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.

**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.

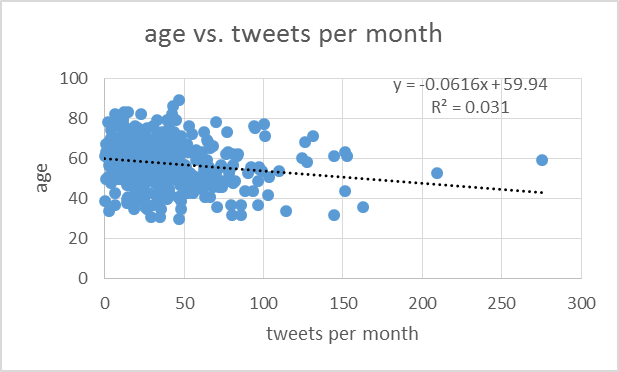
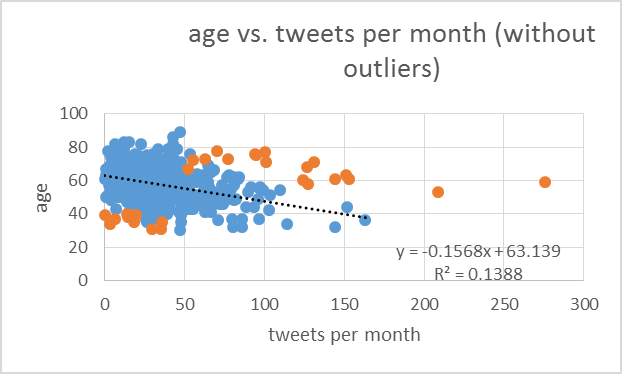
**Analysis**

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 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. In fact, most of the data looks like scatter plots as visible in the graphs presented below.

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. The correlation between congression age and their tweets per month is clearer if we exclude a few outliers as pointed out in orange as the R-squared values rises to 0.1388 making the correlation stronger. The farthest outliers, the 53 year-old who made 209 tweets per month and the 59 year-old who made 275.46 tweets per month, really skew the graph which shows a general downward trend for age vs. tweets per month. It would be interesting to match the outlying data with the actual congressperson’s name to see whether there is a reason for their increased twitter age which we can follow up on for our final report. Otherwise, from preliminary speculation, it seems as though a few younger congresspeople (< 40 years old) who made less than 50 tweets per month and some older congresspeople (> 60 years old) who made more than 50 tweets per month were rather rare in our data set of 472 congresspeople and so excluding their data helps the trend become more significant.



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.

However, some of the outliers are quite interesting if we look at them more carefully. For example, the three congresspeople who follow the most people on twitter are a 46 year old following 22,141 people, a 59 year old following 24,086 people and a 62 year old following 27,825 people as labeled in orange in the graph below. And the congressperson with the most followers is a 79 year old person with 1,798,403 followers. Again, matching the actual

congresspeople to these outliers might give us some more information as to why these outliers came about.

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 number of tweets per month, the number of users followed per capita, the number of followers per capita is statistically different for House and Senate members. And, the number of users followed per capita and the 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 have more 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.

|  |  |
| --- | --- |
|  | **R squared value** |
| Percentage of population 18 to 24 vs. followers | 0.0002 |
| Percentage of population 18 to 24 vs. followers per capita | 0.0004 |
| Percentage of population 18 to 24 vs. following | 0.005 |
| Percentage of population 18 to 24 vs. following per capita | 0.0036 |
| Percentage of population 18 to 24 vs. tweets per month | 0.0016 |
| Percentage of population 65 and over vs. followers | 0.0008 |
| Percentage of population 65 and over vs. followers per capita | 0.0034 |
| Percentage of population 65 and over vs. following | 0.0011 |
| Percentage of population 65 and over vs. following per capita | 0.000007 |
| Percentage of population 65 and over vs. tweets per month | 0.0016 |
| Median age vs. followers | 0.0001 |
| Median age vs. followers per capita | 0.003 |
| Median age vs. following | 0.0021 |
| Median age vs. following per capita | 0.0005 |
| Median age vs. tweets per month | 0.0002 |

From the table above, we saw that there was really no correlation between voter age and their twitter activity in terms of tweets per month, followers, or people following. There aren’t any outstanding outliers either that might help to improve the correlation of the data. This means that voter age does not provide much scope for further analysis. Instead, it will be more critical to look at whether race has any effect on followers, followers per capita, following, following per capita or tweets per month.

**Work to be Completed**

There is still some more data analysis that needs to be done like looking at other age demographics of voters like under 18, 25 to 44, and 45 to 64. These might show some additional information because ages above 25 seem to be most inclined to be politically active in terms of actively following congresspeople or caring about issues. Additionally, a t-test would be useful to see how gender affects twitter activity by voters. Since gender did not seem to affect congresspeople’s twitter activity, it would be interesting to see if there is any difference on the voter’s side of things.

We also have to do some multiple regressions to look at how race (White, Hispanic, Black, Asian, mixed) have an effect on twitter activity so that we can have a better understanding of how different races affect twitter activity. We also intend on doing multiple regressions for congresspeople and see how all the different factors, age, party, gender, and status affect how many twitter followers they have, how many twitter users they are following, and how many tweets they make per month. These multiple regressions will give us a better picture of how all these variables together might affect twitter activity and provide us with better conclusions than what we have from our linear univariate regressions.

Finally, it will be of most use if we can match some of the outliers in our graphs with specific congresspeople and study their political history and see whether we can find ways to explain why they are outliers. These studies can also help us identify why some of the trends are present in our data. It will also be important to see if there is any other way to find better correlational trends. For example, we tried to fit our data to a log scale, but that proved to be worse than a linear scale, and so we might want to go through the data and see if anything else might fit.

Additionally, we might also be able to come up with other things about congresspeople that might affect the way they use twitter. For example, we can use some other measures like how long they have been a congressperson or also look at since when they have started using Twitter or also seeing their presence in the news and using those descriptors to seeing whether they significantly correlate with their twitter usage. We can also look at the location of voters and see whether different states have people using twitter differently. If we can use location to describe voters, then we can also try and see whether being from particular states affects congresspeople’s number of followers and whether certain states have congresspeople who just use twitter more. These conclusions might help us show whether there are more twitter friendly states than others.

These sorts of analyses can also be applied for other studies that are looking at other members of political office other than congresspeople. These studies done on a long run can show just how important social media like twitter can be over a 5 year period rather than the short period that we are studying. Thus, our project provides scope for a wide variety of projects that can be done by us for this class and also in the future for studying the evolution of support of politicians and whether twitter activity correlate with actual outcomes in terms of voter turnout or election victories.

**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**

Ultimately, we have found some important information about some of the trends that might be present in terms of congresspeople’s age, party, status, and gender affecting their twitter activity. It was most interesting to see that there is no effect of congresspeople’s age to the number of followers they had. Also, the fact that there are statistically relevant differences between Democrats and Republicans and between Senators and Representatives in terms of certain aspects of their twitter activity were also quite relevant in our understanding of how Congresspeople use twitter. However, it might also be worthwhile to compare some of our findings to some of the past studies that have been done during election years to see whether these trends might be different. However, the fact that our correlations are so weak make it hard to completely conclude anything for sure, but it does provide us with room for speculation and if we compare our work with previous work and see some of our findings to be supported, then we can at least confirm that our data analysis is meaningful.

**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