to assess if the sentiment analysis technique is appropriate for our study and if there is any correlation between a politician's relationship to Trump and the adjectives Trump uses when tweeting about a politician. This will be achieved by **visualizing the data** in a grouped bar graph as it would be the best visualization to show multiple quantitative components (percentages of emotional sentiment) of categorical points (the politicians) allowing for easy comparison of sentiment by politician. Other exploratory analyses include comparing sentiment of politicians and asking if it makes sense based on Trump's affiliation with that politician. Lastly, analysis of tweets that had outlier sentiment analysis scores will be looked at to understand the sentiment analysis algorithm better and evaluate if it is appropriate for our study.

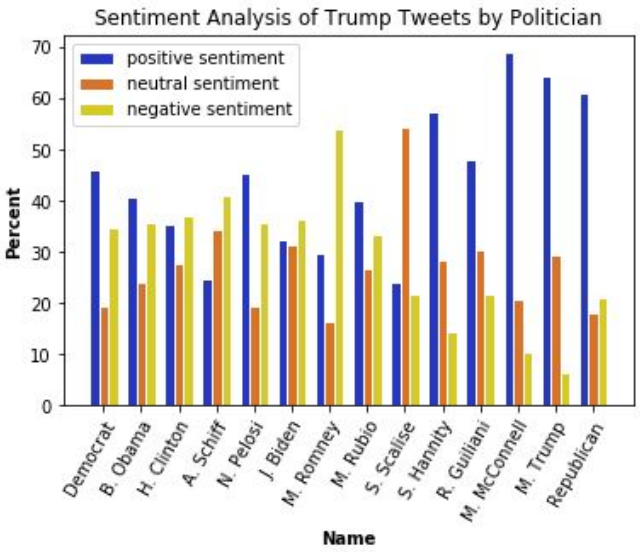

Figure 1. Sentiment Analysis of Trump Tweets by Politician. Trump tweets were grouped by the keyword "Name" and analyzed using TextBlob to determine how many of those tweets were positive, neutral, and negative. Percent was used to account for proportionality. Affiliation to Trump is not included in the figure, but can be partially inferred based on the values.

Eventually, we decided that a stacked bar graph would more accurately show the trends of our data, so we created an explanatory graph:

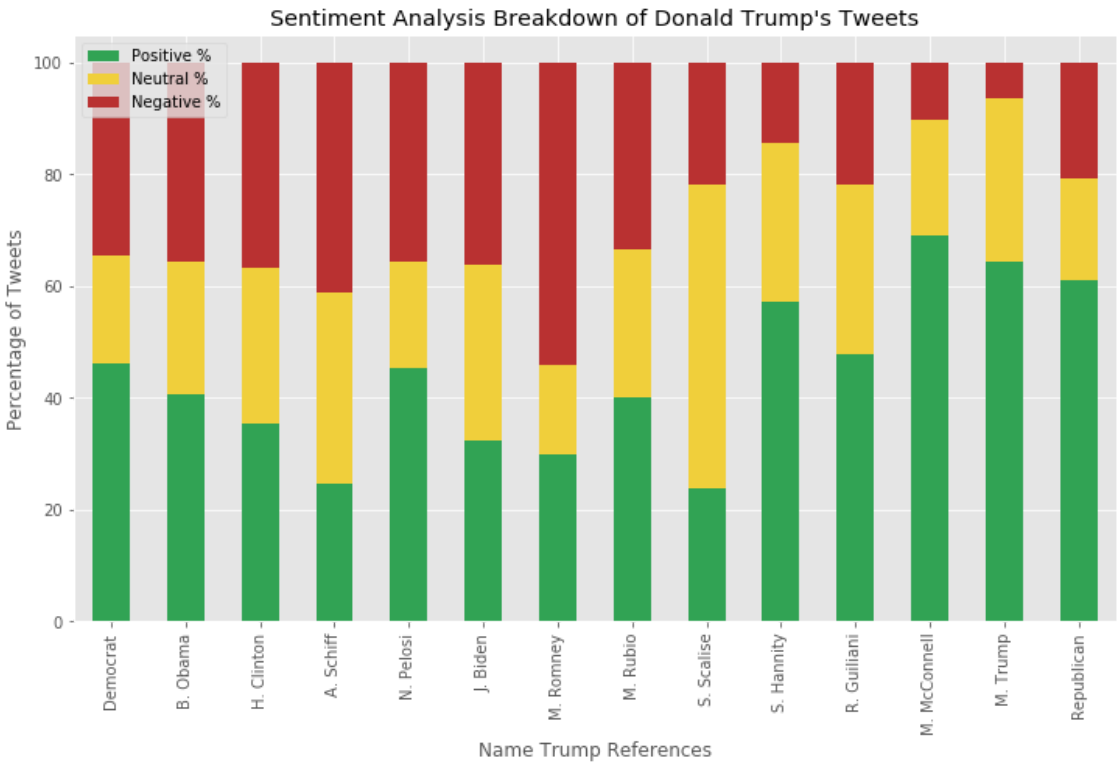

Figure 2. Stacked Bar Graph depicting sentiment analysis of Trump's tweets by keyword. Axes titles were modified from figure 1 to be more descriptive. The underlying dataset is the same for both figure 1 and figure 2.

Discussion

Our proposed analysis will give us an answer to the question: is there a relationship between the sentiment of adjectives in Donald Trump's tweets and the political ideology of the politicians he is tweeting about in his tweets since he announced his run for president in June 2015? However, the result we get must not be considered definite. The sentiment score can vary widely based on the sentiment analysis algorithm used because lexicons can differ between algorithms. Sentiment analysis as a whole is also limited because it is very bad at context, meaning it doesn't catch sarcasm. Tweets such as "We are doing much better than Obama!" will be interpreted as very positive since the token for "much better " is positive. Even if our data points to our H1, we will need to do a more in depth analysis to actually conclude if Trump tweets more positively about his allies due to the subjectivity of sentiment analysis expressed above. A correlation between positive adjectives in tweets and allyship to Trump may be found, but we cannot conclude "Trump tweets more positively about his allies" because our analysis doesn't take context into account. Trump is a great user of sarcasm, and this may create a lot of noise in our data. A statement that is actually negative could appear positive since, in general, negative sarcasm uses positive words. In our study, we group the politicians by their ideologies in reference to Donald Trump's and use this to look at and analyze the results, but it is possible that there is some confounding variable, such as ethnicity, gender, hometown/home country, that would explain the relationship more appropriately.

Our source of data is the Trump Twitter Archive, which catalogs every tweet Donald Trump makes in real time. The site reports that about 4,000 out of 42,000 Donald Trump tweets are missing from the

dataset. Also, the site stores any tweet that Donald Trump deletes, unless it was online for less than 30 minutes. Lastly, Trump is not writing every tweet so we can't say that "Trump tweets more positively about his allies," rather, " The Trump twitter account tweets more positively…" and that is a much less interesting conclusion. These three are the biggest candidates for sources of bias in the dataset. There is also a possibility we missed tweets about a politician in our data collection process due to nicknames, or may have accidentally included tweets not in reference to the politician, but simply happened to include their name. Most of the potential for bias/error in analysis comes from the TextBlob sentiment analysis algorithm. As seen in the graphs above, our sentiment analysis program actually evaluated Trump's tweets about Barack Obama to be slightly positive overall, which was much different from what our group was expecting. We looked into the 'data_2' dataframe containing the sentiment score matched with the tweet's text. We found that some tweets about Obama were as high as 0.8 on a -1 to 1 scale. There were many positive words in the tweet, but they were referring to the Trump administration being better than Obama, or were sarcastic. The TextBlob algorithm was not sophisticated enough to catch these nuances. Because of this and similar results, we determined that our sentiment analysis results should not be used with 100% confidence. To overcome this limitation, any time our research is used, it should be made clear that the sentiment analysis scores merely give us an idea of the trends of the data and are by no means 100% accurate. Our results should not be distributed to external sources since they could be misinterpreted and spread misinformation. We sought to mitigate the no-context bias created by the algorithm by maximizing the size of our dataset per politician. If we were to do this study again, a more sophisticated algorithm specifically trained on sarcasm and subject identification would be used.

From figure 2, the trends roughly support H1. Trump tweeted more positively about the republican party, McConnell, and Melania, and less positively political enemies in the republican party and democratic party. The data is withheld, but the number of tweets about a political enemy seems to correlate with % of positive tweets. We are not sure why this is, but it may have to do with an increase in sarcasm. % of neutral tweets seems to be relatively uniform across the distribution with Scalise as an outlier. The medium article from the background stated that overall, Trump tweets positively. Therefore, % negativity may be a better metric for evaluating how Trump tweets about his allies and enemies. From figure 2 Trump tweets much less negatively about his allies than his enemies. Mitt Romney had the greatest % negative tweets. This was not entirely unexpected since he was a vocal contender against Trump in the 2016 election, making him a dire enemy. By tweeting so negatively, Trump may have been trying to discredit the claims made by Romney and keep his base from considering Romney's point of view.

Group Participation

[redacted] checked initial assignment sections to make sure nothing was missing or incorrect, wrote up the analysis proposal for the final project and did extensive work on the visualization for our actual analysis during the extra credit portion. [redacted] proofread and made minor edits for the visualizations. [redacted] checked initial and final assignment sections to make sure nothing was missing or incorrect. [redacted] Wrote about ethical considerations, added a source to the background information, and wrote in the discussion for the final project. [redacted] came up with the initial project idea and experimental design. [redacted] did the data wrangling and exploratory data analysis, generated figure 1, created question, hypothesis + justification, first paragraph of background information, and made final edits for analysis proposal and discussion.