Michael Paquette and Naomi Nguyen

Prof. Michelle Mouton

A sentiment analysis on New York Times headlines in "US" and "World" sections

1. Overview

The availability of international news nowadays raises a question of how America perceives itself compared to the world. This paper attempts to do so by analyzing the sentiment of news headlines from sections "US" and "World" in the New York Times, one of the most objective news sources. This paper will first compares the sentiment score of these sections and then dives into the difference by looking at the world cloud of neutral, positive, and negative headlines from each section.

The main findings of our analysis are: first, both sections are negative on average even though the biggest portion of each is neutral. Second, US headlines are scored to be more optimistic than World headlines, although this observation may be biased. Third, the month in which election result is released is one of the most "optimistic" month in both sections. In addition, we summarize the most common terms used in each sections in each sentiment category.

2. Methodology

We found a way to extract data from the New York Times online through its developer API's (NYT Developers Network). There are a few ways to do this, but the way we went with was to use the Archive API. We used python to write a program that extracts certain elements from JSON files returned from the New York Times' web services. The program retrieves the JSON files, then iterates through the files writing out the lines containing title, snippet, and date. Once the CSV's were complete we loaded them into RStudio data frames, and the data mining process was complete. RStudio is a statistical scripting environment that is very useful for cleaning and visualizing data. The cleaning that we had to do was very simple in theory. The data frames were composed of three columns. Two of them were strings, and the third was an integer. To obtain the sentiment scores you input a vector of strings (in this case our

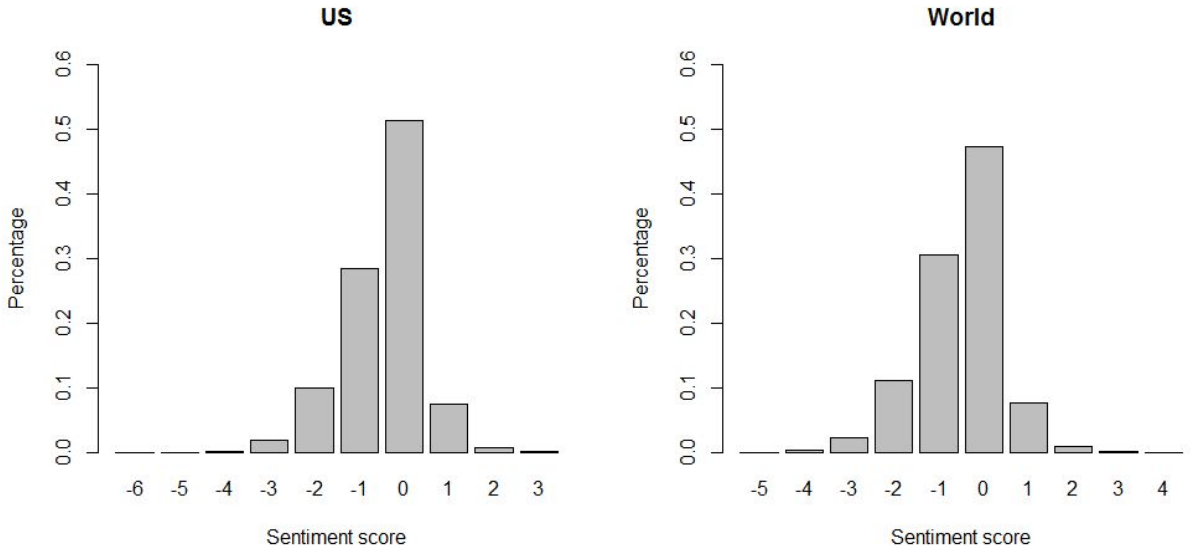
title column), into the function calculate_sentiment() from the R package “RSentiment” developed by Subhasree Bose. This returns a double vector of the inputted strings as one column and the sentiments as the other column. Initially when we tried to run the code it threw an error saying, “Arguments imply differing number of rows.” It doesn't make a whole lot of sense to us but the fix was to extract all special characters from the strings before performing the calculate_sentiment(). R has a wonderful API for strings and we used its string replace function that says to go through the title vector and replace all special characters with the space character. This eliminated the previous error and we were able to now start analyzing the sentiment data.

We then continue our analysis of 25-word word cloud in each section and with each sentiment category with Voyant Tools, an interactive web application for text analysis (Sinclair, S. et. al 2012). We find Voyant Tools is easy for us to customize and has attractive visualizations for word cloud. We used the default stop list in Voyant Tools (words to exclude from the word cloud) and added words that we thought are unuseful: “year”, “years”, “new”, “gets”, “says”,”say”, “talks”, and “latest”.

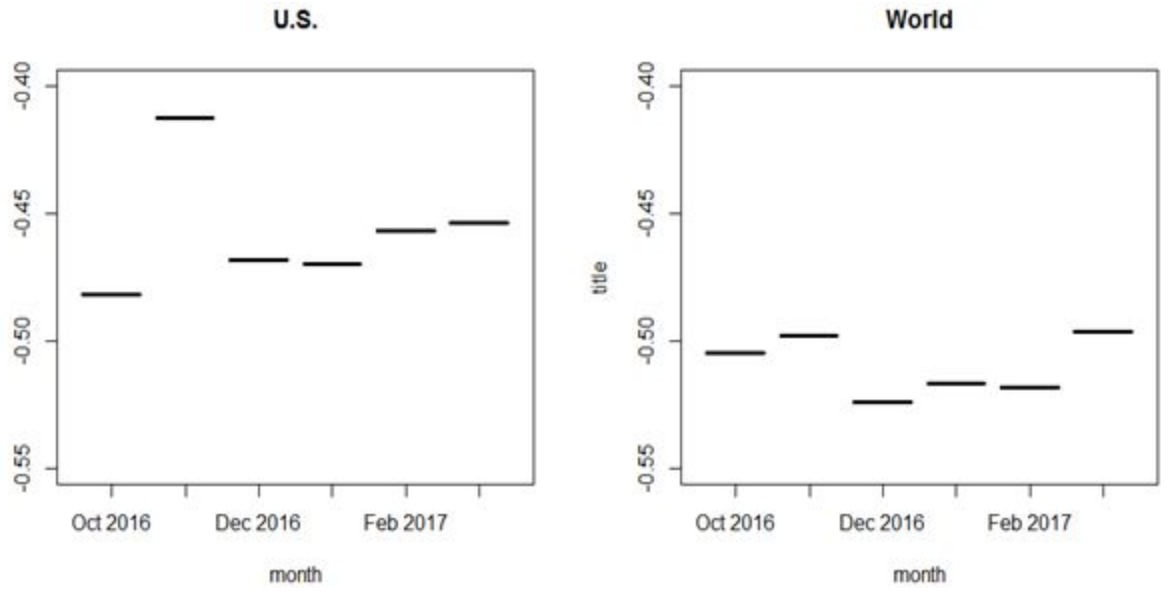
The scoring mechanism of this package is “0 indicates neutral sentiment. Positive value indicates positive sentiment. Negative value indicates negative sentiment. 99 indicates sarcasm” (Bose, S. 2017). However, we exclude headlines with sarcasm due to ambiguity in sentiment scoring. We thought excluding them from our study would not cause bias due to their rarity (1.9% for US articles and 0.8% for World articles). The limitation this project has is looking at and analyzing foreign headlines doesn’t work because foreign languages have more characters compared to English. They require a different character encoding such as utf-16. English is utf-8. If you are unfamiliar with encodings they are essentially what the machine translates to human readable text. We found that there were 63 rows in which the title was not in English. The majority of them were null values and as such when judged by R’s sentiment package returns negative sentiment. Thus we cannot fully analyze from our data, which foreign headlines are seen as more positive or negative because they are all skewed by the null values.

3. Analysis

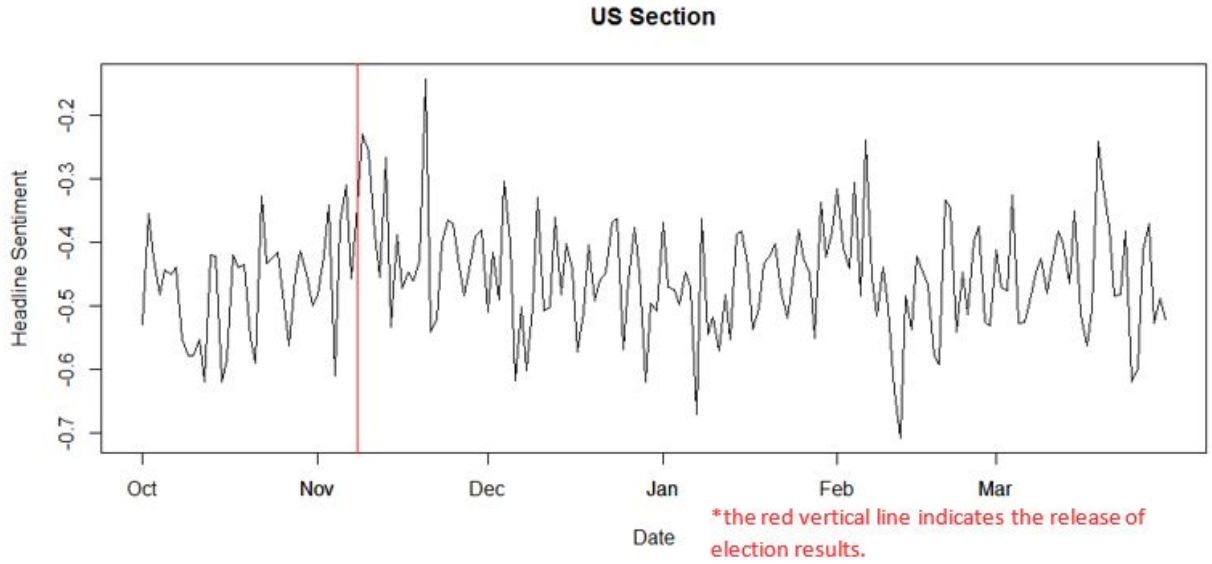
The sentiment score of section “US” and section “World” over the period share much in common. Both has a bell-shape-like distribution, and allocates similar proportions at each score. The biggest portion of each section is neutral, the next biggest is negative, and the least is positive. However, section “US” has about 4% more neutral news.



A further look into monthly aggregate sentiment score reveals that the headlines in “World” section is more pessimistic on average than headlines in “US” section. Furthermore, there seems to be a cutoff separating the monthly sentiment between the two sections: The most negative month for US news is in October 2016, with a value of -0.48 while the least negative monthly value for World news is -0.497, on March 2017.



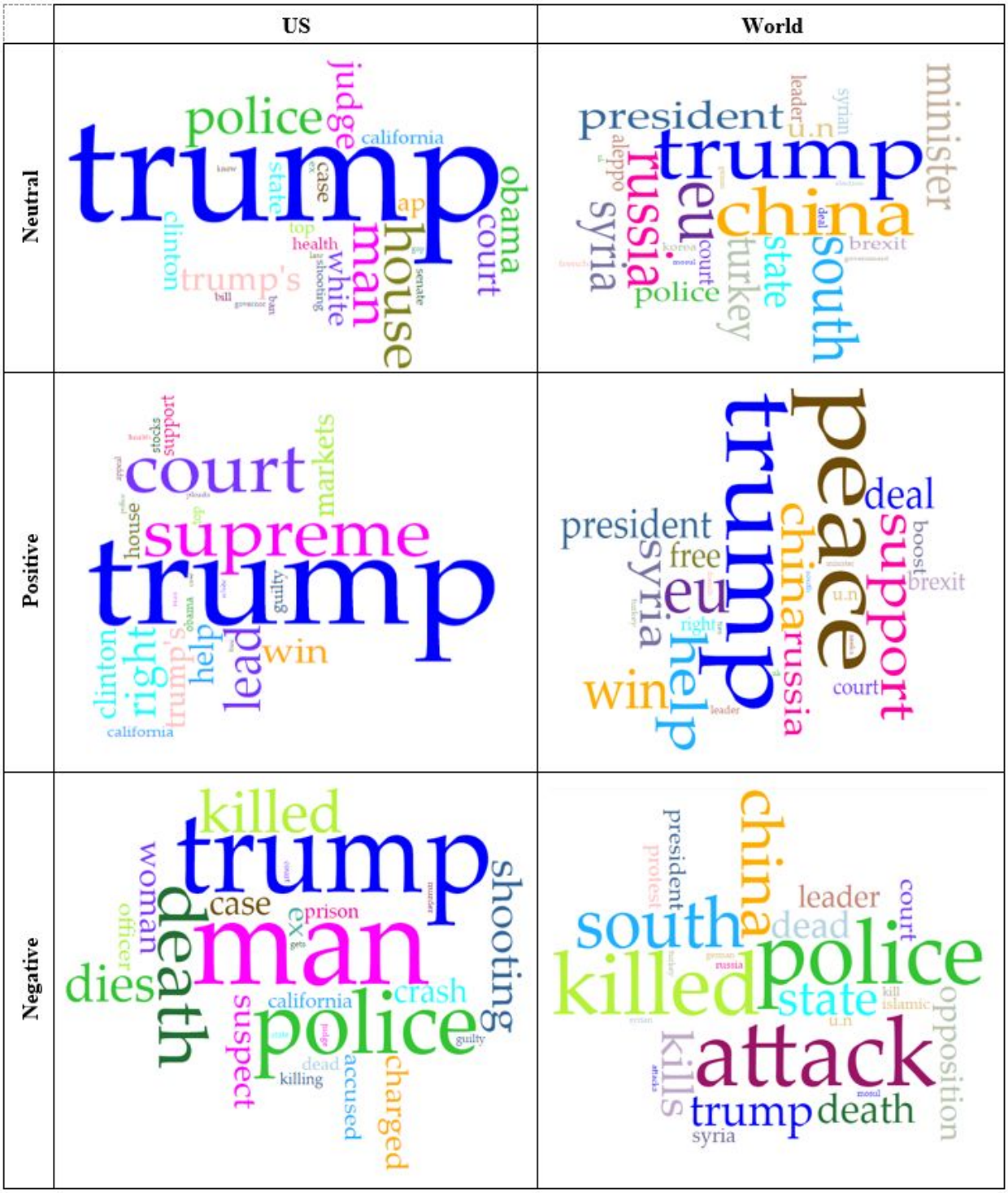
The graph suggests that the most optimistic month for US article headlines, as well as one of the top 2 most optimistic month for World articles headlines, is November 2016, when the election result is released. This is unexpected with respect to New York Times' criticism of Donald Trump. Moreover, the most pessimistic month is October 2016, the month before the election result release and despite all the news about scandals and disorganization on the new government. This is the only month before the release so it does not make a strong evidence for the upwards trend in positivity in "US" news.



A zoom into November 2016's headline sentiment suggests that the November headlines are generally more positive than other months without having any extreme sub-intervals that may drive the monthly aggregate value up. Possible explanations could be the prevalence of news on Trump's victory as the most mentioned positive words (as seen in the word clouds below) for positive articles are about the election: "lead", "win", and "support".

The prevalence words associated with the election result (that is not universally accepted as positive) is also evidence of bias towards the calculation that US article headlines are more positive than World article headlines. In addition, another factor that makes the US headlines calculated to be more positive is the abundance of news on supreme court (the word "supreme" is in the positive word list although it should not be in this context).

Below is a table of word cloud by section and by sentiment. Some interesting facts from the word clouds: the only category where Trump is not the most mentioned is in negative World headlines, Clinton is not in the list of 25 most frequent words in negative US headlines (but does in neutral and positive headlines), "peace" is one of the two most frequent topic for World headlines, and New York Times mention China and Russia much more than other countries in its World section.



4. Discussion

The most potentially interesting finding is that November 2016 is of of two months with the least negativity in headlines in the period from October 2016 to March 2017 in both section US and World. This finding is partly, and undesirably, biased because of the prevalence of words associated with the winning of an election (which decrease the negativity but do not necessarily reflect positivity or negativity). Such bias can be reduced if we can limit the word cloud to only include positive or negative words and intelligently remove the words that cause such bias.

Another interesting finding is that US headlines have less negative sentiment score, indicating that New York Times perceives there is more gloomy events outside America than within America. Again this could be biased by the prevalence of words that are scored positive but do not necessarily indicate sentiment such as “supreme” in “supreme court”. In contrast, looking at the word cloud for positive “World” headlines, we see truly positive words such as “peace”, “support”, “free”, and “help” (even though whether it is positive or not needs to be further examined in context).

This project also opens up discussion for which topics are viewed as positive or negative in New York Times over time. For example, we could see that EU is usually associated with positive headlines and “islamic” with negative headlines. If the span of the dataset increases, we can see the dynamics of the sentiment of each topic, like when a country is struggling and when it flourishes. More importantly, we may see more dramatic values in sentiment for the most significant events, which will help one unfamiliar with the news, of US for example, to get started and know the big events. Expanding the time span could also help us to see how news articles has evolved in terms of sentiment.

References

Sinclair, Stéfan, Geoffrey Rockwell and the Voyant Tools Team. 2012. Voyant Tools (web application).

Subhasree Bose with contributons from Saptarsi Goswami. (2017). RSentiment: Analyse Sentiment of English Sentences. R package version 2.1.2.