PREDICTING ELECTION RESULTS THROUGH SOCIAL NETWORK ANALYSIS

1. ABSTRACT

Social media has exploded as a category of online discourse where people create content, share, bookmark and network at a prodigious rate. Examples include Facebook, MySpace, Digg, Twitter on the academic side. Because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry.

The impact of social media in elections will be studied extensively in this project. Our objective here is to study the social network of the two prominent election candidates- Hillary Clinton and Donald J. Trump. The main aim of the project is to understand the reach of Hillary Clinton and Donald J. Trump among Twitter users and how their social media strategy is helping or affecting their election campaign

2. INTRODUCTION

In recent times, social networking sites have provided a medium through which social identities can be developed and explored. These days the age of internet has changed in the way the people express their opinions and their views on a particular subjects. Social Media has become a new norm of life for billions of people today. With an ever increasing amount of time spent on these online sites sharing information, opinions and emotions, sites like Facebook, Twitter and Google+ have become hot data spots, waiting to be discovered and analyzed. These social media are generating a large amount of data which is rich in sentiments which can be seen in the form of tweets, statues, blogs, reviews, comments etc.

Today many millions of people all over the world are connected by being members of various Internet social networks and Twitter is one of most used social platform. Twitter is becoming an increasingly natural environment for a growing fraction of the world's population. Twitter has become a real-time voice for the public watching along with debates and campaign announcements. Tweets are frequently used to express a tweeter's emotion on a particular subject.

One of the most important moments in every country is presidential elections. Hence in this project we plan to analyze different tweets of people regarding the elections using twitter as a platform to achieve a large, diverse dataset of current public opinions on the candidates for sentiment analysis scraping the data using twitter API and then analyzing tweets for sentiments using R. API (Application Programming Interface) is basically a set of programming instructions and standards which helps to access the web based software applications.

Twitter provides us with live access to opinions about the election across the globe. We are analyzing recent tweets regarding the top candidates in the 2016 election to predict the public sentiments towards each candidate. Our project attempts to solve the real world problem of understanding the current sentiment towards election candidates based off the public's live opinions and emotions rather than off of smaller, localized polls typically done by mainstream media corporation.

Then we come up with different visualisation and resources in order to make valid conclusion about the chances of candidates of winning the 2016 presidential elections.

3. LITERATURE REVIEW SUMMARY TABLE

Authors and Year (Reference)	Title (Study)	Concept / Theoretical model/ Framework	Methodology used/ Implementation	Dataset details/ Analysis	Relevant Finding	Limitations/ Future Research / Gaps identified
Sitaram Asur Bernardo A. Huberman (2010)	Predicting the Future With Social Media	We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be utilized to improve the forecasting	Assess how buzz and attention is created for different movies and how that changes over time. Then focus on the mechanism of viral marketing and pre-release hype on Twitter, and the role that attention plays in forecasting real-world box-office performance.	The dataset that they used was obtained by crawling a regular feed of data from Twitter.com. To ensure that they obtained all tweets referring to a movie, they used keywords present in the movie title as search arguments. They extracted tweets over frequent intervals using the Twitter Search Api 4, thereby ensuring they had the	Shown how social media can be utilized to forecast future outcomes. Specifically, using the rate of chatter from almost 3 million tweets from the popular site Twitter, we constructed a linear regression model for predicting box-office revenues of movies in advance of their release. We then showed that the results outperformed	Using the tweets referring to movies prior to their release, can we accurately predict the box-office revenue generated by the movie in its opening weekend ?

Authors	Title	Purpose	Methodology/Sample	Context	Findings	Future Research
			power of social media.	timestamp, author and tweet text for our analysis. We extracted 2.89 million tweets referring to 24 different movies released over a period of three months.	in accuracy those of the Hollywood Stock Exchange and that there is a strong correlation between the amount of attention a given topic has (in this case a forthcoming movie) and its ranking in the future.	
Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe	Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment	Discuss the use of microblogging message content as a valid indicator of political sentiment and derive suggestions for further research.	We examined 104,003 political tweets, which were published on Twitter's public message board between August 13th and September 19th, 2009, prior to the German national election, with volume increasing as the election drew nearer. We collected all tweets that contained the names of either the 6 parties represented in the German parliament. To extract the	In our study, we use 104,003 tweets published in the weeks leading up to the federal election of the national parliament in Germany which took place on September 27th, 2009. After 4 years in a grand coalition with the social democrats (SPD), Chancellor Angela Merkel member of the conservatives (CDU) - was running for reelection, but favoring a coalition with the liberals (FDP).	Found more than one third of all messages to be part of a conversation indicating that Twitter is not just used to spread political opinions, but also to discuss these opinions with other users. Found the sentiment profiles of politicians and parties to plausibly reflect many nuances of the election campaign.	Further research should test whether text analysis procedures which are more closely tailored to the political debate reflecting both the specifics of microblogging and the political issues can produce even more meaningful results. Researchers

			sentiment of these tweets automatically, we used LIWC2007 (Linguistic Inquiry and Word Count; Pennebaker, Chung, and Ireland 2007), a text analysis software developed to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary.			should also try to capture the context of a particular statement in a more comprehensive manner including threads of conversation and links to information beyond the tweet.
Trammell, K.D., Williams, A.P., Postelnicu, M. and Landreville, K.D., 2006.	Evolution of online campaigning: Increasing interactivity in candidate web sites and blogs through text and technical features. *Mass Communication & Society*, *9*(1), pp.21-44..	In achieving the research goals about interactivity, this study analyzed interactivity in the 10 Democratic primary candidates during the 2004 primary election period. As such, it is helpful to review existing literature regarding interactivity and	This study used the webstyle quantitative content analysis method to review the Web site front pages and blog posts of the 10 Democratic presidential primary candidates in the 2004 election.	Examined the content and features offered in candidate Web sites and blogs regarding interactivity.	*Interaction of user on social media*	This research is limited in that it analyzed candidate posts and did not review the comments left on the blogs in response to the post. Additionally, the analysis only reviewed the front pages of the

		operationalize the concept.				candidate Web sites, thereby limiting the generalization of these findings.
Williams, C. and Gulati, G., 2008.	What is a social network worth? Facebook and vote share in the 2008 presidential primaries. American Political Science Association.	Study undertakes an initial analysis of the impact Facebook had on the 2008 presidential nominating contests.	Estimated Facebook's impact through multivariate regression models explaining each candidate's popular vote share for all state nominating contests	A day or two before a contest, they recorded the percentage of supporters for each candidate by state and use these data to measure relative Facebook strength in each state.	*The impact social media has on elections.*	If future research can confirm these findings and demonstrate a causal connection between online strategies and votes, then Facebook and other social networking sites will be an essential tool in enhancing the democratic process.
Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D., Marlow, C., Settle, J.E. and Fowler, J.H., 2012.	A 61-million-person experiment in social influence and political	Report results from a randomized controlled trial of political mobilization messages	Randomized controlled trial of political mobilization messages delivered to 61 million Facebook users during	Messages to 61 million facebook users	*Examining* real-world voting behaviour of millions of people	Experiments are expensive and have limited external validity.

	mobilization. *Nature*, *489*(7415), pp.295-298.	delivered to 61 million Facebook users during the 2010 US congressional elections. The results show that the messages directly influenced political self-expression, information seeking and real-world voting behaviour of millions of people.	the 2010 US congressional elections.			

4. OBJECTIVE OF THE PROJECT:

In this project, we have considered the twitter social media platform to find out how tweets from the twitter feed can be utilized to perform sentiment analysis. We performed sentiment analysis on two leading candidates of 2016 Elections and R programming language has been extensively used to perform this analysis. The approach followed here is to count the positive and negative words in each tweet and assign a sentiment score. This way, we can ascertain how positive or negative a tweet is. In our study we analysed the tweets before the US elections where we are trying to predict which candidate has a better chance of winning by analysing the tweets collected and classifying them.

Hence our main project boils down to 3 main points.

- Collecting the tweets regarding the candidates using the Twitter REST API.

- To predict the sentiments of public towards each candidate by analyzing the tweets regarding the candidates by using R.

- Come up with resources to arrive at a valid conclusion for each of the candidates.

5. INDIVIDUAL CONTRIBUTION IN THE PROJECT:

Shireen Saini (15BCE0165)- Set up Twitter API, extracted relevant tweets and obtained datasets from Twitter using R.

Tanvi Pareek (15BCE0764)- Performed sentiment analysis on the obtained tweets using R.

Abharika Nayyer (15BCE0779)- Obtained sentiment scores and performed comparative analysis.

7. WORK DONE AND IMPLEMENTATION

a. Methodology adapted:

Software requirements:

- R Software

Steps Followed:

1. Extracting and Analyzing Tweets

TwitterR offers an easy way to extract tweets containing a given hashtag, word or term from a user's account or public tweets.

2. Setting Authorization to Extract Tweets

We needed to get access to the twitter search API which required us to create our own application and generate access token which can be used while querying data from twitter. These access token provide us a method to authorise ourselves when requesting data from twitter.

3. Required Libraries

We loaded the required libraries:

```
library(twitteR) ### for fetching the tweets
library(plyr) ## for breaking the data into manageable pieces
library(ROAuth) # for R authentication
library(stringr) # for string processing
library(ggplot2) # for plotting the results
```

4. Importing Files

We had to import text files containing a dictionary of positive and negative words.

5. Extracting Tweets with Hashtags

To demonstrate sentiment analysis, we analyzed tweets relating to Hillary Clinton and Donald Trump.

7.Processing Tweets

    a. Convert the tweets to a text format.

    b. Calculate the number of tweets for each candidate.

    c. Combine the text of all these candidates.

8. Sentiment Analysis Application

Before we proceeded with sentiment analysis, a function was defined that will calculated the sentiment score for each candidate.

    a. Processed the tweets to calculate the sentiment score.

    b. Created a variable scores$election in the data frame.

    c. Split the data frame into individual datasets for each candidate.

    d. Created polarity variable for each data frame.

9. Generating Graphs

    a. Polarity Plot –Sentiments (Hillary Clinton, Donald Trump)

    b. Sentiment Scores (Hillary Clinton, Donald Trump)

10. Summarizing Scores

We summarized the overall positive, negative and neutral scores

11. Comparison Charts

a. Positive Comparative Analysis

We created a positive comparison pie chart for these two candidates.

b. Negative Comparative Analysis

We created a negative comparison pie chart for these two candidates.

c.Neutral Comparative Analysis

We created a neutral comparison pie chart for these two candidates.

b. Dataset used :

a. Where from you are taking your dataset?
   For our dataset we made use of Twitter API which can provide us with large number of tweets based on the query parameters being passed in the API. First we need to generate access token which can be used while querying data from twitter. These access token provide us a method to authorise ourselves when requesting data from twitter.

b. Is your project based on any other reference project (Stanford Univ. or MIT)?

   Yes, this project is based on a research paper: "Predicting the Future With Social Media Sitaram Asur Social Computing Lab HP Labs Palo Alto, California Email: sitaram.asur@hp.com Bernardo A. Huberman Social Computing Lab HP Labs Palo Alto, California Email: bernardo.huberman@hp.com"

c. How does your project differ from the reference project?[*]
   The research paper focussed on using Twitter to predict how a movie performs at the box office, we will make use of Twitter to predict the results of a General Elections held in any country.

c. Tools used:
   R: R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy. R is the leading tool for statistics, data analysis, and machine learning. It is more than a statistical package; it's a programming

language, so you can create your own objects, functions, and packages. Speaking of packages, there are over 2,000 cutting-edge, user- contributed packages available on CRAN. You can easily use it anywhere. It's platform independent, so you can use it on any operating system. And it's free, so you can use it at any employer without having to persuade your boss to purchase a license. With large number of packages and easy UI we were able to develop our project in time and with great certainty. Thus we used R in our project.

The Twitter API: The Twitter API platform currently offers a few different options for programmatically searching the index of Tweets. These options range from the standard 7-day Search API (search/tweets) to the enterprise Full-Archive Search API. Each option offers a varying level of access and query capabilities. These API provide a public access to tweets which can be further filtered using different query parameters such as name of the person, a particular topic etc. We can provide a time range between which all the tweets should exist. The easy availability, ease to use made us use this tool. This API returns the data in JSON format which is easy to read, understand and analyse.

d. Screenshot and Demo:

Access to Twitter Search API

CODE:

```
reqURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
api_key <- "9TlF5WwcPtHZo6zqVv6SLiAf6"
api_secret <- "K7wobKo7vWeTflqO3uVPpAEnqo58WinZxfpVEpkvvpISUjMVhI"
access_token <- "    909098405514940417-
PsK3w5nWLWgqqR7rQgvEXXz8W5wuxWD"
access_token_secret <- "hoClellc5PGuEC4JUevNSKP7gmeLFxJtw8Qfp0EpYwRQJ"
setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)
```

SCREENSHOT:
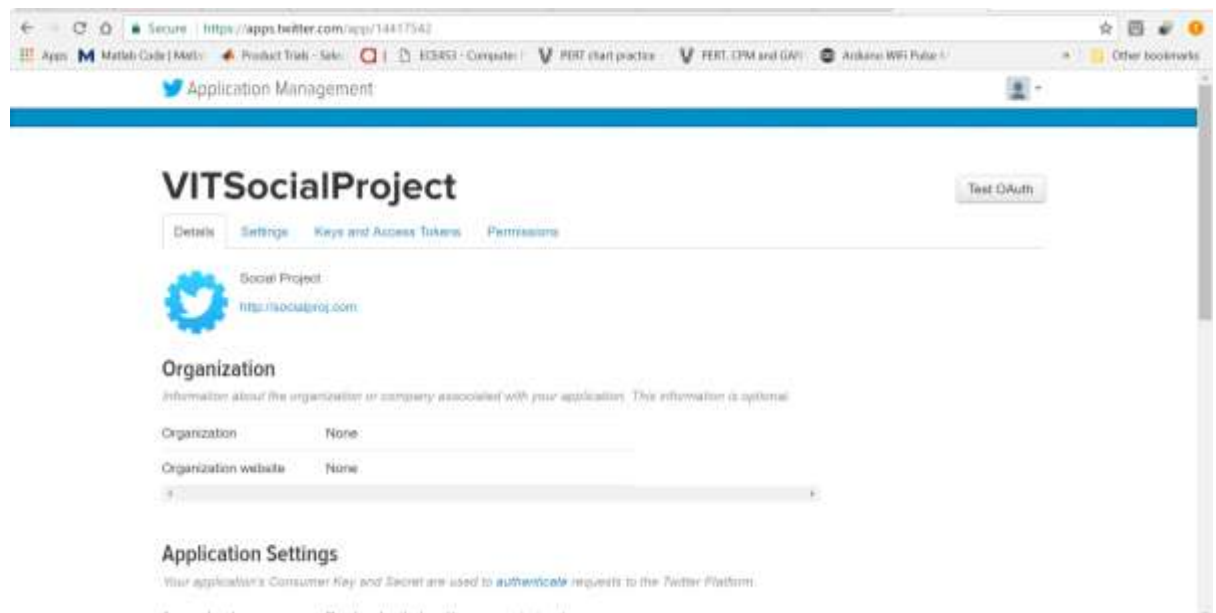
IMPORTING FILES

CODE:

```
posText <- read.delim("C:/Users/user/Desktop/positive-words.txt", header=FALSE,
stringsAsFactors=FALSE)
posText <- posText$V1
posText <- unlist(lapply(posText, function(x) { str_split(x, "\n") }))
negText <- read.delim("C:/Users/user/Desktop/negative-words.txt", header=FALSE,
stringsAsFactors=FALSE)
negText <- negText$V1
negText <- unlist(lapply(negText, function(x) { str_split(x, "\n") }))
pos.words = c(posText)
neg.words = c(negText)
```

Extracting Tweets AND SAVING IN .CSV FILES

CODE:

```
Hillary_tweets = searchTwitter('@HillaryClinton', n=5000)
Trump_tweets = searchTwitter('@realDonaldTrump', n=5000)

tweets.df=ldply(Hillary_tweets,function(t) t$toDataFrame())
summary(tweets.df)
setwd("C:/Users/user/Desktop")
getwd()
write.csv(tweets.df,file="Hillarytweets.csv")

tweets1.df=ldply(Trump_tweets,function(t) t$toDataFrame())
```

```
summary(tweets1.df)
setwd("C:/Users/user/Desktop")
getwd()
write.csv(tweets1.df,file="Trumptweets.csv")
```

SCREENSHOT:

Trumptweets.csv

Hillarytweets.csv
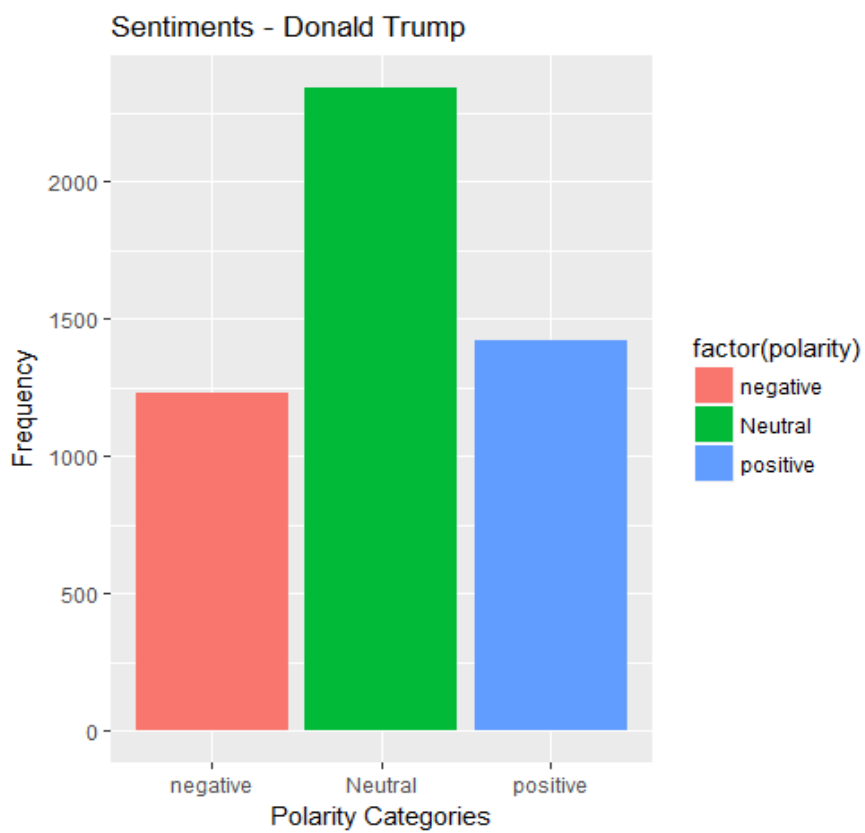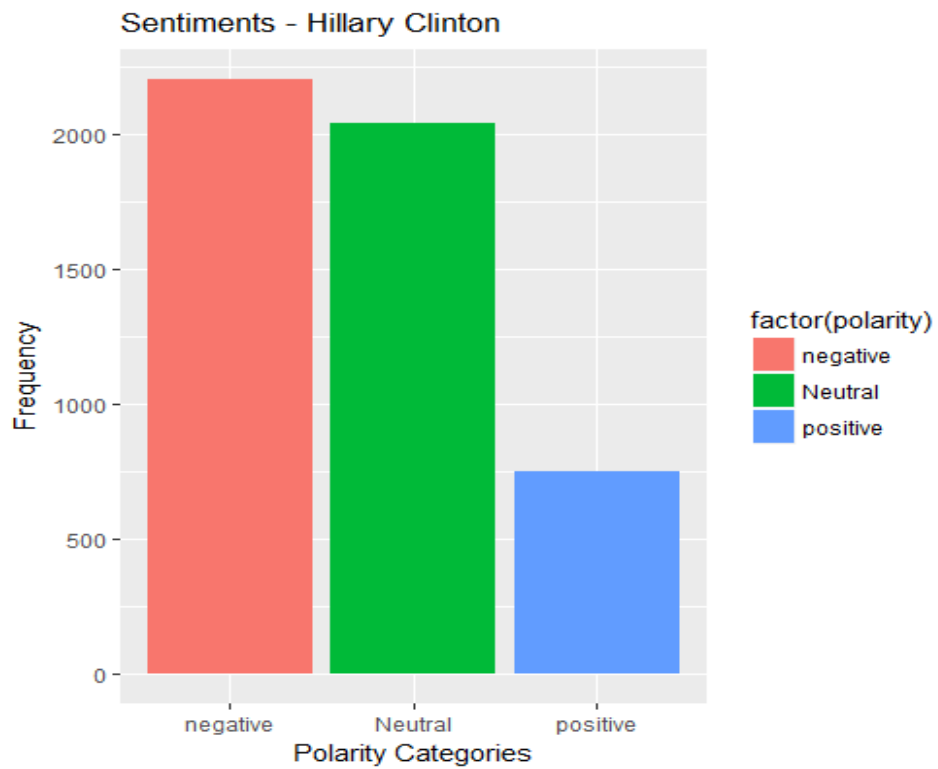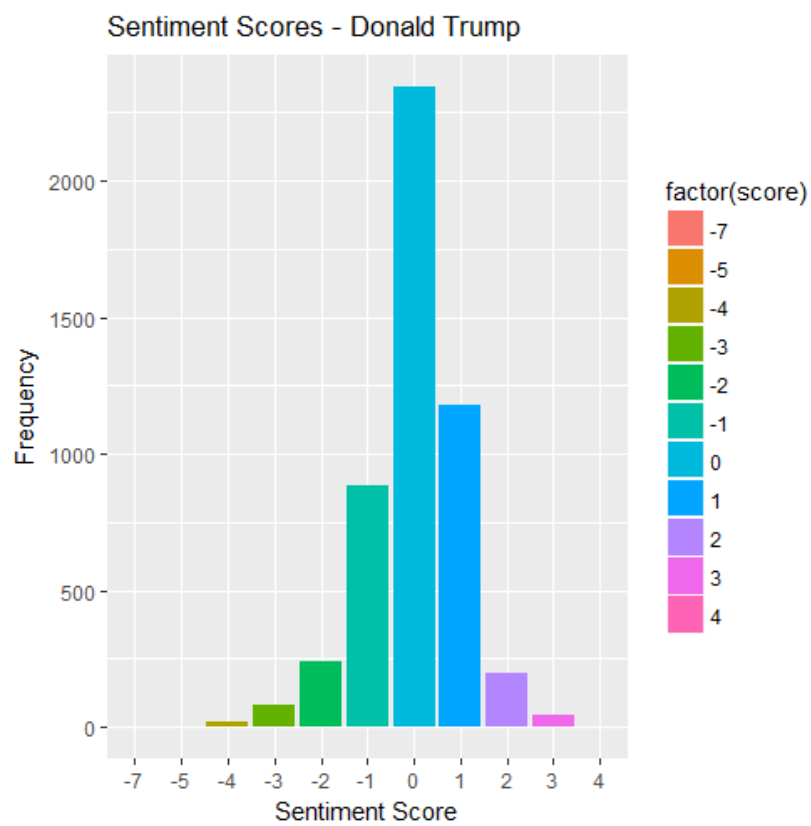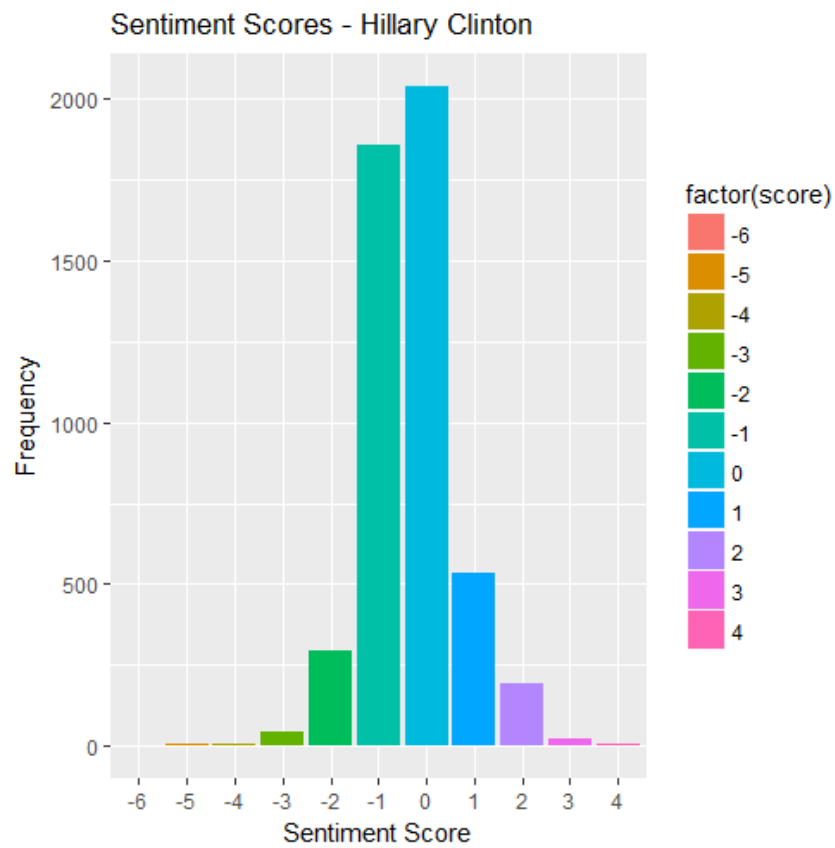
Calculating Scores

CODE:

```
scores$positive <- as.numeric(scores$score >0)
scores$negative <- as.numeric(scores$score <0)
scores$neutral <- as.numeric(scores$score==0)

Hillary_election <- subset(scores, scores$election=="HillaryClinton")
Trump_election <- subset(scores,scores$election=="DonaldTrump")
Hillary_election$polarity                    <-            ifelse(Hillary_election$score
>0,"positive",ifelse(Hillary_election$score                                          <
0,"negative",ifelse(Hillary_election$score==0,"Neutral",0)))

Trump_election$polarity                      <-            ifelse(Trump_election$score
>0,"positive",ifelse(Trump_election$score                                            <
0,"negative",ifelse(Trump_election$score==0,"Neutral",0)))
```
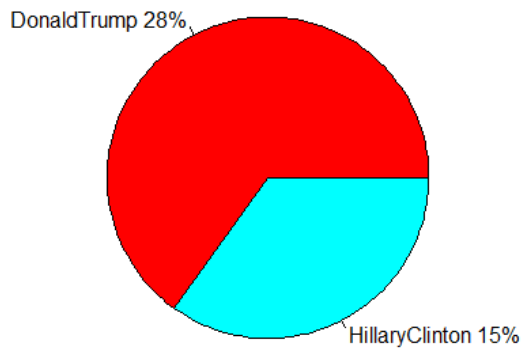
SCREENSHOT

Sentiments – Hillary Clinton

Sentiments – Donald Trump

Sentiment Scores - Hillary Clinton

factor(score)
-6
-5
-4
-3
-2
-1
0
1
2
3
4

Sentiment Scores - Donald Trump

factor(score)
-7
-5
-4
-3
-2
-1
0
1
2
3
4

Positive Comparative Analysis - Elections

DonaldTrump 28%

HillaryClinton 15%

Negative Comparative Analysis - Elections

DonaldTrump 25%

HillaryClinton 44%

Neutral Comparative Analysis - Elections

DonaldTrump 47%
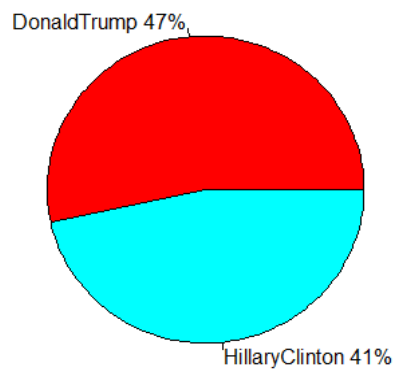
HillaryClinton 41%

8. RESULTS AND DISCUSSION

Based on the comparative analysis graphs we have concluded that Donald Trump is the winner of the elections.

Statistical results obtained by extracting 5000 tweets for each candidate:

Hillary Clinton:

Negative Tweets: 2250(Approximate)

Neutral Tweets: 2000(Approximate)

Positive Tweets: 750 (Approximate)

Donald Trump:

Negative Tweets: 1250(Approximate)

Neutral Tweets: 2250(Approximate)

Positive Tweets: 1400(Approximate)

Thus from the above results we concluded that Donald Trump had received a more positive response than Hillary Clinton.

Also observing the sentiment scores, Hillary Clinton's sentiment scores were tending more towards negative.

9. APPLICATION IN REAL WORLD

As can be seen, sentiment analysis enables enterprises to understand consumer sentiments in relation to specific products/services. Moreover, these insights could be used to improve their products and services by gauging consumers' comments and feedback using sentiment analysis. In the long run, sentiment analysis, if implemented the right way can aid business enterprises in improving the overall consumer experience, enhance brand image and propel business growth.

10. REFERENCES

1. Predicting the Future With Social Media Sitaram Asur Social Computing Lab HP Labs Palo Alto, California Email: sitaram.asur@hp.com Bernardo A. Huberman Social Computing Lab HP Labs Palo Alto, California

2. Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment

3. Trammell, K.D., Williams, A.P., Postelnicu, M. and Landreville, K.D., 2006. Evolution of online campaigning: Increasing interactivity in candidate web sites and blogs through text and technical features. *Mass Communication & Society*, *9*(1), pp.21-44..

4. Williams, C. and Gulati, G., 2008. What is a social network worth? Facebook and vote share in the 2008 presidential primaries. American Political Science Association.

5. Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D., Marlow, C., Settle, J.E. and Fowler, J.H., 2012. A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), pp.295-298.