

CSE/LIN467/567 Project

Trump Twitter Data Analysis

Author(s): Matthew Kirshy | Aniruddha Nandi

The State University of New York at Buffalo.



Abstract

It is our belief that as citizens it is our job to stay informed of issues that are affecting our country. The foundation of American democracy is built upon checks and balances in such a way that ensures no person or institution can rise to the level of tyranny. Democracy flourishes in the hands of informed citizens and our current administration is trying to destroy our ability to be informed. To do this we need to collect and analyze large amounts of data and be able to draw our own conclusions. Therefore, in this project, we have downloaded and analyze all of Trump's tweets (as sourced from <http://trumptwitterarchive.com/archive> which contains over 35,000 entries for us to analyze). We started by pre-processing (tokenizing and creating a corpus) all of the tweets so that we could further analyze the language he uses. We used the corpus to analyze the sentiment of individual tweets against buzz words and sensitive topics such as "guns", "women", "immigrants", etc to see how his sentiment changed as newsworthy events took place in the past. Finally, we produced some interactive graphs based on the results sourced from the above to visualize our findings in an intuitive as well as accessible format

Key words · tokenization, corpus generation and parsing, visualization, sentiment analysis, data science

Data and codes are available at <https://github.com/anandimous/467-tokenization-sentiment-analysis>

Proposed Methods

First we found an archive of all of trump's tweets going back to the start of his twitter account.

1. We downloaded them to a csv file
2. We cleaned up the csv by changing the dates to the correct date format that R can read
3. We forced all text to character type to make it easier for R to handle
4. We forced all numeric non-date columns to integer type

Second we created a column that has a running total that resets every new day so we can properly graph tweets by day. We also calculated the polarity of each tweet and put it into its own corresponding column using the sentimentR R package. The sentiment function returns results for different sentiment lexicons but we choose to go with the standard Jockers 2017 dictionary. Storing the polarity score and graphing it allows us to view his sentiment for each individual tweet.

Third we decided to pull out specific polarity scores for current buzzwords. We settled on the following buzz words:

Liberal | Democrat | Republican | Gun | Immigrant | Women | Latino | North Korea

To find those buzz words we used a regex find on our original tweet csv and returned the row number. Having the row numbers we subsetted the original csv to a specific sentiment data frame. We plotted these as a line graph corresponding to the date and a positive and negative score.

Fourth we copied his tweet text column to a .txt file as required by the TM (text-mining) R package. The TM package creates a corpus for us to use that:

- Removes urls
- Fixes text encoding errors
- Removes special characters
- Removes punctuation
- Removes stop words
- Removes numbers
- Removes white space

It then creates a document term matrix that allows us to access word counts, frequency of words, and co-occurrence if we choose.

Results

We intended to have some interactive graphs to display the following information in a time-series based format:

- Processed version of the tweet (tokenized and pre-processed)
- Buzz words contained in the tweet
- Individual word count corresponding to the tweet
- Overall sentiment of the tweet
- Change in sentiment toward key topics mentioned in his tweets over time (examples of the above are: "democrats", "immigrants", "guns", etc.)

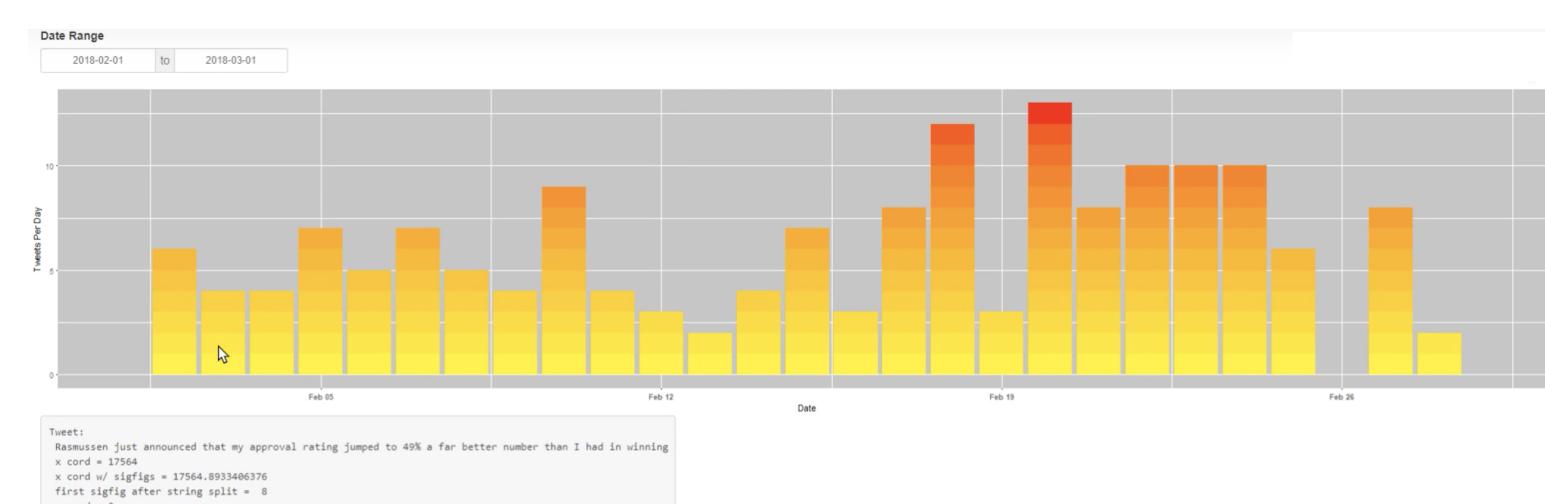


Figure 1: Processed tweet visualized in interactive bar graph

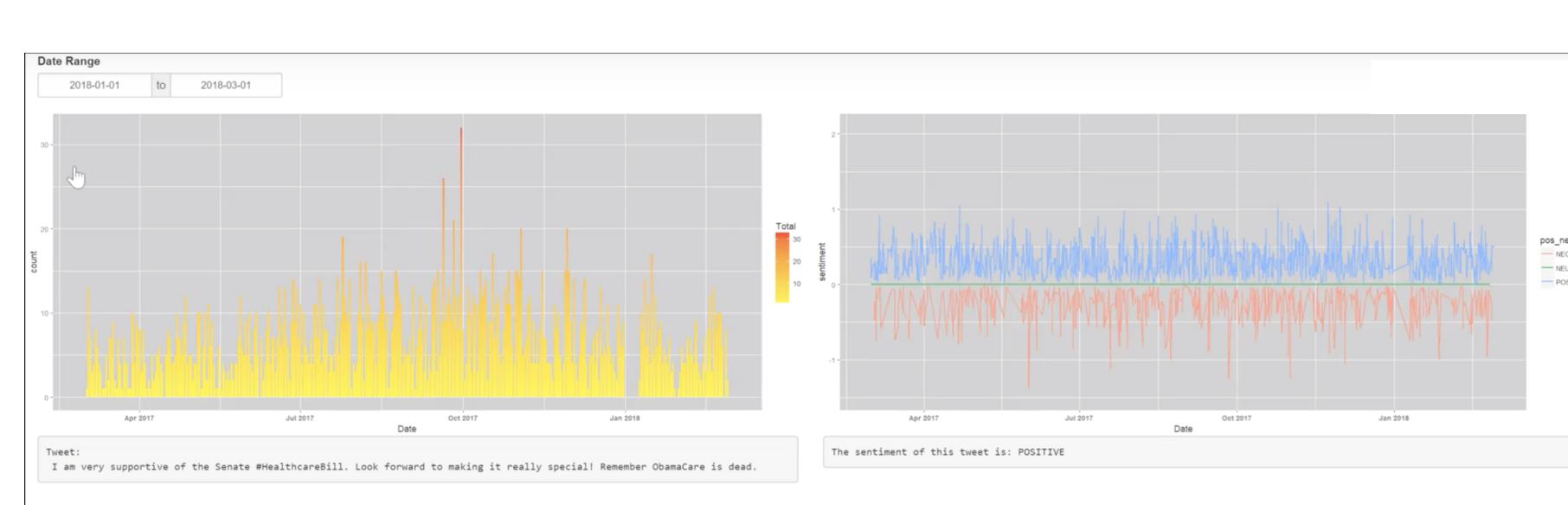


Figure 2: Tweet with POSITIVE sentiment and corresponding visualization

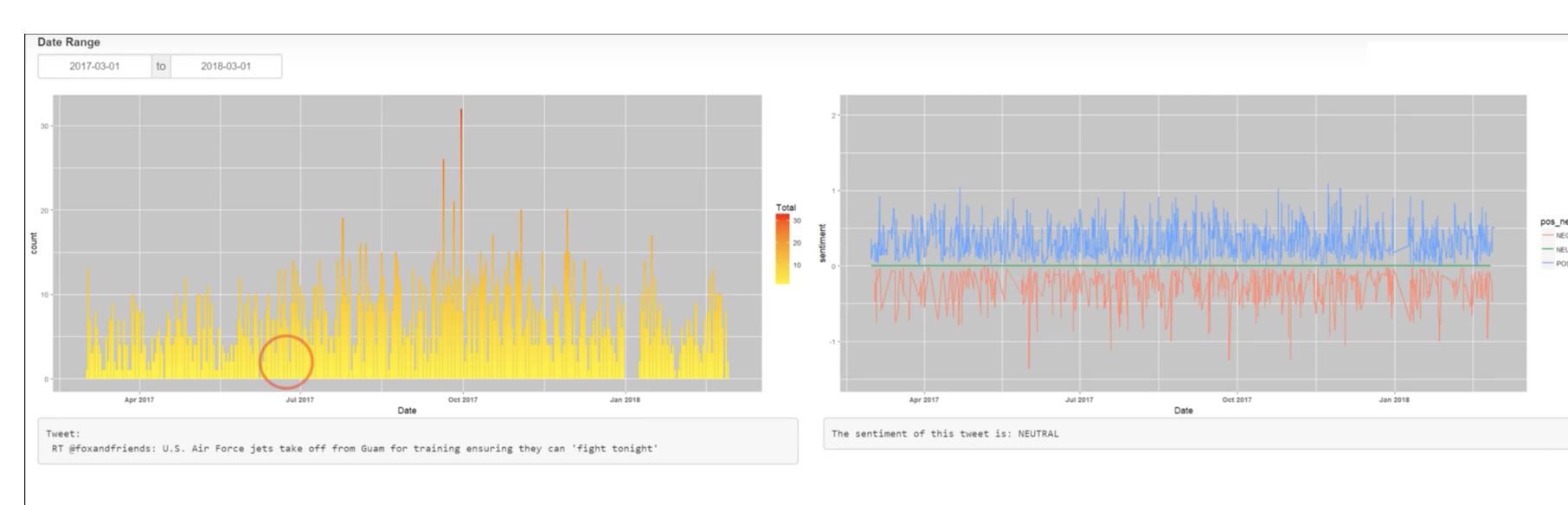


Figure 3: Tweet with NEUTRAL sentiment and corresponding visualization

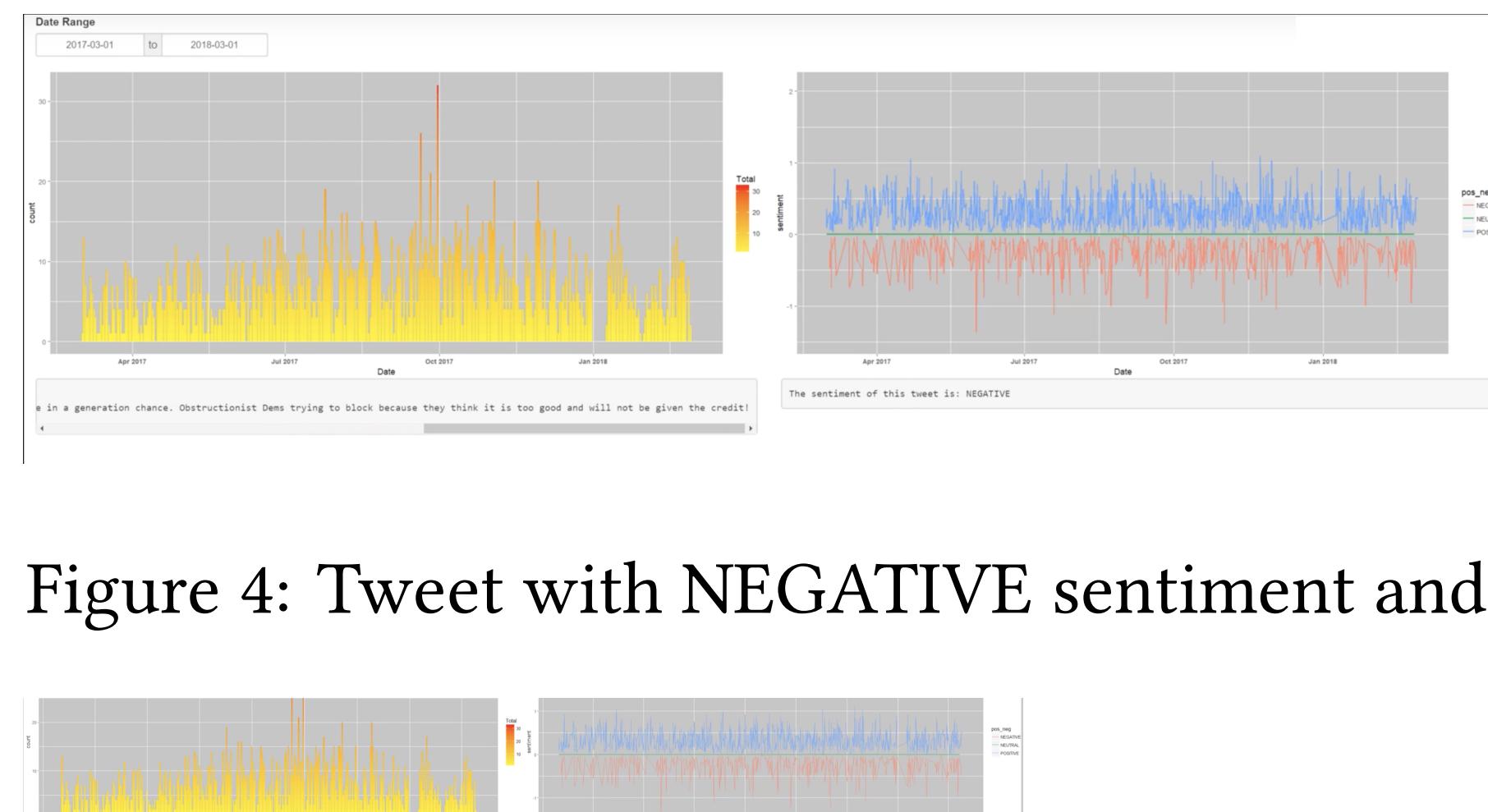


Figure 4: Tweet with NEGATIVE sentiment and corresponding visualization

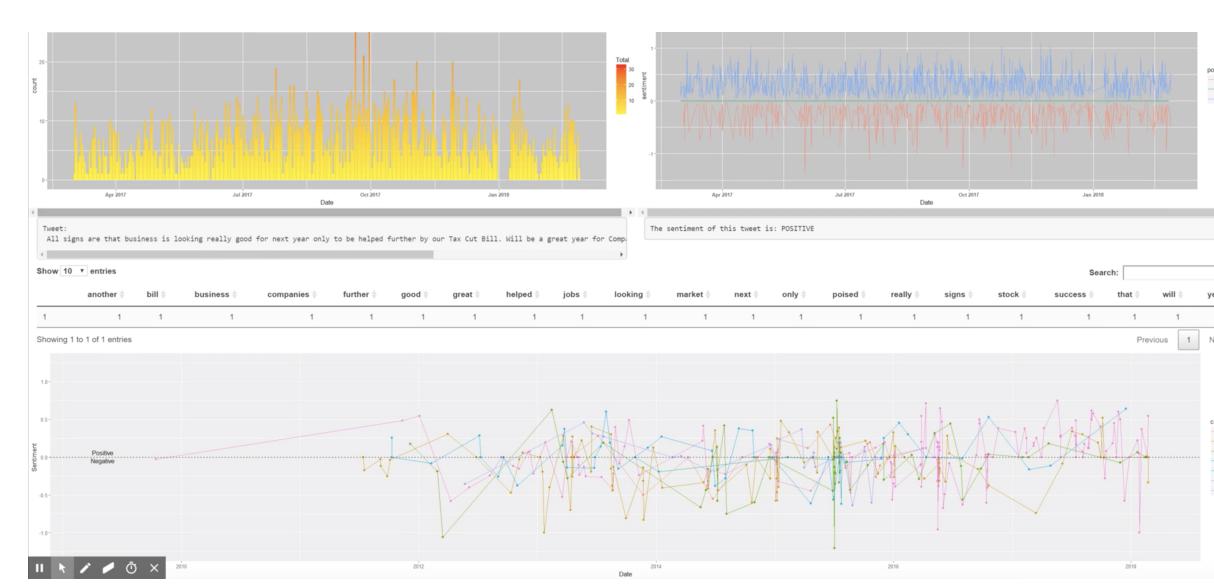


Figure 5: Line graph marking changing sentiments on certain buzz-words