Analysis of Twitter Sentiment by Day of the Week

By Joey Demple

Introduction

This study was inspired by the Textisms assignment in Professor Abby Kaplan's *Language Myths* course at the University of Utah. The assignment involved analyzing "textisms" in tweets (acronyms, number/letter substitutions, emoticons, etc.). For this study, I began with the 100 tweets that I was assigned to analyze for the Textisms assignment (Textisms Dataset in the Flowchart; Figure 1). After removing Twitter handles that restricted access to their data and handles with no available data, 57 handles remained. From here, the Python-Twitter API collected tweets from each handle and separated them by the day of the week they were posted. After this, each tweet was analyzed using the Vader Sentiment Analysis tool.

Hypothesis

My hypothesis was that there will be a statistically significant difference in the proportion of positive and negative tweets between days of the week. For example, that there would be a higher proportion of positive tweets on a weekend day like Saturday than there would be on a mid-work-week day like Wednesday.

Data Collection

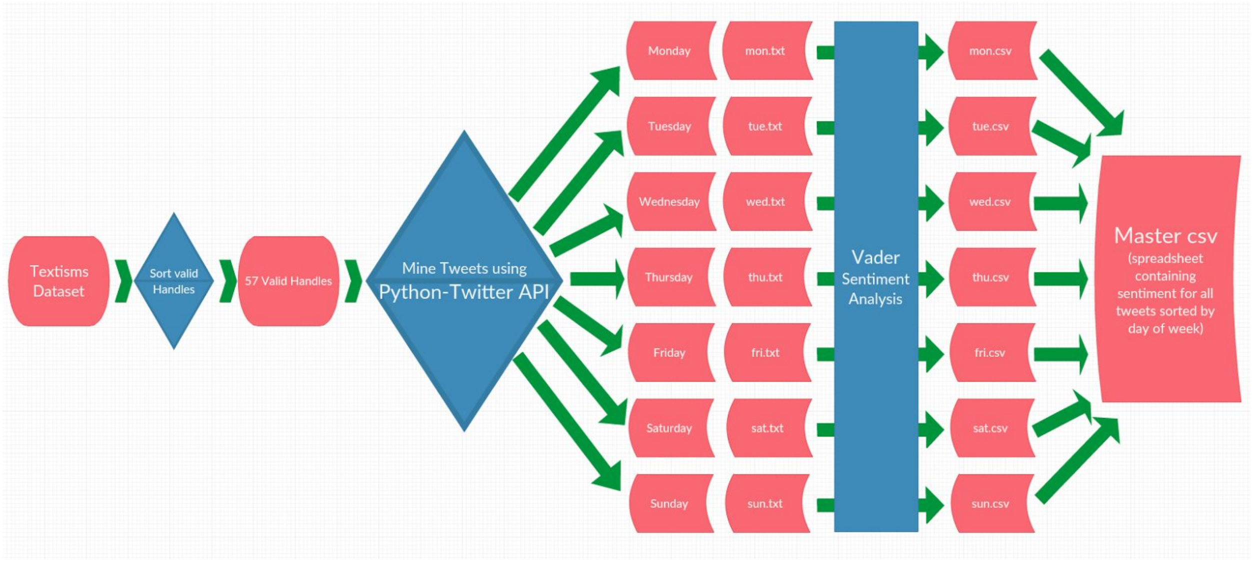

Figure 1: Flowchart for Data Collection

Originally, I had planned on using the data from the textisms assignment and adding to this dataset using the Python-Twitter application programming interface (API) to extract data directly from Twitter. However, I ran into many challenges when attempting to find random tweets from random users (The data from many twitter handles were restricted for me, and others handles returned no data when queried using Python-Twitter). As a result of this, the sample group is narrower than the dataset I used in the textisms assignment (57 users opposed to 100). To compensate for the more narrow pool from which the data is drawn, I have delved deeper into that pool to extract the data (There are 654 tweets from 57 users opposed to 100 tweets from 100 users). What I had hoped to achieve was 500 tweets from 500 different users.

Data Collection Method:
I wrote a Python script to extract tweets from Twitter.com using the Python-Twitter API (see tweets_sort_day.py in my Github repository *sentiment_by_days*). For each twitter handle in my sample group, between 1 and 20 tweets were collected. The script copies each of these tweets to a corresponding list named for the day of the week it was tweeted (e.g., “mon_tweets”). These lists are then saved permanently as .txt files corresponding to the day of the week they were tweeted (e.g., mon.txt).

After the tweets were mined using the API and stored in .txt files, they were Analyzed using the sentiment analysis tool Vader and written to .csv files. These .csv files contain the each tweet for that day of the week along with their corresponding sentiment scores: positive, negative, neutral, and compound.

Analysis
The analysis of the data takes place at two stages: Sentiment Analysis through Vader and the analysis of those results.

Sentiment Analysis through Vader:
Each tweet was analyzed by the Vader Seniment-Analysis tool. Vader is a “lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media” (Readme file in vaderSentiment repository; cjhutto, 2014). Vader has been tested and verified by humans to ensure effectiveness. According to the documentation, It takes into account factors such as punctuation, emoticons, word order, and word choice to give a positive, negative, and neutral score for each tweet. For each tweet Vader also calculates a “compound score” on a scale from -1 (as the most negative score) and +1 (as the most positive score)[1] (ibid).

Analysis of Vader’s Results:
For my analysis of the Vader’s results, the compound scores were the primary focus. According to the authors of Vader, the “compound score” is the best measurement “if you want a single unidimensional measure of sentiment for a given sentence” (cjhutto, 2014).

For the analysis of these data, I performed the Wilcoxen Two Sample Test. For each day of the week: 1) I tested its compound scores against each other day’s compound scores (see Table 1), and 2) tested each day’s compounds against the compounds for every other day of the week (Table 2).

Below is the table showing the results of the Wilcoxen Two Sample Test comparing each day’s compound scores. Note that the only day to day comparison that is statistically significant is Wednesday to Sunday.

[1] From the README.rst file in the vaderSentiment repository:
“The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric if you want a single unidimensional measure of sentiment for a given sentence. Calling it a 'normalized, weighted compound score' is accurate.”

P-Values Comparing Composite Sentiment Score, Day to Day (p <=...)
** Indicates statistical significance*

	Mon	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*	*Sun*
Mon		**0.9989**	**0.9395**	**0.4253**	**0.2194**	**0.1318**	**0.1034**
Tue	**0.9989**		**0.999**	**0.3618**	**0.1916**	**0.103**	**0.06937**
Wed	**0.9395**	**0.999**		**0.3458**	**0.128**	**0.06805**	**0.04291***
Thu	**0.4253**	**0.3618**	**0.3458**		**0.6865**	**0.4638**	**0.377**
Fri	**0.2194**	**0.1916**	**0.128**	**0.6865**		**0.7416**	**0.6988**
Sat	**0.1318**	**0.103**	**0.06805**	**0.4638**	**0.7416**		**0.9694**
Sun	**0.1034**	**0.06937**	**0.04291***	**0.377**	**0.6988**	**0.9694**	

Table 1: P-Values Comparing Composite Sentiment Score, Day to Day

The next table (2) shows the results of the Wilcoxen Test on each day of the week when compared to the other six days of the week. Note that none of these results are statistically significant.

P-Values Comparing Day of the Week to the other Six Days (p <=...)

Mon	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*	*Sun*
0.2782	**0.2181**	**0.1458**	**0.8837**	**0.4154**	**0.1811**	**0.1277**

Table 2: P-Values Comparing Day of the Week to the other Six Days

Figure 2 is a comparison of the average compound scores for each day as evaluated by Vader. Note that every day has an average rating in the positive range (above 0). In addition, note that this graph shows our only statistically significant result--that the compound scores for Wednesday are higher than those for Sunday.

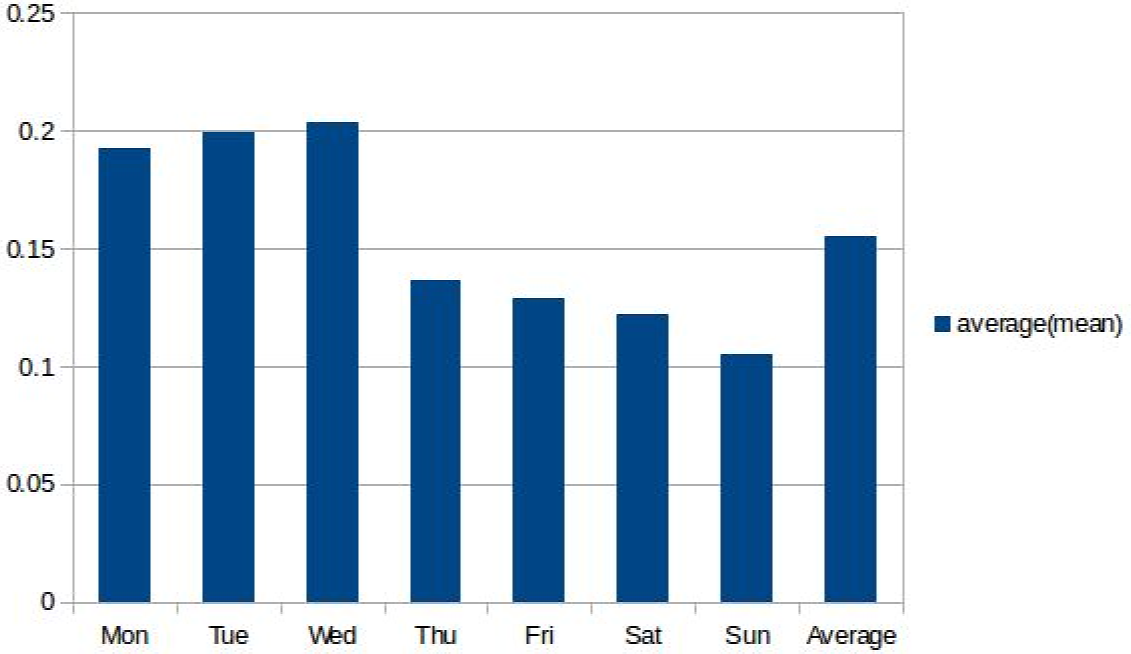

Figure 2: Comparison of Compound Scores by Day (range -1 to +1)

This next Figure (3) shows the number of tweets in each sentiment category by day they were tweeted. There are a large amount of neutral tweets. This is because Vader assigns tweets as neutral that do not appear to lean either positive or negative. However, when combing over these neutral results myself, I found that I would have rated a small portion of these neutral tweets as either positive or negative. Though, if I were to re-rate these neutral tweets, I would keep the vast majority of them in the neutral category, and judging from the neutral tweets I have examined, I would likely assign about as many to the positive category as I would to the negative one. Therefore, I do not believe that combing through the neutral tweets manually would substantially alter the results.

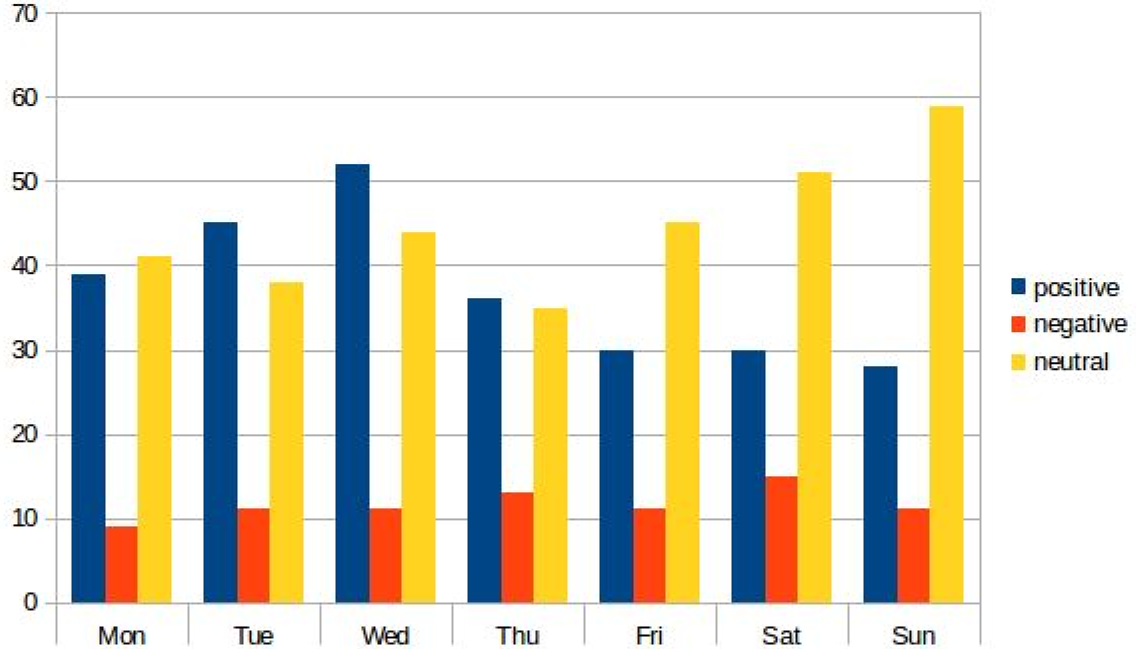

Figure 3: Number of Tweets by Sentiment

The last Figure (4 below) represents the proportion of positive to negative tweets per day. For this graph and the rest of my analysis a Vader compound score greater than 0 is categorized as *positive* and a compound score less than 0 is categorized *negative*. Compound scores of exactly 0 are categorized as *neutral.* Figure 3 displays the number of tweets in each category per day.

I found the regular curve in Figure 4 to be particularly interesting. Despite only finding statistical significance when comparing the compound scores for the most positive day (Wednesday, Figure 2) to those of the most negative day (Sunday, Figure 2), this curve may suggest that a larger dataset could yield a greater number of statistically significant findings. One caveat to this projection is that this dataset is from 57 users and may not be representative of the larger Twitter environment.

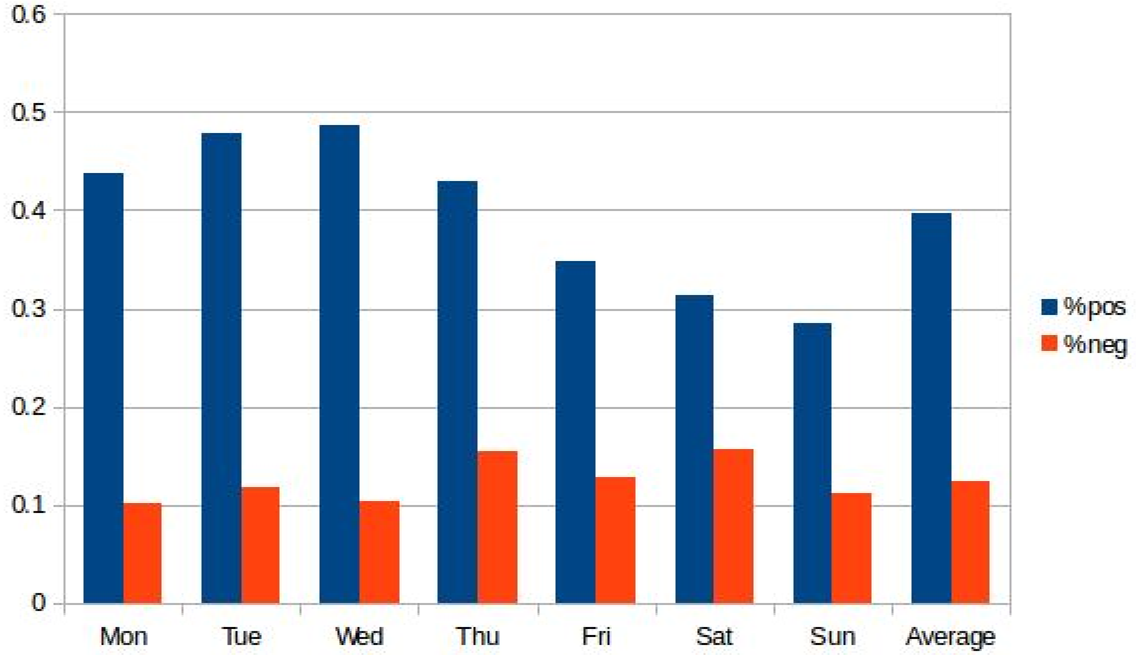

Figure 4: Percent Positive and Negative Tweets

Statistical Significance of Results

The only statistically significant result found in the analysis is that the tweets on Wednesday were more positive than the tweets on Sunday. This can be seen in Figure 2 and Table 1.

Discussion

Weaknesses of this study:

There are a few weaknesses to this study that I would have liked to reduce:

1) The dataset is comprised of the tweets from 57 handles. Therefore the results from this sample set may not be representative of larger twitter demographics (e.g., English speakers, Americans, North Americans). However, it should be noted, that despite having fewer twitter users in the study, there are 654 tweets from these users, and this is a step up from having only analyzed 100 tweets in the textisms assignment.

2) There is another weakness, but this is also a strength in one respect, using an automatic sentiment analyzer. Vader is a sentiment analysis tool "that is specifically attuned to sentiments expressed in social media" (Hutto & Gilbert, 2014). This is the tool that was used to analyze the 654 tweets in the dataset. Its weakness is that some of the tweets it evaluates as neutral might be better categorized as positive or negative, but its strengths are in its uniformity; Because Vader is automatic, a) these results are recreatable, and b) human bias is removed from the analysis of each sentence.

Conclusion, What I Learned

This study was my first opportunity to use the Python-Twitter API. For further studies of this kind I might consider using an extant Twitter dataset, or if I wanted to use the API again, I might consider using its “stream” function where tweets are streamed in real-time as they are posted. Leaving the API to collect data from the stream for an entire week is another possible method of collecting the data.

The statistically significant difference between the sentiment for Wednesday and Sunday was the opposite of what I expected to find. I had expected that tweets originating on the weekend would be more positive than those originating during the workweek. This statistically significant result, and the apparent curve in Figure 4 may suggest that a larger dataset will yield more statistically significant findings.

Citations and Resources

bear (2007). *python-twitter.* Github repository, https://github.com/bear/python-twitter. Last viewed December 2016.

cjhutto (2014). *vaderSentiment.* Github repository, https://github.com/cjhutto/vaderSentiment. Last viewed December 2016.

Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

jsdemple (2016). *sentiment_ by_ days.* Github repository, https://github.com/jsdemple/sentiment_by_days.

jsdemple (2016). *sentiment_analysis_of_tweets.* Github repository, https://github.com/jsdemple/sentiment_analysis_of_tweets.