Donald Trump (DT) Tweets and

Sentiment Analysis

James Taylor, Jeffery Watson, Scott Vigil, & Selamawit Tesfaye

University of Maryland University Global Campus

DT Tweets and Sentiment Analysis

Historically, speeches have been essential to the leadership role in order to capture a message directed towards a national/ global audience. Initially, the delivery of a speech required people to be present to hear it physically. As time has gone on with the increase of modern communication, the value of physical presence decreased and shifted towards the written word, followed by the printing press, radio, and television. Technology has changed and expanded how humans communicate. Through modern communication and technological advances, the speed and reach of speech have drastically increased.

The internet is considered a new frontier to public speaking. The majority of Americans receive news from social media, with younger generations disproportionately represented. As time continues to progress, tech-savvy citizens will significantly outnumber those who receive news from traditional channels. An official or political message can be received without the presence of speechwriters and public-relations managers crafting responses. In 2016, DT was recognized as bending the established rules of campaigning, utilizing social media to deliver an "underdog" victory during the presidential campaign (Boyle, 2016).

Politicians will continue to use social internet platforms to facilitate campaign efforts, gain support, and manage their reputations. This paper will analyze the sentiment of the tweets of then-candidate DT. A detailed discussion of policy may not be what excites people around a candidate. It is relatively impossible to untangle the politics of-the-day from how politicians are using emerging social media. Also, specific techniques are likely to have worked better for DT in comparison to other political or government officials. Other political hopefuls may learn from the discoveries to maximize their social media presence.

This report will evaluate the dataset through pre-processing, exploration, analysis, and assessment. Select packages, such as dplyr, tidytext, tm, tidyr, word cloud, RColorBrewer, Syuzhet and ggplot2 are used in R to assist in performing the analysis. The report will create a variable with text as a character, create a data frame, convert words into tokens, remove stop words, graph words with high frequency, identify positive/ negative words, gather sentiment count, and graph sentiment count. Furthermore, the report will sort words, separate bigrams, identify word usage, and explore the dataset through text mining visualizations.

Analysis and Model Demonstration

In the Analysis section, the DT tweet data goes through exploratory analysis, perform necessary pre-processing activities, provide insight into the algorithm and core parameters, demonstrate the model building steps along with parameter tuning, and explain all assumptions.

Data Information and Preprocessing

The DT tweet dataset has 7375 observations of 10 variables. Attributes for this dataset include Date, Time, Tweet_Text, Type, Media_Type, Hashtags, Tweet_Id, Tweet_URL, Favorite Tweet Count, and Retweet Count. This DT dataset consists of

Factor, Numeric, and Integer types of variables, including several sub-levels, and continuous measurements. This dataset includes posts to and from DT. Upon additional review, the dataset includes the conjunction of unique characters, punctuation, capitalization, and grammatical error. Type breaks down Tweet_Text into two sub-variables, text and links. In select situations, the text includes the URL. Excluding information found within the Hashtags column, Hashtags can be found within Tweet_Text. During Data Preparation, the decision point will be to determine which variables are removed and which variables will remain and be evaluated.

First, stop words, a character vector, is engaged in preparation for data evaluation. This stage includes the removal of all non-essential attributes, such as Date, Time, Type, Media_Type, Hashtags, Tweet_ID, Tweet_URL, Retweets, and additional characteristics, reducing the dataset to Tweet_Text.

Upon review of tokenized words and removal of all stop words, new custom stop words are added. The purpose of adding custom stop words is to remove words joined with ASCII symbols, such as URL, HTTPS, HTTP, and Twitter Responses. In the process, tokenized words are sorted, resulting in count and sentiment analysis using Bing and the NRC Word Library. Additional steps will be conducted to prepare a dataset for trend, pattern, and association analysis, such as frequency and inclusion of date and time.

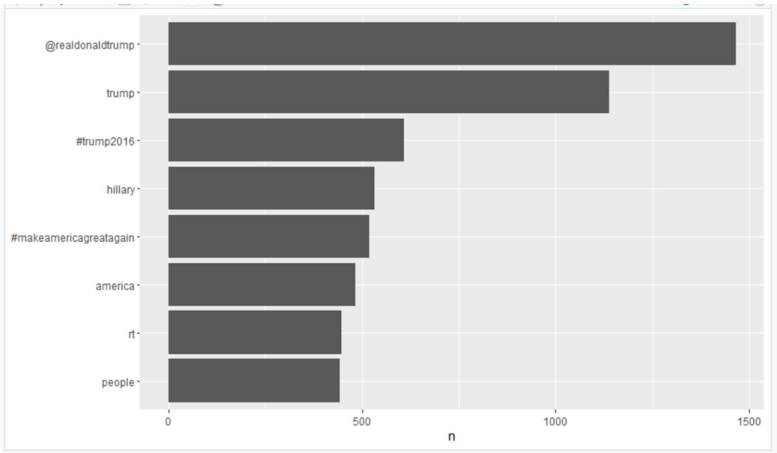
Exploratory Data Analysis

The data pre-processing resulted in the conversion of the text of DT tweets to a data frame with 124,289 observations with two variables (i.e., word and count using 'tweets' token). The tokenized set of words assigned to the dtWords variable was further processed to remove stop words with custom stop words created. The removal of custom stop words resulted in 62,161 words in the dataset.

Eight words are identified with the highest frequency. Figure 1 shows the distribution of the frequency of words. According to Figure 1, "@realdonaldtrump," "trump," and "#trump2016" account the top three, followed by "hillary" and “#makeamericagreatagain" as the fourth and the fifth highly used words.

Figure 1

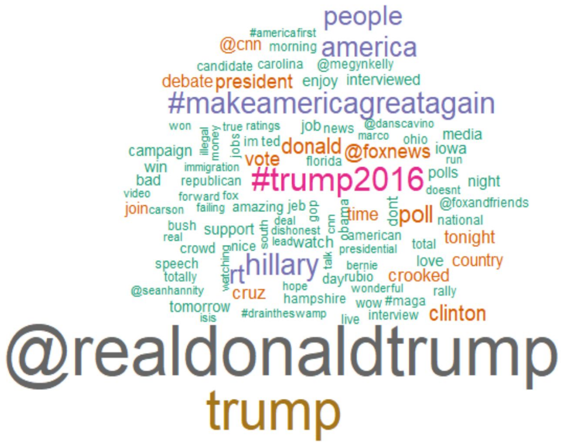
Frequency of Words



The top 100 repeated words are shown in the word cloud in Figure 2. The “@realdonaldtrump" is bold, and the repeated use of the word is shown within the word cloud and the tweet text. Then the next frequent word "trump" is shown with a golden orange color to be followed by "#trump2016" using pink color. Words that have similar count are shown with the use of the same color.

Figure 2

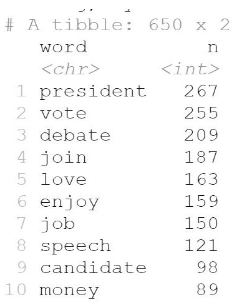
Word Cloud of Top 100 Words



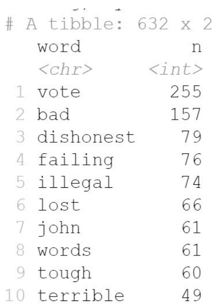
Further exploration was done using NRC, and Bing word lexicon libraries. The NRC lexicon identified 650 common positive words, and out of those words' "president" accounts for the highest frequency followed by "vote," "debate," "join," and "love." There are 632 common words in the NRC lexicon library with the tweet words indicating negative words. The count on Figure 3 shows that "vote" and "bad" have the highest count. The word "vote" appears in both groups, with the same count indicating the word as positive and negative at the same time.

Figure 3

Positive (Left), Negative (Right) Words based on NRC Lexicon Library



```
# A tibble: 650 x 2
   word          n
   <chr>     <int>
 1 president   267
 2 vote        255
 3 debate      209
 4 join        187
 5 love        163
 6 enjoy       159
 7 job         150
 8 speech      121
 9 candidate    98
10 money        89
```

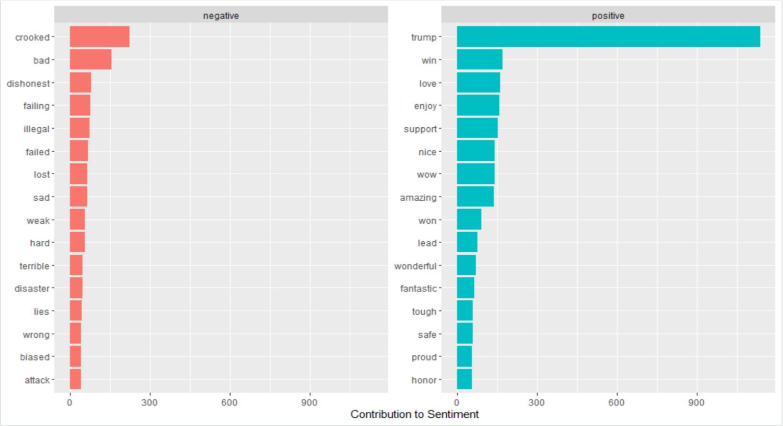


```
# A tibble: 632 x 2
   word          n
   <chr>     <int>
 1 vote        255
 2 bad         157
 3 dishonest    79
 4 failing      76
 5 illegal      74
 6 lost         66
 7 john         61
 8 words        61
 9 tough        60
10 terrible     49
```

The use of the Bing lexicon library resulted in a little different set of words compared to the NRC lexicon library. Words such as "crooked," "bad," "dishonest," "failing," and "illegal" words account for the top five negative words, and the word "crooked" is significantly more frequent than the others. The highly significant five positive words on the DT tweet dataset with Bing lexicon are "trump," "win," "love," "enjoy," and "support." However, the word "trump" is more than four times more frequent than the other positive words in Figure 4.

Figure 4

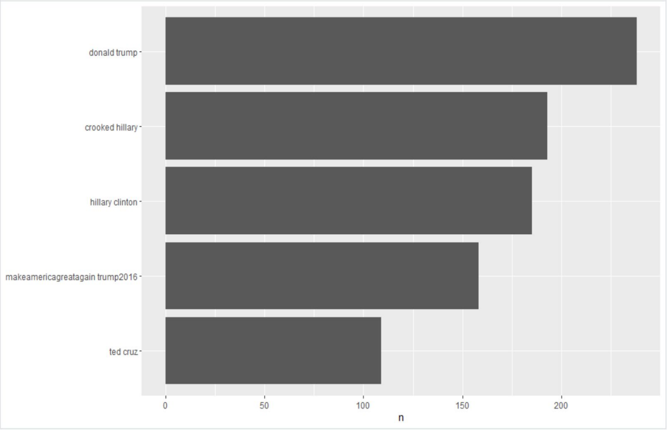
Positive & Negative Words using Bing Lexicon



The bigram words on the tweet dataset are shown in Figure 5. The top five bigrams tweet words are "donald trump," "crooked hillary," "hillary clinton," "makeamericagreatagin trump2016”, and “ted cruz." Four out of the five have some names associated with a word. The most famous name in the bigrams is "hillary” by appearing on “crooked hillary” and “hillary clinton” bigram words.

Figure 5

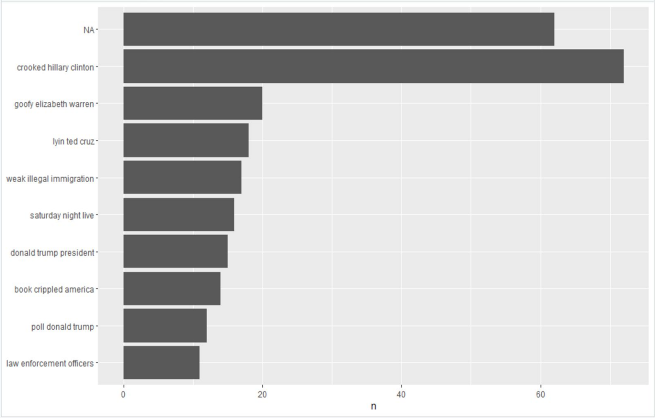
Bigrams Frequent Word Count



The trigram for words shows names associated with the words "crooked, goofy lyin." The top five frequent trigrams are “crooked hillary clinton," "goofy elizabeth warren," "lyin ted cruz," "weak illegal immigration," and "Saturday night live."

Figure 6

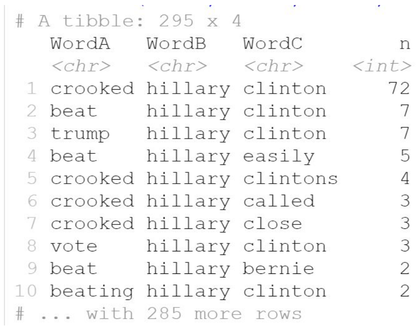
Trigram Frequent Word Count



In both the bigram and trigram, the use of “hillary” is significantly more than others. The trigram was separated and sorted, making the middle word “hillary” to see the different ways “hillary” was used. The result in Figure 7 shows that the word “crooked hillary clinton” is used 72 times. The other words in the top 10 range that accompany the word “hillary" are "beat," "trump" and "vote."

Figure 7

Trigram on “hillary” Word Use



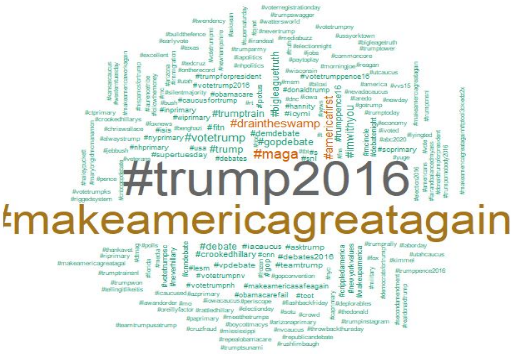
```
# A tibble: 295 x 4
   WordA   WordB   WordC        n
   <chr>   <chr>   <chr>    <int>
 1 crooked hillary clinton     72
 2 beat    hillary clinton      7
 3 trump   hillary clinton      7
 4 beat    hillary easily       5
 5 crooked hillary clintons     4
 6 crooked hillary called       3
 7 crooked hillary close        3
 8 vote    hillary clinton      3
 9 beat    hillary bernie       2
10 beating hillary clinton      2
# ... with 285 more rows
```

Note. Further review of the bigrams and trigrams is found in Figure B and C, Appendix 1, which includes the bigram and trigram word count with the associating IDF values.

A hashtag is a word or collection of them that link tweets that contain that specific hashtag together. For example, if a user were to look at the hashtag #ElectionNight, they would see the most trending tweets (likes and retweets recently) that contain the same hashtag. These could be from celebrities, candidates, journalists, or regular people who had a clever tweet. It joins a particular tweet to an event or conversation that is ongoing. Trump made use of these. Below is a wordcloud of the most common hashtags he used, where the size represents the frequency it appears.

Figure 8

#Tag Word Cloud



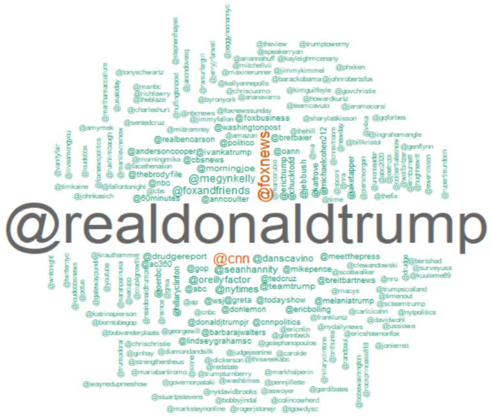
While DT has been criticized for running an unusually negative campaign, this word cloud does not contain overwhelming evidence for it. The hashtags appear to reveal much self-promotion. The top five were all campaign slogans. Other hashtags, such as #votetrump, #trumptrain, and #trumpforpresident, are clearly about the election. A significant proportion the hashtags used was referencing the Republican primaries and campaigning in states to win the nomination. The most prominent hashtag with a negative, attacking tone (also called mudslinging) was #draintheswamp, in conjunction with #crookedhilary and #obamacarefail. However, these do not dominate the other hashtags used. Of course, DT could have been very aggressive in attacking other individuals without using hashtags, which would not be captured in Figure 8.

A majority of counts are linked to DT sharing supporting people and their posts, to include capturing retweets. The messages shared capture, both witty messages, and support for DT. As a result, these messages are often retweeted. It appears that DT is continually reviewing his twitter feed, liking, and retweeting select messages or posts.

There is more evidence of DT retweeting positive messages about himself in the frequencies of account names that appear. His twitter handle appears far more than any other. Again, DT is likely scrolling through his mentions on the platform, liking and retweeting. The next tier of handles is the news media and their personalities. These counts are going to be an amalgamation of tweets about a news piece they reported (both about him and not), retweeting something they said on their account, and DT talking directly to them about something they said. A large portion of DT's campaign was spent on media outlets and how he feels they are biased and pursuing an agenda. Although the content of messages sent to these accounts is not represented, it is evident the media is on DT's mind frequently.

Figure 9

@tag Word Cloud



The retweeting of endorsements is traditional campaigning, showing voters are the right group for the job by the positive things people are saying on the subject. Trump certainly does this because his twitter handle would not appear unless he retweeted others mentioning him.

An additional set of variables was created to understand the frequency of the DT tweets in an hour and month. Figure 10 shows that the frequency of the tweets increased during the early hours. The time around 2 pm shows more than 500 frequency, and the frequency drops during the day just to pick up in the afternoon. The lowest frequency of tweets is observed between 5 pm, and 10 am. The month count of the tweets in Figure 11 shows that there is a higher number of tweets in October, to be followed by July. April, May, and June show the lowest frequency of the tweets.

Figure 10

Tweet Count per Time

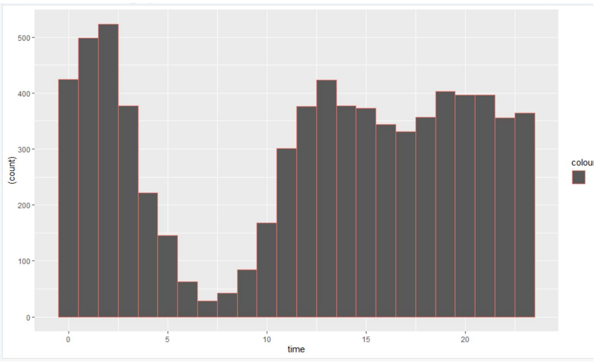
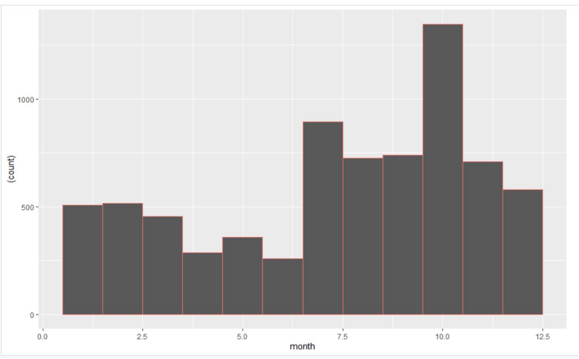


Figure 11

The Monthly Count of Tweets



The daily count of the tweets can be seen on appendix figure A. The tweet count shows dispersed distribution. The 7th, 20th, and 21st days of a month have the highest count of tweets while the last days of the month receive the lowest counts.

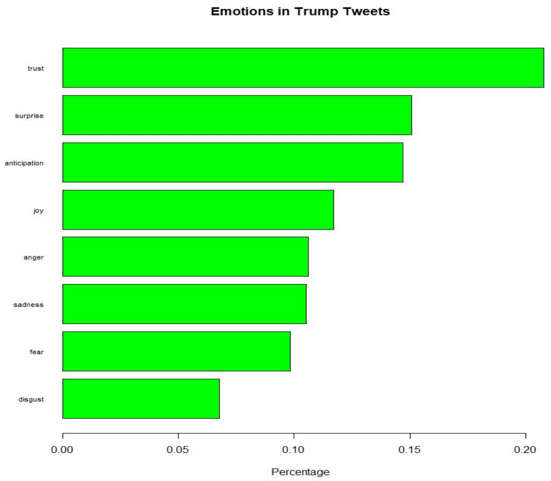
Results and Model Evaluation

Before diving into detail, it is imperative to note, the following analysis and results are not a representation of opinions. Instead, they are based on the emotional sentiment definitions set by an abstracted team of individuals who have no political influence nor customization bias to this analysis. These individuals produce libraries for the emotions that are, in turn, used for analysis. That said, these 'emotional' sentiments may be very flawed and taken with a grain of salt. For example, the terms "miss" and "John" are negative sentiments. It is reasonable to presume on anecdotal evidence that these algorithms contain significant bias (Michael Wiegand, 2018). The sexist interpretation evidences that "miss" is a negative emotion, as there is no context to clearly show "miss," a pronoun for a female, as a socially harmful construct. Of particular note and reference is the word "trump," which both references the individual being analyzed, but is also defined in Merriam Webster as, “a decisive overriding factor or final resource.” (Merriem-Webster, 2020)

Having analyzed the files using the NRC lexicon library, many statistical insights and context gleamed, but the emotional expressions and sentiment were missing. This gap was closed using the Syuzhet Library, which provides eight emotions as well as a positive and negative sentiment (Jockers, 2020). Figure 12 below breaks down these emotions, and it is immediately apparent that DT is using three anxiety-related emotions as ~51% of his twitter communication. Accurately, "anticipation" and "surprise" represent ~31% of emotions, which are only superseded by the emotion 'trust' at 30%. One interpretation of this result is that DT generates a sense of anxiety caused by a combination of known ('anticipation') and unknown ('surprise') sensations followed by 'trust' emotions. Given his name is used 5.5x more than any other word, it is reasonable to associate this sense of 'trust' with DT himself and the "Trump brand." Also, 'fear' is the least used emotion, which can be interpreted as both not disseminating 'fear' to listeners as well as not being fearful of himself. With the emotion of 'trust' as highest and 'fear' as the lowest, DT is projecting confidence either to himself or his audience (Saif M. Mohammad, 2018). Reference Figure G, Appendix 1.

Figure 12

Syuzhet Emotional Sentiment Analysis



Analysis above is reinforced with the overall positive sentiment of DT's tweets, as defined by NRC and Syuzhet libraries. One visualization of this shows a density correlation of the NRC and Syuzhet libraries and noted by the positioning and density in figure 13. Another visualization of the positivity was noted earlier in figure 4 and again in figure 14 using the Syuzhet library. Of note, however, are two anomalies: 1) the term 'trump' is an outweighed positive count by a factor greater than 5 of all other words and the spike timeline of positivity of the limited data set, as noted in figure 15. Over time, trump is "negative," but there is a spike in positivity towards the end of the data set. If removing the "trump" word from the data set, the overall trend changes to a slightly negative weighted value. This again highlights the outsized use and effect of the word/brand "trump." Reference Figures D and E, Appendix 1, for additional plots of emotional variance overtime.

Figure 13

Syuzhet Vector over NRC Vector Valence Density Plot of Positive and Negative Emotion

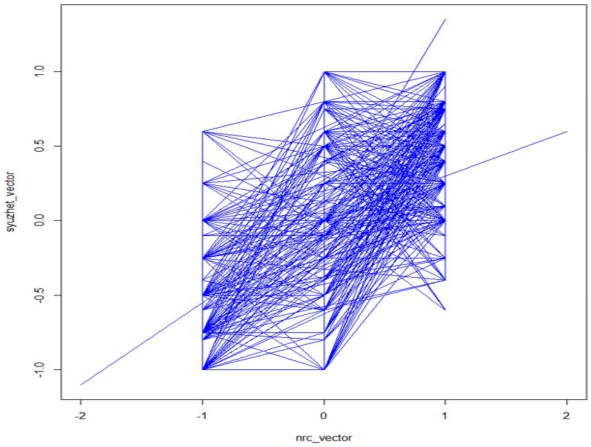


Figure 14

Positive/Negative Sentiment- Syuzhet Library

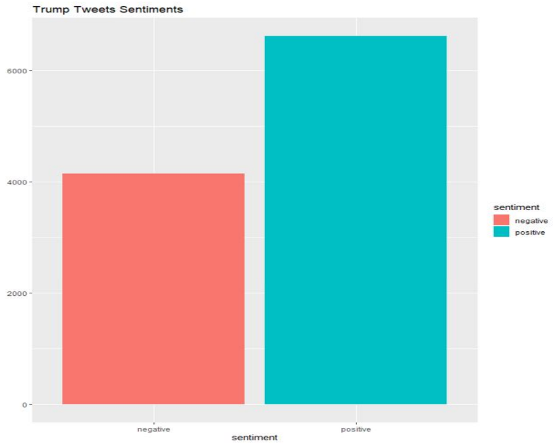
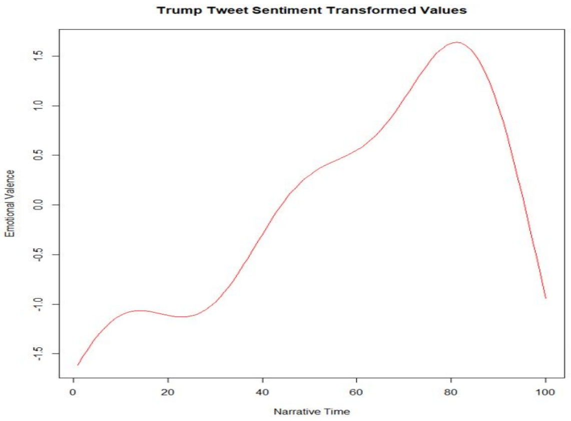


Figure 15

Emotional Valence Sentiment



Given the positivity of the results, a review of emotional statements was made as proof of concept toward emotions. Specifically, what hashtags "#" and what entities "@" are used when "angry tweets" are sent to the world. Using a word cloud to visualize these words, we see the same @realdonaldtrump pattern, but top secondary entities are media outlets, reporters followed by political opponents.

Emotionally angry tweets maintain the top 2 # noted earlier, but the term ‘#draintheswamp' is the third largest # that was missing before. Curiously, common campaign slogans are present in angry tweets. This may be because these tweets are about an event/person that DT's campaign and supporters do not like, then mentioning DT or his campaign. It may be a way of pointing out the things that need to change, then inferring DT will make those changes.

For example, there is a retweet in the dataset stating, "Our leaders are dummies. They dont know whats going on. Its true, theyre incompetent. #Trump #MakeAmericaGreatAgain". They are speaking angrily. However, the message is not directed towards DT or his campaign.

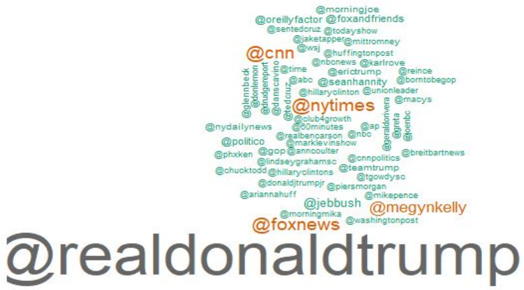
These findings show the value of text mining and pose future research vectors for detailed emotional analysis using the existing code base as a bootstrap. Not only can emotional values be derived, but emotional comparative analytics, time series analysis, and the world even correlations are possible as well.

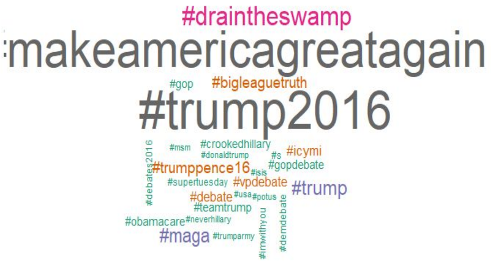
Figure 16

Angry Entities with @ Symbol

Figure 17

Angry Entities with # Symbols





Conclusion

The above analysis explains how the DT tweet dataset was used to understand the sentiment of the tweets. To achieve that, different libraries were used to create word counts or frequencies and different visuals. This helped in identifying the frequent words, the sentiment associated with each word. In terms of how words are perceived or classified, limitations exist between each library used above. Although the negative sentiment is found in several instances, the positive side of sentiment overtakes posts or responses that are assessed to be negative. All in all, the sentiment captured within DT's tweets were guided through effective planning and preparation, which hone in on different levels of sentiment, from multiple perspectives that drive understanding of opinion and emotion, resulting in a large and extensive popularity trend.

Appendix 1

Figure A

Tweet Count per Day of the Month

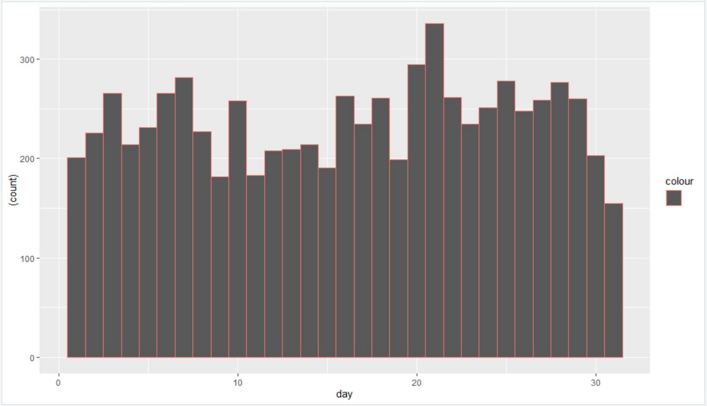
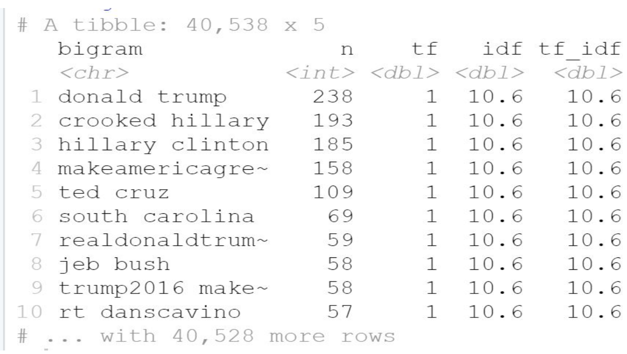


Figure B

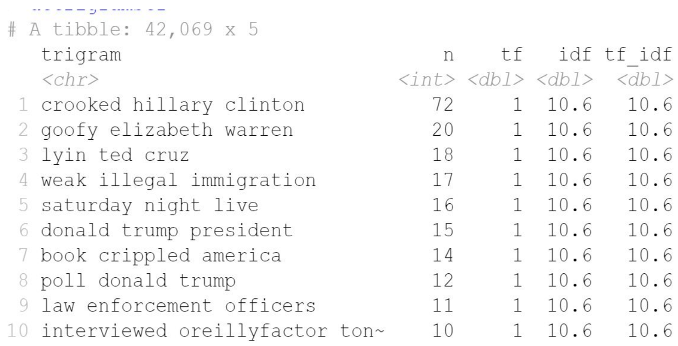
Bigram Words, Count and IDF Values



```
# A tibble: 40,538 x 5
   bigram                n    tf   idf tf_idf
   <chr>             <int> <dbl> <dbl>  <dbl>
 1 donald trump        238     1  10.6   10.6
 2 crooked hillary     193     1  10.6   10.6
 3 hillary clinton     185     1  10.6   10.6
 4 makeamericagre~     158     1  10.6   10.6
 5 ted cruz            109     1  10.6   10.6
 6 south carolina       69     1  10.6   10.6
 7 realdonaldtrum~      59     1  10.6   10.6
 8 jeb bush             58     1  10.6   10.6
 9 trump2016 make~      58     1  10.6   10.6
10 rt danscavino        57     1  10.6   10.6
# ... with 40,528 more rows
```

Figure C

Trigram Words, Count and IDF Values



```
# A tibble: 42,069 x 5
   trigram                             n    tf   idf tf_idf
   <chr>                           <int> <dbl> <dbl>  <dbl>
 1 crooked hillary clinton            72     1  10.6   10.6
 2 goofy elizabeth warren             20     1  10.6   10.6
 3 lyin ted cruz                      18     1  10.6   10.6
 4 weak illegal immigration           17     1  10.6   10.6
 5 saturday night live                16     1  10.6   10.6
 6 donald trump president             15     1  10.6   10.6
 7 book crippled america              14     1  10.6   10.6
 8 poll donald trump                  12     1  10.6   10.6
 9 law enforcement officers           11     1  10.6   10.6
10 interviewed oreillyfactor ton~     10     1  10.6   10.6
```

Figure D

Emotional Variance Over Time using Syuzhet Library

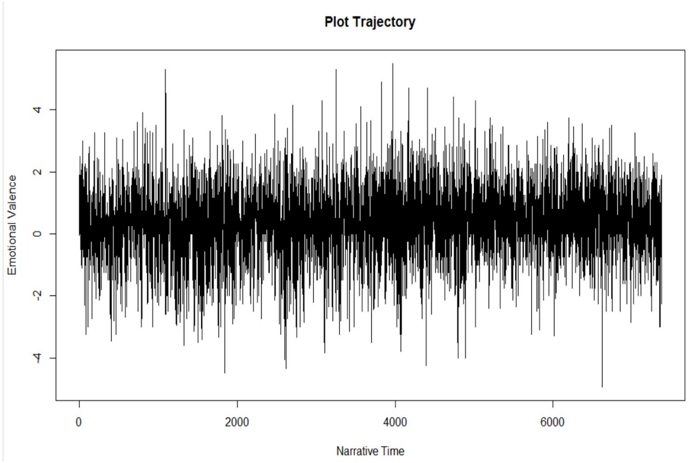


Figure E

Trajectory of Emotion Over Time

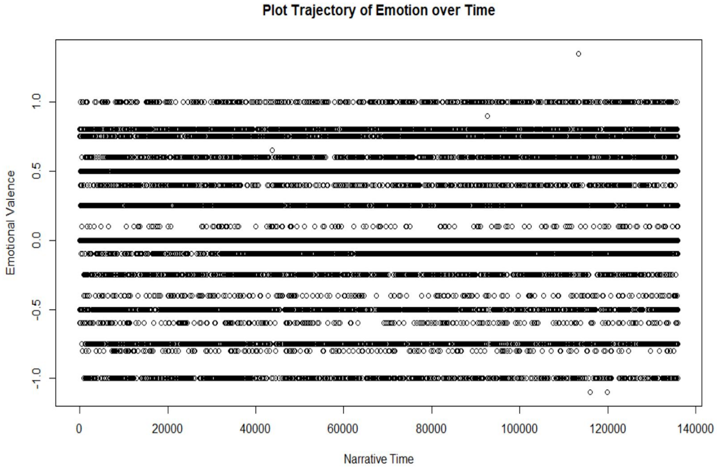


Figure F

The Cos Function using Syuzhet

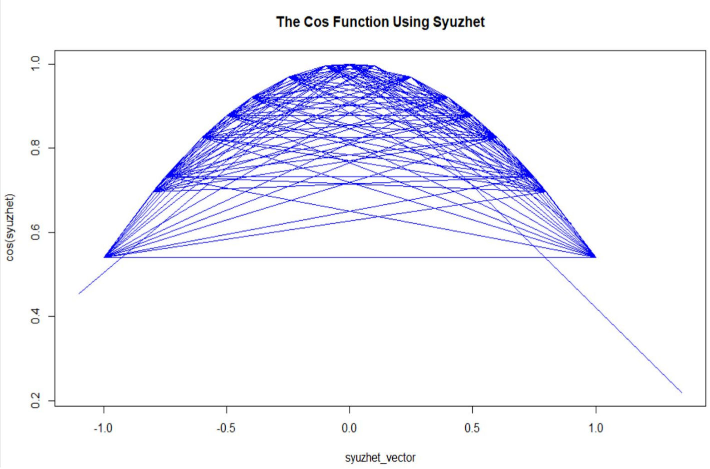
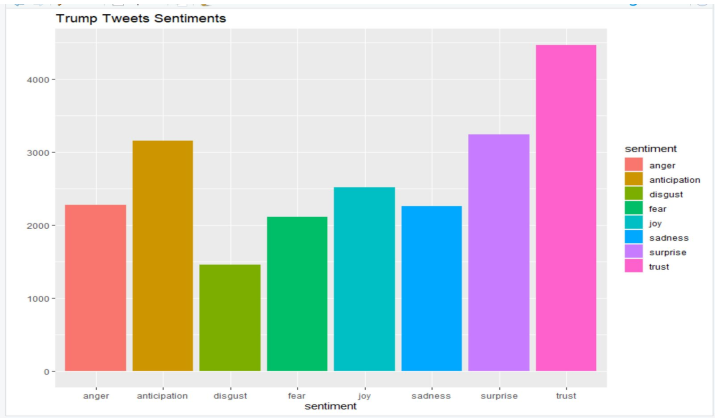


Figure G

Trump Tweets Sentiments



References

Boyle, M. (2016). Exclusive - 'American Comeback Story': Watch Story of How Trump Changed Boxer William Campudoni's Life. Retrieved from https://www.breitbart.com/politics/2016/11/05/exclusive-american-comeback-story-watch-unbelievable-story-trump-changed-boxer-campudonis-life/

Jockers, M. (2020). *Introduction to the Syuzhet Package*. From https://cran.r-project.org/: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

Merriem-Webster. (2020). *dictionary/trump*. From https://www.merriam-webster.com/: https://www.merriam-webster.com/dictionary/trump

Michael Wiegand, A. S. (2018). *Inducing a Lexicon of Abusive Words – A Feature-Based Approach.* Mannheim: Institute for German Language. From https://www.aclweb.org/anthology/N18-1095.pdf

Saif M. Mohammad, S. K. (2018). Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories*. Montreal: National Research Council Canada*.

Silge, J., & Robinson, D. (2017). Text Mining with R: A Tidy Approach. *Sebastopol, CA: OReilly Media*.

Text Mining and Word Cloud Fundamentals in R : 5 Simple Steps You Should Know. (n.d.). Retrieved from http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know

Ward, B. (2019). A light Introduction to Text Analysis in R. Retrieved from https://towardsdatascience.com/a-light-introduction-to-text-analysis-in-r-ea291a9865a8