

Title: Sentimental Analysis on President TRUMP.

First Name	Last Name	Online Students? (Y or N)	Monday or Tuesday	Shared with ITMD 525? (Y or N)
Abhishek	Dutta	N	Tuesday	N
Anup	Kulkarni	N	Tuesday	N
Shreyas	Patil	N	Tuesday	N

Group Number: 18

Table of Contents

1. Introduction and Motivations.....	3
2. Data Description.....	4
3. Research Problems and Solutions	5
4. KDD	6
4.1. Data Processing.....	6
4.2. Data Mining Tasks and Processes	7
5. Evaluations and Results	9
5.1. Evaluation Methods	9
5.2. Results and Findings.....	11
6. Conclusions and Future Work.....	11
6.1. Conclusions	14
6.2. Limitations.....	14
6.3. Potential Improvements or Future Work	14

1. Introduction and Motivations

Twitter is a beautiful and most used platform for spontaneous and raw emotional expression. Unlike the other social media platforms Twitter is a user friendly and simple platform which attracts many users (including celebrities) to post their opinions which may have a global impact as it has a wide audience and user-base. Therefore, pulling data from such a tremendous source is a good base for any analysis.

Therefore, we have taken up the task to perform a Sentimental Analysis on such a personality who happens to be the President of one of the most influential country, the United States of America. The main reason behind taking up this topic is that Mr. Trump being the president of America holds a very crucial position and after 2016 election, people around the world have started following news related to him very closely and started sharing their views on him frequently. People from various parts of the world have opinions and views on each decision that he makes or bill that his team proposes. Therefore, it gives us a fantastic opportunity to analyze what people think about him and about the decision that he makes. Based on the tweets, we are planning to extract the sentiments of the people related to president Trump.

Sentiment analysis is the detection of attitudes or emotions conveyed in a body of text. It is a subset of a larger field of study called Natural Language Processing (NLP). Examples: Movie reviews: Is this review positive or negative? How many stars?

Twitter sentimental analysis is the process to determine emotional tone behind a series of words used to gain understanding of attitudes, emotions and opinions.

Why Sentimental Analysis? Shifts in sentiments of the social media has shown correlation with the shift in trend of the stock market. Example: Obama administration team applied and used sentimental analysis to grasp the public opinion and accordingly made some policy announcements and campaign topics just ahead of the 2012 elections.

2. Data Description

We are planning to extract data from Twitter using **#President Trump**. For this we would have a data set of 5000 entries out of which 80 % would be falling under training and 20% test. We have download the tweets from twitter by setting up and API on <http://apps.twitter.com> based on few keys like Customer key, Customer Token, Access token and Secret Token key.

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key) dXcTnA9X21qje7ZU6hgwlycgL

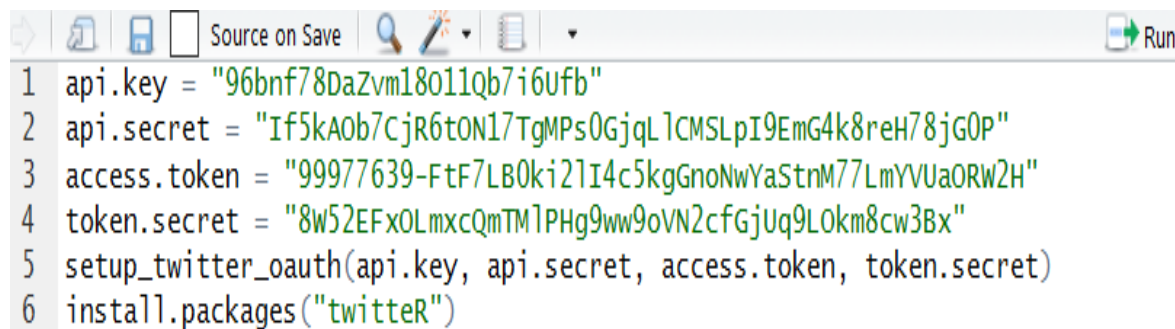
Consumer Secret (API Secret) UA4cJyRH36bLNpIm9WiVrzAacmyXxjwnjv4XxVpijVnF2JR3HI

Access Level Read and write ([modify app permissions](#))

Owner abhishek3107191

Owner ID 741367186518740992

R commands to fetch data based on the unique keys generated.



```
1 api.key = "96bnf78DaZvm18011Qb7i6Ufb"
2 api.secret = "If5kAOb7CjR6tON17TgMPs0GjqL1cMSLpI9EmG4k8reH78jG0P"
3 access.token = "99977639-FtF7LB0ki21I4c5kgGnoNwYaStnM77LmYVUaORW2H"
4 token.secret = "8W52EFxOLmxcQmTM1PHg9ww9oVN2cfGjUq9LOkm8cw3Bx"
5 setup_twitter_oauth(api.key, api.secret, access.token, token.secret)
6 install.packages("twitterR")
```

3. Research Problems and Solutions

With the data that we have extracted from twitter we will perform a sentimental analysis. After cleaning the data (tweets) i.e. by removing all the unwanted words and keeping list of positive and negative words present to calculate the sentiment score and based on that we will proceed with our analysis to determine that what are the sentiments of the people about Trump i.e. are there more positive or more negative things being said about him.

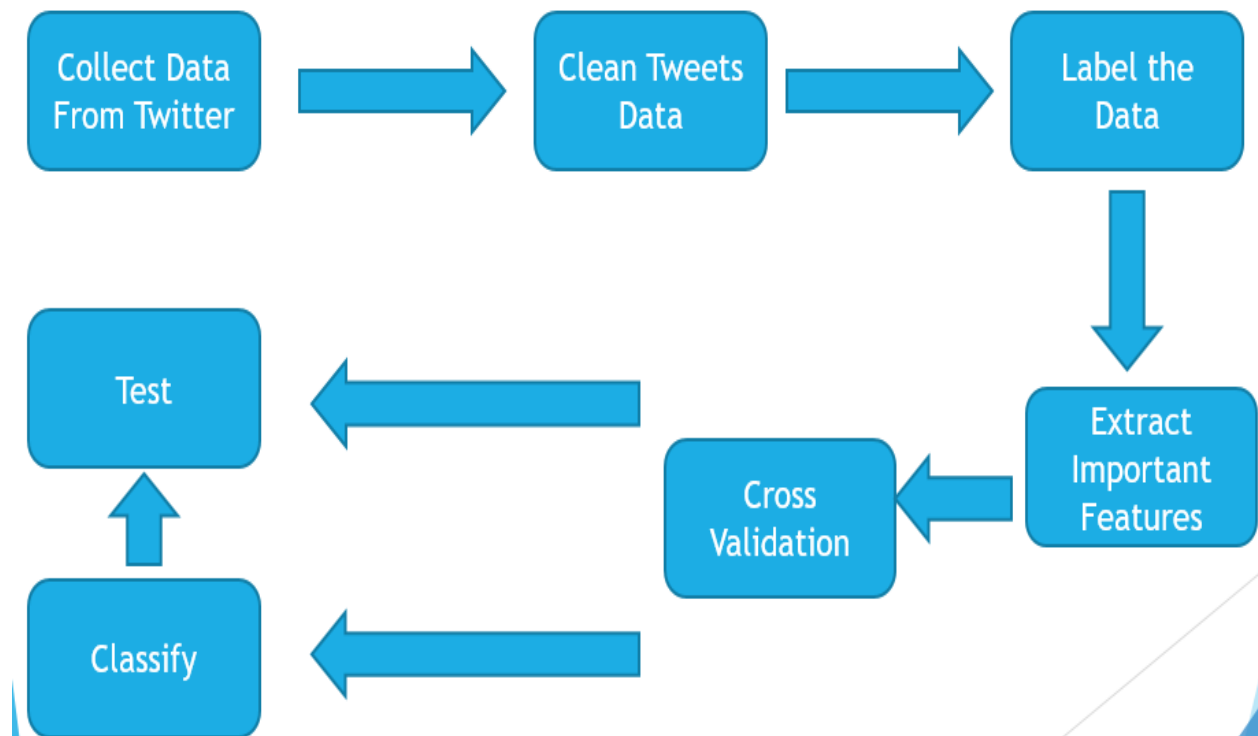
We will calculate the score and based on the score being positive, negative we will define labels to it as positive /negative/neutral. Neutral label is assigned when the sentiment score is either 0 or both positive and negative are equal.

Below mentioned are the steps they we have followed to achieve the solution that we are looking for.

Based on our analysis and sentiment score calculation we will find and analyze the following things:

- i. Top 5 Positive tweets abut TRUMP.
- ii. TOP 5 Negative tweets about TRUMP.
- iii. Mostly used words amongst the 5000 tweets dataset.
- iv. Has more number of people tweeted positive tweets or negative tweets for TRUMP over a span of time.

- **Process Flow step by step:**



4. KDD

4.1. Data Processing

Data Cleaning

The 5000 tweets that we extracted from twitter had many words, characters and images that were not required for our sentimental analysis. For example, a link will not help us find a positive or negative sentiment or calculate a sentiment. Also, there are many words in a tweet that don't have a positive or negative meaning and hence are of no use to us and therefore we remove them using Corpus function available in R BY STEMMING and stopping them.

```
# Remove Numbers
tc_tm <- tm_map(tc_tm, removeNumbers)

# Remove Stop words
trumpCorpus = tm_map(tc_tm,removePunctuation)

# Remove Punctuations
trumpCorpus = tm_map(trumpCorpus,removeWords,stopwords("english"))

trumpCorpus = tm_map(trumpCorpus,stripWhitespace)

trumpCorpus <- tm_map(trumpCorpus, removeWords, c("presidenttrump", "trump","donald", "rally", "rting", "first","media","cnn","f

install.packages("SnowballC")

library(SnowballC)
trumpCorpus = tm_map(trumpCorpus,stemDocument, language = "english")
inspect(trumpCorpus[[1]])

trumpTDM = TermDocumentMatrix(trumpCorpus)

trumpTDM

trump.mat = as.matrix(trumpTDM)
length(trump.mat)
trump.mat[226:238,1:30]

word.freq = sort(rowSums(trump.mat),decreasing=T)

word.freq[1:20]

#To figure out the most common words.
```

To remove links (http://, https://), emoticons, punctuations and non-other important words and expressions or numbers we have used GSUB function available in R.

```
#remove the characters
df$text <- sapply(df$text,function(row) iconv(row, "latin1", "ASCII", sub="")) #remove emoticons
df$text = gsub("(f|ht)tp(s?):/(.*)[.][a-z]+", "", df$text) #remove URL
df$text = gsub("(RT|via)((?:\\b\\w*@[\\w+)+)", "", df$text)
df$text = gsub("@\\w+", "", df$text)
df$text = gsub("[[:punct:]]", "", df$text)
df$text = gsub("[[:digit:]]", "", df$text)
df$text = gsub("http\\w+", "", df$text)
df$text = gsub("https\\w+", "", df$text)
df$text = gsub("[ \\t]{2,}", "", df$text)
df$text = gsub("[ \\n]{2,}", "", df$text)
df$text = gsub("^\\s+|\\s+$", "", df$text)
df$text = gsub("amp", "", df$text)
df$text = gsub("[ \\n]{2,}", "", df$text)

sample <- df$text
```

4.2. Data Mining Tasks and Processes

- ▶ Search through the words in the body of text and find POSITIVE (+1) and NEGATIVE (-1) words. Many/most of the words in the text will not receive a score or be scored as zero. The sentiment score could simply be the sum of the scored words.
- ▶ For Positive /Negative words calculation we have used the below mentioned link <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon>
- ▶ But we also understand that in social media there are many people who use urban-lingo i.e. words like gr8, Awsum, Kewl and even they fall under a specific positive or negative category and can influence our sentimental analysis therefore we have also considered and added a few such mostly used words in our R-code. But again, since words are endless therefore it totally depends what all to include and what basis.

Once we calculate the score therefore if the score of positive words is high then negative words then it would fall under POSITIVE LABEL. If the negative score is greater than Positive score then it would fall under NEGATIVE LABEL. If scores are 0 or both POSITIVE and NEGATIVE scores are equal then it would fall under the NEUTRAL LABEL.

Using function in R and the calculating the score we see that the Text gets the score i.e. positive /Negative corresponding to the Text

Text	Score	Positive	Negative
1 Dont ever forget the lifestyle PresidentTrump gav	-1	0	1
2 RT ANOMALY PresidentTrump POTUSWe will ma	-1	0	1
3 FirstDaysThis short video shows some of the exe	0	0	0
4 PresidentTrump We will make AmericaSafeAgain	0	0	0
5 PresidentTrump is the same man today as he wa	1	1	0
6 Another greatclipshowing INCREDIBLE line for Pr	1	1	0
7 PresidentTrump We will make AmericaSafeAgain	0	0	0
8 Another greatclipshowing INCREDIBLE line for Pr	1	1	0
9 PresidentTrump We will make AmericaSafeAgain	0	0	0
10 FirstDaysThis short video shows some of the exe	0	0	0
11 PresidentTrump is the same man today as he wa	1	1	0
12 KoolAidMan AIEatsSteak PresidentTrump WhiteH	0	0	0
13 The yr assault on your nd Amendment freedoms	0	2	2
14 Thank You PresidentTrumpan amazingdays MAG	2	2	0
15 PresidentTrump We will make AmericaSafeAgain	0	0	0
16 McCain Said He Died In A Failed MissionPresiden	0	2	2
17 The yr assault on your nd Amendment freedoms	0	2	2
18 PresidentTrump may have denigrated HRCs actic	-1	0	1

Data with the Labels Positive/Negative/Neutral based on the Score. Therefore, Labels are assigned to the tweets next to it. Here If the score is positive then positive label, if score is negative then negative label and if the score of positive and negative are equal or if the score =0 for both positive and negative then the score comes out to be Neutral.

As this is a classical classification problem we need to just determine whether we have the sentence as positive or negative in totality, we then need to use the Naïve Bayes Algorithm, Decision tree and Random Forest. Here the process will be: Classifier: Input: The Tweets which the user posts on twitter. Also, called as the “bag of words” are used as the features as we use the unique words as a feature. So, we would capture the frequency of the words. Output: Prediction of the class/labels if the tweet is Positive/Negative/Neutral.

stat.Text	stat.label
1 Dont ever forget the lifestyle PresidentTrump gave	negative
2 RT ANOMALY PresidentTrump POTUSWe will make	negative
3 FirstDaysThis short video shows some of the execu	neutral
4 PresidentTrump We will make AmericaSafeAgainw	neutral
5 PresidentTrump is the same man today as he was	positive
6 Another greatclipshowing INCREDIBLE line for Pres	positive
7 PresidentTrump We will make AmericaSafeAgainw	neutral
8 Another greatclipshowing INCREDIBLE line for Pres	positive
9 PresidentTrump We will make AmericaSafeAgainw	neutral
10 FirstDaysThis short video shows some of the execu	neutral
11 PresidentTrump is the same man today as he was	positive
12 KoolAidMan AIEatsSteak PresidentTrump WhiteHo	neutral
13 The yr assault on your nd Amendment freedoms h	neutral
14 Thank You PresidentTrumpan amazingdays MAGA	positive
15 PresidentTrump We will make AmericaSafeAgainw	neutral
16 McCain Said He Died In A Failed MissionPresidentT	neutral
17 The yr assault on your nd Amendment freedoms h	neutral
18 PresidentTrump may have denigrated HRCs action:	negative
19 Congratulations PresidentTrump on your	positive
20 The yr assault on your nd Amendment freedoms h	neutral
21 Dont ever forget the lifestyle PresidentTrump gave	negative

5. Evaluations and Results

5.1. Evaluation Methods

We have used Cross Validation as the Evaluation method for our model as our dataset is comparatively small and therefore we have proceeded with it and used classification techniques such as Naïve Bayes Classifier, J48 Decision Tree and Random Forest Classifier. Below are the attached screen-prints for the same.

- **On using Naïve Bayes Classification, we find out that the Currently classified instances or the accuracy comes out to be 88.94% as seen in the figure.**

The screenshot displays the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' section shows a list of test instances with their predicted and actual class labels. The 'Result list' on the left shows the execution time for various classifiers, with NaiveBayes being the fastest. The 'Summary' section provides a detailed breakdown of the classifier's performance, including accuracy, error rates, and a confusion matrix.

Classifier output

Instance	Actual Class	Predicted Class
America loves our PresidentTrump/VicePresidentPence thank you for working so very very hard to MAGA God bless you!	negative	negative
PresidentTrump at the NRA Convention gldckubuy	negative	negative
Boosted by m During Caign Trump Will Thank NRA By Speaking at Convention iHogWynI presidenttrump	negative	negative
This is completely disgusting Im willing to bet Soros has his paws in this Time to grow upgive PresidentTrump a cha	negative	negative
Americans TrustTrump PresidentTrump WhiteHouse Tells TruthVeryFakeNews FakeNews Political NationalMedia FBjLANWQ	negative	negative
MAGA SuperHeroOfTheDay is our PresidentTrump wvawQmGL	negative	negative
PresidentTrump said he thought the job would be easier Funnya woman could do the job no problem ERC could do it and wouldn't whine	negative	negative
Obama threatened to out federal funds from Indiana and North Carolina PresidentTrump does it and its the end of the world Hypocrisy	negative	negative
Well that makes since since PresidentTrump has an America first agenda while the political media has any lie any t uJcEKxXf	negative	negative
PresidentTrump absolutely the finest President ever JustBreakUpThebCircuitCorruptCourt OnlyYouCanDoIT J	negative	negative
The eight year assault on your second amendment freedoms has come to a crashing end PresidentTrump Trump TrumpDays	negative	negative
[total]	2015.0	3303.0

Summary

Metric	Value
Correctly Classified Instances	4447
Incorrectly Classified Instances	553
Kappa statistic	0.8234
Mean absolute error	0.1121
Root mean squared error	0.218
Relative absolute error	26.6059 %
Root relative squared error	47.492 %
Total Number of Instances	5000

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.811	0.000	0.999	0.811	0.895	0.890	0.982	0.946	negative
0.896	0.076	0.905	0.896	0.900	0.820	0.981	0.979	neutral
0.923	0.107	0.831	0.923	0.874	0.800	0.983	0.974	positive
Weighted Avg.	0.889	0.073	0.896	0.889	0.890	0.824	0.982	0.971

Confusion Matrix

a	b	c	<-- classified as
770	72	108	a = negative
0	2005	233	b = neutral
1	139	1672	c = positive

- **J-48 Classifier Accuracy=89.7%**

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **J48-U-M 2**

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds **10**
☐ Percentage split % **66**

More options...

(Nom) stat:label

Start Stop

Result list (right-click for options)

00:47:56 - bayes.NaiveBayes
00:48:09 - trees.J48
00:48:16 - trees.RandomForest
00:48:59 - bayes.NaiveBayes
00:49:57 - trees.J48

Classifier output

```

stat.Text = Boosted by m During Caign Trump Will Thank NRA By Speaking at Convention iWogMFunI presidentrump: positive (4.0)
stat.Text = This is completely disgusting Im willing to bet Soros has his paws in this Time to grow upgive PresidentTrump a cha: neutral (1.0)
stat.Text = Americans TrustTrump PresidentTrump WhiteHouse Tells TruthVeryFakeNews FakeNews Political NationalMedia FBjLAMIQ: neutral (1.0)
stat.Text = MAGA SuperHeroOfTheDay is our PresidentTrump xvAwQomGL: neutral (1.0)
stat.Text = PresidentTrump said he thought the job would be easier Funnya woman could do the job no problem HRC could do it and wouldnt whine: negative (1.0)
stat.Text = Obama threatened to cut federal funds from Indiana and North Carolina PresidentTrump does it and its the end of the world Hypocrisy: negative (1.0)
stat.Text = Well that makes since since PresidentTrump has an America first agenda while the political media has any lie any t uGtOfXxxY: neutral (1.0)
stat.Text = PresidentTrump absolutely the finest President ever JustBreakUpTheCircuitCorruptCourt OnlyYouCanBolt J: positive (1.0)
stat.Text = The eight year assault on your second amendment freedoms has come to a crashing end PresidentTrump Trump TrumpDays: neutral (1.0)

Number of Leaves : 1070
Size of the tree : 1076
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 4465 89.7 %
Incorrectly Classified Instances 515 10.3 %
Kappa statistic 0.8333
Mean absolute error 0.0779
Root mean squared error 0.199
Relative absolute error 18.4795 %
Root relative squared error 43.3455 %
Total Number of Instances 5000

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.811	0.001	0.995	0.811	0.893	0.878	0.981	0.917	negative
	0.997	0.182	0.816	0.997	0.897	0.814	0.982	0.970	neutral
	0.819	0.003	0.995	0.819	0.898	0.858	0.982	0.957	positive
Weighted Avg.	0.897	0.083	0.915	0.897	0.897	0.842	0.982	0.955	

```

=== Confusion Matrix ===
a b c <-- classified as
770 178 2 | a = negative
1 2231 6 | b = neutral
3 325 1484 | c = positive

```

- **Random Forest shows Accuracy=89.8%**

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **RandomForest -P 100 -i 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1**

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds **10**
☐ Percentage split % **66**

More options...

(Nom) stat:label

Start Stop

Result list (right-click for options)

00:47:56 - bayes.NaiveBayes
00:48:09 - trees.J48
00:48:16 - trees.RandomForest
00:48:59 - bayes.NaiveBayes
00:49:57 - trees.J48
00:50:28 - trees.J48
00:50:55 - trees.RandomForest

Classifier output

```

Instances: 5000
Attributes: 3
X
stat.Text
stat.label
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===
RandomForest

Bagging with 100 iterations and base learner
weka.classifiers.trees.RandomTree -X 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.32 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 4490 89.8 %
Incorrectly Classified Instances 510 10.2 %
Kappa statistic 0.8348
Mean absolute error 0.0831
Root mean squared error 0.2001
Relative absolute error 19.7207 %
Root relative squared error 43.5901 %
Total Number of Instances 5000

=== Detailed Accuracy By Class ===

```

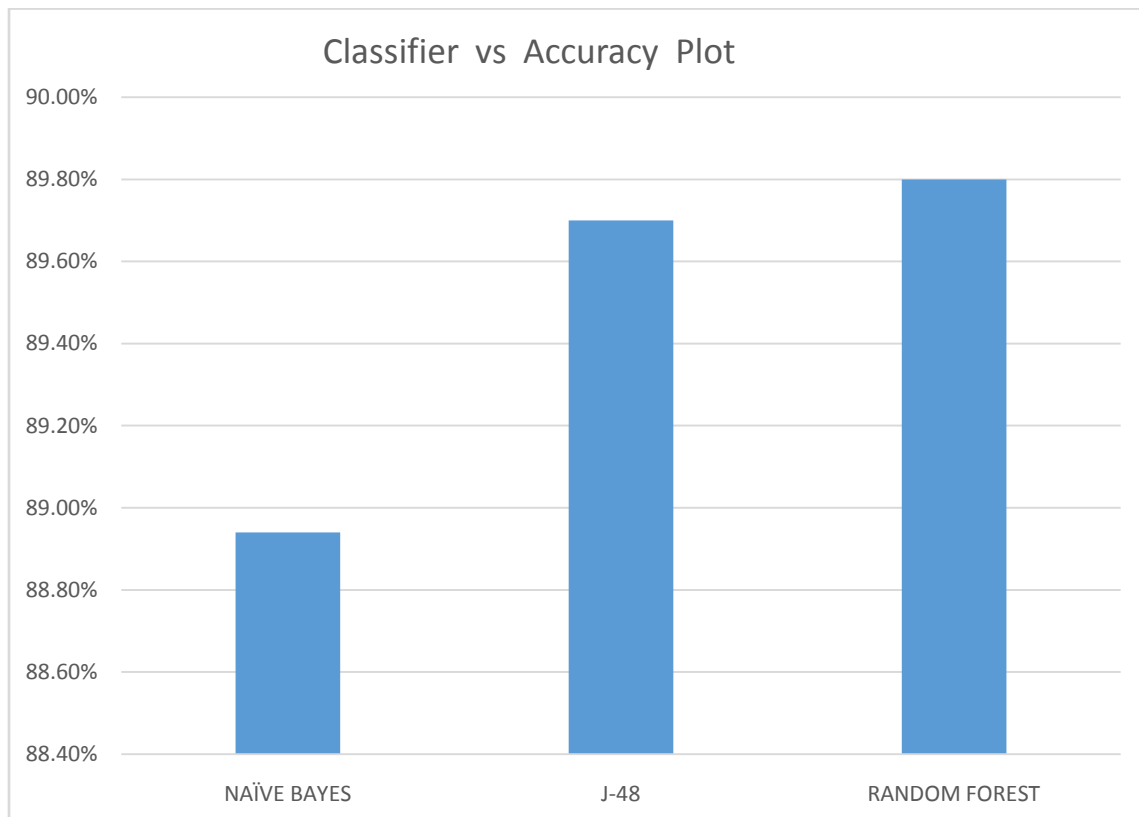
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.811	0.000	0.997	0.811	0.894	0.879	0.981	0.940	negative
	0.998	0.182	0.816	0.998	0.898	0.815	0.983	0.979	neutral
	0.821	0.002	0.996	0.821	0.900	0.860	0.983	0.972	positive
Weighted Avg.	0.898	0.082	0.916	0.898	0.898	0.844	0.983	0.969	

```

=== Confusion Matrix ===
a b c <-- classified as
770 178 2 | a = negative
1 2233 4 | b = neutral
1 324 1487 | c = positive

```

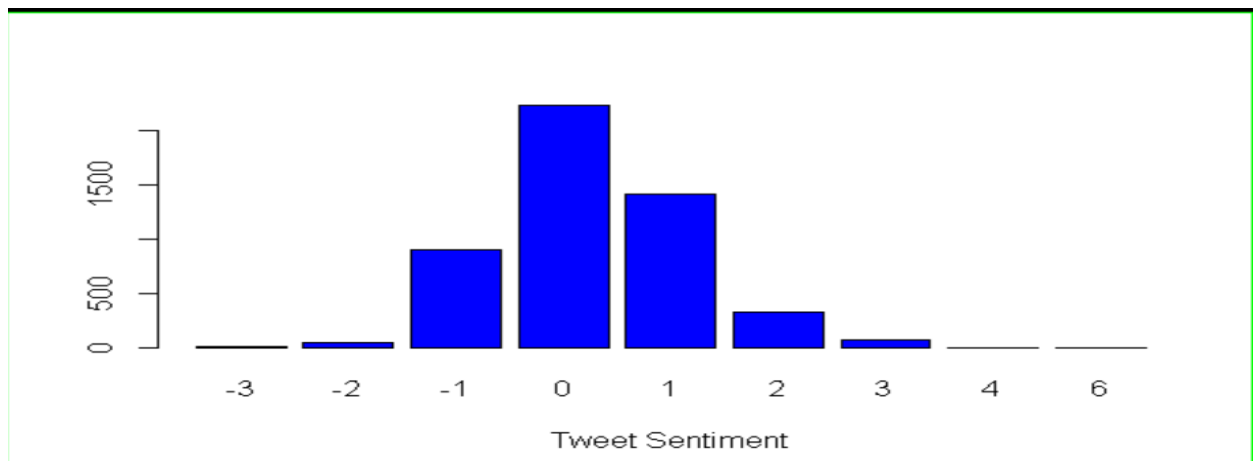
- Plot showing Accuracy of the classifiers.



Therefore, as the plot suggests Random Forest showed us the highest Accuracy i.e. 89.8% in comparison to others which are J48 decision tree and Naïve Bayes

5.2. Results and Findings

This shows us the graph of the No. of tweets to the score of the tweets. From this graph we can interpret that there are a lot of neutral tweets as the value on 0 score is high and then come positive no. of tweets with score 1.



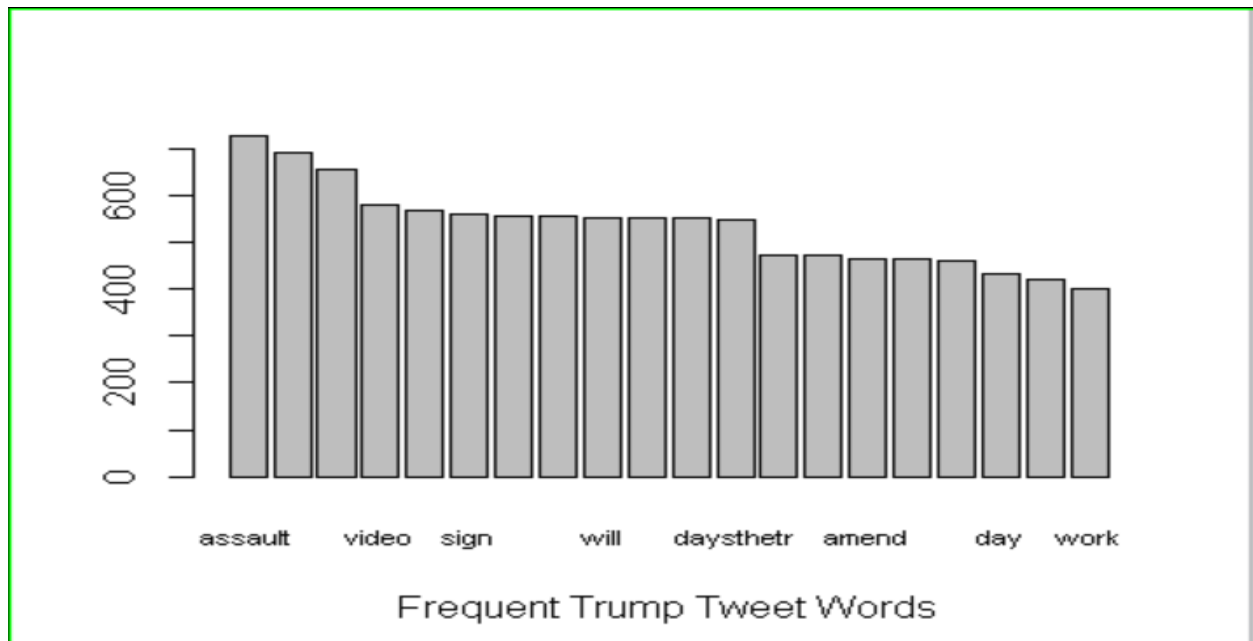
- This gives us the list of the best (most positive) among the tweets that are extracted from twitter.

```
> best_tweet = table_final$Text[table_final$score==3]
> best_tweet
[1] Congratulations PresidentTrump I can sleep better knowing my Sailor will serve under your leadership Its an honor
[2] Be Proud to be an American Be Proud to Stand Tall during Our NationalAnthemBe Proud that PresidentTrump wants to
[3] We LOVE eduAUBDedUBUE ourhe is OneofAKind Keep up the GOOD WORK PresidentTrump FirstDays TrumpRally HarrisburgPA htt
[4] We LOVE eduAUBDedUBUE ourhe is OneofAKind Keep up the GOOD WORK PresidentTrump FirstDays TrumpRally pOKAPQPDBC
[5] But we know you are doing an amazing job Keep it up we love everything youre doing Ronald Reagan would be proud zIluoBgZfd
```

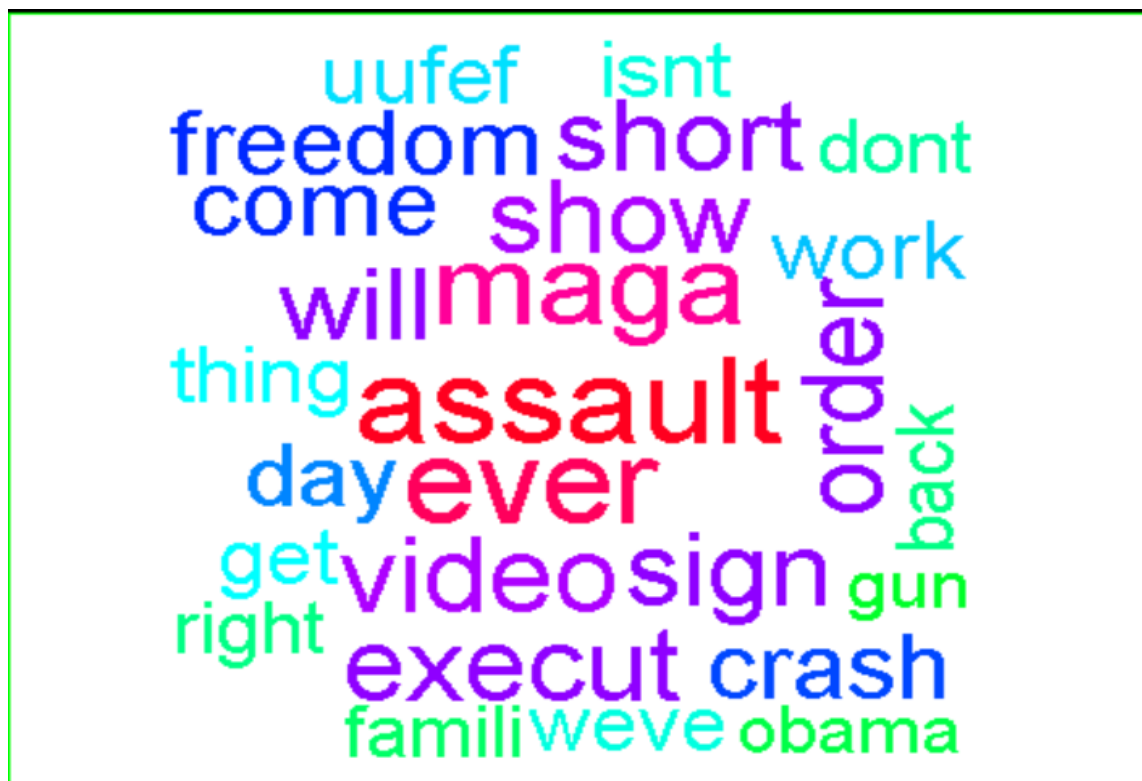
- This gives us the list of the worst tweets from the 5000 tweets extracted by us.

```
> worst_tweet = table_final$Text[table_final$score== -3]
> worst_tweet
[1] No but you sure as all hell are biased MAGA PresidentTrump fakenews
[2] The left will never accept PresidentTrump no matter how hard he tries Hillary lost and theyre deva z iGqxjioc
[3] Im disappointed w PresidentTrumps actions sick of pandering to israel and what does he do BuildTheWall arou boiRGxZ
[4] why dont YOU fucking Cowardly Americans Save PresidentTrumpeduAUBDedUBUAfrom the enemy
[5] FoxNews ppl asked about PresidentTrumpstdays someone said hes unpredictable wtf how so tired of hearing this eduAUBDedUBUB
[6] Wasted time wasted effort Wasted money PresidentTrump swzzlrjT
```

- This is visually showing us the Frequently used words in the tweets.



- This visually shows us the maximum no. of words used in form of WORD CLOUD.
- From here we can see that assault is the word used mostly in the tweets.



6. Conclusions and Future Work

6.1. Conclusions

After extracting and performing our analysis we have figured out various points which helped us analyze many things about the current thoughts and opinions people have about TRUMP.

It helped us get the information that there are recently many positive things being said about the president. This is also supposed to be practically possible as we must not forget the fact that he has been elected by the people of USA and therefore it shows that there is a good section of people that believes in him and his policy making and support him.

The second this we analyzed was that maximum no. of times the word assault was used during the extraction of the recent 5000 tweets. This is practically possible as recently he has proposed a policy for making GUNS legal in USA and therefore the assault tweets were all related to it as many people shared their opinion over it.

6.2. Limitations

1) Nowadays, most of the tweets have emoticons to express their emotions and therefore emoticons, GIF, Pictures can also be a valuable tool for sentimental analysis and in our case, we have considered only positive and negative words in calculating our sentimental analysis. Therefore, if we perform our sentimental analysis considering the above-mentioned expressions then analysis would become more effective and higher accuracy can be achieved.

2) The Positive and Negative bag of words that we have added to help us get the positive and negative score is though large but still does not contain all the words in the dictionary therefore this is a limitation. But still we have added few more words which are present in our positive or negative bag of words and considered few widely used urban-lingos as they are mostly tweeted these days, words like gr8, awwsom, aww, Sex C etc.

6.3. Potential Improvements or Future Work

1) We can introduce the emoticons, GIFs and try to implement a function so that we can also consider it in our sentimental score analysis in future that would increase our scope for getting more accuracy and we must also not forget that today's generation likes to keep things short and simple therefore use of GIFs, Urban-lingos and emoticons are too much.

2) We can also work on the area of timely comparison and growth of TRUMPs popularity or his downfall. The results of the sentimental analysis would help us predict some solutions which would help him in his future campaigns and policy announcements ahead of future elections as sentimental analysis tells us about the recent mindset of the public about whom they tweet.