Congressional Use of Twitter

Midterm Report

By Jessica Su and Sohini Khan

**Introduction**

Twitter is a microblogging site with 500 million users. [1] Users may post “tweets,” which are messages of 140 characters or less. They may post “retweets,” which are copies of others’ tweets. They may also “follow” users, which causes their tweets to show up in their newsfeed.

Since Twitter has such a large userbase, some members of Twitter have amassed a large number of followers. Among these are politicians such as Mitt Romney (1.5 million followers) and Barack Obama (30 million followers). [2] Each follower is part of their audience; they read their tweets, retweet their messages, and take appropriate actions in real life. For this reason, having large numbers of followers gives them political power. On March 19, Barack Obama posted the tweet “RT if you stand with same-sex couples around the country fighting for the right to marry who they love.” This tweet got 39,000 retweets, which means 39,000 people spread his message to all of their followers. [2]

It is obvious that politicians use Twitter to communicate with the electorate and achieve their political goals. However, we would like to formalize this by quantifying the ways that politicians use Twitter. Specifically, we would like to see how congressional Twitter use correlates with (1) personal characteristics of the congressperson (2) voter demographics of their state. In particular, we can look at variables such as (1) the number of tweets posted per month (2) the number of followers a congressperson has (3) the number of users he is following. We chose these variables because they are straightforward to measure using the Twitter API, but our future work will look at other variables.

In the second half of the class we plan to do a sentiment analysis of congressional tweets and determine how personal characteristics and voter demographics correlate with the fraction of happy tweets. We will use a naïve Bayes classifier to classify tweets into “happy,” “sad,” and “angry” (or maybe “happy,” “angry,” and “other”). We chose this topic because sentiment analysis is interesting and I wanted an opportunity to practice machine learning.

**Previous work**

**Data**

So far we have gotten the following data: (1) Personal information for each congressperson on Twitter, including their name, age, party, title, gender, state, and Twitter handle. (2) Twitter data for each congressperson, including their join date, number of tweets, number of followers, and number of users they are following. (3) Demographics for each state, including the number of people in each age bracket and the fraction of people of each race. (4) The total population of each state.

We took the personal information from the Sunlight Foundation. [3] The table was available for direct download, and we wrote a Python script to remove information that was not pertinent to our analysis. The state demographics came from Wikipedia [4] and the 2010 US Census [5]. The total population of each state also came from the US Census. [8] In each case the data was available for direct download.

We gathered the Twitter data using the Twitter API [6] and python-twitter [7]. The Twitter API is a set of tools Twitter built to allow developers to scrape information from the site, and python-twitter is a tool that makes it easy to use the Twitter API from Python scripts. To use these, we first created a Twitter account, “jsuPS120,” and generated an authentication key. Then we found a “user object” for every congressperson in our list. Each object contained (among other things) the join date, number of tweets, number of followers, and number following. We used standard Python tools to export this information to a file. (Instead of writing the join date and number of tweets separately, we combined them to get the number of tweets posted per month.)

The Twitter data was slightly challenging to get because the Twitter API is rate limited. That means you can only get data for 180 users every 15 minutes. Since there are only 471 congresspeople on Twitter, this was not prohibitive to our analysis. However, we did have to write the code in a way that respected the rate limits. Originally, we tried using the “sleep” command to force the program to wait for 15 minutes every time it processed 170 users. But we kept getting strange bugs, so we gave up and split the congresspeople into three lists. While this was fine for our current purposes, our approach would not scale to larger datasets.

After getting all this data, we used a MySQL database to correlate the tables. We needed to do this because the congressional data, the Twitter data, and the state demographics were all in different tables, and integrating the tables by hand would have been impossible. We created four MySQL tables: “twitter” (containing Twitter data), “statepop” (containing state populations), “race” (containing race demographics information), and “age” (containing age demographics information). Then we correlated these tables using queries of the form

SELECT age, (num\_following / population) AS followingPC FROM

twitter INNER JOIN statepop

ON twitter.state = statepop.state

INTO OUTFILE 'ageVSfollowingPC.csv'

FIELDS TERMINATED BY ','

ENCLOSED BY '"'

LINES TERMINATED BY '\n';

Although the queries were standard SQL, executing these queries was slightly challenging because (1) we originally didn’t have permission to write data to any files (2) when we did get permission, MySQL wrote the data to a directory I didn’t know about (3) the directory was inaccessible to non-root users.

We prepared the following tables for analysis: (1) congressional age vs. number of followers, (2) age vs. number of followers per capita, (3) age vs. number of users followed, (4) age vs. following per capita, (5) age vs. tweets per month, (6-10) gender vs. all of the above, (11-15) party vs. all of the above, (16-20) title (House/Senate) vs. all of the above, (21-25) age, party, gender, and title vs. all of the above, (26-30) percentage of voters who are between ages 18-24 vs. all of the above, (31-35) percentage of voters 65 and over vs. all of the above, (36-40) median voter age vs. all of the above, (41-45) percentage of voters who are of various races vs. all of the above.

**Analysis**

We then have been using the twitter data that has been aggregated in order to look at whether we can find any correlations. So far, we have done linear univariate regressions of congressional age vs. (1) number of followers, (2) age vs. number of followers per capita, (3) age vs. number of users followed, (4) followed per capita, and (5) tweets per month. The R squared or coefficient of determination for the linear regressions are stated below in the table.

|  |  |
| --- | --- |
| **Congressional age vs. \_\_\_\_\_\_** | **R squared value** |
| Number of followers | 0.0079 |
| Number of followers per capita | 0.01 |
| Number of users following | 0.0008 |
| Number of users following per capita | 0.0005 |
| Tweets per month | 0.031 |

From this data, we were able to conclude that the strongest linear correlation was between congressional age and the average number of tweets made per month while the weakest one was between congressional age and number of users following per capita. However, it should also be noted that none of the linear relationships are that strong in the first place and that is probably because of the very large number of congresspeople who have made so few tweets, or have so few followers or number following that it skews the linear relationship. But the result that congressional age correlates linearly with the tweets per month makes sense since they are probably the group of people who are most comfortable with using social media in order to spread their voice on issues. It is also interesting to see that the correlation with congressional age and the number of followers per capita is not that strong; this means that regardless of age, people must follow congresspeople in equal numbers. But, just getting followers clearly does not seem to be indicative of active use of twitter. Also, there seems to be no correlation whatsoever between congresspeople’s age and the number of users they follow which probably indicates that this is not a way that congresspeople use twitter in order to further their political agenda.

Then we wanted to test whether their gender (male/female), party (Democratic/Republican), or title (House/Senate) affect congressional use of twitter in terms of the number of followers, following, or number of tweets per month. From this analysis, we calculated the following results using two-tail t-tests assuming different standard deviations for both sets of each pair of data.

|  |  |
| --- | --- |
|  | **p-value** |
| **Status vs. tweets per month** | **0.02454** |
| **Status vs. following per capita** | **0.093848** |
| Status vs. following | 0.467911 |
| **Status vs. followers per capita** | **0.024612** |
| Status vs. followers | 0.143284 |
| Party vs. tweets per month | 0.238769 |
| **Party vs. following per capita** | **0.079642** |
| **Party vs. following** | **0.038098** |
| Party vs. followers per capita | 0.11412 |
| Party vs. followers | 0.150374 |
| Gender vs. tweets per month | 0.501575 |
| Gender vs. following per capita | 0.305149 |
| Gender vs. following | 0.989266 |
| Gender vs. followers per capita | 0.353873 |
| Gender vs. followers | 0.554183 |

If we consider the significance level to be 0.10, then we see that very significantly, differences in twitter usage are due to status (House/Senate) and party (Democratic/Republican). Specifically this means that the mean number of tweets per month, the mean number of users followed per capita, the mean number of followers per capita is statistically different for House and Senate members. And, the mean number of users followed per capita and the mean number of users followed is statistically different for Republicans and Democrats. Further analysis showed that senators on average produce more tweets per month than representatives, follow more users per capita, and followers per capita. Additionally, Republicans have a higher average number of users they follow per capita and more users they follow.

Then we moved on to look at voter demographics

**Work to be Completed**

**Future directions – sentiment analysis**

One thing we could do is classify congressional tweets into different categories and see what types of congresspeople post certain types of tweets. For example, we could divide them into “happy,” “sad,” and “angry” tweets. Then we could draw conclusions like “female congresspeople post more happy tweets” or “congresspeople in states with a higher fraction of youth post more angry tweets.” This is called sentiment analysis, and while many people have applied sentiment analysis to Twitter, [9, 10] few have applied it to congressional tweets.

We would determine the emotional valence of the tweets using a naïve Bayes classifier. First we would identify “features,” or words that contribute to the emotional valence. These might be words like “congratulate,” “stop,” “war,” or “victory.” If the tweet we are analyzing was “stop the war on terror,” we would try to find the probability that the tweet was angry given that it contained the words “stop” and “war.” By Bayes’ theorem, we can write

The naïve Bayes classifier assumes that features are independent of each other given the class variable, so .

How can we compute these quantities? To compute P(angry), we would hand-process a small number of tweets and report the fraction that we thought were angry. To find P(stop | angry), we would report the fraction of these angry tweets that contained the word “stop.” To find P(war | angry), we would report the fraction of angry tweets that contained the word “war.” We would not have to find the denominator P(stop, war) because our objective is to find the category with the greatest probability. That is, we have to compare P(angry | stop, war), P(happy | stop, war), and P(sad | stop, war), and the denominator is the same for each of these.

To find the features, we would probably just use the most common words in our tweet sample. This is because we can’t assume a word doesn’t contribute to emotional valence just because it’s not obviously political. Maybe a word like “I” softens the tone of a tweet or makes it less likely that the tweet is about an angry subject. And the penalty for using a nonemotional word as a feature is small because the numbers more or less cancel out. So we may as well use all the words that appear a large number of times.

To collect the data, we would simply use the Twitter API as before. This time, we would send a GET statuses/user\_timeline request, which would allow us to return up to 3,200 tweets from each congressperson. Rate limits should not be a problem since they allow us to process 180 congresspeople every 15 minutes, and there are only 471 congresspeople.

To find our tweet sample, we could pick two tweets at random from every congressperson. However, this may introduce a partisan bias, since there may be more congresspeople in one party than another. A better idea might be to lump each party’s tweets together, then choose fifty from each. We will decide on the exact methodology when we actually do this experiment.

Once we’ve computed the probabilities using the tweet sample and tested our classifier on a different sample, we will run the classifier on the rest of the data and see which congresspeople post the happiest tweets.

**Conclusion**

**Works Cited**

[1] <http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html>

[2] http://www.twitter.com

<http://blog.site-seeker.com/who-uses-twitter-demographic/>

[3] http://sunlightlabs.github.io/congress/

[4] http://en.wikipedia.org/wiki/Demographics\_of\_the\_United\_States

[5] Source: U.S. Census Bureau, "Demographic Profiles: Census 2010."

[6] https://dev.twitter.com/docs/api

[7] https://code.google.com/p/python-twitter/

[8] "Table 1. Annual Estimates of the Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2012 (NST-EST2012-01)"

[9] <http://arxiv.org/pdf/1010.3003.pdf>?

[10] http://arxiv.org/pdf/0911.1583.pdf