Project 1 Report - Twitter Celebrity Trends and Interactions

Abigail Shchur (aks236), Edward Mei (ezm4), Kristi Lee (ksl72)

March 7, 2017

1 Description of Data

Our primary data source for this project was the collection of all tweets from 20 handpicked popular Twitter users from January 1^{st}, 2015 to roughly March 1^{st}, 2017. We did query Twitter data from many other users before eventually settling on the group of 20 seen in the visualization. The chosen users were sampled from 4 categories (celebrities, talk show hosts, politicians, athletes). These categories were selected because we thought that they were an adequate generalization of the groups that different famous people could fall into. We ultimately chose particular Twitter users by seeing who had the most followers in each category and by how many tweets they sent in the past two years.

Obtaining Main Source of Data

To obtain all the tweets from the past 2 years of our 20 Twitter users, we utilized the Twitter API (*tweepy*). In the data folder of our submission, there is an *ipynb* file titled "Gathering Twitter Data". It should be noted that all of the *ipynb* files also have a corresponding *html* file for easy viewing. In this file, we call the Twitter API as the documentation for it recommends. For each user, we store the name of the user, the text in the tweet, and the time-stamp of the tweet. We did run into issues with the rate limit for the Twitter API (which is why there is currently error output in the file) but we did manage to accumulate all the desired data. The data for all of the tweets is split between two files ("all_tweets" and "all_tweets_half"). The former of these files was obtained when data processing was done earlier in the assignment before we settled on our final group of people. The second file has the people that the first twitter file was missing. In all of the data processing segments of code we simply join these two files on the 20 selected Twitter users.

1.1 Bars and Word Cloud Visualization Data

The bar plot utilizes two data sets derived from the list of all tweets. All of the code for the calculation of both of these data sets can be found in the "Bar Chart Data Creation" file.

Categorical Counts

The file "*categorical_counts.csv*" contains a matrix where each row represents one of the Twitter users and each of the columns represents the number of tweets that this user had in a particular category. We split the tweets into 5 different categories (trends, celebrity, politics, sports, news). These categories were chosen because we thought that they would provide interesting insight into the types of tweets that different Twitter users have.

We classify a tweet based on the hashtags contained in the tweet. For instance, if a particular tweet contains "#voteTrump" and "#tbt" then the "#voteTrump" component would classify the tweet as *political* and the "#tbt" hashtag would classify the tweet as *trends*. This would increase the count for this user for both of these categories. That being said, the vast majority of the tweets in our data set fall into just one category. In the event that a tweet has at least one hashtag but that hashtag does not map to a category, we place that tweet in a 6^{th} "*other*" category. These tweets are not shown in our final visualization since we assume that they follow a similar distribution to the tweets that were classified.

The challenge with properly categorizing tweets was categorizing the specific hashtags into categories in the first place. This was done manually. We first sorted all of the hashtags used by the number of times they appeared in the data and categorized those until we got to the point that the hashtags left over appear in less than 10 data entries. After that, we looked at individual users and categorized at least their top 20 hashtags. Manually categorizing hashtags was rather subjective and it was done to our best judgment given the definitions that we proposed for the different categories. A few of the corner cases for categorization are outlined below:

- Hashtags involving sponsorship deals for athletes were categorized as "*celebrity*"
- Self promotional hashtags with the exclusion of politicians were categorized as "*celebrity*"
- If a hashtag truly had no clear category then it was placed under "*trends*"

User Hashtag Counts

The file "*user_hashtag_counts.json*" essentially stores a dictionary indexed by user and stores their top 50 hashtags, the number of times that hashtag appeared in any of their tweets, and the category of each hashtag. The python code for this was fairly straightforward - it involved iterating through all of the tweets and utilizing the category mappings described earlier. This data was used for the word cloud that appears above each user.

1.2 Heatmap Visualization Data

Reply Sentiment Counts

The file "*reply_sentiment_counts.json*" stores a list of sender, receiver pairs and a count for the number of positive, neutral, and negative tweets from the sender to the receiver. All of the code for this can be found in the "*HeatmapDatasetCreation.ipynb*" file in the data folder.

To create this dataset, we once again started with the set of all tweets. The next step was to add a field to each tweet depicting the recipient if one exists. We classified a recipient of a tweet as anyone who is both "@" mentioned in the tweet (aside from the user themselves) and is one of the other 20 queried Twitter users. This does mean that some tweets had more than one recipient.

We also ran sentiment analysis on each of the unique tweets that had a recipient that was one of our 20 users. This was done by utilizing a python sentiment analysis library that was trained on tweets and returned classifications of -1 for negative, 0 for neutral, and 1 for positive. Some more details regarding the logistics of this library and citations can be found in the code. To ensure that the classifications were accurate, we manually went through all of the unique tweets and changed the classification if it did not seem correct. There are a few caveats in the this process that are outlined below:

- Political tweets that clearly have a positive or negative connotation but "@" 2 people on opposing parties were marked as neutral since these were usually directed at only one of those individuals.
- Jokes about another user were taken to have negative sentiment.

Once we had the sentiment of each tweet the creation of the data set became rather straightforward. We looped through all of the tweets to aggregate the total number of tweets for each sentiment categorization for every sender and receiver pair.

Note on Data Selection

For all intents and purposes, our main data set is the list of all tweets for our chosen Twitter users for the past 2 years. This data set was segmented into the 3 subsets described above. By placing tweets into the 5 hashtag categories and doing sentiment analysis, we added extra dimensions to the data. We chose to exclude the time-stamp data since it did not provide interesting data for the majority of users. The one exception is shown as an outlier in the Heatmap. We also took elements of the data and largely simplified them for the purpose of the visualization. For instance, we took the text of a tweet and mapped it to a sentiment value.

2 Description of Mapping from Data to Visual Elements:

2.1 Bars and Word Cloud Visualization

The data from "*categorical_counts.csv*" was used in creating the bars. The basic calculation for the percentage of tweets for a given user in a given category is as follows: % tweets in category $A = \frac{\text{count category A}}{\sum_{i \in \text{Categories}} \text{count category i}}$. As mentioned earlier, the "other" category was ignored since we assumed the unlabeled hashtags follow a similar distribution to the labeled hashtags for a particular user.

The data from "*user_hashtag_counts.json*" was used to create the word clouds. The dataset had 50 words and we did not limit that. The word cloud library deals with excluding lower priority words if they do not fit in the size constraint given for the cloud. We are aware that the lower bound for the size of the font is low and thus some words are difficult to read. We did this on purpose. This is one aspect of our project that people had polarized opinions on. We ultimately decided that it was more aesthetically pleasing to see many words in the word cloud. This makes it clear that each user uses a wide array of hashtags, and viewers aren't necessarily meant to read every single one of them. Since at least the top 5 hashtags are clearly visible for each user, the viewer of the visualization can still easily get a sense of what each person generally tweets about.

We mapped each category of tweets to both a color and an emoji, using objects which mapped the name of the category to its corresponding color or emoji. While it may seem redundant to use both color and emoji, we did so because the emoji's themselves are more intuitive. Once viewers scroll down, they might not remember the color mapping, but the emoji mapping is more obvious. We did not do exclusively emoji's because the coloring is more aesthetically pleasing as well as more quickly

distinguishable. We also mapped the percentage of tweets in a given category to the horizontal length of the corresponding bar (created with line elements). For the word cloud, we mapped the number of times the given person used the given hashtag to the font size of the word in the cloud, using a log scale. We used the d3 cloud library to handle the x and y placement of the words. The category of each hashtag is mapped again to the same color scheme used for the bars.

2.2 Heatmap Visualization

The data from "*reply_sentiment_counts.json*" was used to create the heatmap. The sizes of the circles were scaled sums of all the positive, negative, and neutral interactions for each sender, receiver pair.

We mapped each person to a row and a column, so that a circle at (x1, y1) represents that person x1 tweeted at person y1. In other words, a circle represents an interaction between two people, and its x and y positions indicate which two people the circle represents. We also mapped the number of tweets to the size of the circle, so the larger the circle at (x1, y1), the more x1 tweeted y1.

Finally, we mapped the average sentiment of these tweets to the color of the circle. The average sentiment is calculated by subtracting the number of negative tweets from the number of positive tweets and dividing that number by the total tweets. This accounts for how prevalent sentiments were tweeted. So the more red a circle is, the more negative the overall sentiment was, and the more green a circle is, the more positive the overall sentiment. We wrestled over the color scale for quite some time - we recognize that the red-green color scale may not be comprehensible to individuals who are color blind; however, the scale of colors was the most intuitive way to model positive-negative feelings. Viewers don't need to be coached to understand that green is positive and red is negative. We felt that the scale's intuitiveness outweighed the obstacle to color-blindness. The color scale is linear and the size scale is a square root scale (we assumed this is better suited for circles).

3 Story of Visualization:

3.1 Bars and Word Cloud Visualization

The primary goal of this visualization was to see how different categories of celebrities compare in terms of the content of their tweets for the past two years. The first, and somewhat trivial observation is that members of the same category seem to generally tweet about similar topics. There are of course a few exceptions, but the athletes tend to tweet about sports, the politicians tend to tweet about politics or law and news, the talk show hosts tend to tweet about celebrities, and the celebrities also tend to tweet about celebrities. More interesting observations can be seen by comparing members of the same general category. Politicians are specifically interesting because the 4 politicians that were running for office tweeted primarily about just the election. However, former President Obama tweeted much more about news and legislation. This implies that a lot of the tweets from the politicians running for office lacked substance. They were mainly advertising their campaigns without focusing on specific issues. It is also interesting to note that Bernie Sanders and Ted Cruz, who dropped out of the election following the primaries, went on to tweet the most about news (presumably because they went back to their jobs). By looking at the word clouds, one can see that the vast majority of top tweets for people of any category are some form of self-advertisement. A few of the athletes, Obama, and Oprah are really the only exceptions to this. Lebron, for example, uses many motivational hashtags, presumably to inspire or motivate his followers.

3.2 Heatmap Visualization

The Heatmap conveys an overall sense of how celebrities, politicians, athletes and talk show hosts feel about each other. From the Heatmap, one can see that the vast majority of tweets have positive or neutral sentiment, with the exception being politicians. This is likely because politicians benefit from confronting other politicians in public, while celebrities might suffer the adverse effect to their public image. This is especially true for talk show hosts, who would presumably like to maintain positive relationships with most other celebrities so that their reputation among stars would allow them to keep guests coming to their shows.

One outlier that is rather interesting is the Bernie to Hillary sentiment. From looking at the actual tweets, it was clear that Bernie sent many negative tweets towards Hillary when he was running against her in the primaries; once she won the Democratic ballot, however, the sentiment of his tweets almost immediately switched to positive. For this reason, it seems like he is generally neutral. This was the only occurrence in the data with such an obvious switch. The most notable negative relationship in the plot not involving exclusively politicians is the Jimmy Kimmel to Trump sentiment. This is negative because Jimmy Kimmel tends to tell jokes about Trump and we classified those as negative. In terms of the number of interactions between users, one interesting and fitting observation is that the talk show hosts are the only group that really interact with every other group.

Combined Interpretation

Combining the two charts can lead to interesting interpretations of the data for every group of people. Looking specifically at politicians, one can understand the topics that they tweet about and their general interactions. They are clearly the most combative group in the sense that they have the most negative tweets and they direct them at each other. By looking at athletes, one can see that they are all rather unique in terms of the hashtags that they use and they tweet the other each other less than any other group tweets their own group. The celebrities and talk show hosts are similar in the sense that practically all of their tweets are self promotional with the exception of Oprah. The difference between these two groups is that talk show hosts tweet at other groups a lot more.