Group Gamma:

Text Mining and Sentiment Analysis of Donald Trump’s Tweets

Jasmine Hawkins, Rishi Hebbar, Raheed Khan, Kelly Khare, Tanushree Kumar, and Paul Linza

University of Maryland Global Campus

DATA630: Machine Learning

Dr. Ami Gates

August 6, 2024

Introduction

With the advent of social media platforms such as Facebook, Twitter, Instagram, and TikTok, politicians have been able to influence and reshape public opinion as well as participate in political discourse. Platforms such as Twitter have been instrumental in creating immediate, direct, and unfiltered forms of political communication, allowing politicians to share their thoughts and policies with their followers instantaneously.

However, few political figures have utilized social media as effectively and controversially as Donald Trump. Throughout his presidential campaign and victory in 2016 to his tenure in the White House, Trump's usage of Twitter became a mainstay and defining feature of his political career. His social media engagement allowed him to communicate directly with his followers, influence public opinion, respond to media critics and other politicians, and even announce significant domestic and foreign policy decisions. Despite the controversial nature of his campaign and his tweets, Trump has been an effective anchor for political engagement and discourse due to his unfiltered, and often provocative style of tweeting (Bulman, 2016).

Trump's tweets have had significant reach beyond his followers, often becoming the subject of many news panel segments and mainstream media outlets, further amplifying the level of political engagement between panelists, critics, and the general public. Due to such widespread reach, Trump has been able to influence key narratives in American political history and change public sentiment on key issues facing Americans (Dimock & Gramlich, 2021)

The objective of this project is to utilize machine learning to analyze a collection of Donald Trump's tweets from July 2015 to November 2016 and understand prevalent themes, sentiments, and topics conveyed before and after winning the 2016 Presidential Election. By applying text mining techniques, several key areas can be explored such as sentiment trends and communication strategies, as well as levels of public engagement and sentiment through the number of likes and retweets each tweet received. This analysis can provide valuable insights for many political scientists and sociologists on how social media can sway and shape thoughts on political narratives.

Analysis

Data Information and Preprocessing

The dataset is the Donald-Trump_7375-Tweets-Excel.csv from Kaggle user Liam Larsen (Figure 1). The file was produced to offer the greatest compilation of tweets from Donald Trump's Twitter account between July 16, 2015, to November 11, 2016 (LiamLarson, 2017). The information was obtained by scraping Twitter and Reddit. The dataset is comprised of 7,375 entries and 10 columns:

- Date: A character variable indicating when the tweet was sent.
- Time: A character variable indicating what time the tweet was sent.
- Tweet_Text: A character variable enriched with the content of the tweet.
- Type: A character variable indicating if the tweet was plainly text or if there was an embedded link.
- Media_Type: A character variable denoting if the tweet contained a corresponding image.
- Hashtags: A character variable populated with the hashtags corresponding to each tweet.
- Tweet_Id: A numerical variable of the tweet's identifier.
- Tweet_Url: A character variable populated with the hyperlink URL to the tweet.
- twt_favorites_IS_THIS_LIKE_QUESTION_MARK: A numerical variable containing the number of likes each tweet received.
- Retweets: A numerical variable containing the number of retweets each tweet received.
- X: An empty column present within the dataset.
- X.1: An empty column present within the dataset.

There are no missing values within the dataset but there are empty characters present within the Media_Type and Hashtag columns, indicating the tweet did not have a corresponding image or a hashtag respectively.

Data preprocessing was done individually for each model. As a result, different methods were used for preprocessing, however, the general idea was the same. Each model used similar preprocessing criteria with some models choosing to perform further preprocessing. It was decided to drop the Tweet_Id and the Tweet_Url columns since these are essentially identifiers for each tweet and have no true significance in any machine learning model. The X and X.1 columns are dropped as they are entirely empty. Additionally, the decision was made to drop the Media_Type column as this column only denotes whether the tweet included an image. The clustering and neural network models had further preprocessing done with natural language processing, the words in each tweet and hashtag will be stripped of any stop word, embedded hyperlink URLs, punctuations, emojis, or any unwanted characters. Additionally, each word will be lemmatized to return return the root word in each tweet. The apriori algorithm had unnecessary columns removed along with renaming come columns; changing some columns to factors, and adding more columns for better description. Some instances also included scaling the variables where necessary. The naive-bayes and decision trees algorithms used discretization to convert some variables into categorical variables. Wherever necessary, the the dataset will be split into a 70% training set used to train the machine learning model and a 30% test set to evaluate and validate the performance of the model.

Exploratory Data Analysis

An exploratory data analysis was conducted to examine the distribution and characteristics of key variables. The following visualizations were created:

Top 100 Terms Word Cloud

A word cloud compiled of the top 100 most frequent words from the corpus of tweets can be seen below and is available in Figure 2 in Appendix A.

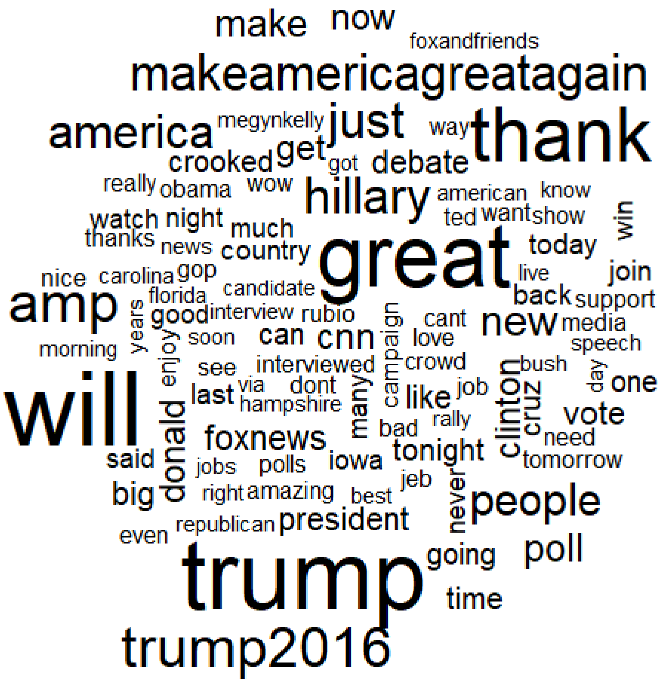


Figure 2. *Word Cloud of Top 100 Terms.*

From the results of the word cloud, it appears the most common words were “great”, “thank”, “makeamericagreatagain”, “will”, “trump”, “america”, “people”, and “trump2016”. Other words to note include “crooked”, “hillary”, and “polls” These words were resonant leading into Trump’s 2016 presidential campaign as he was amassing a large group of supporters and followers with his unique and often controversial use of slogans, hashtags, and political themes.

Top 25 Words Bar Chart

A bar chart consisting of the top 25 words in the collection of tweets is shown in Figure 3 below and in Appendix A.

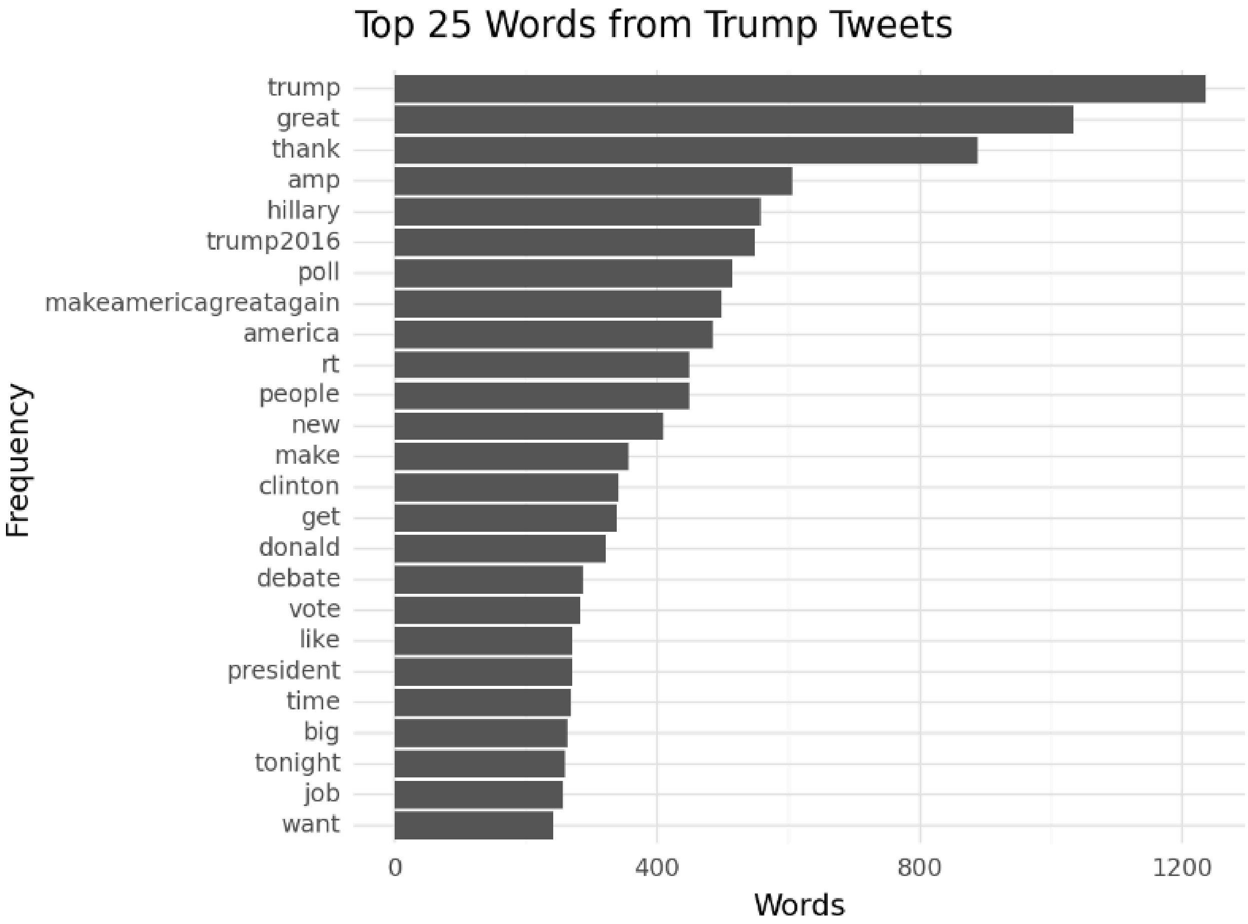


Figure 3. Top 25 words from Donald Trump’s Twitter Account..

From the results of this bar chart, it appears Trump often refers to himself whenever he tweets and very frequently makes tweets against Hillary Clinton as she was his political opponent during the 2016 presidential debate. Additionally, Trump often tweeted his political slogan #makeamericagreatagain as a hashtag which likely resulted in it becoming frequent within the corpus of text.

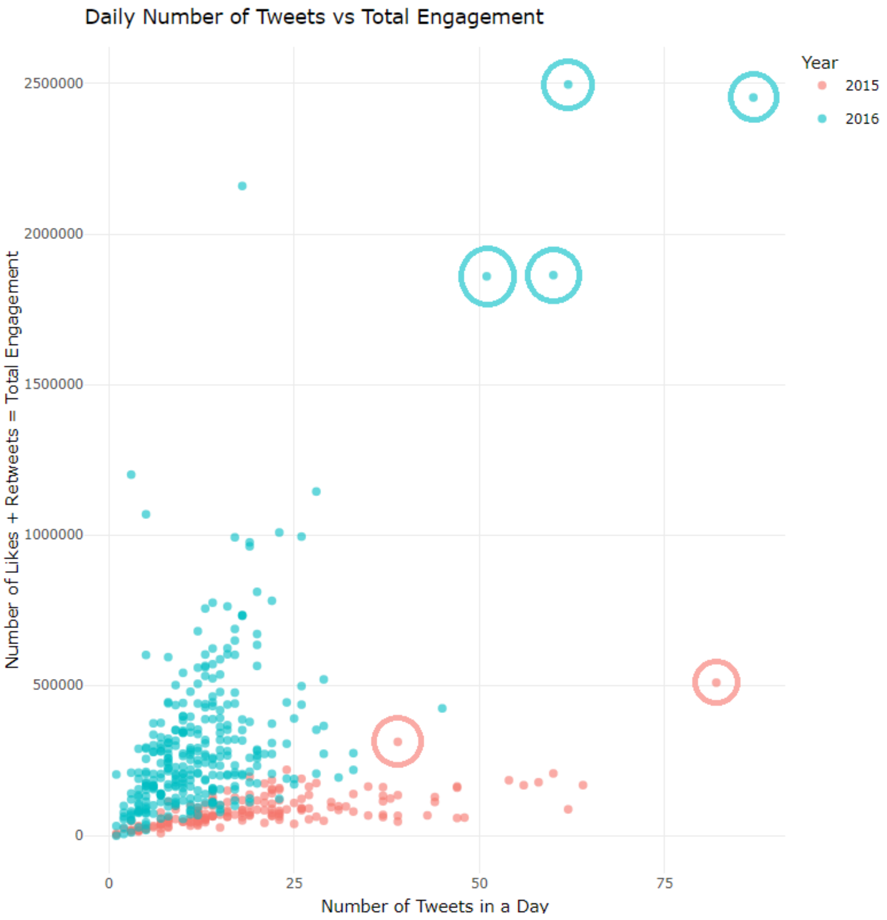


Figure 4. Scatter Plot of Daily Number of Tweets vs. Total Engagement.

The “Daily Number of Tweets vs. Total Engagement” scatter plot was generated to see how the number of tweets posted affects the total number of likes and retweets (Figure 4). Each dot represents an entire day’s tweet history. The x-axis shows the number of numbers in a day and the y-axis shows the amount of total engagement (sum of likes + sum of retweets). The dots are color-coded by year. The red dots represent days in 2015 and the blue dots represent days in

2016. As the number of tweets increases, the amount of engagement should also increase. However, tweets in 2016 consistently received more engagement than tweets in 2015! The scatter plot shows the cluster of blue dots has higher engagement levels than the red dots.

Outliers circled in Figure 4, correspond with the dates of significant events. For example, the 2015 dot with the highest engagement level was posted during the Virginia Trump Rally. The 2015 dot with the second-highest engagement level was on the day when Trump announced a “total and complete shutdown of Muslims entering the United States”. The 2016 dots with the high engagement levels were posted on major debate days, which include the "Hillary vs. Sanders” and “Hillary vs. Trump” debates. This plot shows a right-skew distribution. Most tweets receive a small number of likes and retweets. However, public interest and engagement increase with the proximity to an election day.

Model Methods

The objective of this analysis is to understand Trump’s media campaign. As a result, six different models will be used with each model being evaluated for its accuracy. The methods being used are Hierarchical Clustering, K-Means Clustering, Apriori Association Rules, Naive Bayes, Supervised Algorithm, Logistic Regression, Decision Tree, and Neural Networks.

Pattern Discovery & Sentiment Analysis

Clustering (Raheed Khan)

Clustering is an unsupervised machine-learning algorithm that discovers patterns and relationships within unlabeled data. Specifically, hierarchical clustering attempts to group words or phrases of texts with the most contextual similarities into groups for further analysis. After the tweets have been lemmatized and stripped of stop words, they will be part of a corpus and analyzed as a document term matrix. Once converted, the data will be clustered through a Hierarchical Clustering Algorithm with the number of clusters arbitrarily set to 10. The resulting clusters will be analyzed for common political themes and messages and potentially labeled based on uniquely selected categories.

Apriori Association Rules / K-Means Clustering (Jasmine Hawkins)

Association rules mining is a method used to find relationships between variables in a dataset and use those relationships to make predictions. The Apriori algorithm scans the dataset, finds the most frequent itemsets, and generates association rules. Four sets of association rules were generated from the Donald Trump Twitter dataset: The first set of rules pertains to all the variables in the dataset, including variables such as date, type of tweet, hashtags, number of likes, and retweets. This analysis revealed which factors influence engagement with Donald Trump's tweets. The second set of rules focuses on the tweets that contain hashtags. This analysis will show the impact of hashtags on engagement levels. The third set of rules was applied to the tweet text column. The identification of frequently used terms and their associations can provide insight into important themes for the Trump campaign. The fourth set of rules was applied for the clusters created from K-means clustering. This analysis shows the characteristics of each group of tweets.

The models were built using the apriori function from the 'arules' package. The text was lemmatized, stripped of stop words, and converted into a transactions format. All variables were included in the first general set of rules, but only tokenized words were analyzed in the second text-specific set of rules. The apriori algorithm was given parameters to find rules that were in

40% of the transactions and were true 70% of the time. A grouped matrix plot was generated to visualize the most significant rules from the dataset. Additionally, cluster analysis was conducted to group tweets based on similar characteristics. A silhouette model confirmed that the dataset is best split into five clusters (Figure 12). The five clusters represent the types of tweets and their engagement levels: Cluster 1 tweets have low likes and retweets. Cluster 2 tweets have very low likes and medium retweets. Cluster 3 has a medium engagement for likes and retweets. Cluster 4 has high likes and medium retweets. Cluster 5 has the highest engagement for likes and retweets (Figure 11).

By identifying patterns and frequently used terms, we gain a greater insight into the messaging strategies used by the Donald Trump Twitter team. For example, specific catchphrases are linked to higher engagement with political campaigns.

Predictive Modeling

Naïve Bayes (Kelly Khare)

The Naïve Bayes method is a technique that uses categorical data to predict the probability that a given case belongs to a specific class. This method can be useful to identify the likelihood of Trump’s tweets relating to specific popular topics. This would also be a useful method to capture the different overall sentiments of Trump’s tweets by focusing on word patterns that might express joy, anger, fear, etc.

The Naive Bayes algorithm uses labeled data to train and use probability to solve classification problems. The two packages "e1071" and "arules" are loaded in memory. The

Tweet_Id, Tweet_Url, X.1, and X columns are removed. The column Type is the variable of interest. There are 8 columns or variables left, with 6 in factor form and 2 in integer form.

The columns twt_favourites_IS_THIS_LIKE_QUESTION_MARK and Retweets are discretized. The next step is to split the trump tweets dataset into a train set and a test set with 70% of the data in the training set and 30% in the test set. This is to ensure that there is enough test data to produce an accurate model. The command “model<-naiveBayes(Type~., train.data)” is used to create a Naive Bayes model with the dependent variable being Type. This model will be used to predict the probable type of Trump’s future tweets.

Decision Tree Methods (Paul Linza)

Decision trees are a useful model for supervised learning that can be used for the analysis of Donald Trump’s tweets. Trees split the data based on the levels of the independent variables at each node to predict the level or value of the target variable. The records are split until no more statistically significant divisions can be made. The target variable for this analysis is the like/retweet ratio, which can be turned into levels and used for classification models. This is an important metric because it can help show the level of engagement on a given post, which people like politicians and influencers want to maximize.

Three methods, in particular, bagging, boosting, and random forests, will be used to best classify tweets and understand underlying predictors. These models all aggregate a large number of trees in order to reduce variance and improve accuracy. Bagging and random forests also use bootstrapping, a process of producing multiple samples from a training set to avoid the need for multiple sets. These two models can be evaluated by using the out-of-bag error rate or testing predictions using a confusion matrix. The key difference between bagging and random forests is the number of variables considered by each, since bagging models use all the variables while random forests only take a subset the size of the square root of the cardinality Boosting models are different because they learn based on previous trees they constructed and do not use bootstrapping. The output from boosting models is also continuous numeric data.

In this analysis, the decision tree models were all constructed using a discretized version of the dataset. The target variable was also constructed and clustered to have three levels. A second dataset was also constructed to predict tweets using up to one commonly appearing hashtag. The datasets were then broken into training and testing sets, which were used by all three model types. Finally, each model was used to predict the target variable.

Logistic Regression (Rishi Hebbar)

Logistic regression is used for binary classification problems, where the outcome or dependent variable is categorical with two possible outcomes. Logistic regression models will be used to predict controversial versus non-controversial tweets based on a key variable, retweets. This will help understand how engagement is impacted by the controversial nature of the tweet. In order to do this, data cleanup will be necessary in order to reduce features in the dataset. Next, by manually inputting controversial keywords, a column, is_controversial, will be created from the tweet_text column and classified as noncontroversial (0) and controversial (1). This will be the target variable. The dataset is then broken into test and training sets which will be used for this model.

Neural Network (Tanushree Kumar)

A neural network is a computational model inspired by the biological neural networks of the human brain. It consists of interconnected layers of nodes, or "neurons," which process and transmit information. Each connection between neurons has an associated weight that is adjusted during training to minimize the error between predicted and actual outputs. Neural networks are widely used in various machine learning tasks due to their ability to learn complex patterns and relationships from data.

To effectively utilize text data in a neural network, it must be converted into a numerical format that the model can process. The Document-Term Matrix (DTM) is a widely used approach for this purpose as it represents text data as a matrix where each row corresponds to a document (tweet) and each column represents a term (word) from the entire corpus. The values in the matrix indicate the frequency or presence of each term in the corresponding document.

The DTM transforms raw text into a structured numerical format. Each entry in the matrix reflects the occurrence of a term in a tweet, which is essential for inputting data into a neural network. By representing tweets as vectors of term frequencies, the DTM allows the neural network to learn which words are important for distinguishing between different types of tweets or sentiments. The DTM can be refined by removing sparse terms which reduces the dimensionality of the data and enhances the efficiency of the neural network by focusing on the most relevant terms.

Once the text data is converted into a DTM, it serves as the input for the neural network. The neural network is trained to classify tweets based on their content. During training, the network adjusts the weights associated with each connection to minimize the error between predicted and actual classifications. After training, the neural network can predict the type of new, unseen tweets based on their DTM representation. The performance of the model is evaluated using accuracy metrics which measure the proportion of correctly classified tweets.

Results

Clustering (Raheed Khan)

Output. Figure 15 in Appendix A is a hierarchical dendrogram of the Term Document Matrix based on the corpus of tweets with sparse terms removed. By removing sparse words, the hierarchical clustering will not be as crowded, and only the most frequent and important words are shown. From the resulting dendrogram, it appears words such as "makeamericagreatagain", "rt", "new", and "america" are clustered into one group which may be the result of Trump retweeting supporters, public officials, or news media hosts that provide him with high favorability ratings and adding his signature hashtag #makeamericagreatagain towards the end of his retweet. Other clusters to note include the words "hillary" since she was frequently mentioned while Trump was running for president. Additionally, the words "trump" and "realdonaldtrump" were often noted and grouped closely together since Trump often referred to himself whenever retweeting.

Apriori Association Rules / Text Mining (Jasmine Hawkins)

Output. The four sets of association rules were generated to show significant connections between the variables. The following insights were gathered:

1. The Full Dataset: The full association rules showed the strong relationships between the likes, retweets, hashtags, and time of day variables (Figure 5). The top rule stated that

tweets with 501-1,000 likes would at least have 100-500 retweets. This had a confidence of 91% and a high lift of 11.87. The second top rule states that tweets with 50,0001 - 100,000 likes would get around 10,001 - 50,000 retweets. The confidence was 99% with a lift of 9.5. A high number of likes is strongly correlated with a high number of retweets.

2. Tweets with Hashtags: The hashtag association rules also had several top rules relating to the number of likes matching the number of retweets (Figure 7). However, the #DemDebate and #DrainTheSwamp hashtags had strong connections to the month of October. The confidence levels were 91% and 84% respectively. If paired with the "Trump2016 and October equals 1,000 - 5,000 likes" rule, analysts can surmise that the Trump campaign focused on trashing the opponent's credibility while promoting Trump as the best option.
3. Tweet Text: The analysis of the 'tweet_text' column revealed interesting associations with prominent figures (Figure 9). The association between the adjective "Crooked" and the name "Hillary" had a confidence of 19.75% and a lift of 13.36. The connection between President-Elect Trump and his Communications Advisor Dan Scavino was also highly ranked. Similarly, the rules featured permutations of "make", "great", and "america". This highlights the fact that the Trump campaign focused on catchphrases and terms.
4. Clusters: Clustering the dataset allowed for the generation of rules specific to each cluster (Figure 13). Cluster 1 was related to 1,001-5,000 likes. Cluster 2 was connected to 5001-10,000 likes. Cluster 3 was linked to 501-100 likes. Cluster 4 is related to 10,001-50,000 likes. Surprisingly, Cluster 5's relationship with 0-100 likes and 101-500

likes had a confidence level of 100%. Each cluster could be examined with another machine learning method to determine why they had such different engagement levels.

Evaluation. The apriori algorithm is evaluated using the confidence, lift, and support metrics. The high confidence values indicate strong relationships with high reliability. While high support and lift values indicate meaningfulness. Additionally, grouped matrix plots were generated to illustrate the strength of the top rules (Figure 6, Figure 8, Figure 10, Figure 14). For K-means clustering, the cluster sum of squires and silhouette analysis were generated to evaluate the model. Both showed that the clusters had 84.8% total variance. The clusters were well separated with unique characteristics. The combination of Apriori association rules and K-means clustering gave an initial look into the factors that influence public engagement and the characteristics of different tweets in the dataset.

Naive Bayes (Kelly Khare)

Output. Once the model is created, it shows that the Apriori probability of the type of Trump's tweets is approximately 0.12 or 12% for a link in the tweet, 0.05% for a video, or 0.87 or 87% for text. There is a 21% conditional probability that tweets with a link type would have over 5,000 favorites. There is a 16% conditional probability that tweets with a text type would have over 5,000 favorites.

Evaluation. The command "mean(predict (model, train.data, type="class")== train.data$Type)" is used to determine the accuracy of the training set, which is approximately 86%. The classification error for the training set is approximately 14%. The model created has good accuracy and is capable of making accurate predictions on the type of tweets Trump may have in the future. Based on the Apriori probabilities, Trump may mostly have text-type tweets but his videos and links have a high popularity, shown by the number of favorites. His videos and links are more likely to be popular and go “viral”. Finally, a confusion matrix is created to determine the classification accuracy of the test data. The confusion matrix shown in Figure 16 in Appendix A results in an 85% classification accuracy of the test data used to create the Naive Bayes model. This suggests that the model has a high rate of accuracy in classifying future tweets accurately.

Decision Tree Models (Paul Linza)

Output. There were a total of four decision tree models built, which all attempted to classify the like-to-retweet ratio based on the three clusters obtained from discretization. The accuracy of the models ranged from roughly 66% to roughly 72%, as seen in Figure 17.

Evaluation. The boosting model performed worst due to the nature of the output being different than needed and the difficulty converting the continuous output to levels. The bagging model was the best by a narrow margin, beating out the two random forest models. This could possibly be attributed to the low number of predictive variables making it difficult for the random forest to be as effective. The contrasting error rate graphs of the random forest models can be seen in Figure 18 of Appendix A. Using hashtags as a predictive variable did not lead to higher accuracy for random forests but did hurt the other two models when tested, which is why those were excluded from the final analysis. A variable importance chart can be seen in Figure 19, where “Hashtags” were only more important than media type for reducing the Gini coefficient.

Due to the distribution of the original data/clusters, the models all tended to classify data into the middle cluster too frequently. This can be seen by the sensitivity and specificity values along with the positive and negative prediction values in the confusion matrix in Figure 20 of Appendix A. This trend is also visualized for easier understanding, with an accompanying histogram of the actual values (Appendix A Figures 21 & 22). Based on this knowledge, the engagement of a tweet or the virality of it will likely fall somewhere along a normal bell curve distribution.

Logistic Regression (Rishi)

Output. The model formula is is_controversial ~ ., meaning that the model is predicting the is_controversial variable using all other available variables in the tweetsTrain dataset. Figure 23 in Appendix A details the model. The coefficients indicate the relationship between each predictor variable and the likelihood of a tweet being classified as controversial. A positive coefficient increases the log odds of a tweet being controversial, while a negative coefficient decreases it. The significance codes (***, **, *) next to each p-value provide a shorthand for the level of statistical significance: *** indicates a p-value < 0.001, ** indicates a p-value < 0.01, * indicates a p-value < 0.05.

Evaluation. The logistic regression model achieved a reasonable level of overall accuracy but displayed a marked bias towards predicting non-controversial tweets. Figure 24 in Appendix A shows the predicted probabilities of being controversial versus retweets. As the number of retweets increases, the predicted probability of the tweet being controversial also increases. This is consistent with the idea that more widely shared tweets might be more likely to be controversial. However, the smooth curve shows a nonlinear relationship, where the probability increases with retweets but then starts to decrease after a certain point. This suggests that extremely high retweet counts might be associated with less controversial content, possibly because very popular tweets might be more general and less polarizing. With around 100,000 retweets, the predicted probability begins to decline, indicating that tweets with exceptionally high engagement might be perceived as less controversial. This could be due to the nature of viral tweets, which often appeal to a broader audience. Figure 25 of Appendix A shows the confusion matrix (which was used to predict accuracy, Recall, Specificity, and Balanced Accuracy. While it performs well in recognizing non-controversial content, it struggles to accurately classify controversial tweets, as shown by the low specificity and balanced accuracy. These results suggest the need for further model refinement to enhance its capability to identify controversial content.

Neural Network (Tanushree Kumar)

Output. The neural network model developed for classifying the tweets produced a complex network with multiple layers and connections, as seen in Figure 26 in Appendix A. The network includes an input layer with nodes representing keywords from the tweets, two hidden layers, and an output layer. The output nodes correspond to the types of tweets: text and link. The neural network was trained with a total of 24,350 steps, achieving a final error value of 553.97. The training process reached a minimal error threshold of approximately 0.01, indicating convergence. The plot illustrates the architecture of the neural network, which includes an input layer with nodes representing keywords from the tweets, two hidden layers, and an output layer that classifies tweets into three types: 0, 1, and 2.

Evaluation. The model had an accuracy of 0.87799 or approximately 87.8%. This metric indicates that the neural network correctly classified tweets 87.8% of the time in the test dataset. An accuracy of 87.8% suggests that the model is performing well, correctly predicting the class of the tweets in a significant majority of cases. This is a strong result, indicating that the model has effectively learned patterns in the data and can generalize well to unseen examples. This high accuracy also means that the neural network is robust in distinguishing between different types of tweets, based on the features extracted from the text data.

The ROC curve as seen in Figure 27 in Appendix A shows the trade-off between sensitivity and specificity for the model. The Area Under the Curve (AUC) is 0.609, which indicates a relatively low ability of the model to distinguish between the different tweet types. The overall accuracy of the model is 87.8%, which is primarily driven by its performance in classifying Class 1 tweets. The high accuracy is misleading because the model fails to classify Class 0 and Class 2 tweets effectively. The Kappa statistic is very low, -0.0018, indicating a poor agreement between the predicted and actual tweet types beyond chance.

Conclusion

Summary

The 45th President of the United States once said, “I think that social media has more power than the money they [Opponent Hilary Clinton] spent”. Social media provided Trump with a cheaper and more effective form of campaigning (McCormick, 2016). In this analysis, several different machine-learning models (both supervised and unsupervised) were used to examine how Donald Trump used Twitter in 2015 and 2016 leading up to and just after winning the presidential election. The most common tweets had a few commonalities. First, they featured threats. The opponent Hillary Clinton was described as “crooked” and the political sphere was represented as a “swamp” that needed to be drained. Second, the concept of America’s future was promoted. America isn’t a good country right now, but it could be. Third, Donald Trump was exalted as the solution to the country's ails. If you vote for Trump, America will become great again. This cycle of threat, possibility, and solution was highlighted in their Twitter social media campaign. This empowered Donald Trump to create alt-right echo chambers that significantly amplified his messages and helped him win the presidential election.

Limitations

The analysis was limited by the dataset. On Twitter, a user's tweet can receive likes, retweets, and comments. When the tweet's number of comments exceeds the number of likes and replies, it suggests that more people disagree with the tweet's information. This occurrence is known as "getting ratio'd" because the ratio of comments is much higher than the number of likes and retweets. Additionally, The dataset did not have a column for the number of comments for each tweet. The dataset only represents a very short and specific time period which does not fully capture the broader patterns in political discourse and the related social media behavior. Additionally, the complexity of human language means that this approach might miss nuances and context crucial for understanding the true meaning behind the words.

Improvement Areas

Three areas could improve the analysis of Donald Trump's Twitter campaign. First, further research should be conducted on Donald Trump's entire Twitter account. The dataset used was the subset of Donald Trump's tweets from the 2016 Presidential Campaign period. A comparison between tweets made during the campaign and tweets made during the presidency could show changes in communication strategies and shifts in themes. Second, future studies could include posts from other social media platforms (ex: Facebook). Researchers could explore how different social media platforms impact different segments of the population. Interviews and news segments could also be factored in to add a more holistic approach. Third and finally, these posts should be analyzed with real-world events. A contextual analysis can examine the effects of Donald Trump’s social media usage on public opinion and political discourse. Incorporating additional context, such as images, news articles, and links in the tweets, could provide a deeper understanding of the sentiment and intention behind the messages. This will not be the last time that Twitter will affect the outcome of an election. In fact, it has introduced a new wave of political social media campaigns that focus on the most extreme voters.

References

Bulman, M. (2016, November 28). This man analyzed Donald Trump's tweets and found something that might explain why he won. The Independent. https://www.independent.co.uk/news/world/americas/donald-trump-twitter-account-election-victory-presidentelect-david-robinson-statistical-analysis-data-scientist-a7443071.html

Dimock, M., & Gramlich, J. (2021, January 29). *How america changed during Donald Trump's presidency*. Pew Research Center. https://www.pewresearch.org/politics/2021/01/29/how-america-changed-during-donald-trumps-presidency/

Lind, D. (2015, December 7). *Donald Trump proposes "total and complete shutdown of Muslims entering the United States."* Vox. https://www.vox.com/2015/12/7/9867900/donald-trump-muslims

LiamLarsen. (2017, April 16). (better) - donald trump tweets!. Kaggle. https://www.kaggle.com/datasets/kingburrito666/better-donald-trump-tweets?resource=download

McCormick, R. (2016, November 14). Donald Trump says Facebook and Twitter "helped him win." The Verge. https://www.theverge.com/2016/11/13/13619148/trump-facebook-twitter-helped-win

Seth A. Richardson, cleveland. com. (2021, January 17). New study examines how Donald Trump used Twitter to craft an alternate reality for his followers. cleveland.

https://www.cleveland.com/open/2021/01/new-study-examines-how-donald-trump-used-twitter-to-craft-an-alternate-reality-for-his-followers.html

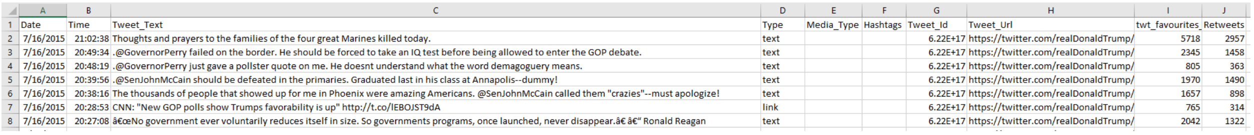
Trump, D. (n.d.). X.com. X (formerly Twitter). https://x.com/realdonaldtrump?lang=en

Appendix A

All figures, tables, and visualizations mentioned in the report can be seen below. The R code is attached as a separate file.

Figure 1

A figure of the Trump Dataset.



	A	B	C	D	E	F	G	H	I	J
1	Date	Time	Tweet_Text	Type	Media_Type	Hashtags	Tweet_Id	Tweet_Url	twt_favourites_	Retweets
2	7/16/2015	21:02:38	Thoughts and prayers to the families of the four great Marines killed today.	text			6.22E+17	https://twitter.com/realDonaldTrump/	5718	2957
3	7/16/2015	20:49:34	.@GovernorPerry failed on the border. He should be forced to take an IQ test before being allowed to enter the GOP debate.	text			6.22E+17	https://twitter.com/realDonaldTrump/	2345	1458
4	7/16/2015	20:48:19	.@GovernorPerry just gave a pollster quote on me. He doesnt understand what the word demagoguery means.	text			6.22E+17	https://twitter.com/realDonaldTrump/	805	363
5	7/16/2015	20:39:56	.@SenJohnMcCain should be defeated in the primaries. Graduated last in his class at Annapolis--dummy!	text			6.22E+17	https://twitter.com/realDonaldTrump/	1970	1490
6	7/16/2015	20:38:16	The thousands of people that showed up for me in Phoenix were amazing Americans. @SenJohnMcCain called them "crazies"--must apologize!	text			6.22E+17	https://twitter.com/realDonaldTrump/	1657	898
7	7/16/2015	20:28:53	CNN: "New GOP polls show Trumps favorability is up" http://t.co/lEBOJST9dA	link			6.22E+17	https://twitter.com/realDonaldTrump/	765	314
8	7/16/2015	20:27:08	â€œNo government ever voluntarily reduces itself in size. So governments programs, once launched, never disappear.â€ â€" Ronald Reagan	text			6.22E+17	https://twitter.com/realDonaldTrump/	2042	1322

Figure 2

Word Cloud of Top 100 Terms.

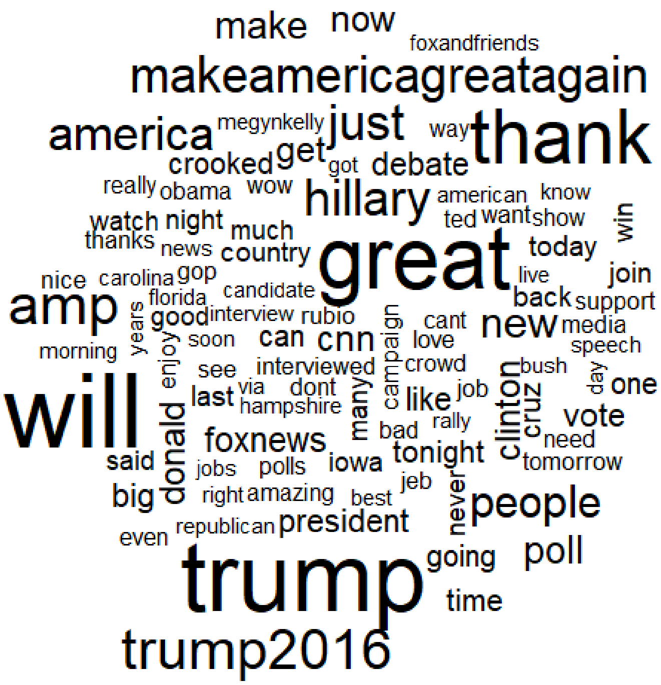
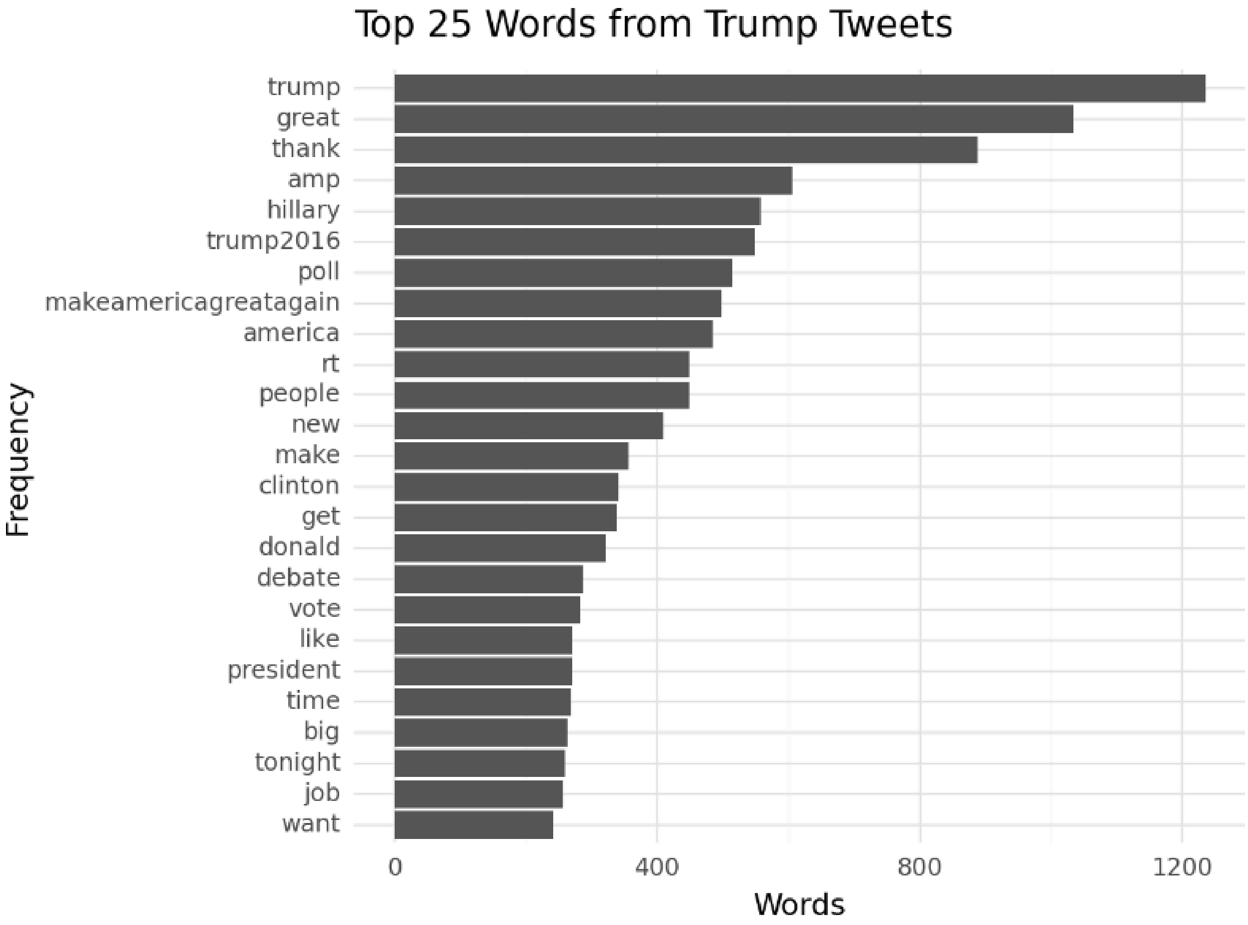


Figure 3

Top 25 words from Donald Trump's Twitter Account.



.

Figure 4

Scatter Plot of Daily Number of Tweets vs. Total Engagement.

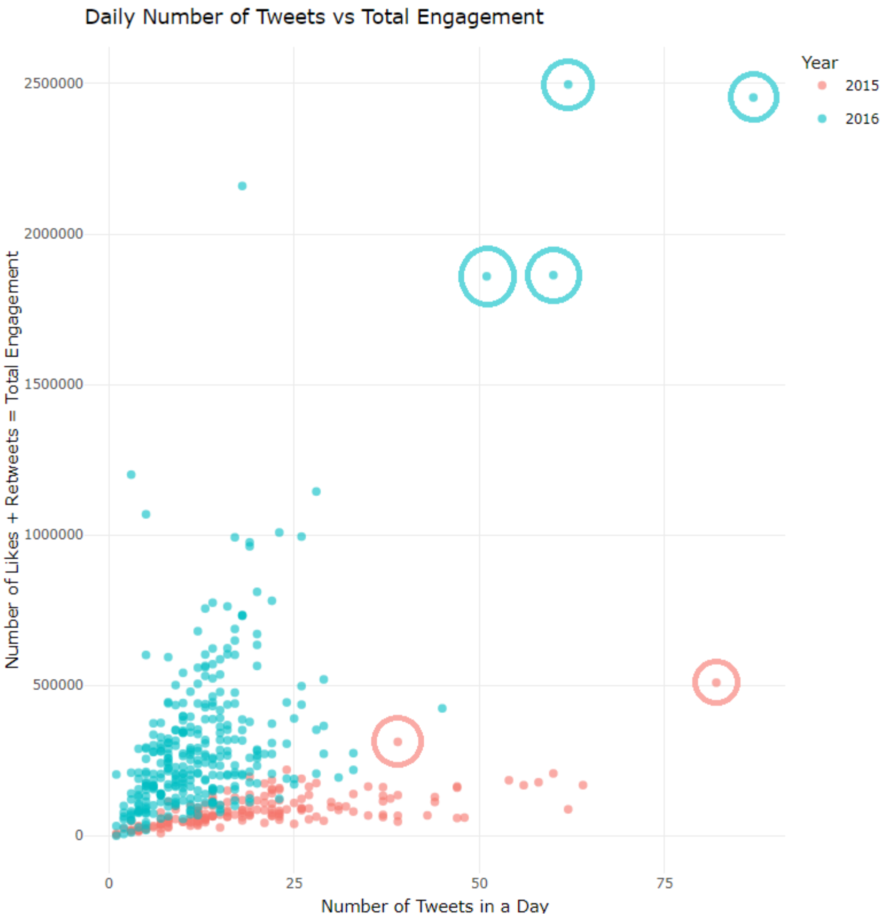
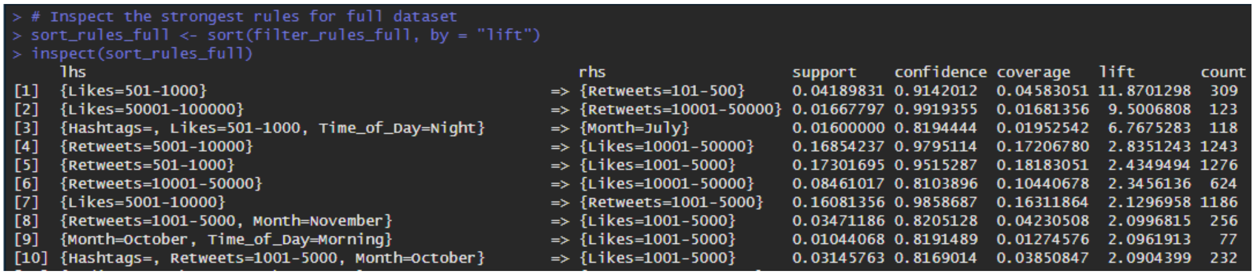


Figure 5

Full Dataset Strongest Rules.



```
> # Inspect the strongest rules for full dataset
> sort_rules_full <- sort(filter_rules_full, by = "lift")
> inspect(sort_rules_full)
     lhs                                                     rhs                    support    confidence coverage   lift       count
[1]  {Likes=501-1000}                                     => {Retweets=101-500}     0.04189831 0.9142012  0.04583051 11.8701298  309
[2]  {Likes=50001-100000}                                 => {Retweets=10001-50000} 0.01667797 0.9919355  0.01681356  9.5006808  123
[3]  {Hashtags=, Likes=501-1000, Time_of_Day=Night}       => {Month=July}           0.01600000 0.8194444  0.01952542  6.7675283  118
[4]  {Retweets=5001-10000}                                => {Likes=10001-50000}    0.16854237 0.9795114  0.17206780  2.8351243 1243
[5]  {Retweets=501-1000}                                  => {Likes=1001-5000}      0.17301695 0.9515287  0.18183051  2.4349494 1276
[6]  {Retweets=10001-50000}                               => {Likes=10001-50000}    0.08461017 0.8103896  0.10440678  2.3456136  624
[7]  {Likes=5001-10000}                                   => {Retweets=1001-5000}   0.16081356 0.9858687  0.16311864  2.1296958 1186
[8]  {Retweets=1001-5000, Month=November}                 => {Likes=1001-5000}      0.03471186 0.8205128  0.04230508  2.0996815  256
[9]  {Month=October, Time_of_Day=Morning}                 => {Likes=1001-5000}      0.01044068 0.8191489  0.01274576  2.0961913   77
[10] {Hashtags=, Retweets=1001-5000, Month=October}       => {Likes=1001-5000}      0.03145763 0.8169014  0.03850847  2.0904399  232
```

Figure 6

Full Dataset Matrix Plot.

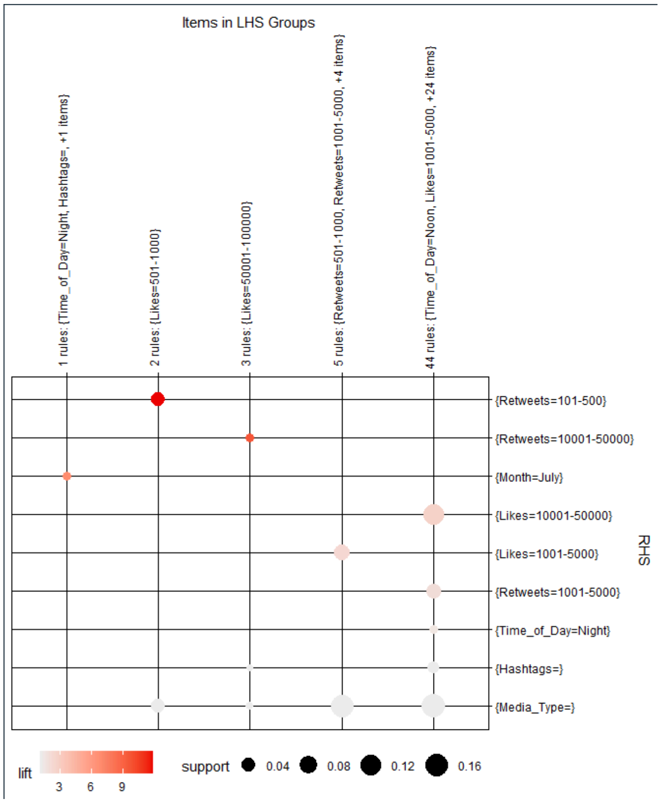
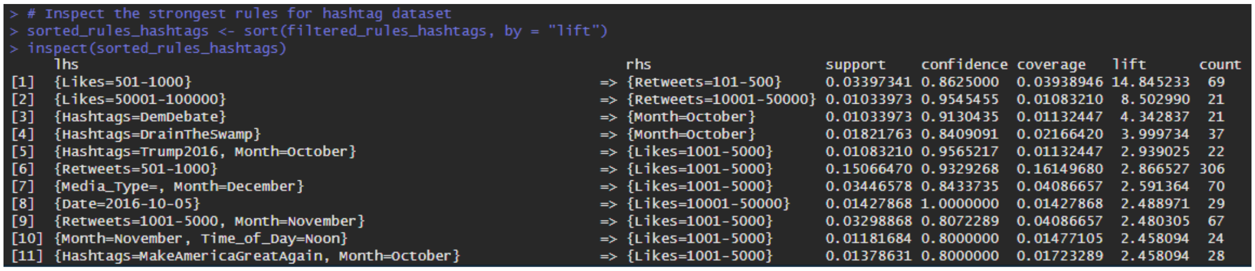


Figure 7

Hashtag Strongest Rules.



```
> # Inspect the strongest rules for hashtag dataset
> sorted_rules_hashtags <- sort(filtered_rules_hashtags, by = "lift")
> inspect(sorted_rules_hashtags)
     lhs                                                rhs                     support    confidence coverage   lift      count
[1]  {Likes=501-1000}                                => {Retweets=101-500}     0.03397341 0.8625000  0.03938946 14.845233  69
[2]  {Likes=50001-100000}                            => {Retweets=10001-50000} 0.01033973 0.9545455  0.01083210  8.502990  21
[3]  {Hashtags=DemDebate}                            => {Month=October}        0.01033973 0.9130435  0.01132447  4.342837  21
[4]  {Hashtags=DrainTheSwamp}                        => {Month=October}        0.01821763 0.8409091  0.02166420  3.999734  37
[5]  {Hashtags=Trump2016, Month=October}             => {Likes=1001-5000}      0.01083210 0.9565217  0.01132447  2.939025  22
[6]  {Retweets=501-1000}                             => {Likes=1001-5000}      0.15066470 0.9329268  0.16149680  2.866527 306
[7]  {Media_Type=, Month=December}                   => {Likes=1001-5000}      0.03446578 0.8433735  0.04086657  2.591364  70
[8]  {Date=2016-10-05}                               => {Likes=10001-50000}    0.01427868 1.0000000  0.01427868  2.488971  29
[9]  {Retweets=1001-5000, Month=November}            => {Likes=1001-5000}      0.03298868 0.8072289  0.04086657  2.480305  67
[10] {Month=November, Time_of_Day=Noon}              => {Likes=1001-5000}      0.01181684 0.8000000  0.01477105  2.458094  24
[11] {Hashtags=MakeAmericaGreatAgain, Month=October} => {Likes=1001-5000}      0.01378631 0.8000000  0.01723289  2.458094  28
```

Figure 8

Hashtag Confusion Matrix.

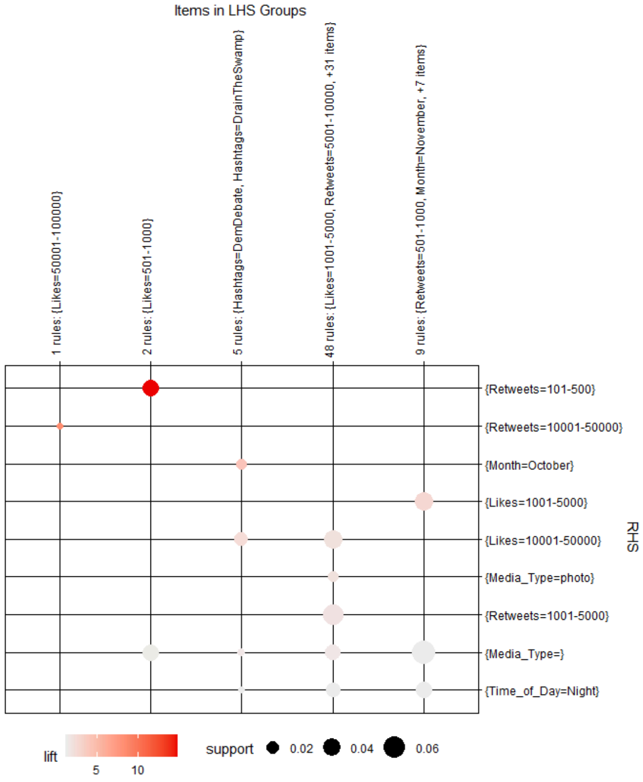
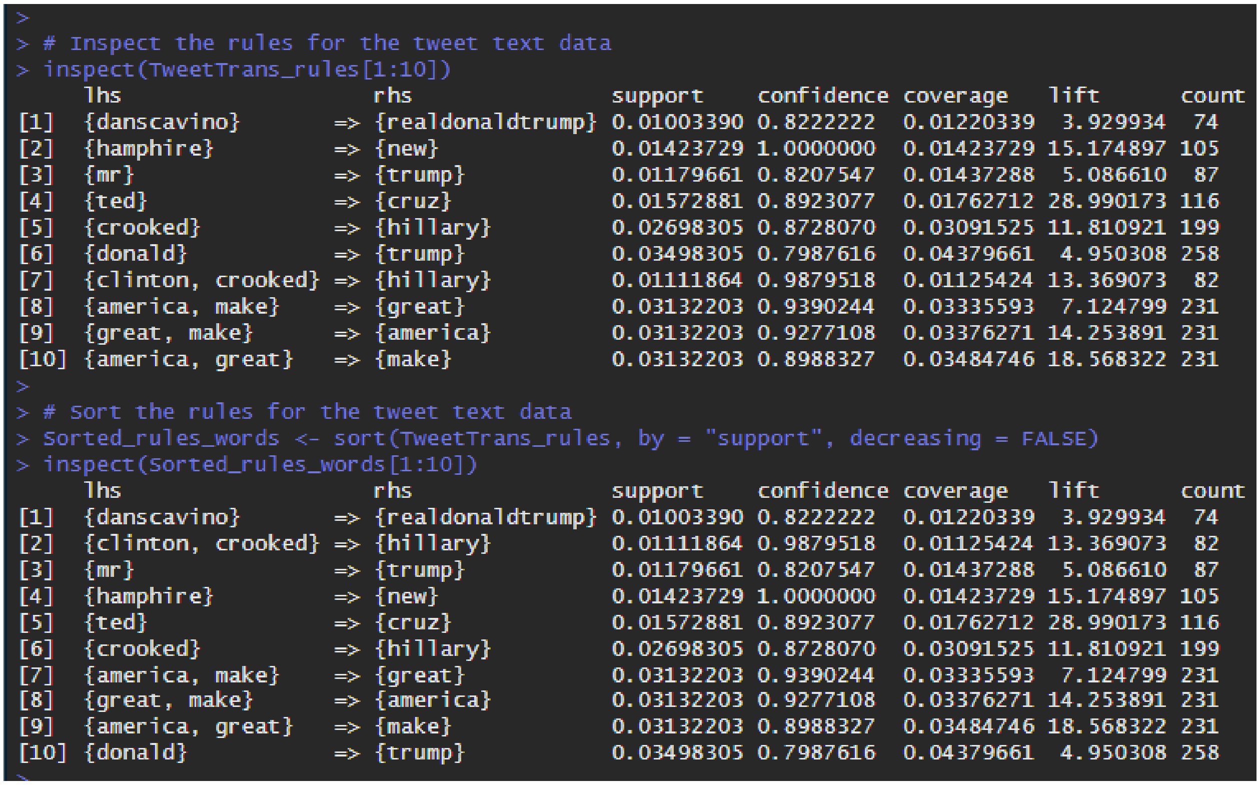


Figure 9

Tweet Text Strongest Rules by Support.



```
>
> # Inspect the rules for the tweet text data
> inspect(TweetTrans_rules[1:10])
     lhs                   rhs               support    confidence coverage   lift      count
[1]  {danscavino}       => {realdonaldtrump} 0.01003390 0.8222222  0.01220339  3.929934  74
[2]  {hamphire}         => {new}             0.01423729 1.0000000  0.01423729 15.174897 105
[3]  {mr}               => {trump}           0.01179661 0.8207547  0.01437288  5.086610  87
[4]  {ted}              => {cruz}            0.01572881 0.8923077  0.01762712 28.990173 116
[5]  {crooked}          => {hillary}         0.02698305 0.8728070  0.03091525 11.810921 199
[6]  {donald}           => {trump}           0.03498305 0.7987616  0.04379661  4.950308 258
[7]  {clinton, crooked} => {hillary}         0.01111864 0.9879518  0.01125424 13.369073  82
[8]  {america, make}    => {great}           0.03132203 0.9390244  0.03335593  7.124799 231
[9]  {great, make}      => {america}         0.03132203 0.9277108  0.03376271 14.253891 231
[10] {america, great}   => {make}            0.03132203 0.8988327  0.03484746 18.568322 231
>
> # Sort the rules for the tweet text data
> Sorted_rules_words <- sort(TweetTrans_rules, by = "support", decreasing = FALSE)
> inspect(Sorted_rules_words[1:10])
     lhs                   rhs               support    confidence coverage   lift      count
[1]  {danscavino}       => {realdonaldtrump} 0.01003390 0.8222222  0.01220339  3.929934  74
[2]  {clinton, crooked} => {hillary}         0.01111864 0.9879518  0.01125424 13.369073  82
[3]  {mr}               => {trump}           0.01179661 0.8207547  0.01437288  5.086610  87
[4]  {hamphire}         => {new}             0.01423729 1.0000000  0.01423729 15.174897 105
[5]  {ted}              => {cruz}            0.01572881 0.8923077  0.01762712 28.990173 116
[6]  {crooked}          => {hillary}         0.02698305 0.8728070  0.03091525 11.810921 199
[7]  {america, make}    => {great}           0.03132203 0.9390244  0.03335593  7.124799 231
[8]  {great, make}      => {america}         0.03132203 0.9277108  0.03376271 14.253891 231
[9]  {america, great}   => {make}            0.03132203 0.8988327  0.03484746 18.568322 231
[10] {donald}           => {trump}           0.03498305 0.7987616  0.04379661  4.950308 258
```

Figure 10

Tweet Text Confusion Matrix.

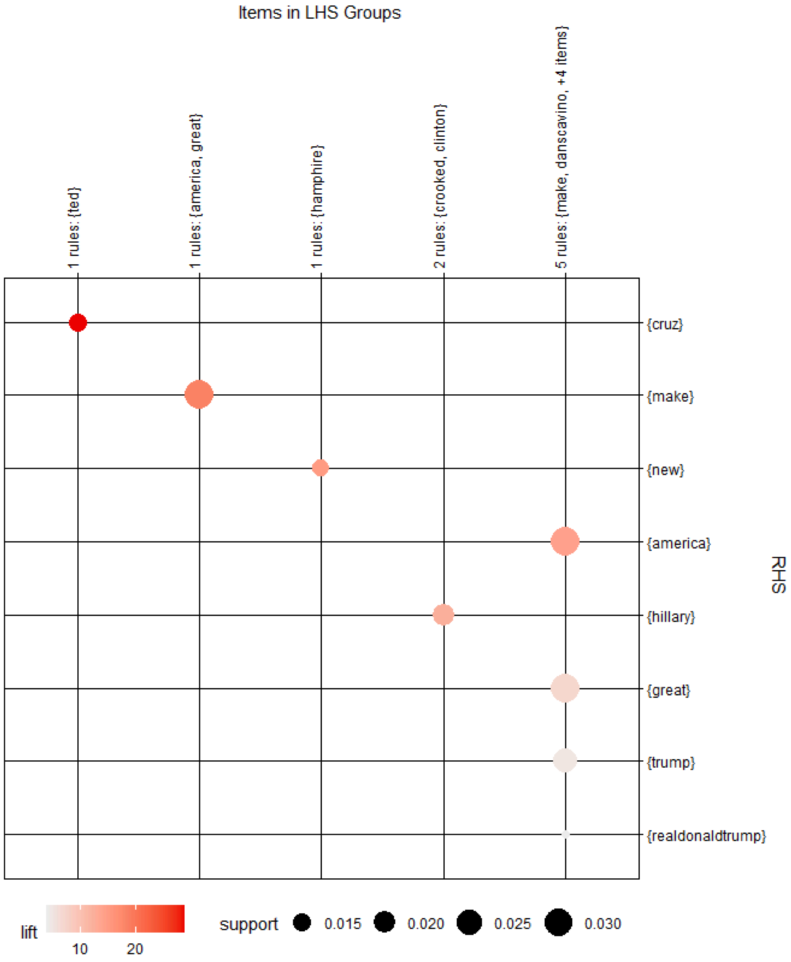


Figure 11

K-Means Clustering Plot.

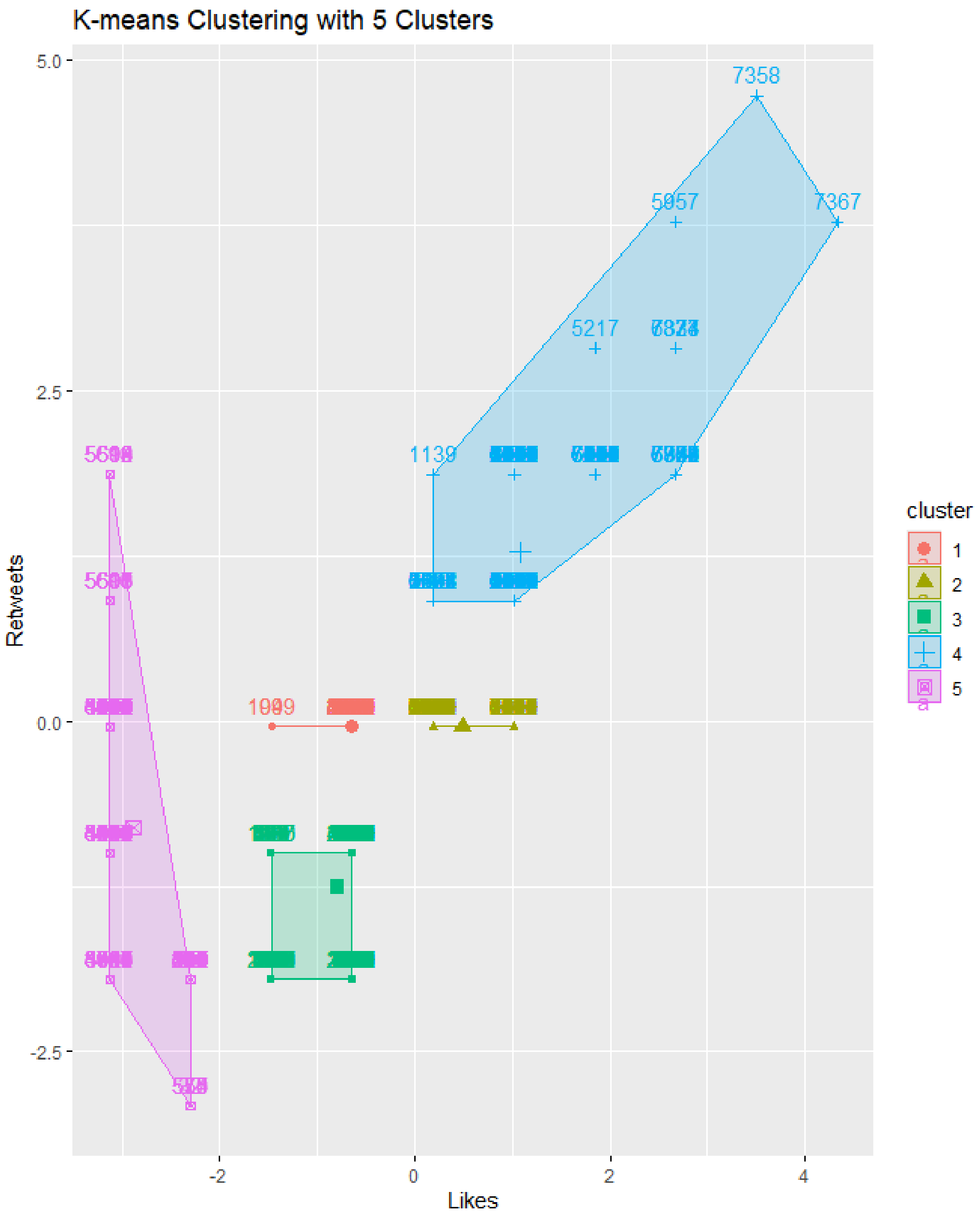


Figure 12

K-Means Clustering Silhouette.

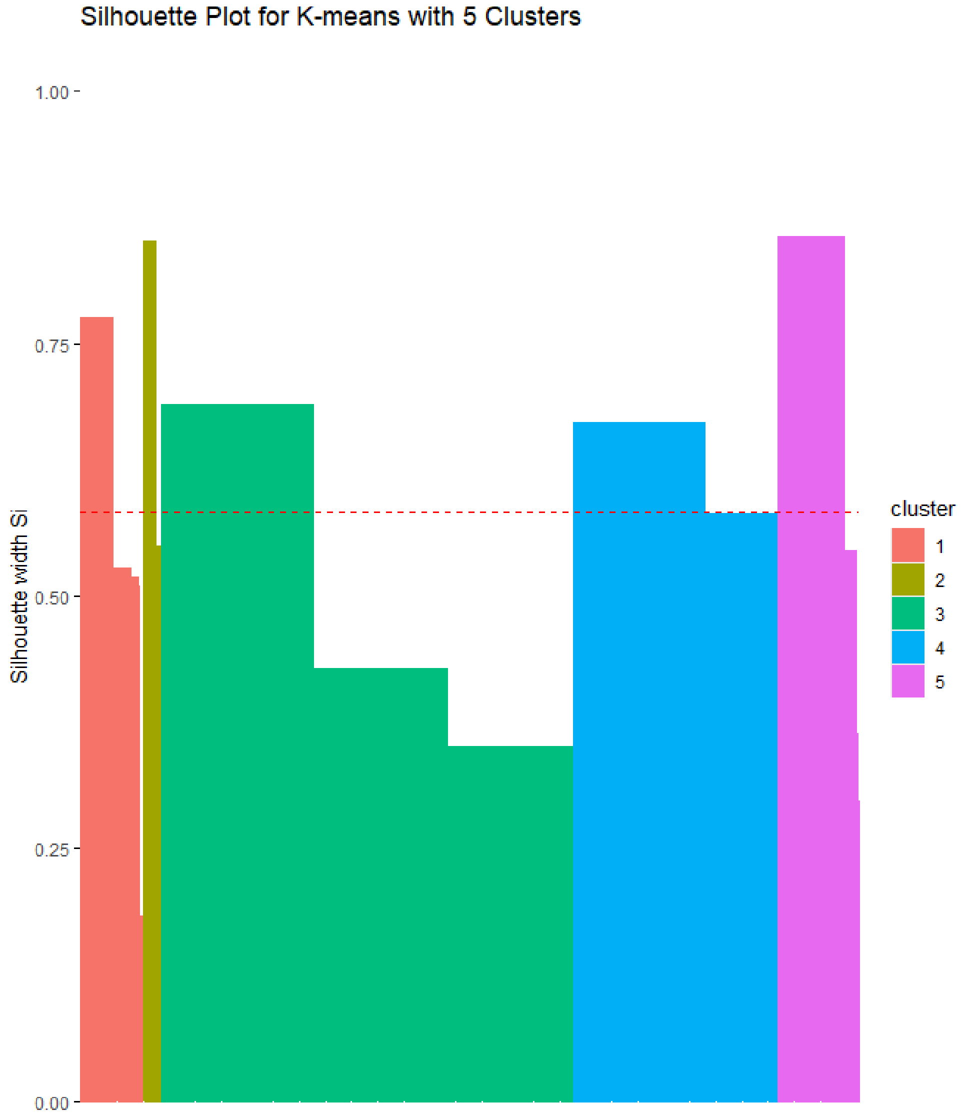
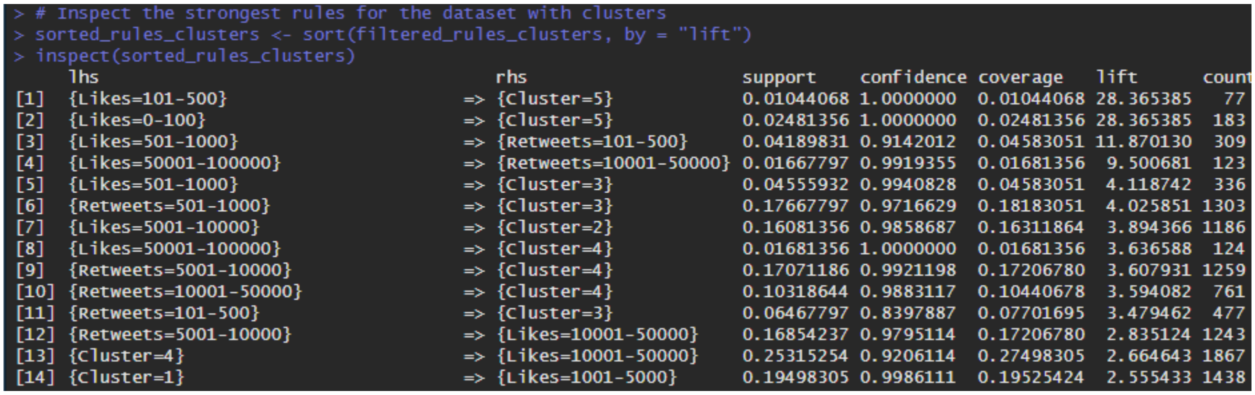


Figure 13

K-Means Clustering Strongest Rules.



```
> # Inspect the strongest rules for the dataset with clusters
> sorted_rules_clusters <- sort(filtered_rules_clusters, by = "lift")
> inspect(sorted_rules_clusters)
     lhs                                 rhs                        support    confidence coverage   lift      count
[1]  {Likes=101-500}                  => {Cluster=5}                0.01044068 1.0000000  0.01044068 28.365385   77
[2]  {Likes=0-100}                    => {Cluster=5}                0.02481356 1.0000000  0.02481356 28.365385  183
[3]  {Likes=501-1000}                 => {Retweets=101-500}         0.04189831 0.9142012  0.04583051 11.870130  309
[4]  {Likes=50001-100000}             => {Retweets=10001-50000}     0.01667797 0.9919355  0.01681356  9.500681  123
[5]  {Likes=501-1000}                 => {Cluster=3}                0.04555932 0.9940828  0.04583051  4.118742  336
[6]  {Retweets=501-1000}              => {Cluster=3}                0.17667797 0.9716629  0.18183051  4.025851 1303
[7]  {Likes=5001-10000}               => {Cluster=2}                0.16081356 0.9858687  0.16311864  3.894366 1186
[8]  {Likes=50001-100000}             => {Cluster=4}                0.01681356 1.0000000  0.01681356  3.636588  124
[9]  {Retweets=5001-10000}            => {Cluster=4}                0.17071186 0.9921198  0.17206780  3.607931 1259
[10] {Retweets=10001-50000}           => {Cluster=4}                0.10318644 0.9883117  0.10440678  3.594082  761
[11] {Retweets=101-500}               => {Cluster=3}                0.06467797 0.8397887  0.07701695  3.479462  477
[12] {Retweets=5001-10000}            => {Likes=10001-50000}        0.16854237 0.9795114  0.17206780  2.835124 1243
[13] {Cluster=4}                      => {Likes=10001-50000}        0.25315254 0.9206114  0.27498305  2.664643 1867
[14] {Cluster=1}                      => {Likes=1001-5000}          0.19498305 0.9986111  0.19525424  2.555433 1438
```

Figure 14

K-Means Clustering Confusion Matrix.

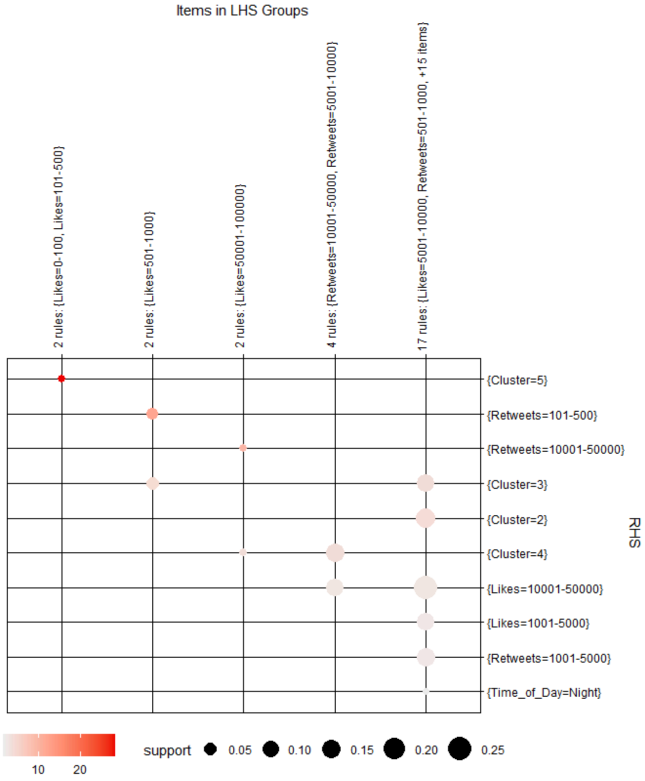


Figure 15

Hierarchical Dendrogram.

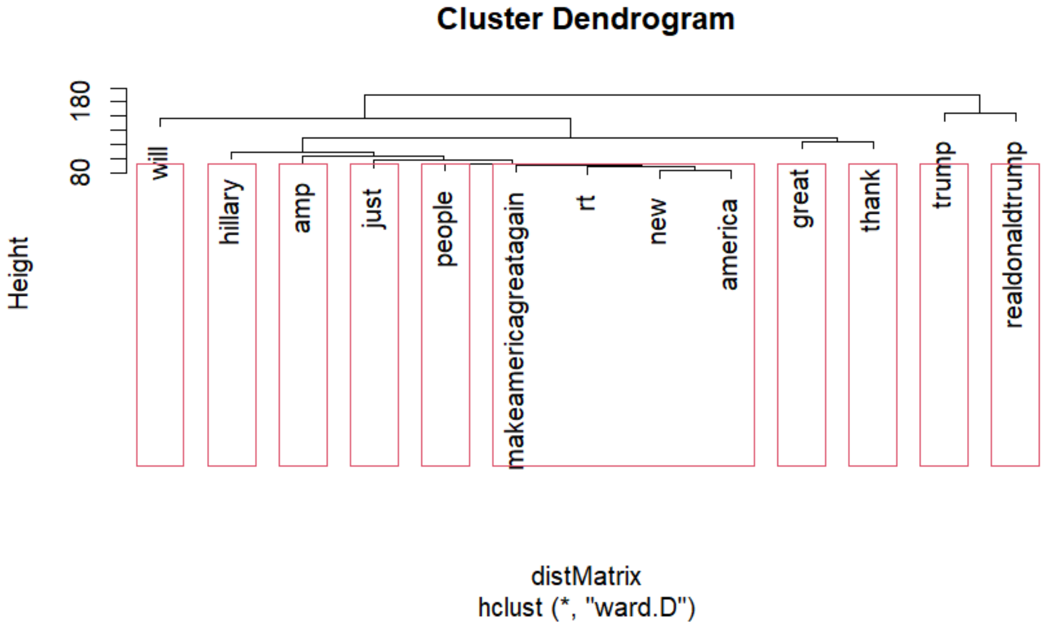
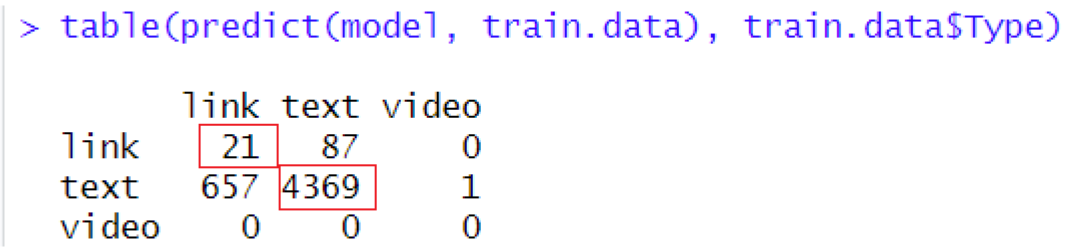


Figure 16

The confusion matrix to determine the classification accuracy of the Naive Bayes model.

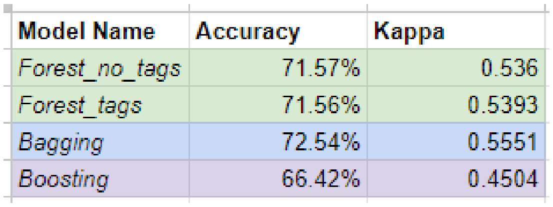


```
> table(predict(model, train.data), train.data$Type)

        link text video
  link    21   87     0
  text   657 4369     1
  video    0    0     0
```

Figure 17

Model Accuracy for the decision trees.



Model Name	Accuracy	Kappa
Forest_no_tags	71.57%	0.536
Forest_tags	71.56%	0.5393
Bagging	72.54%	0.5551
Boosting	66.42%	0.4504

Figure 18

Error rate versus number of trees for both random forest models.

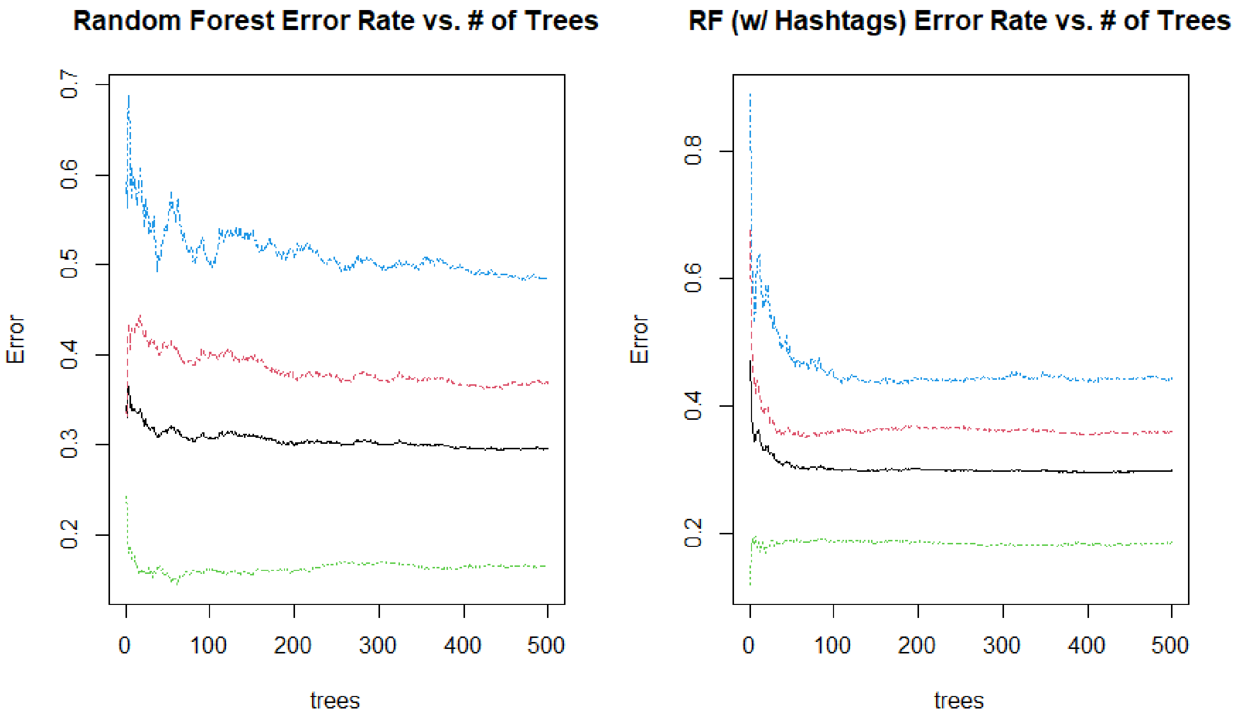
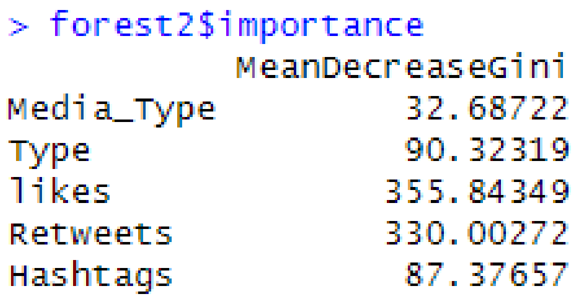


Figure 19

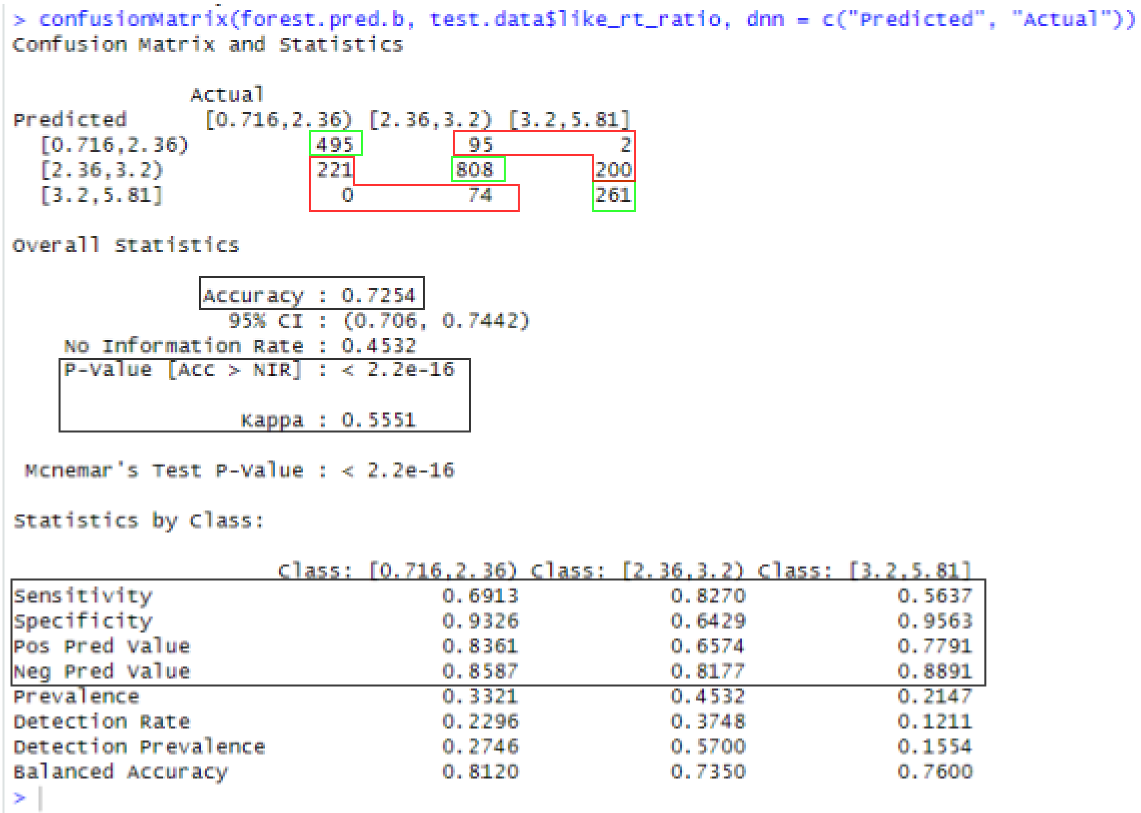
Variable importance chart.



```
> forest2$importance
           MeanDecreaseGini
Media_Type         32.68722
Type               90.32319
likes             355.84349
Retweets          330.00272
Hashtags           87.37657
```

Figure 20

Confusion matrix for bagging model.



```
> confusionMatrix(forest.pred.b, test.data$like_rt_ratio, dnn = c("Predicted", "Actual"))
Confusion Matrix and Statistics

              Actual
Predicted      [0.716,2.36) [2.36,3.2) [3.2,5.81]
  [0.716,2.36)          495         95          2
  [2.36,3.2)            221        808        200
  [3.2,5.81]              0         74        261

Overall Statistics

               Accuracy : 0.7254
                 95% CI : (0.706, 0.7442)
    No Information Rate : 0.4532
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5551

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: [0.716,2.36) Class: [2.36,3.2) Class: [3.2,5.81]
Sensitivity                       0.6913            0.8270            0.5637
Specificity                       0.9326            0.6429            0.9563
Pos Pred Value                    0.8361            0.6574            0.7791
Neg Pred Value                    0.8587            0.8177            0.8891
Prevalence                        0.3321            0.4532            0.2147
Detection Rate                    0.2296            0.3748            0.1211
Detection Prevalence              0.2746            0.5700            0.1554
Balanced Accuracy                 0.8120            0.7350            0.7600
>
```

Figure 21

Predicted versus actual levels of engagement for the bagging model.

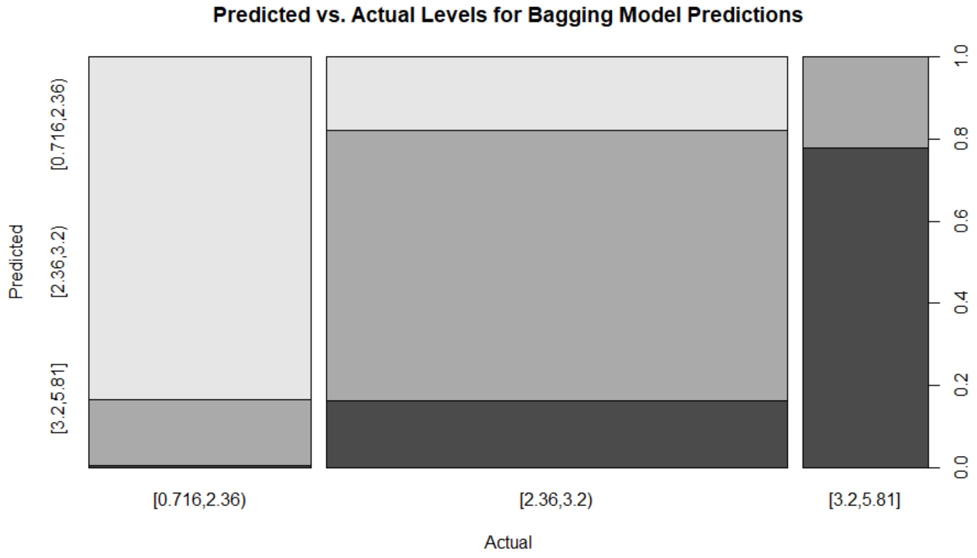


Figure 22

Histogram of like-to-retweet ratio values.

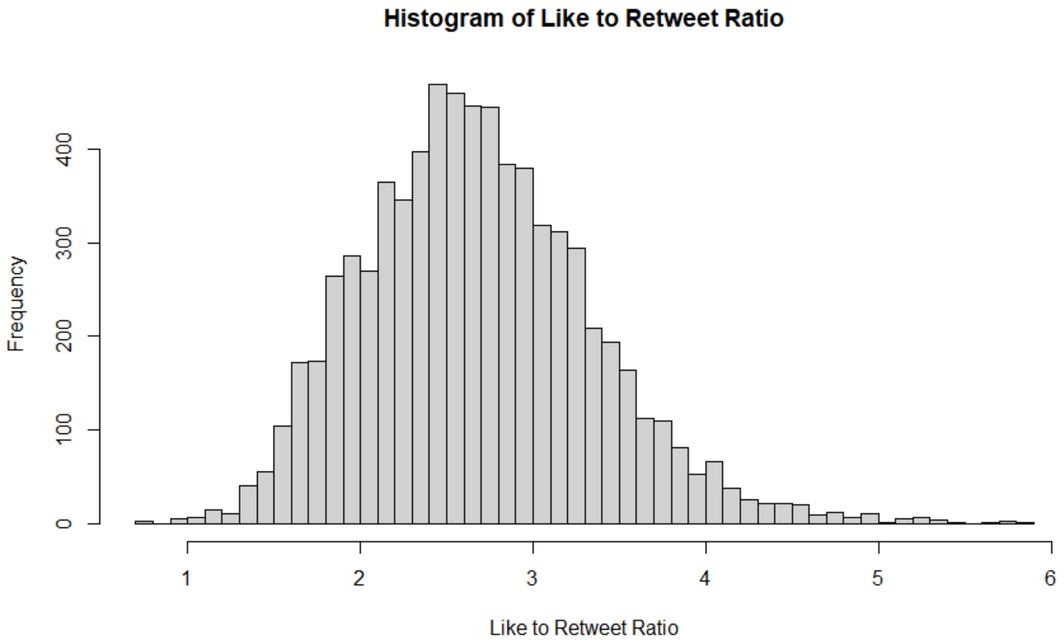
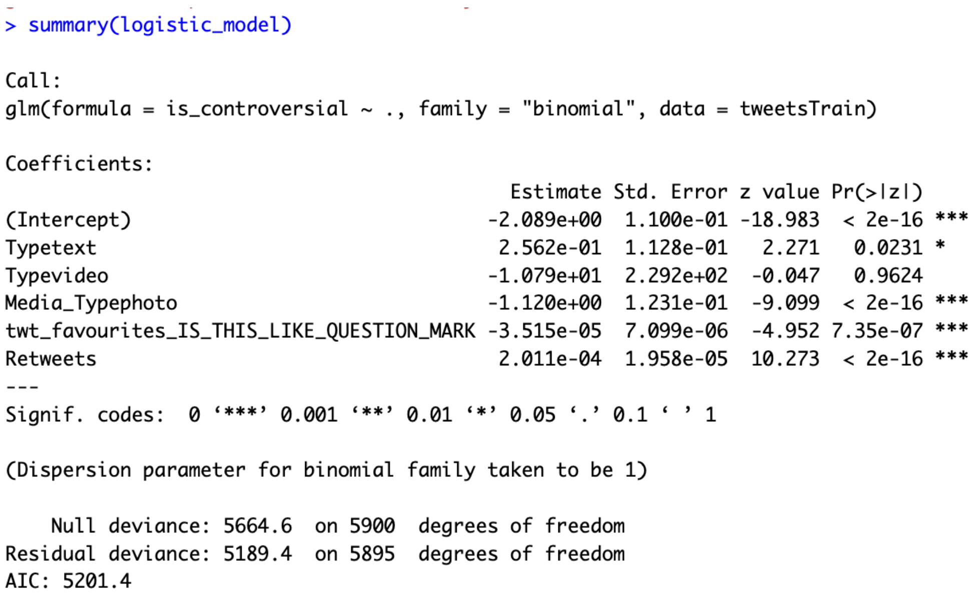


Figure 23

Summary of the Logistic Regression model.



```
> summary(logistic_model)

Call:
glm(formula = is_controversial ~ ., family = "binomial", data = tweetsTrain)

Coefficients:
                                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                               -2.089e+00  1.100e-01 -18.983  < 2e-16 ***
Typetext                                   2.562e-01  1.128e-01   2.271   0.0231 *  
Typevideo                                 -1.079e+01  2.292e+02  -0.047   0.9624    
Media_Typephoto                           -1.120e+00  1.231e-01  -9.099  < 2e-16 ***
twt_favourites_IS_THIS_LIKE_QUESTION_MARK -3.515e-05  7.099e-06  -4.952 7.35e-07 ***
Retweets                                   2.011e-04  1.958e-05  10.273  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5664.6  on 5900  degrees of freedom
Residual deviance: 5189.4  on 5895  degrees of freedom
AIC: 5201.4
```

Figure 24

Predicted Probability of Being Controversial vs. Retweets.

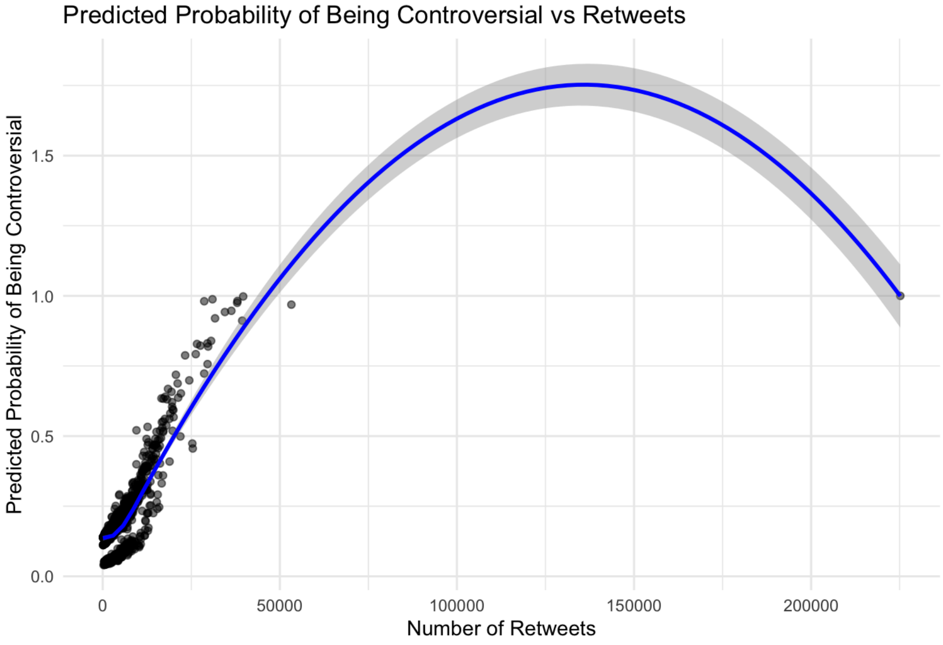
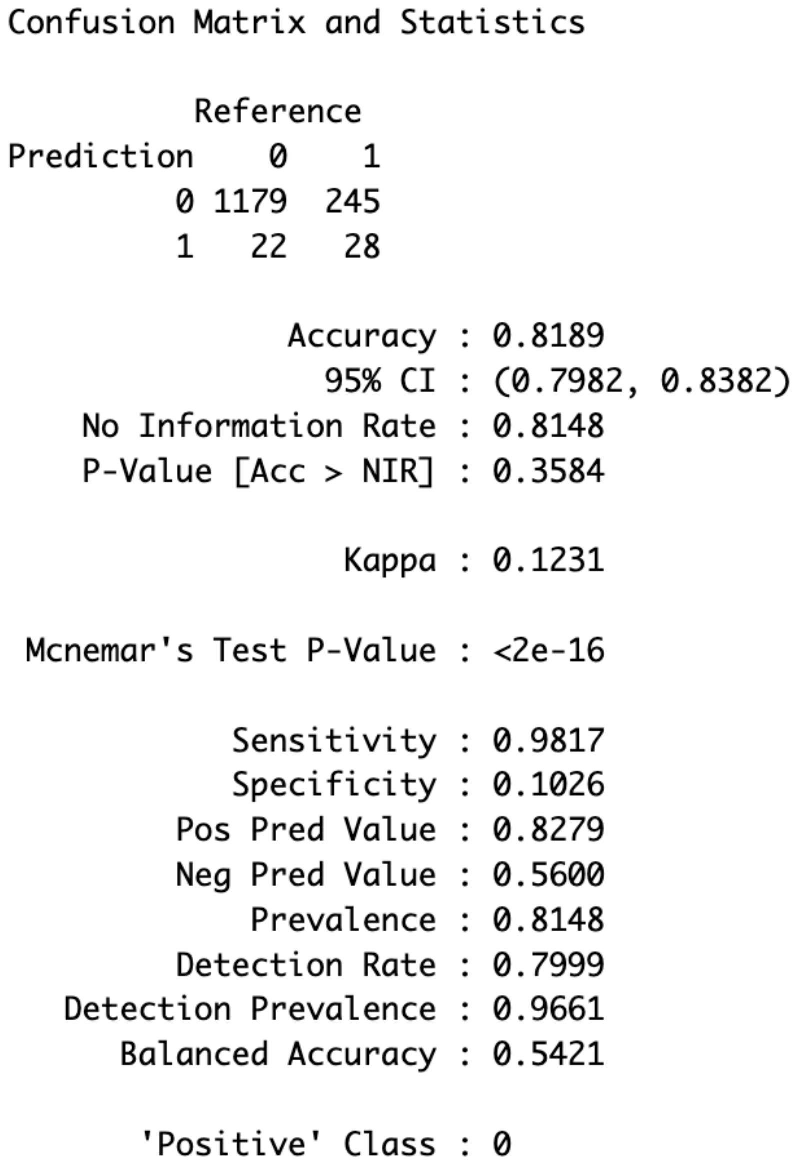


Figure 25

Confusion Matrix and Statistics for the Model.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1179  245
         1   22   28

               Accuracy : 0.8189
                 95% CI : (0.7982, 0.8382)
    No Information Rate : 0.8148
    P-Value [Acc > NIR] : 0.3584

                  Kappa : 0.1231

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9817
            Specificity : 0.1026
         Pos Pred Value : 0.8279
         Neg Pred Value : 0.5600
             Prevalence : 0.8148
         Detection Rate : 0.7999
   Detection Prevalence : 0.9661
      Balanced Accuracy : 0.5421

       'Positive' Class : 0
```

Figure 26

Neural network plot of the tweets.

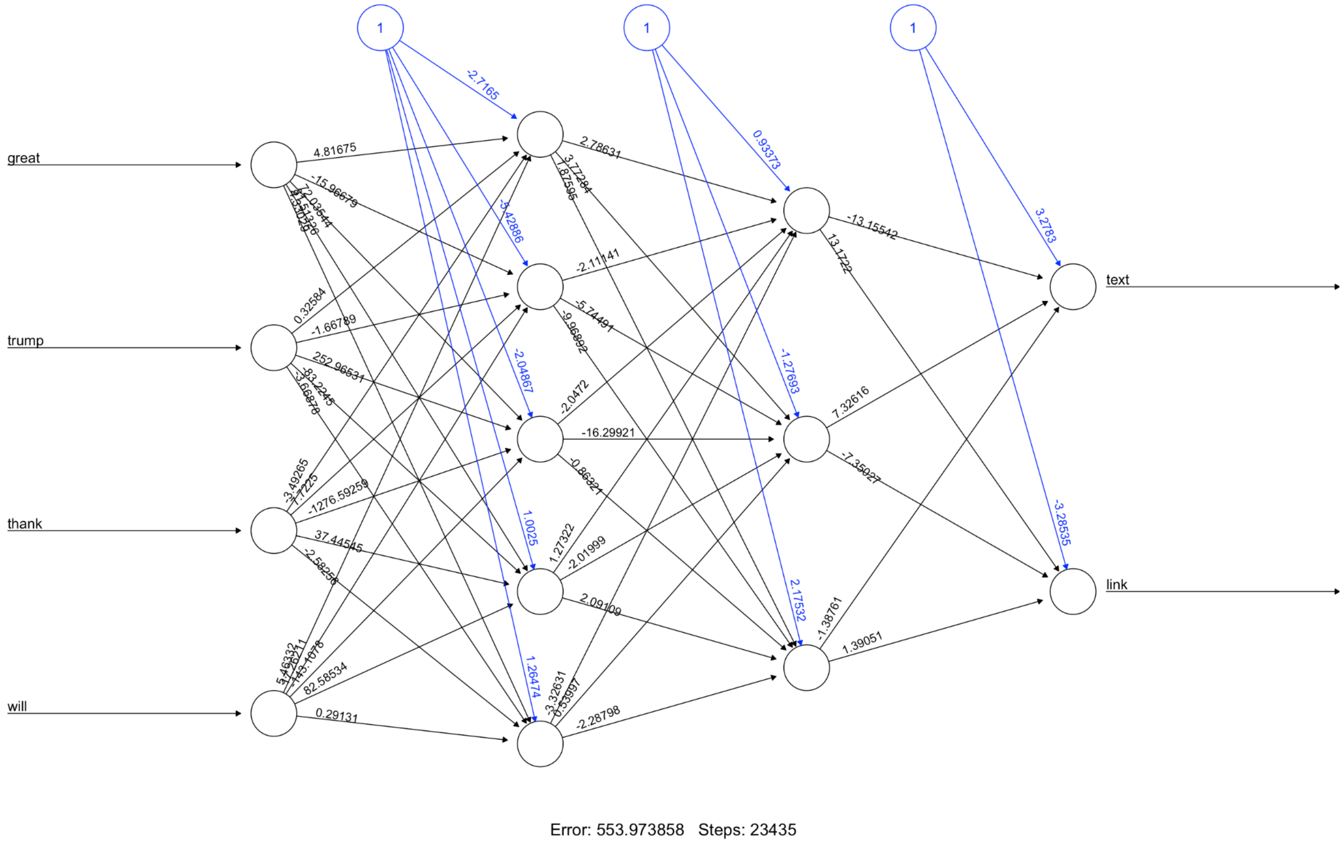
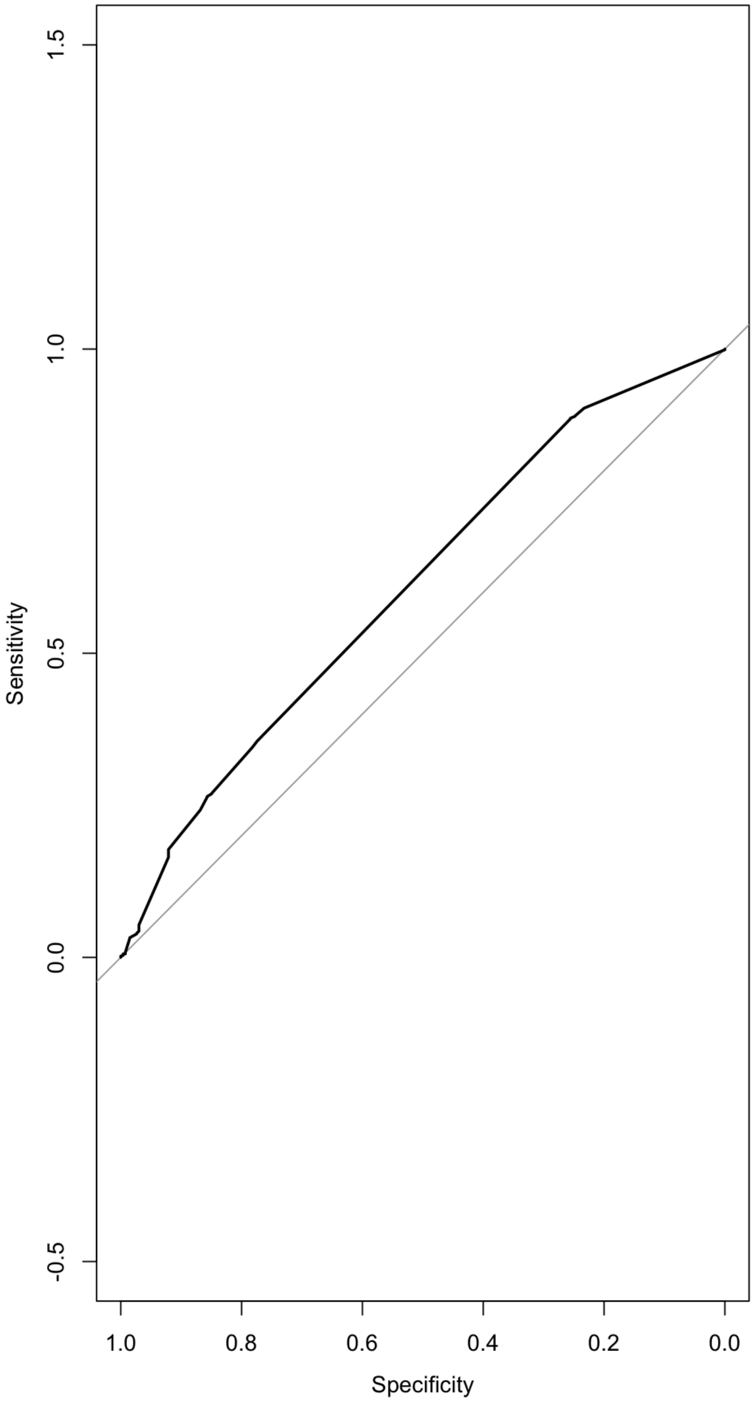


Figure 27

ROC Curve of the neural network plot of the tweets.


Sensitivity
Specificity