Final Project Proposal - STA 199

due Friday, Oct 9 at 11:59p

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

Introduction

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).

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. The research question we will be exploring is how President Trump's sentiment in his Tweets and its reception by his large Twitter following (popularity) varies across people, organizations, and policies. Our hypothesis regarding our research question is that Trump's tweets have negative sentiments towards opposing groups and policies and positive sentiments towards groups and policies that align with his beliefs. In addition, we believe the tweets with negative sentiments will be slightly more popular in terms of favorites and retweets.

Data description

The dataset 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 dataset, which is made available to the public. This dataset is updated every minute, which also means that deleted Tweets would most likely also appear in this dataset.

This data set entitled trumpTweets includes 53,697 observations. Each individual observation is one of

President Donald Trump's tweets. The original dataset 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 - text, timestamp, number of retweets / replies / favorites, geolocation, etc.

Glimpse of data

```r
library(tidyverse)
trumpTweets <- read_csv("data/trumpTweetCSV.csv")
glimpse(trumpTweets)
```

```
## Rows: 53,697
## Columns: 7
## $ source         <chr> "Twitter for iPhone", "Twitter for iPhone", "Twitter...
## $ text           <chr> "https://t.co/g5lGbRG4aE", "Will be interviewed by @...
## $ created_at     <chr> "10/8/20 12:08", "10/8/20 11:47", "10/8/20 2:57", "1...
## $ retweet_count  <dbl> 10524, 7434, 71908, 22541, 26360, 28868, 12852, 2362...
## $ favorite_count <dbl> 41050, 38386, 470060, 98083, 89625, 76854, 44386, 10...
## $ is_retweet     <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
## $ id_str         <dbl> 1.31418e+18, 1.31417e+18, 1.31404e+18, 1.31404e+18, ...
```

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%7Ctwgr%5Eauthor