**ALY 6040 Data Mining Applications**

**Week 5 Assignment**

Northeastern University

Professor Justin Grosz

A picture containing black, photo, red, sitting

Description automatically generated

College of Professional Studies, Northeastern University, Boston, MA 02115

This report was created as part of Week 5 Assignment report for MODULE 5. The report uses data from Trump Tweets for analyzing with the use of the text mining framework “tm” in R.

Submitted by:

Madhav Khanzode

Meenakshi Krishan

Pui Ling Ada Sung

Siddharth Mani

Sirisha Kompella

Vinoothna Peruri

**INTRODUCTION**

This report focuses on the analysis of Trump’s Tweets with the use of the text mining framework “tm” in R. 11 text files of sizes ranging from 4KB to 40KB comprising of Trump’s Tweets in 2016-17 (Murphy, 2017) were imported to create a corpus. Data preprocessing, including the removal of numbers, capitalization, common words and punctuation, was performed on the corpus for the preparation work. Stemming was also used to combine words to a common root to aid in the process of analysis.

**ANALYSIS**

A document-term matrix containing 11 documents and 3659 terms with sparsity of 79% was subsequently generated. Sparse terms were then removed to decrease the number of terms in the matrix down to 87. The histogram in Figure 1 shows that the most common words used in Trump’s Tweets are all very simple in meaning. It suggests not only Trump’s personal style on using direct and simple language, but also his strategy of reaching out to the general public in social media for effective communication. It is observed that Trump used action words (e.g., want, thank, know, take and work) to rally his supporters in his tweets. The most common word “will” was used 428 times, over 4 times more often than the 10th frequently-used word “now”. It indicates that Trump puts a lot more emphasis on the future and the plan of actions than the current situation. The next two most common words “people” and “going” with frequencies over 260 reflect Trump’s motives of engaging general public for his future plan or policy. The word cloud in Figure 2 displays the common words in different sizes based on their frequencies. The top 10 common words include will, people, going, country, great, want, one, know, america and now. Other common words are immigration, law, united, reality and jobs. All these common words apparently coincide with the theme in Trump’s 2016 presidential campaign whose slogan was “Make America Great Again”. It is a direct but somewhat emotional message, especially for those forgotten Americans. Trump tried to draw attention on securing American jobs and illegal immigration (Wikipedia, 2020). Refer Figure 1 and 2.

The association between these common words were further examined by using findAssocs(). The most common word “will” has 65 other highly-correlated words from a wide variety of topics including crime, laws, workers, policy, gunned, Latino, Syria, terror, unemployment, Louisiana , nuclear, Clinton and China. Trump wanted to express his stands on these topics so as to gain trust from the voters. These topics include how to reduce the crime rate, protect American from terrorists, support law enforcements and improve employment. Trump also expressed his opinion on the crisis in Syria and his strong opposition to the trade policy with China as a strategy to win votes from the patriotic American. Trump’s attempt to reach out to Latino conservative Christians turned out to be very successful in getting the majority of Latino votes in 2016 Presidential Election (Khalid, 2016).

The second most common word “people” have high correlation with other words, such as money, corporations, terrorists and worker. In Trump’s mind, he cared about money, economy, business and homeland security when he mentioned “people”. The third most common word “going” has 27 other correlated words including hell, unbelievable, mess and wrong. It indicates that Trump has a sense of negativity when mentioning the media or the fact checks. We checked the words Immigration and Visas with a high correlation limit of 0.95 and both resulted in having strong association with almost same words such as amnesty, death, deported, enforced, illegal, abolishing, criminal, policies, abuse etc. This clearly shows his resentment towards the immigration policy of the country.

It indicates that Trump has a sense of negativity when mentioning the media or the fact checks. we checked the words Immigration and Visas with a high correlation limit of 0.95 and both resulted in having strong association with almost same words such as amnesty, death, deported, enforced, illegal, abolishing, criminal, policies, abuse etc. This clearly shows his resentment towards the immigration policy of the country.

We tried to determine optimal number of clusters and used the sum of intra-cluster distances (withinss) as a function of k and plotted the results on an elbow plot. The plot depicted the optimal value as 5.

Hierarchical clustering was performed by hclust() with the use of Ward’s method to categorize words into different clusters based on the Euclidian distance. The dendogram in Figure 3 shows that the common words are categorized into 5 cluster groups. The first four clusters show the most common words Will, Country, Great, Want, Going, People. The fifth cluster has the remaining common words. It indicates that Trump’s main theme in his tweets is about the future plans of his interests (as discussed previously) and about uniting Americans to build a great country. The fifth group represents the topics less important to Trump. Although there is a popular perception that Trump’s victory is because of his portrayal of immigrants as potential threat to American worker class, the words Visas and Immigration being in the fifth cluster and the words Will, People, Going, Great, Want, One, Country in the first four clusters depicts that he prioritized and talked more about building the country than the rest.

The grouping in K-means clustering and Hierarchical clustering are not identical but somewhat similar, representing Trump’s motives of advertising his presidential campaign via his tweets. Upon comparison of the algorithm with k values as 3, 4 and 5, all of them resulted in the same 96.18% variability with no difference between k being 3 or 5. Finally, k=5 was selected to show a clear separation of 5 distinct components. The first component contains the word Will and second component contains the next two common words People and Going. All other words are contained in the remaining 3 components.

Despite the controversial allegation of his false statements, Trump knows how to utilize online social media to amplify his messages. (Reference in Figure 3 and Figure 4)

**CONCLUSION**

Sentiment analysis provides a way to understand the attitudes and opinions expressed in texts. On trying to analyze Trump’s tweets through sentiment analysis we get an idea if his messages are constructive or not. It also helps us know the state of mind he is in by gauging the usage of the words if they are more positive or negative. What kind of an influence are his tweets to the nation. On comparison of the two we can easily weigh the positive words on the higher end indicating his tweets to be a positive sentiment , meaning his tweets could possibly have a positive influence on people and that Trump wants to take a positive approach when it comes to welfare and nation. (Refer Figure 5)

Another approach of bringing out the positive sentiment is by giving the color green to positive words and color red to the negative words and it clearly brings out how much the green out weighs the red. (Refer Figure 6)

**APPENDIX**

Figure 1: Histogram showing the occurrences of the words (>50) in descending order

A close up of a logo

Description automatically generated

Figure 2:Word cloud showing words in different sizes based on their frequencies

A close up of a piece of paper

Description automatically generated

Figure 3: Dendogram showing 5 clusters generated from hierarchical clustering of corpus

A close up of a piece of paper

Description automatically generated

Figure 4: Principal component plot using K-means clustering (k=2)

A close up of a map

Description automatically generated

Figure 5: Positive and Negative Sentimental analysis of top 30 words excluding stop words

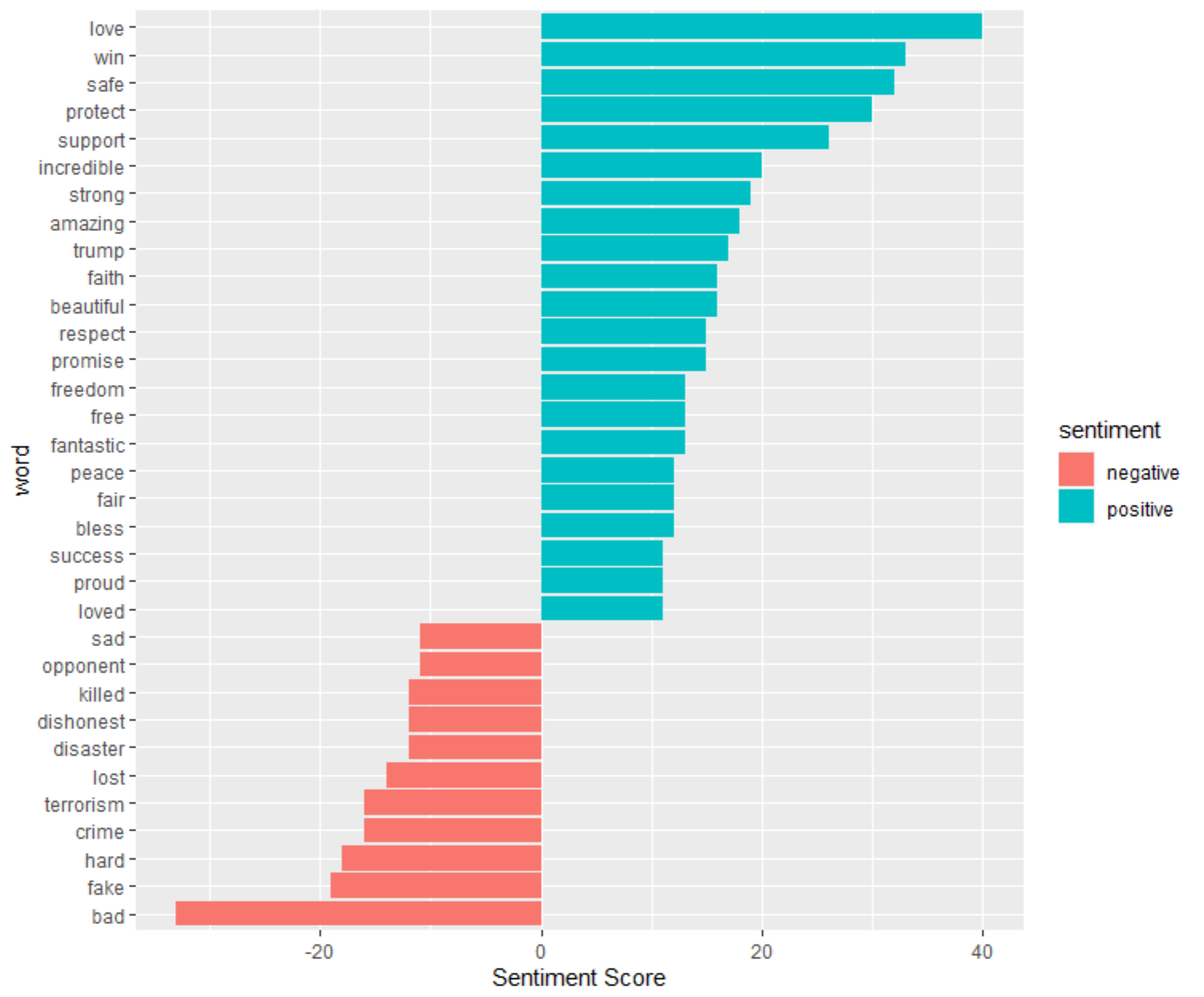


Figure 6: Word cloud of positive and negative sentiments



**Reference**

Khalid, A. (2016). Latinos Will Never Vote For A Republican, And Other Myths About Hispanics From 2016. Retrieved from https://www.npr.org/2016/12/22/506347254/latinos-will-never-vote-for-a-republican-and-other-myths-about-hispanics-from-20

Murphy P. (2017). Trump’s Tweets. Retrieved from <https://drive.google.com/drive/folders/0B914dXn0AXvlaWhmVXdPclZrTjA>

Silge, J. Robinson, D. (2020). Text Mining with R. Retrieved from <https://www.tidytextmining.com/index.html>

Wikipedia (2020). Donald Trump 2016 presidential campaign. Retrieved from <https://en.wikipedia.org/wiki/Donald_Trump_2016_presidential_campaign>