Final Project Draft - STA 199

due Thursday, Oct 29 at 11:59p

Ten Out of Ten: Arushi Bhatia, Luke Vermeer, Kevin Wang, Lauren May

Introduction and Data

Background and Motivation

Relative to other forms of media, social media plays a much greater role in how people consume news in today's technology-driven society. A study conducted in 2016 by Pew Research points out how 62% of people get news on social media (Gottfried & Shearer, 2016). As a result, social media plays an integral role in politics, as the presence of a specific subset of information on a user's feed can influence the way they categorize and see the world around them. With more and more individuals on social media such as Twitter, the role that this platform can play is significantly greater than before.

Social media, and Twitter in particular, have become an increasingly large part of the political landscape in the wake of Donald Trump's 2016 election. Trump has been active on Twitter prior to and during his presidency and uses the platform as a tool to communicate with his constituency in real time, posting updates about policy, campaigning, and his feelings on everything from members of Congress to celebrities. He is one of the first politicians to use social media this frequently and has personally referred to his use of Twitter as "modern day presidential" (Trump, 2017).

Donald Trump's Twitter also has "unpresidented" reach among the American public, boasting a follower base of over 87 million. This makes him the second most-followed political personality and sixth most-followed overall account on Twitter (Wikipedia, 2020). On top of this, Trump's Twitter also receives significant attention in the media. Over 850,000 news articles have referenced his Twitter use since 2016 and 31% of his Tweets since then have received individual media coverage (Real Clear Politics, 2019).

Research Questions

Because Trump uses Twitter to convey his political agendas in short blurbs, analyzing his Tweets can give a unique insight into the way that he thinks. Of principal interest were the following research question: 1) How does Donald Trump's sentiment in his Tweets vary across people? and 2) How does Donald Trump's sentiment in his Tweets vary across "hot topics?"

To answer these questions, our team obtained a dataset of Donald Trump's Tweets (n = 53,697). In order to assess how his sentiment varies across people and various hot topics, we created indicator variables for different people and topics to indicate that a Tweet referenced a specific person/topic. We then used the sentimentr package to calculate the overall sentiment of each Tweet using the sentiment_by() function. Using the indicator variables, we were then able to distinguish the Tweets into discernable categories, allowing us to conduct hypothesis tests and create interesting visualizations.

Our hypothesis regarding our research questions is that Trump's Tweets about individuals in his "outgroup" have a greater proportion of Tweets with a negative sentiment. Similarly, we hypothesize that his Tweets about individuals in his "ingroup" will have a greater proportion of Tweets with a positive sentiment. We will define ingroup and outgroup in different ways, and evalate it from different perspectives as well.

Our Data

The data set was extracted from a website - TrumpTwitterArchive.com. The original curator of the data created their own Twitter scraper in order to obtain the data. They utilized Python, Selenium (which is a software suite that allows the automation of tests utilizing web browsers), and Tweepy (a Python library for accessing the Twitter API). Since Twitter makes it challenging to scrape all of a user's Tweets in one go, the way to get around this is to individually search for a specific day and extract all the Tweets from that user on that specific day. To do this manually would take ages, but the scraper that the curator built allows for automated accessing for any desired day and also a range of days. The scraper then obtains the Tweet ID, which contains all of the metadata of the Tweet, and then uses the metadata to obtain all the other information about the Tweet (such as the text, timestamp, number of favorites, etc.). This other information is then compiled into a data set, which is made available to the public. This data set is updated every minute, which also means that deleted Tweets would most likely also appear in this data set.

This data set includes 53,697 observations. Each individual observation is one of President Donald Trump's Tweets. The original data set contains 7 variables: source, text, created_at, reTweet_count, favorite_count, is_reTweeted, id_str. The descriptions of each of the original variables is given below.

- source: Original source where Tweet was posted
- text: text of the Tweet
- created_at: Date and time the Tweet was posted/created, provides context
- reTweet_count: number of reTweets
- favorite_count: number of favorites
- is_reTweeted: whether or not the Tweet was originally posted on a different account and Trump reTweeted
- id_str: The scrape.py script collects Tweet ids. If you know a Tweet's ID number, you can get all the information available about that Tweet using Tweepy

Setting Up Our Data Our first step in analyzing the relationships between subject and sentiment of President Donald Trump's Tweets was to manipulate the data set, creating some new variables that we could use for analysis. We started by creating a set of identifier variables for different people that his Tweets might be about. These variables were created using multiple mutate commands to set them as either 1 or 0, 1 if the person or topic was mentioned in a Tweet, and 0 if they weren't.

We created these variables for a total of 11 relevant political figures: Barack Obama, Joe Biden, Hillary Clinton, Alexandria Ocasio-Cortez, Nancy Pelosi, Kamala Harris, Mike Pence, Mitch McConnell, Amy Coney Barrett, and Nikki Haley. For each of the person identifier variables, we searched for people's full names, their Twitter handles, and commonly-used nicknames to determine whether or not a particular Tweet was about them.

We also repeated the same process to create topical identifier variables for a range of subjects that are prevalent in the nation's political discourse. The topics were as follows: COVID-19, climate change, abortion, the Black Lives Matter movement, guns, news, immigration, Russia, and the United States. For each topic we searched for the name of the thing outright (such as "BLM" or "climate change"), as well as words and phrases that are commonly used in conjunction with these topics. For example, Tweets that mentioned ICE or the term "border wall" were categorized under immigration, and Tweets about CNN, FOX News and "media" all went under the "news" topic.

Next, using the pivot command in R, we were able to create a variable called "person," and manipulate our original data set into a data set that allowed us to count Tweets that had multiple people mentioned as an observation that counted towards both/all of the mentioned people, as it was a categorical variable that told us which of the identifier variables for people were equal to "1" for each Tweet. As a result, some of the Tweets were the same in this new pivoted data set, but it allowed us to do our analysis in a much more effective and representative manner.

From the new "person" variable, we created two new variables called party and gender. The party variable separated the 11 people we were looking at into either Democrats or Republicans and assigned Tweets about them to the appropriate party category. We did the same for the gender of each of these 11 people, and added another "gender" variable categorizing the Tweet as either about a "male" or "female."

We did a similar pivoting process for the topic variables as well.

Next, we used the package sentimentr to identify the sentiment of each of Trump's Tweets. For each Tweet, we created a new variable called ave_sentiment that contained the raw, numeric sentiment score of the text. This sentiment score was calculated using the sentiment_by() function. The function finds the mean sentiment score of the entire Tweet by averaging the individual sentiment scores of the words in the Tweet. We then used the mutate command to create a new variable called posNeg that grouped sentiment scores into three categories: positive, neutral, or negative. Finally, we used the separate command in r to break up the variable "created_at" into two separate variables for date and time. After this, each observation had a date variable in the mm/dd/yy format and a time variable in the 24-hour time format. After these manipulations, the data set contained the following variables:

- source: the version of Twitter used to share the Tweet. For example "Twitter for iPhone."
- text: the full text of the Tweet.
- created_at: gives the date (in mm/dd/yy format) and time (in 24-hour format) at which the Tweet was published.
- reTweet_count: number of reTweets that the Tweet received.
- favorite_count: number of favorites that the Tweet received.
- is_reTweet: assigns a value of either true or false for whether the Tweet is originally written by Trump or is a reTweet of someone else.
- id_str: The scrape.py script collects Tweet ids. If you know a Tweet's id you can get all the information available about that Tweet using Tweepy - text, timestamp, number of reTweets / replies / favorites, geolocation, etc.
- obama: 1 or 0 for whether the Tweet talks about Barack Obama.
- biden: 1 or 0 for whether the Tweet talks about Joe Biden.
- pelosi: 1 or 0 for whether the Tweet talks about Nancy Pelosi.
- kamala: 1 or 0 for whether the Tweet talks about Kamala Harris.
- hillary: 1 or 0 for whether the Tweet talks about Hillary Clinton.
- aoc: 1 or 0 for whether the Tweet talks about Alexandria Ocasio-Cortez.
- pence: 1 or 0 for whether the Tweet talks about Mike Pence.
- mcconnell: 1 or 0 for whether the Tweet talks about Mitch McConnell.
- fauci: 1 or 0 for whether the Tweet talks about Dr. Anthony Fauci.
- amy: 1 or 0 for whether the Tweet talks about Amy Coney Barrett.
- nikki: 1 or 0 for whether the Tweet talks about Nikki Haley.
- covid: 1 or 0 for whether the Tweet talks about COVID-19.
- climateChange: 1 or 0 for whether the Tweet talks about climate change.
- abortion: 1 or 0 for whether the Tweet talks about abortion.
- blm: 1 or 0 for whether the Tweet talks about the Black Lives Matter movement.
- guns: 1 or 0 for whether the Tweet talks about guns.

- news: 1 or 0 for whether the Tweet talks about news media.
- usa: 1 or 0 for whether the Tweet talks about the United States.
- russia: 1 or 0 for whether the Tweet talks about Russia.
- immigration: 1 or 0 for whether the Tweet talks about immigration.
- person: categorical variable for the name of the person that the Tweet is about (from among the 11 looked at in this analysis).
- party: political party of the person who is talked about in the Tweet (either democrat or republican).
- gender: gender of the person who is talked about in the Tweet (either male or female).
- ave_sentiment: numeric sentiment score of the Tweet.
- posNeg: categorical sentiment of the Tweet (either positive, neutral or negative).
- date: date that the Tweet was posted (mm/dd/yy).
- time: time that the Tweet was posted (24-hour format).

Methodology

The main variables in the data set that we used to address our research questions were the following:

- text: the full text of the Tweet.
- obama: 1 or 0 for whether the Tweet talks about Barack Obama.
- biden: 1 or 0 for whether the Tweet talks about Joe Biden.
- pelosi: 1 or 0 for whether the Tweet talks about Nancy Pelosi.
- kamala: 1 or 0 for whether the Tweet talks about Kamala Harris.
- hillary: 1 or 0 for whether the Tweet talks about Hillary Clinton.
- aoc: 1 or 0 for whether the Tweet talks about Alexandria Ocasio-Cortez.
- pence: 1 or 0 for whether the Tweet talks about Mike Pence.
- mcconnell: 1 or 0 for whether the Tweet talks about Mitch McConnell.
- fauci: 1 or 0 for whether the Tweet talks about Dr. Anthony Fauci.
- amy: 1 or 0 for whether the Tweet talks about Amy Coney Barrett.
- nikki: 1 or 0 for whether the Tweet talks about Nikki Haley.
- covid: 1 or 0 for whether the Tweet talks about COVID-19.
- climateChange: 1 or 0 for whether the Tweet talks about climate change.
- abortion: 1 or 0 for whether the Tweet talks about abortion.
- blm: 1 or 0 for whether the Tweet talks about the Black Lives Matter movement.
- guns: 1 or 0 for whether the Tweet talks about guns.
- news: 1 or 0 for whether the Tweet talks about news media.
- usa: 1 or 0 for whether the Tweet talks about the United States.
- russia: 1 or 0 for whether the Tweet talks about Russia.
- immigration: 1 or 0 for whether the Tweet talks about immigration.

- person: categorical variable for the name of the person that the Tweet is about (from among the 11 looked at in this analysis).
- party: political party of the person who is talked about in the Tweet (either democrat or republican).
- gender: gender of the person who is talked about in the Tweet (either male or female).
- ave_sentiment: numeric sentiment score of the Tweet.
- posNeg: categorical sentiment of the Tweet (either positive, neutral or negative).
- date: date that the Tweet was posted (mm/dd/yy).
- time: time that the Tweet was posted (24-hour format).

We created multiple visualizaitons to help us explore our data. In Figure 1, we created a word cloud of the most common words found in his Tweets. The most common words are displayed with the most frequent words being displayed larger. Given the results, we can see that some of the words he most commonly uses are "great", "Trump", "president", "people", "country", "America", and "time." While these words don't necessarily indicate the full scope of what he discusses on his Twitter account, we notice that most of his Tweets are in regard to his role as the President of the United States. We believed it would be interesting to see if a common theme would develop among his most common words, which is partially indicated in our results.

Similarly, Figure 2 is a frequency chart of his top 20 most used words in his Tweets. It is interesting to see how much more the word "great" is used in comparison to the other words displayed on the chart.

Figure 3 shows the relative proportions of Tweets with positive, neutral and negative overall sentiment that Donald Trump has shared about each of the people covered in our data set. By creating a percent stacked bar plot, we were able to actively compare the sentiments Trump tends to have towards specific people, making it easier to visualize the results. Overall, it appears that Alexadria Ocasio-Cortez, Hillary Clinton, and Nancy Pelosi recieved the highest proportion of Tweets with negative sentiment. Amy Coney Barrett had the lowest proportion of negative Tweets, followed by Vice President Mike Pence. Looking at positive Tweet proportions, Senate Majority Leader Mitch McConnell had the highest proportion of positive Tweets while Amy Coney Barrett, Mike Pence, and Nikki Haley also received a significant majority of Tweets with a positive sentiment. These results were particularly interesting because it shows that Trump tends to Tweet about people of the same party with positive sentiment at a greater proportion compared to the people of the Democratic party.

The purpose of Figure 4 is to explore the different sentiment Trump displays towards male versus female politicians. We decided to look at this based on party affiliation, because we believed that the results could be different depending if we were considering a Republican female politician versus a Democrat female politician (we just wanted to learn if he had a different sentiment towards strictly males or females, or whether it had to also do with politican affiliation). This visualization shows that when Donald Trump discusses Democratic politicians, he has a similar median sentiment towards male and female, which is fairly close to "neutral". We also noticed that for this visualization, the male Democrats had a larger range of sentiments than the female Democrats, but the lowest score was given to a female Democratic politician. In the visualization of male and female Republicans, the females had a slightly higher average sentiment compared to their male counterparts. This could indicate that he talks about males in his party in a less positive manner as compared to the females. We also recognized that, again, the male Republicans had a larger range of sentiments than the female Republicans, showing that his sentiments cover a larger range of positive and negative sentiments when talking about Republican males.

Figure 5 allows us to see the change in Donald Trump's sentiments towards specific "hot button topics" over time. We compiled all the Tweets into a line plot with time as the x-axis and average sentiment as the y-axis. By taking the average sentiment of each topic by year, we were able to create this visualization and see how his sentiments in his Tweets changed when discussing some specific topics.

Figure 1: Word cloud of the top 100 words Donald Trump uses in his Tweets. Some of the words that were not part of the actual text of the Tweet or were artifacts of Tweets (such as URLs, account names, and years) were removed to clean the data.

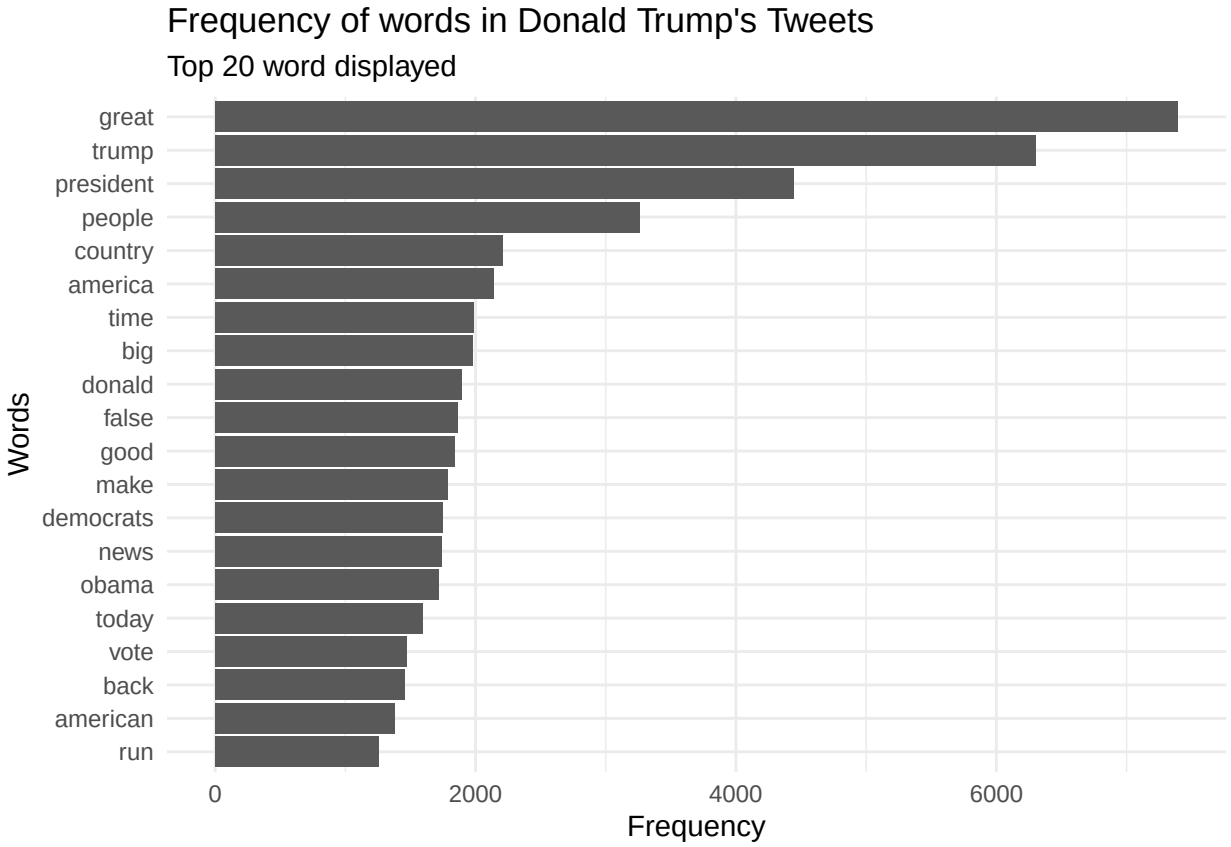

Figure 2: Frequency of Donald Trump's top 20 words that are used in his Tweets. Just like in Figure 1, some

of the words that were not part of the actual text of the Tweet or were artifacts of Tweets (such as URLs, account names, and years) were removed to clean the data.

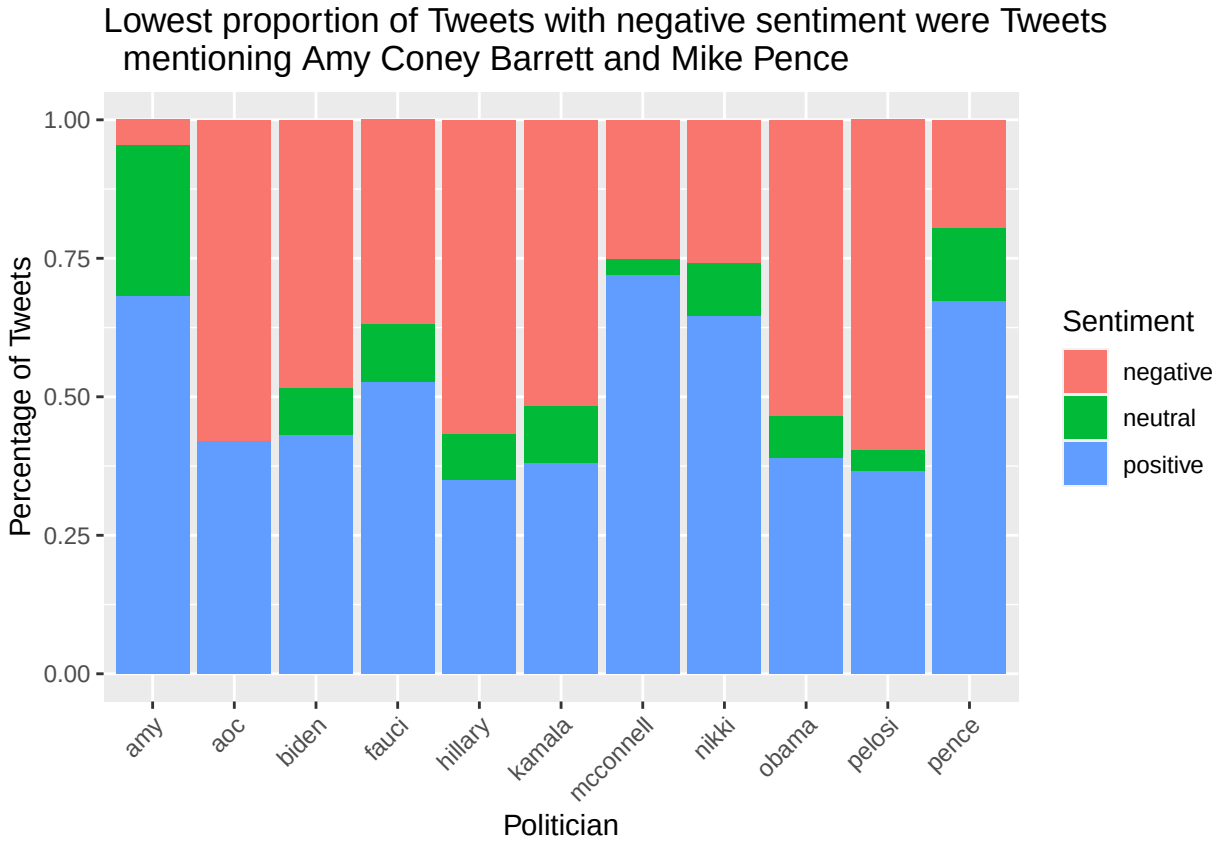

Figure 3: Illustrates the proportion of Tweets with average negative (<0), neutral (=0), and positive (>0) sentiment that mention a specific politician.

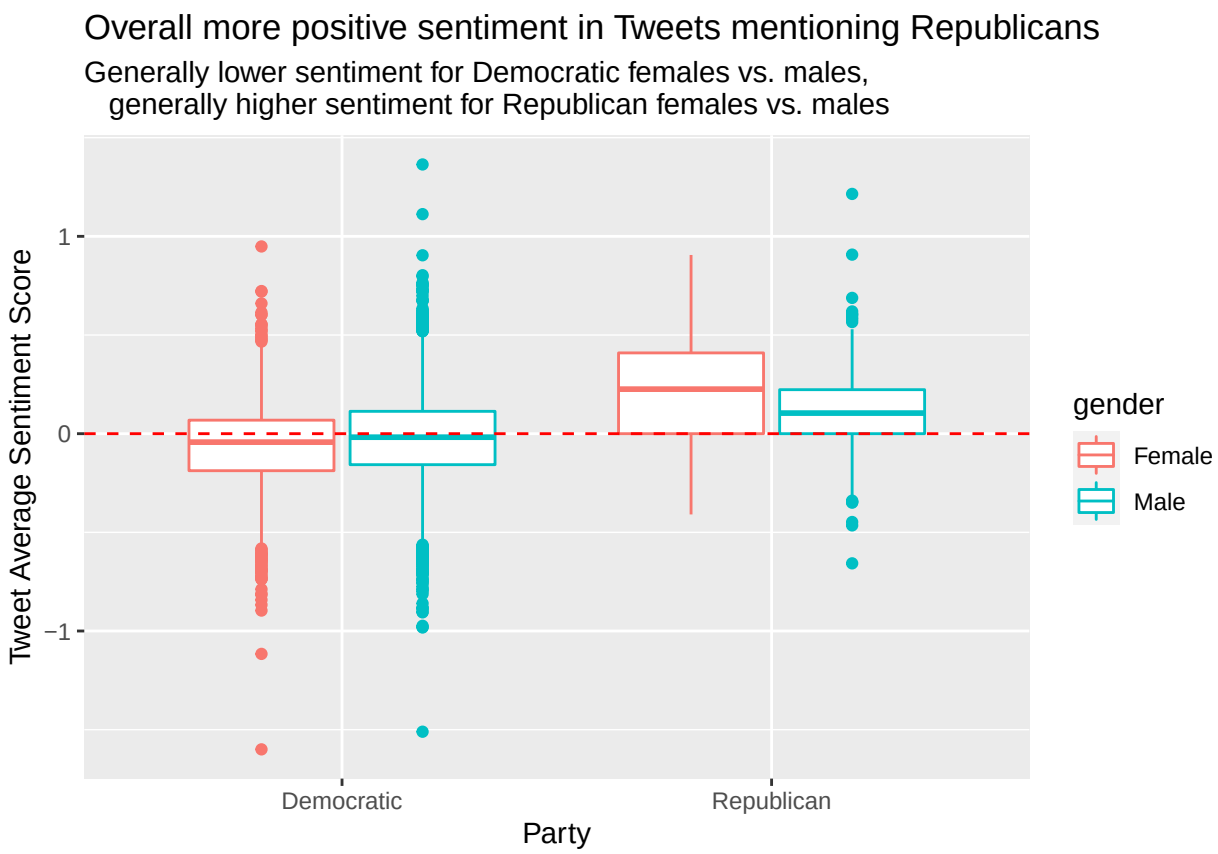

Figure 4: Boxplot showing the average sentiment scores for Tweets, divided by political party of the person mentioned in the Tweet, and split within party to show any differences in how he talks about politicians of different genders within either the Democratic or Republican party.

```
dems <- peopleData %>%
  filter(party=="Democratic")

repubs <- peopleData %>%
  filter(party=="Republican")

t.test(dems$ave_sentiment,
       repubs$ave_sentiment,
       mu = 0,
       var.equal = FALSE,
       alternative = "less",
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  dems$ave_sentiment and repubs$ave_sentiment
## t = -14.678, df = 469.92, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1414063
## sample estimates:
##   mean of x   mean of y
## -0.03922378  0.12006845
```

```r
male <- peopleData %>%
  filter(gender=="Male")

female <- peopleData %>%
  filter(gender=="Female")

t.test(male$ave_sentiment,
       female$ave_sentiment,
       mu = 0,
       var.equal = FALSE,
       alternative = "greater",
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  male$ave_sentiment and female$ave_sentiment
## t = 6.6948, df = 3900.6, p-value = 1.234e-11
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.03061738        Inf
## sample estimates:
##   mean of x   mean of y
## -0.01642442 -0.05701745
```

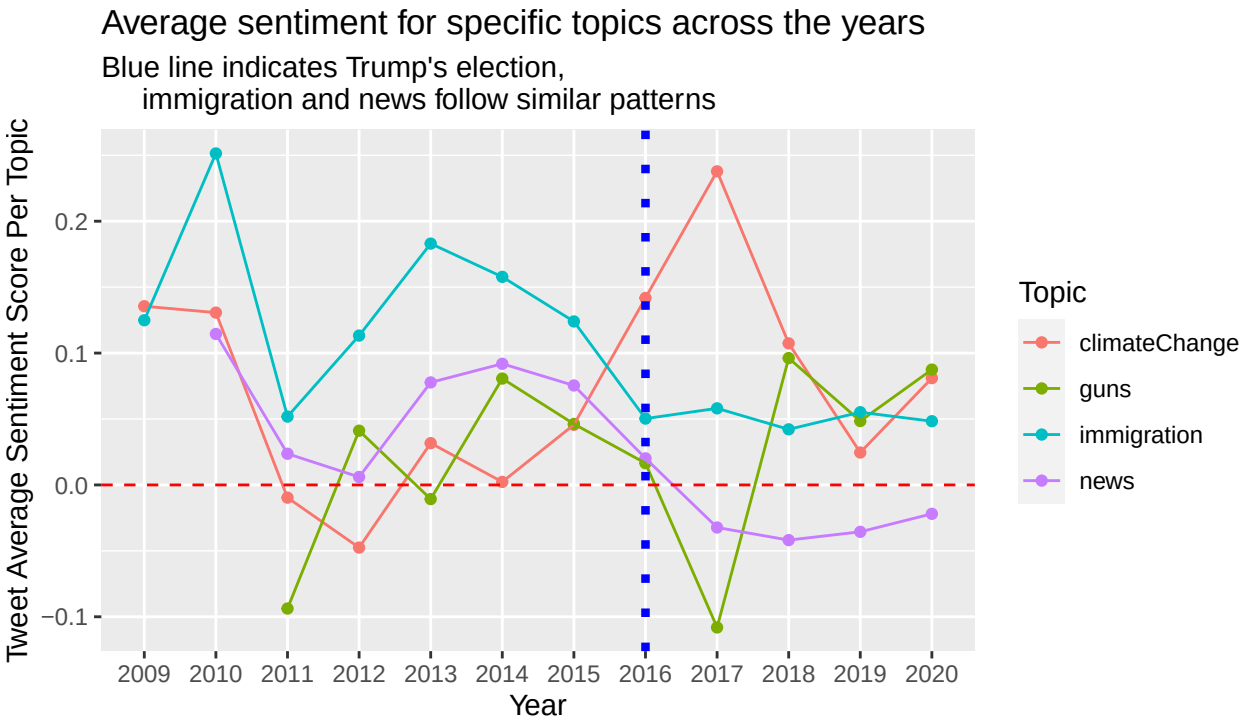

Figure 5: Shows change of average sentiment of Tweets referring to a specific topic across the years. Only showing a few topics, we selected the ones that have been Tweeted about for a long time. The vertical blue line indicates when Trump first became elected.

```r
yearVar <- TweetsWithSentimentDateTimeFormat  %>%
   mutate(year = year(date))
```

```r
#CLIMATE CHANGE
pre2016CC <- yearVar %>%
  filter(year==2016 & topic=="climateChange")

post2016CC <- yearVar %>%
  filter(year>=2016 & topic=="climateChange")

t.test(pre2016CC$ave_sentiment,
       post2016CC$ave_sentiment,
       mu = 0,
       var.equal = FALSE,
       alternative = "two.sided",
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  pre2016CC$ave_sentiment and post2016CC$ave_sentiment
## t = 0.72595, df = 26.822, p-value = 0.4742
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06757291  0.14153277
## sample estimates:
## mean of x mean of y
## 0.1417982 0.1048183
```

```r
#GUNS
pre2016guns <- yearVar %>%
  filter(year==2016 & topic=="guns")

post2016guns <- yearVar %>%
  filter(year>=2016 & topic=="guns")

t.test(pre2016guns$ave_sentiment,
       post2016guns$ave_sentiment,
       mu = 0,
       var.equal = FALSE,
       alternative = "two.sided",
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  pre2016guns$ave_sentiment and post2016guns$ave_sentiment
## t = -1.157, df = 27.246, p-value = 0.2573
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1423921  0.0396828
## sample estimates:
##  mean of x  mean of y
## 0.01647441 0.06782909
```

```r
#IMMIGRATION
pre2016imm <- yearVar %>%
  filter(year==2016 & topic=="immigration")
```

```r
post2016imm <- yearVar %>%
  filter(year>=2016 & topic=="immigration")

t.test(pre2016imm$ave_sentiment,
       post2016imm$ave_sentiment,
       mu = 0,
       var.equal = FALSE,
       alternative = "two.sided",
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  pre2016imm$ave_sentiment and post2016imm$ave_sentiment
## t = 0.021915, df = 297.88, p-value = 0.9825
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03226623  0.03299297
## sample estimates:
##  mean of x  mean of y
## 0.05030422 0.04994085
```

```r
#NEWS
pre2016News <- yearVar %>%
  filter(year==2016 & topic=="news")

post2016News <- yearVar %>%
  filter(year>=2016 & topic=="news")

t.test(pre2016News$ave_sentiment,
       post2016News$ave_sentiment,
       mu = 0,
       var.equal = FALSE,
       alternative = "two.sided",
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  pre2016News$ave_sentiment and post2016News$ave_sentiment
## t = 3.3319, df = 559.89, p-value = 0.0009195
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01832116 0.07094526
## sample estimates:
##   mean of x   mean of y
##  0.02015388 -0.02447932
```

Results

(Chi Square Test) The variable, polarization, is defined as how intense the sentiment of Donald Trump's Tweet is. This hypothesis test will analyze if the polarization of the Tweet and the amount of reTweets are dependent or independent of each other. We chose to analyze this because we believed it to indicate the public's response to his Tweets. We can draw the conclusion that the public reaction is dependent on how intense of a sentiment Trump expresses if we reject the null hypothesis.

We are performing a hypothesis test using a chi-square test. We are performing this hypothesis test at the 0.05 significance levels. Here are our null and alternative hypotheses:

H_0: The polarization of the Tweet and the amount of reTweets are independent; there is no association between the two variables.

H_1: The polarization of the Tweet and the amount of reTweets are NOT independent; there is an association between the two variables

Under the assumption that the null hypothesis is true, the chi-square test statistic follows a Chi-square distribution with degrees of freedom equal to 39554.

The value of our test statistic, chi-square, is 41580.

The p-value of our chi-square test was 6.949e-13, which is less than the significance level of 0.05. This means we reject the null hypothesis. We are concluding that there is sufficient evidence to suggest that the presence of throat pain and ASA classification are NOT independent or that there is sufficient evidence to suggest that there is an association between the two variables. More specifically, this hypothesis test shows that the amount of reTweets are dependent on how polarizing President Trump's Tweets are.

Sources

http://www.trumptwitterarchive.com/about

https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms- 2016/

https://www.realclearpolitics.com/articles/2019/09/11/numbers_show_how_trumps_Tweets_drive_the_news_cycle_141217.html

https://en.wikipedia.org/wiki/List_of_most-followed_Twitter_accounts

https://twitter.com/realDonaldTrump?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7 Ctwgr%5Eauthor