Twitter Data Analysis - Final Report

Emre Yurtbay, Shashank Mahesh, Ryan Carlson

4/12/2019

Introduction

Our dataset is a collection of tweets we collected during the midterm election cycle in 2018. While much of our analysis focuses on the Florida senator election race between Ron DeSantis (R) and Andrew Gillum (D), we also collected tweets from the Kansas and Texas senator elections to bolster our dataset. The fields we have included user count, screen name, description, tweet text, like count, retweet count, source, friend count, and many more. We obtained the data ourselves using the twitter API and a continuous crawler. All we have to do is specify a set of hashtags and text, and we then recieve a JSON object that contains various information about tweets.

There are a number of questions we would like to answer about our data. First of all, we would like to learn about how Twitter works and how people tweet. What can we say about the relationships between retweets, favorites, and followers? What is the distribution of follower counts of twitter users? Do people who tweet more also have more followers? Do tweets follow Zipf's law? Next, we wanted to explore what kind of political questions we could answer using tweets. Can we find which topics dominate the discussion? Can we do time series anaysis to discover the times people tweet the most? Who are major influencers in the twitter network?

To supplement our election related tweets, we also used web scraping to collect polling data from RealClearPolitics.com. We gathered polling data from multiple states, including Florida, Georgia, Nevada, and Arizona. By tracking the polls, we can see how candidates are performing at different points in time, seeing how the momentum shifts and changes.

Twitter Analytics: How do Twitter Users interact with the Platform?

A Linear Regression between Follower Count and Tweet Count

The first thing we wanted to check was whether or not there was a relationship between follower count and tweet count. If we know how many tweets a user has sent out, is there any way we can predict how many followers they have? A good way to do this is to make a linear regression model, using the *lm* command in R. We subsetted our data to only include users with less than 500,000 followers, since very famous twitter accounts skew results quite a bit. By subsetting the data, we can get a more representative slice of the platform.

term	estimate	std.error	statistic	p.value
(Intercept)	3045.0135065	82.9148692	36.72458	0
Number of Tweets	0.0313364	0.0008175	38.33153	0

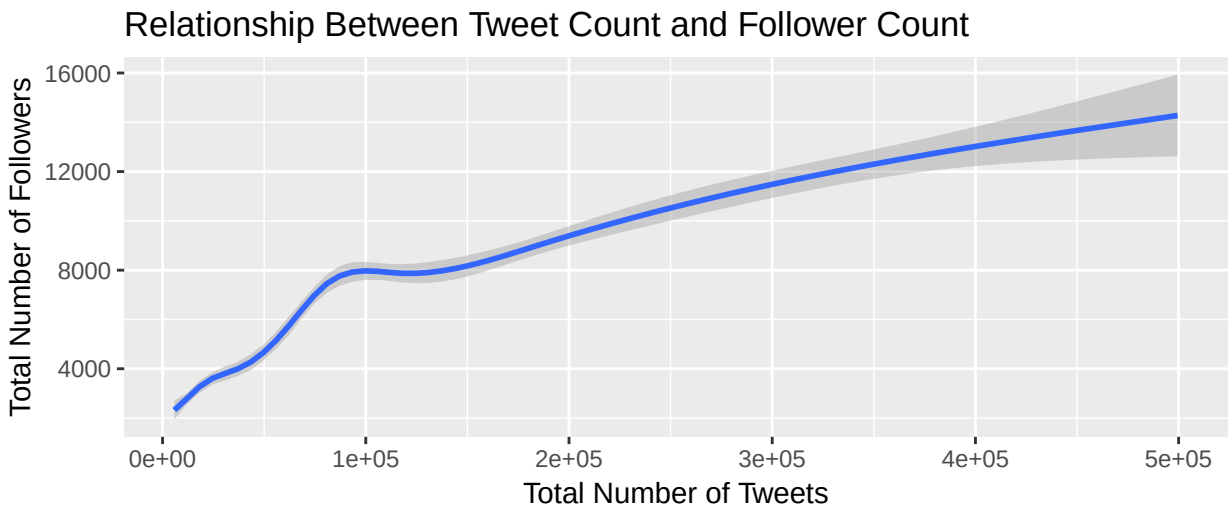
Hence, we could fit the model

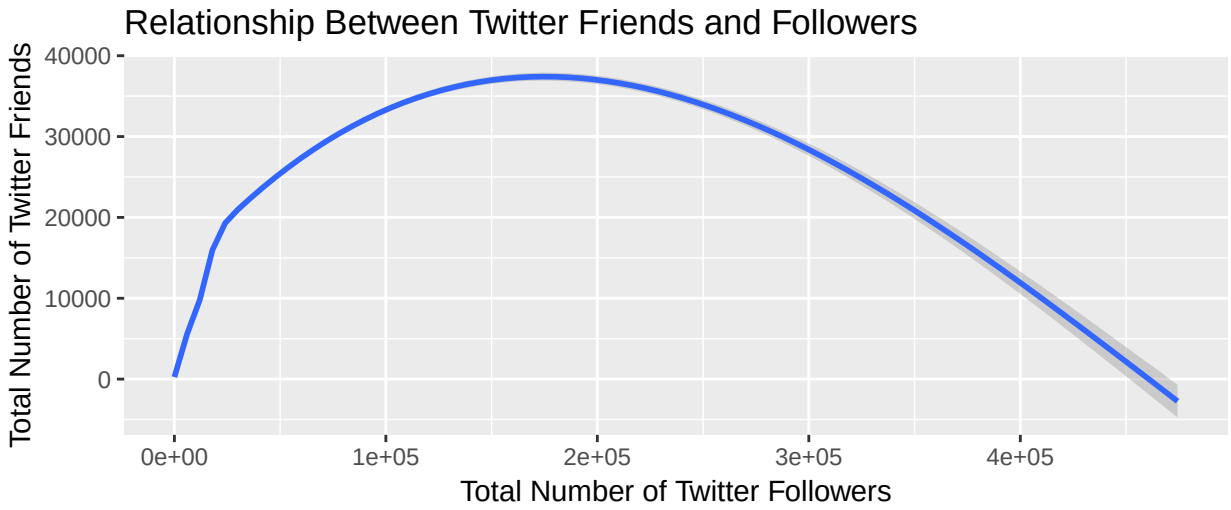
$$y = 0.03x + 3045.012$$

to our data, where x represents the number of tweets and y represents the number of followers. As one would expect, those who tweet more are also more likely to have more followers. With every tweet sent out, the user can expect more engagement and hence, they will reach a large audience. A larger audience suggests the

potential for more followers, regardless of the contoversality of the content. A graphical representation of this relationship can be found on the next graph. Note that R plots a smoothing spline instead of a line, which is a slightly more complex, yet related, linear model.

What is the Relationship Between Friends, Followers, and Number of Tweets

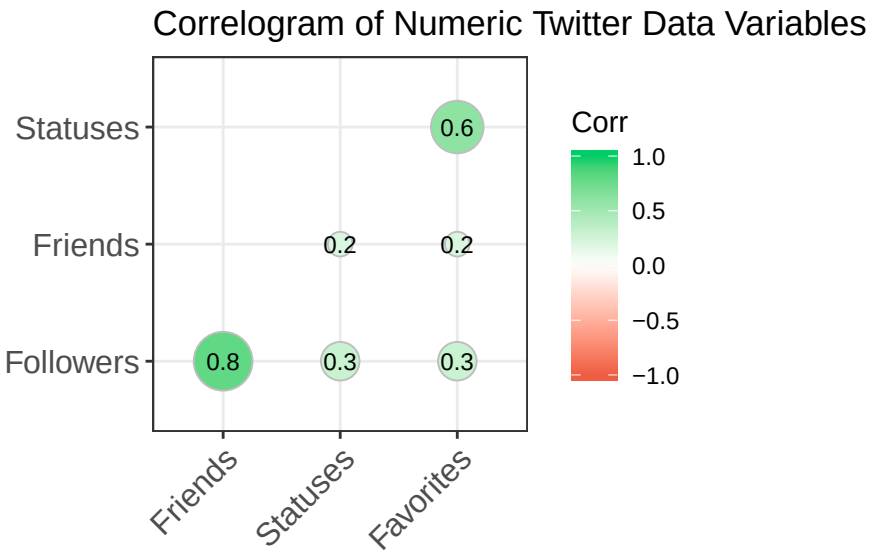




As was explained earlier, the more tweets somebody sends out, the more foloowers they are likely to have. Intrestingly, we see a different relationship between the number of followers a twitter user has and the number of "friends" they have. You are a "friend" with a twitter user if you follow them and they follow you back. As you gain more followers, your number of friends also tends to increase, up to a point. Then, as your follower count increases, the number of friends you have decreases. This implies that the ultra-famous follow very few of their followers back.

How are certain Variables Correlated?

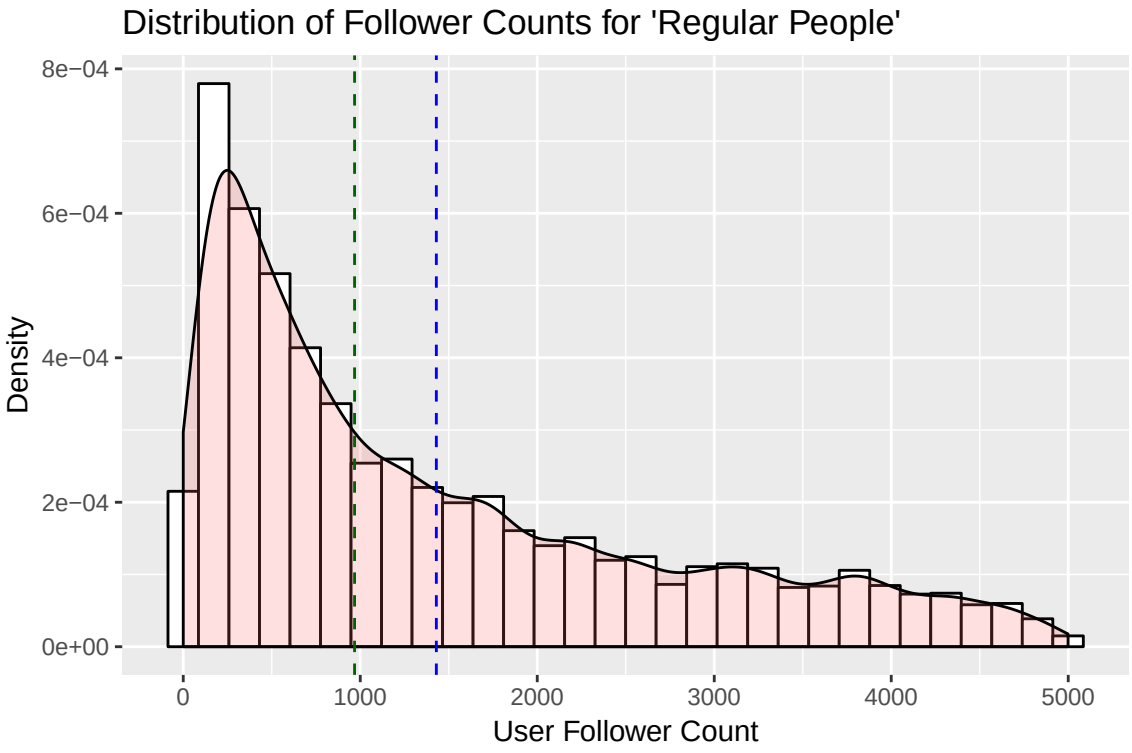
A quick way to see relationships between variables in our dataset is to plot a correllegram. If two variables have a high correlation, that means that they vary together strongly. As we can see in the plot below, the variables Friends and Followers have a very high correlation (0.8), while favorites and friends are hardly correlated at all (0.2)



Distribution of Follower Counts

The vast majority of twitter users are not famous people, and the majority of political "tweeters" are normal people making their opinions heard on a public forum. In fact our analysis shows that the median twitter user in our dataset has just a shade over 1460 followers. What does the distribution of followers for regular people look like? Below, we show a denisty plot of follower counts for twitter users with less 5000 followers. The dark green line represents the median and the blue line represents the mean. The mean being greater than the median shows that the distribution is heavily right skewed.

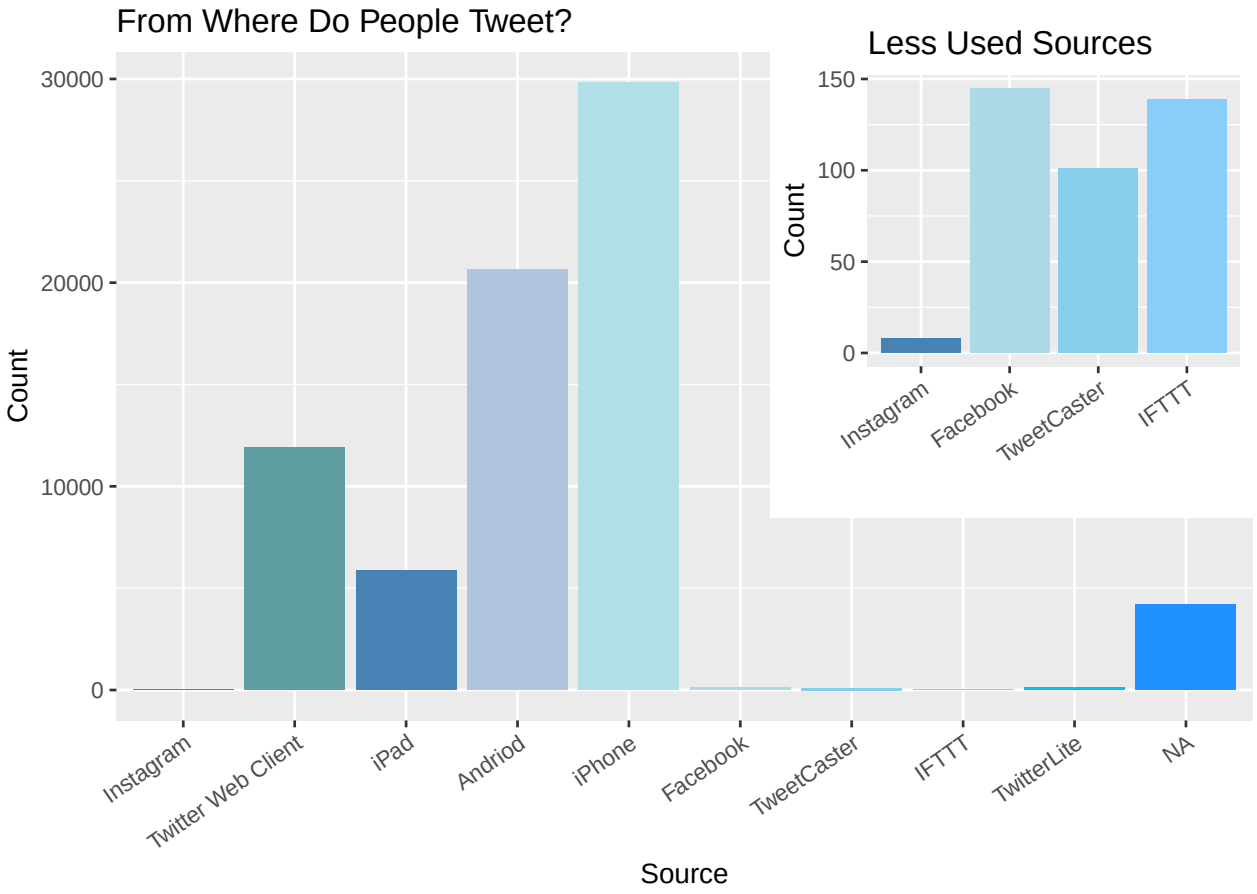
.



From What Apps and Devices do People Tweet?

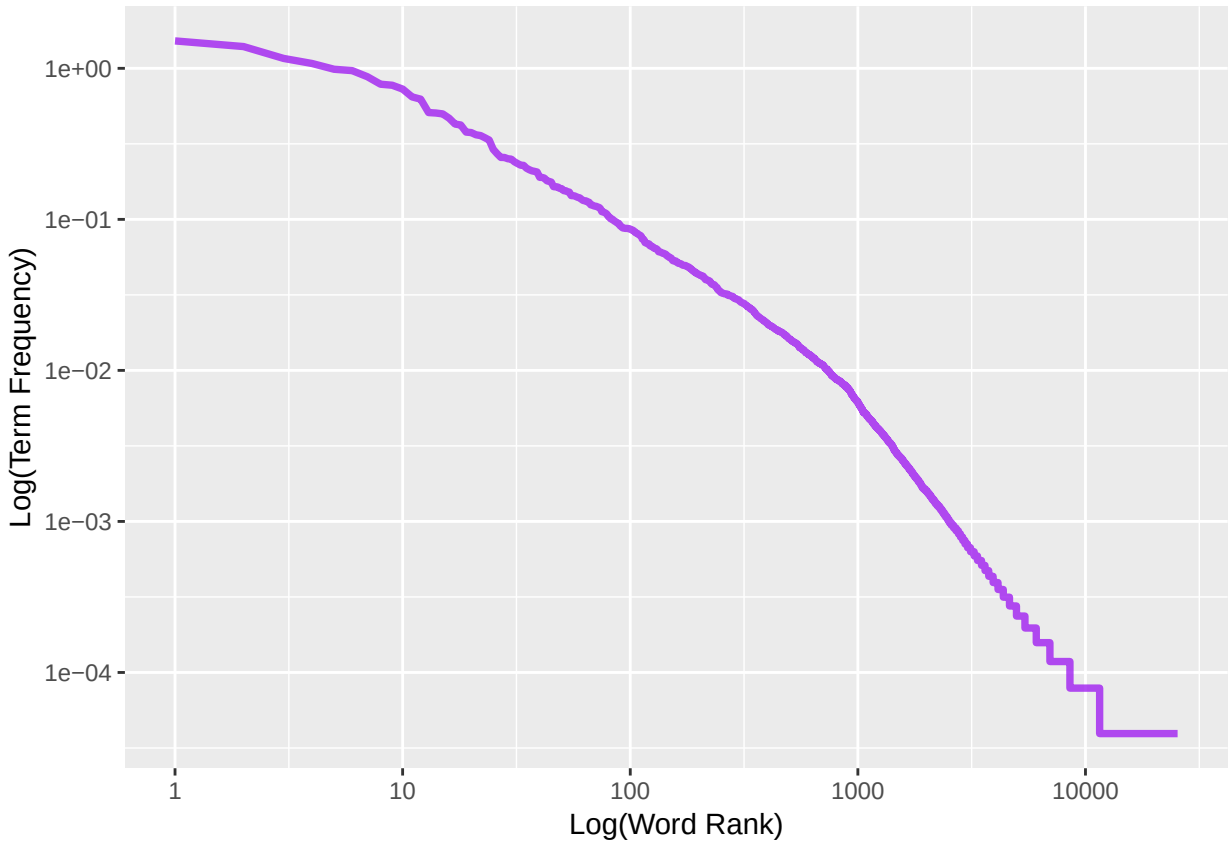
Here, we see a barplot showing the sources from which people tweet the most. It appears most people are tweeting from their iPhones, which is probably to be expected. The next most used is the Andriod Twitter App, while a smaller percentage are using iPads and the Twitter Web Client. The various other sources are Andriod specific twitter clients. This chart confirms the popularity of the iPhone in the United States, but it shows that other media sources are not far behind.

.



Do Tweets Follow Zipf's Law?

A common thing to check when dealing with natural language is to see if your corpus folows Zipf's law. Zipf's law states that the frequency that a word appears is inversely proportional to its rank. That is, the the second most common word in a corpus should appear half as much as the most common word, the third most common word should appear about a third as much as the most common, and so on. On a log-log scale, we should see an approximately straight line when we plot word rank vs term frequency, since an inversely proportional relationship will have a constant, negative slope. The corpus we put together consists of all the text of the tweets we collected. The purple line is the term frequency/rank relationship our data actually show. The deviations we see here at high rank are not uncommon for many kinds of language; a corpus of language often contains fewer rare words than predicted by a single power law. It seems that our tweets follow Zipf's law.



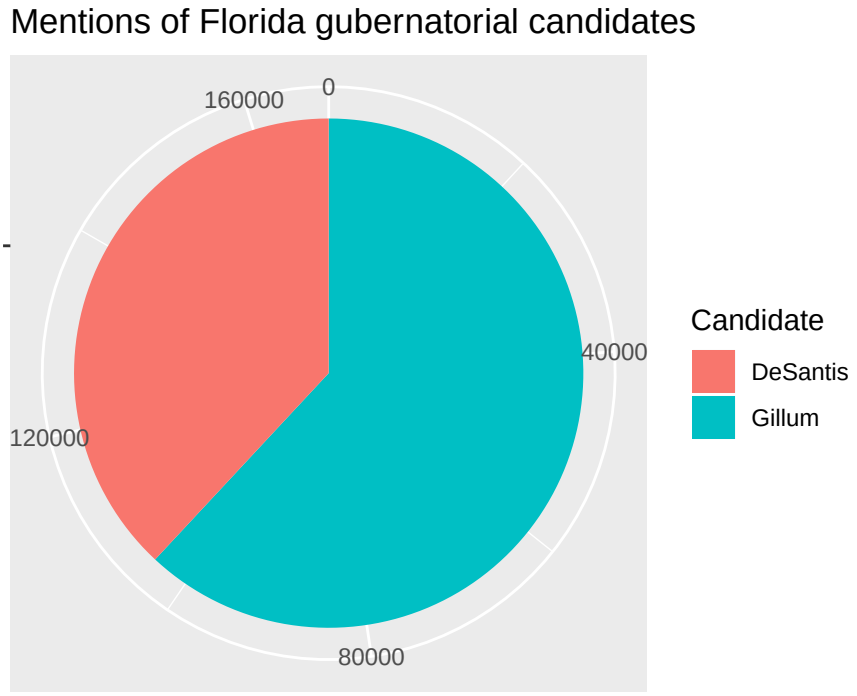
Answering Political Questions Using Twitter and Polling Data

The majority of our twitter analysis focuses on the Florida gubernatorial race between Ron DeSantis (R) and Andrew Gillum (D). The election was hotly contested and the subject of much national discourse. Using twitter and polling data, we wanted to answer some questions about the elections.

Who has a larger Twitter presence: DeSantis or Gillum?:

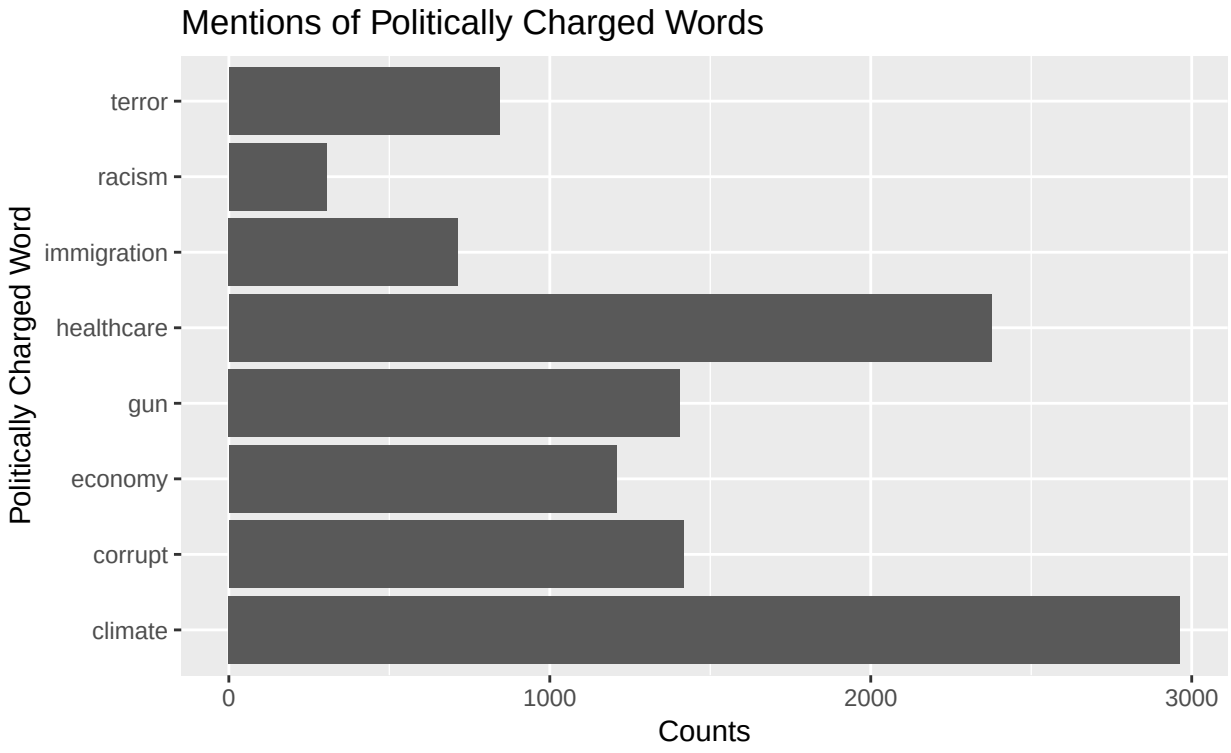
A quick gauge of how much support a gubernatorial candidate may be garnering can be seen in how often they are mentioned in tweets. We counted the amount of times the strings "Gillum" or "DeSantis" occurred in our dataset. From the pie graph below, it seems that the democratic candidate, Gillum, was far more mentioned than his opponent DeSantis. This may also be due to twitter's well documented left-leaning bias. Twitter skews young, urban, and educated - these groups also tend to vote liberal. Because of this, we may expect to see many more tweets talking about the democratic candidate as compared to the Republican. Our analysis seems to confirm this bias.

Candidate	Mentions
DeSantis	63958
Gillum	104064



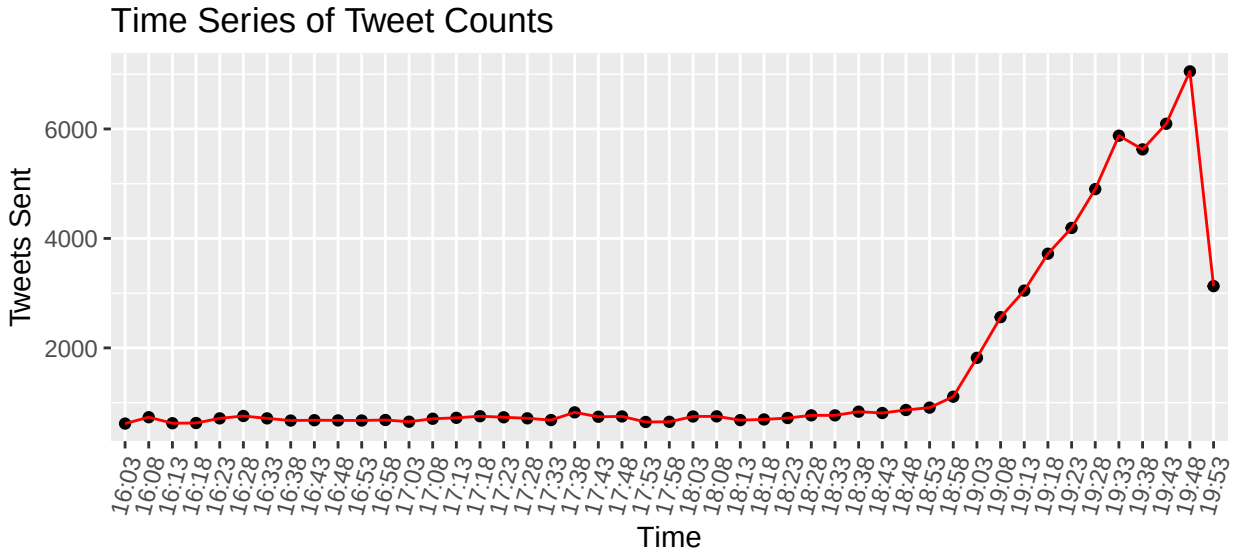
Which Topics Dominate the Discussion

Below is a collection of some politically charged words, summarized by how often they are mentioned in the twitter corpus. In Florida, climate change and healthcare seem to be some of the biggest issues. Intrestingly, corruption seems to be a major issue, which perhaps reflects Andrew Gillum's corruption scandal when he was mayor of Tallahassee.



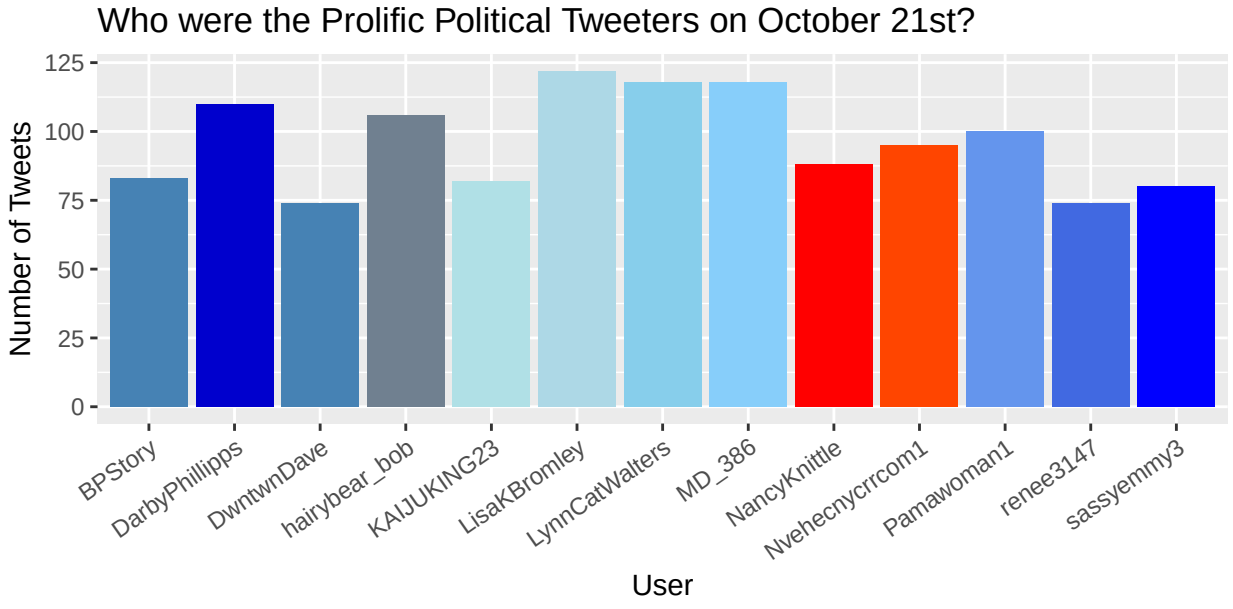
When are people tweeting?

For this plot, we solely focus on tweets we collected on tweets collected on October 21st, the day of the debate between Ron DeSantis and Andrew Gillum. This time series shows the number of tweets fired out in 5 minute periods in the hours leading up to the debate. Political tweet traffic is pretty slow in the hours leading up to the debate, but skyrocket as the debate approaches closer, and hits a peak right before the debate starts.



Does anybody Tweet Disproportionately?

In this partcular set, we are looking at some of the most prolific political tweeters on October 21st, the day of the Florida Midterm Election Debates. Certain users are dominating the political conversation, most notablely @LisaKBromley and "@LynnCatWalters". By looking at the accounts, we can try to map the user's political allegiance. Most of the accounts have their political leanings right in their bio, making these users rather easy to label. @BPstory tweets very liberal content, so we labeled their bar blue. @NancyKnittle tweets pro-Trump memes regularly, so it is safe to label them as a Republican. As we can see, pro-Gillum tweeters dominated the platform, with the majority of high volume users identifying as liberal.

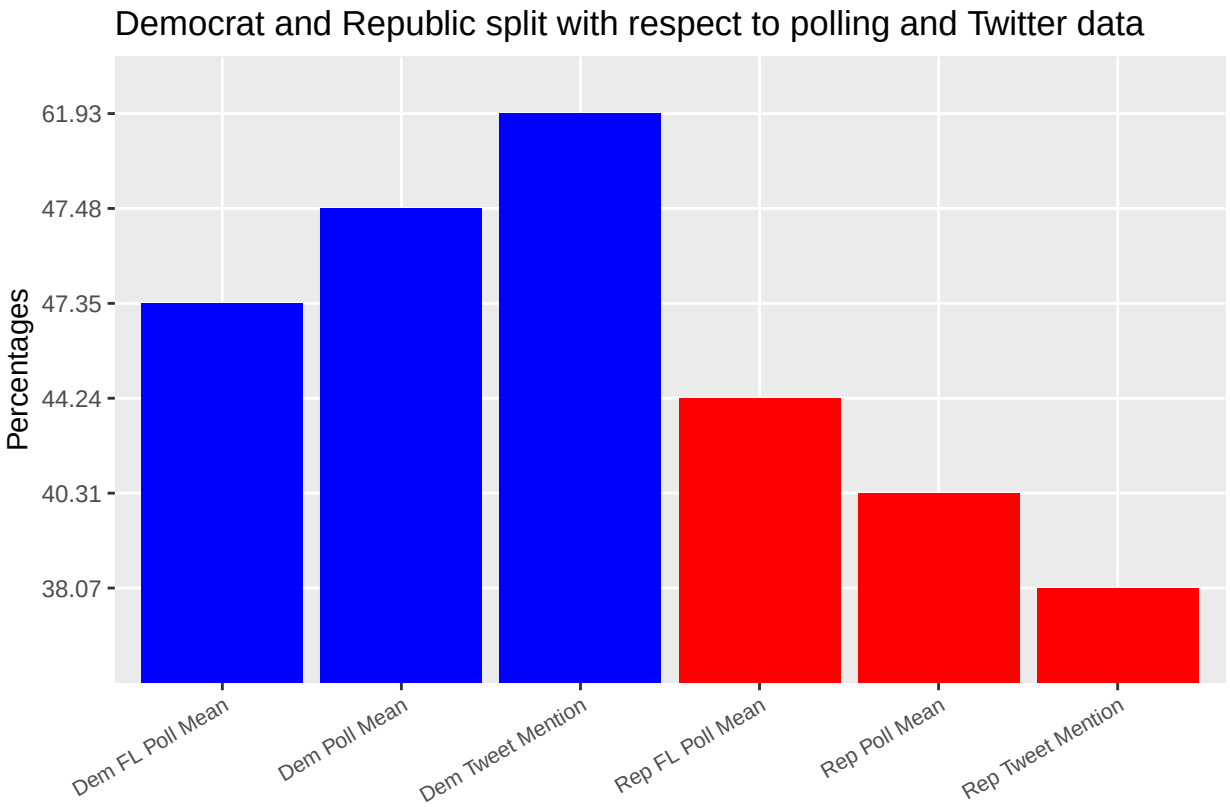


What can we learn from the polling data

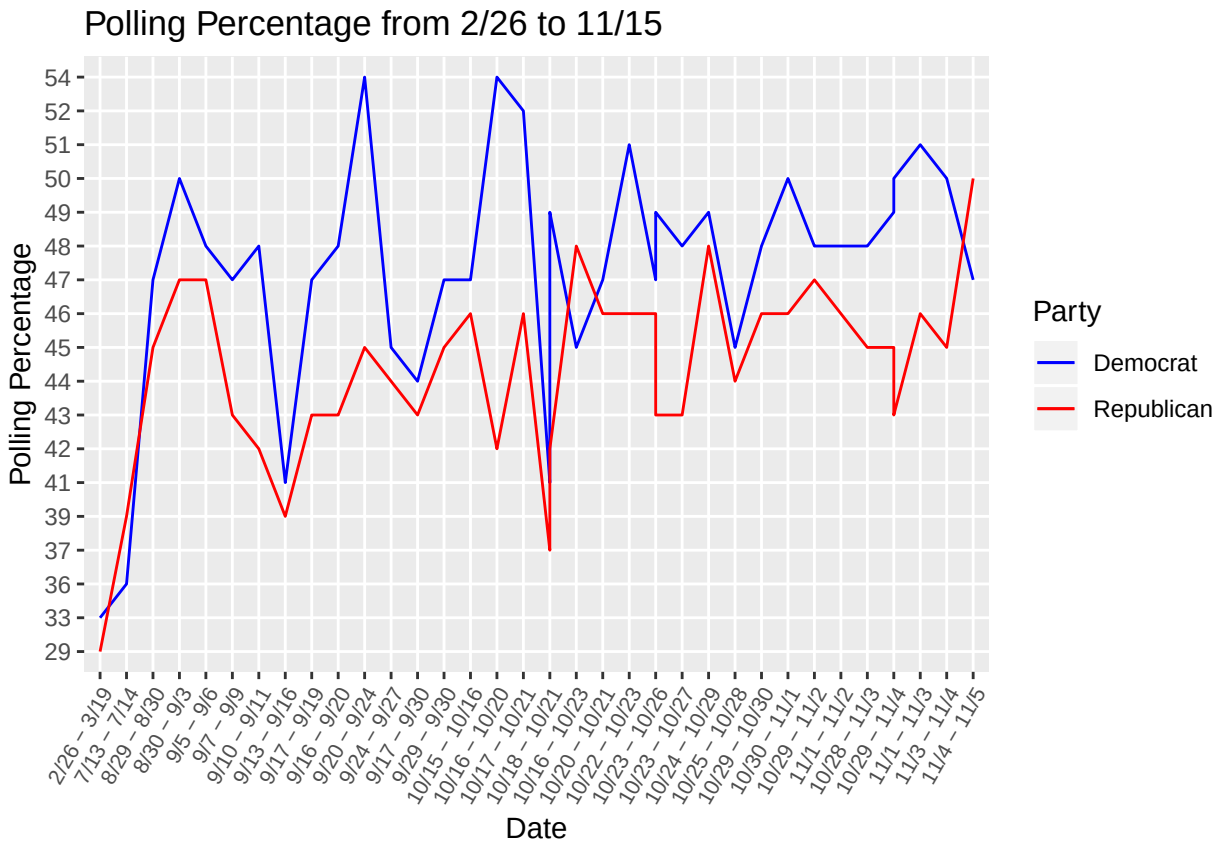
One interesting question we can ask with our data is how tweets around political candidates represent election results and polling results. Can twitter be used as a data source similar to a poll for election prediction? It is difficult to tag a particular tweet as "republican" or "democrat", so we looked at the mentions of the democratic and republican candidate and took the average number of tweets mentioning these candidates when candidates were mentioned. Here, we are assuming activity surrounding a politcal candidate can be considered equivalent to endorsement of said candidate, or at least can be compared to polls.

We plot the average polling results for republicans and democrats from the data we scraped and the average polling results just for Florida. Alongside these, we also plot the average mentions of candidates for the Florida election. We can see from a quick visual that the tweet mentions of democrats and republicans are very different from the national and state polling averages. Florida election polling also seems to be different from the national average, with less support for the democrats. This may mean that twitter has an outsized democratic support or that our assumption about activity surrounding a candidate is an incorrect measurement technique. Either conclusion is interesting and both warrant further investigation.

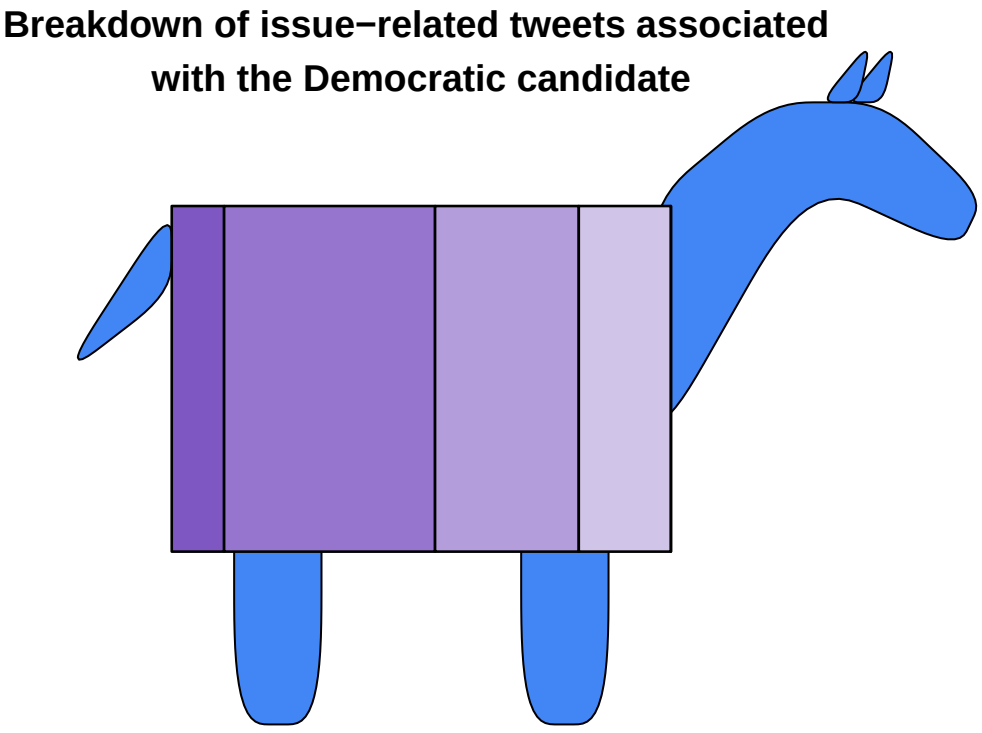
.

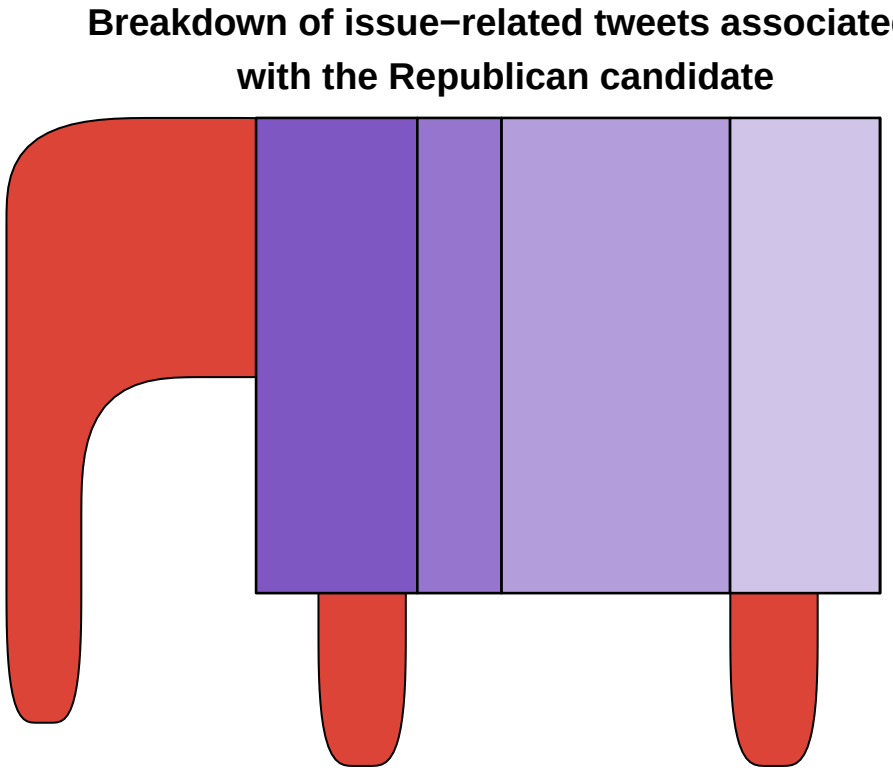


Just looking at the polls also can tell us quite a lot about how the election would go. We collected polling data from February the 26th to November 5, a few days before the election. As we can see, the Democrats had a substantial lead, but the gap closed completely in the days leading up to the elections. The strong democrat lead reflects the large amount of Andrew Gillum tweets we see in our dataset. However, the polling numbers related to the Republican's strong surge at the beginning of November suggests that they may have the edge in the election.



Killer Plot - Showing breakdown of Issues by Party





In our killer plot, we wanted to show the breakdown of issues by party. The democratic party is represented by the donkey and the republican party is represented by the elephant. The size of the box represents the importance a certain topic had in the political convresation - if it was mentioned more, its correpsonding box has more size. We classified tweets as republican if they mentioned "DeSantis" and democrat if they mentioned "Gillum." From left to right, the issues we looked at are gun control, climate change, healthcare, and the economy. For the democrats, climate change is much more important than for the republicans. In contrast, guns play a much bigger roll in the conversation for the republicans than the democrats. This shows is that the importance of certain issues are very different depending on the party. To make the plot more "killer", we have also added shiny compatibility to show how the conversation changes per day.

Conclusions

By completing this analysis, we were able to answer many of our questions regarding how people interact with Twitter. As we saw, there is strong positive correlations between twitter metrics like followers, favories, statuses, and friends. Using this fact, we were able to create a model to predict the number of followers a twitter user has given the number of statuses the user has. We learned that follower distribution is highly right skewed. We also saw the tweets follow Zipf's law, meaning tweets have many of the properties of conventional natural language.

Also, we wanted to answer some political questions with our twitter data. By showing that democratic Twitter traffic was much higher than that of the republicans, we suggest that Twitter has a left-leaning bias. However, we saw that the topics that dominate the political conversation include healthcare and climate change. Further, we saw that much of the prolific tweeters were left leaning.

For future consideration, we might want to consider how many of these tweets are produced by bot accounts and how they influence the discussion. Furthermore, we may want to perform some senitiment analysis in order to better classify tweets as democratic or republican.