FACULTY OF ARTS AND SOCIAL SCIENCES

GROUP COURSEWORK COVERSHEET

Coursework Details			
Module Name and Code (please check front of module handbook)	Introduction to Marketing Analytics MANM317		
Coursework Title	Nike's Customers Conversations on Twitter Analysis		
Date of Submission	10/06/19	**Word Count**	2364

Students Details			
Students URN (7 digit number on Uni card)	6562752 6556972 6264027 6560833 6345112	**Student Names**	Amol Dixit Arun Loyal Johan Thomas Stefan Dimitrov Stoyanov Vartan Zahorodnykov
Programme	Business Analytics MSc - 2018/9		

Students Declaration

To be agreed by all students

Please refer to the **University of Surrey Regulations for the Conduct of Examinations and Other Forms of Assessments** and your departmental **Student Programme Handbook** for more information on Academic Misconduct and Plagiarism.

Declaration:

I confirm that the submitted work is my own work and that I have clearly identified and fully acknowledged all material that is entitled to be attributed to others (whether published or unpublished) using a referencing system. I agree that the University may submit my work to means of checking this, such as the plagiarism detection service Turnitin® UK. I confirm that I understand that assessed work that has been shown to have been plagiarised will be penalised.

By completing and submitting this form, we confirm that:

- We have read and fully understand the University's Regulations and guidance on Academic Misconduct and Plagiarism
- This submission is our own work
- All quotes and sources have been fully and properly attributed and referenced
- This work has not been previously submitted, in full or in part, for the purpose of assessment at this or any other institution
- No effort has been made to subvert plagiarism detection processes of the University
- This submission may be transferred to and stored in the Turnitin Plagiarism Detection database for the purpose of plagiarism detection now and in the future
- We understand that all required work must be received within the published deadline
- We understand that work received after the published deadline will be penalised in line with University Regulations
- We understand that any request for mitigating circumstances must be made formally, using the appropriate form and including evidence; the application and associated evidence must be received by the stipulated date

BY SUBMITTING THIS FORM, YOU AGREE TO THE STUDENT DECLARATION IN FULL

Executive Summary

Purpose – .

The purpose of this report is to analyse customer tweets using new technologies such as text mining and social media mining to provide a short overview about Nike customer conversations to the line manager.

Design/methodology/approach – .

This exploratory analysis will be done by first scrapping Nike customer tweets from Twitter. Analyse the data using text mining to produce visualisations in R. Finally use the results to provide insights which the company can use to make better decisions in their marketing team.

Findings – .

The findings from the analysis showed that the sentiment towards Nike was mostly neutral with a slight skew towards negative. Most of the frequent words were associated with the products that are sold by Nike. There are a few to do with the social causes Nike participates in.

Practical implications – .

Practical implications are Nike and other companies should use text mining and other new technologies to get a better understanding of their current and potential customers. It can help provide feedback at a faster rate which will help make better and faster decisions.

Keywords: Sentiment Analysis, Nike, Text Mining, Web scraping,

Paper type: A student research report

Contents

1. Introduction

Due to the advent of social media platforms like Twitter, Instagram, the traditional approach to marketing have been made inefficient [Scott, 2015]. These platforms allow customers and potential customers to share their views on products and services provided by companies. These social media platforms are now a treasure trove of data which if used by companies can help them improve their marketing strategy and therefore their sales [Barker, 2017]. One such method companies are using to analyse this data is known as text mining. Text mining is a method to extract valuable information from randomly organised data [Fan et al., 2006]. It can be used to gauge the sentiment towards the brand and its products which will help companies make better marketing decisions. As customer opinions can influence both current and potential customers.

This report will analyse customer tweets about the brand Nike. We will be focusing on Twitter as Nike has 7,826,497 followers [Socialbakers.com, 2019]. The analysis will be done by web scraping tweets, conducting basic text mining analysis, visualising the results into charts and finally conclude by providing actionable insights. this will be done in R which is an open source platform for the analysis of data.

2. Premises / Theoretical Foundations

2.1 Problem Definition – Effect of Social Media on Company Performance

Social media is an ever-growing activity which is known to be one of the defining phenomena in the 21st century and is reshaping business due to worldwide accessibility on the internet. Social media has many forms which include blogs, forums, business networks, photo sharing platforms, social gaming, chat applications, but most importantly social networking. The number of worldwide social media users is expected to reach 3.02 billion by 2021 [Statista, 2017].

As social media platforms like Twitter, Facebook and Instagram are increasing at a rapid rate, communication on these applications are seen to be the desirable option rather than the traditional communication of emails or face to face interaction. Not only has communication tweaked over time, but purchasing habits have also changed as social media is now a key component of an organization's marketing strategies [Barker, 2017]. As the emergence of online shoppers increases, brands need to invigilate their social media, i.e. Twitter to see what customers review about their products and services.

With such intense user activity, thousands of customer interactions with brands occur every day. Whilst this may seem beneficial to certain companies, there are also major threats of heavy conversations occurring on rapidly increasing social media platforms like Twitter. A barrage of large quantities of messages which contain negative word-of-mouth and complaint behaviour against a company in social media networks is a persistent problem, this is known as a social media firestorm [Ebner, 2014]. Online firestorms affect financial performance, if customers aren't satisfied then the brand's stock price will decrease.

2.2 Social Media Listening

Social media listening is looking for specific keywords in a brand's social media channels. This helps get customer feedback, information on competitors etc. Leading to analysis to gain insights and act on those opportunities [Amaresan, 2018].

Brands opt to use social media listening as we are emerging into a digitalisation era where thousands of user interactions are occurring across the internet rather than face to face. The commercial value of big international companies are heavily impacted by the development of social media, so they must find solutions to these issues early on otherwise there will be problems with maintaining their profitability relationships affecting their predicted growth.

Social media listening is a process which is similarly related to google as it performs a search on the internet, but instead focuses on information retained from social media channels. Information isn't just gathered from social media however, news channels and non-social channels like websites are also obtained. Major brands require social media listening to achieve their business goals, these are; customer segmentation, customer feedback on product/service, customer care and competition.

2.3 Brand Sentiment Analysis

Sentiment analysis builds systems that try to identify and extract opinions within text. It is a field within Natural Language Processing (NLP). These systems extract attributes of the expression by 3 given examples which are polarity, subject and opinion holder. Polarity identifies whether the individual expresses a positive or negative opinion, subject is the thing that is being talked about and opinion holder, the person, or entity that expresses the opinion [monkeylearn.com].

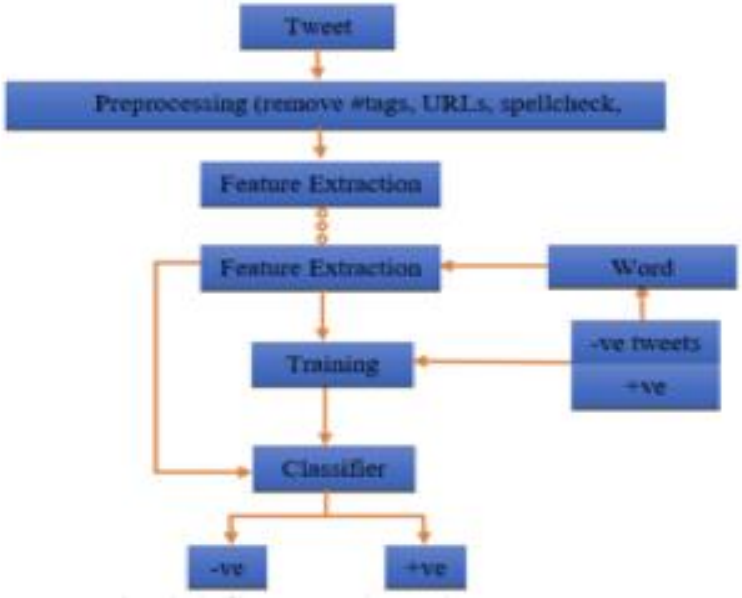

Fig. 1 System
Architecture of Twitter Sentiment Analysis [Shobana, Vigneshwara, Maniraj, 2018]

The diagram above shows the key components of how twitter sentiment analysis process works. Tweets are extracted from twitter using the API which is then pre-processed. Feature extraction recognises the patterns within the tweets, this is then obtained by the training set which regulates which set of tweets are positive or negative. These are analysed by machine learning algorithms, and the classifier outputs the polarity of tweet.

There are three different levels of sentiment analysis. The first level is document level, which is the simplest form of classification [Kolkur et al, 2015]. The aim of this is to classify whether the person expresses a positive or negative sentiment towards a brand, an example of this would be a product review of Nike, the system filters whether the review is positive or negative about a Jordan shoe.

The second level is sentence level. The task at level determines the sentence is positive, negative or neutral. The sentence level sentiment analysis has two tasks which are subjectivity classification and sentiment classification. Subjectivity classification means whether a sentence can be either subjective or objective sentence [Kolkur et al, 2015].

The last level is aspect level sentiment analysis. This is the stage where fine grained analysis is undertaken. The goal is to determine the aspects of each entity which is used in indicating the sentiments [Sarawgi, Pathak, 2017].

2.4 Customer Conversations

Different marketing strategies drive customer purchase decisions via customer conversations on social media platforms like Twitter whether it's offline or online engagement. An example of this was the Seattle-based retailer, Nordstrom. Donald Trump was newly elected as president of the USA, took to Twitter and criticised Nordstrom for dropping Ivanka Trump's clothing line. This reaction sparked a significant increase of 1700% weekly twitter mentions of the Nordstrom brand. Despite sentiment analysis indicating the tone of messages went slightly negative, the overall sentiment was positive. Nordstrom benefitted with a 2.5% increase of sales over the 2017 holiday season [Fay, Keller, Larkin, Pauwels, 2019]. This shows with a sheer volume of customer engagement on social media, whether they're offline or online conversations, it could impact heavily on profitability of the company.

The most impactful social media platform for your (company) brand is Twitter. Users who are active on twitter daily: 72% publish blog posts at least once a month [Macale, 2011]. Due to its influential nature of customers publishing reviews, the importance of brands companies paying more attention to Twitter followers is vital. Tweets are not only exposed about the brand on Twitter, but a simple Google or Bing search can also index the tweets about the brand. Conversations between the brand and customer won't ever be erased from history, therefore the margin of error for brands to promote their product or service is miniscule.

The power of social media online conversations could also impact the stock market. This was investigated by [Ranco et al] as they tested the effects of twitter sentiment on stock price returns. A timescale of 15 months was analysed between the volume of tweets and financial markets in 30 stock companies. They concluded that a relatively low Pearson correlation corresponded over the entire time period, however there was a significant correlation between Twitter sentiment and abnormal returns during peaks of twitter volume. From this investigation, sentiment polarity can impact stock price of companies, with such high volumes of positive and negative tweets in a time frame.

3. Methodology

The current project is performed using the R statistical software. The following R packages were used: twitteR, ROAuth, httr, tidyverse, quanteda, tm and word cloud. The steps and operations in implementing the proposed computational text analysis research model are as follows:

3.1 Step. 1 Web scraping

Initially, a Twitter development account was set up to access and use Twitter APIs. 1000 tweets containing the word 'Nike' were extracted from the web. They were stored in a CSV file.

3.2 Step. 2 Explore and prepare the data.

1. Importing the text into R

The raw data from Twitter was imported as a data frame into the R environment. It consisted of 17 variables and 1000 observations (tweets). The tweets' text which was the main interest of the research was stored in the second variable 'text'. The other variables were: 'X' (number), favorited, 'favoriteCount', 'replyToSN', 'created', 'truncated', 'replyToSID', 'id', 'replyToUID', 'statusSource', 'screenName', 'retweetCount', 'isRetweet', 'retweeted', 'longitude' and 'latitude'.
Each tweet is limited by 140 characters. However, the Twitter API includes the usernames in the tweets. Therefore, as a result, some tweets are truncated, and some tweets text is lost.

2. Data cleaning (with string operations)

The data from the 'text' variable was converted into a string vector with 1000 elements. Hexadecimal characters, URLs, new line and carriage return, and twitter names were removed.

3. Data preprocessing (tokenization and normalisation)

The data was put into the Volatile corpus. The text was converted to lowercase. The typical English stop words were removed, e.g. "a", "the", "of", "in", "and", "or", etc. The punctuation and the numbers from the text were removed. The space between words was also removed.

The stemming technique was performed. The inflected forms of words are considered as equivalent because of their close semantic relation. Therefore, they were converted into their base forms (stems). This helped to reduce the feature space.

4. Creating a Document-term matrix (DTM)

The tweets text corpus was presented in one of the most common bag-of-words formats, namely the document term matrix (DTM) format. DTM counts the frequency of each word used in the document. The corpus was transformed into a single document of all tweets.

3.3 Step. 3 Text mining analysis ("bag of words" approach)

The bag-of-words text analysis approach was applied. Only the frequencies of words per text were used and word positions were ignored.

1. Counting and dictionary

A deductive approach that uses patterns (like simple keywords or complex Boolean queries and regular expressions) to count how often certain concepts occur in texts. Dictionaries are also a popular approach for measuring sentiment.

2. Statistics

The most frequent words were mined and analysed. The frequencies of all words within the entire corpus, i.e. all tweets, were estimated. Then the frequencies of the 5 most popular words were highlighted.

3.4 Step. 4 Results Visualisation

1. Word cloud

A word cloud with the most frequent words was created.

2. Most frequent words barplot

A bar plot with the most frequent words was created.

The following figure represents the logical links between the steps of the applied methodology.

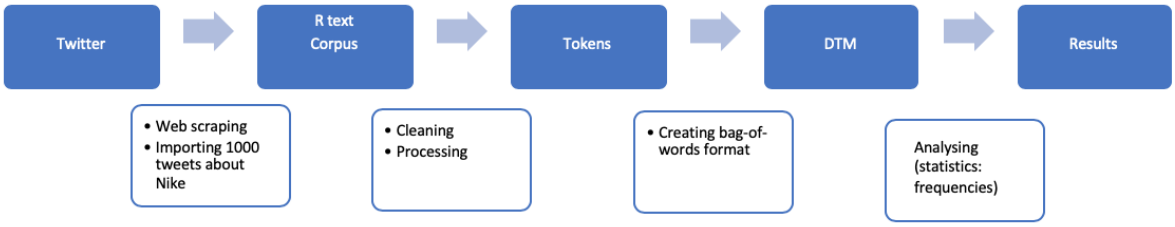

Fig. 2 Flowchart of the proposed model of text analysis operations for data preparation and analysis

4. Data Analysis and Results

4.1 Most frequent words

```
# get the freq of 10 most popular words
print(head(ordered,n = 10))
 size   know  think jersey continu    now    new    eat agenda    air
   98     82     82     77     76     75     73     72     71     71
```

Fig. 3 The 10 most popular words from the scrapped tweets

```
> print(tm::findFreqTerms(dtm,lowfreq =30))
 [1] "agenda"   "air"      "arent"    "avail"    "can"      "continu"  "custom"   "cut"      "day"
[10] "digit"    "dust"     "eat"      "feet"     "fit"      "get"      "jersey"   "just"     "know"
[19] "latest"   "max"      "mbhokodo" "men"      "new"      "now"      "offer"    "one"      "onlin"
[28] "open"     "peopl"    "registr"  "shoe"     "size"     "sport"    "star"     "thank"    "think"
[37] "use"      "version"  "wear"     "women"    "world"    "wrong"
```

Fig. 4 Word cloud formation with most popular words

Fig. 5 Word Cloud of most popular words

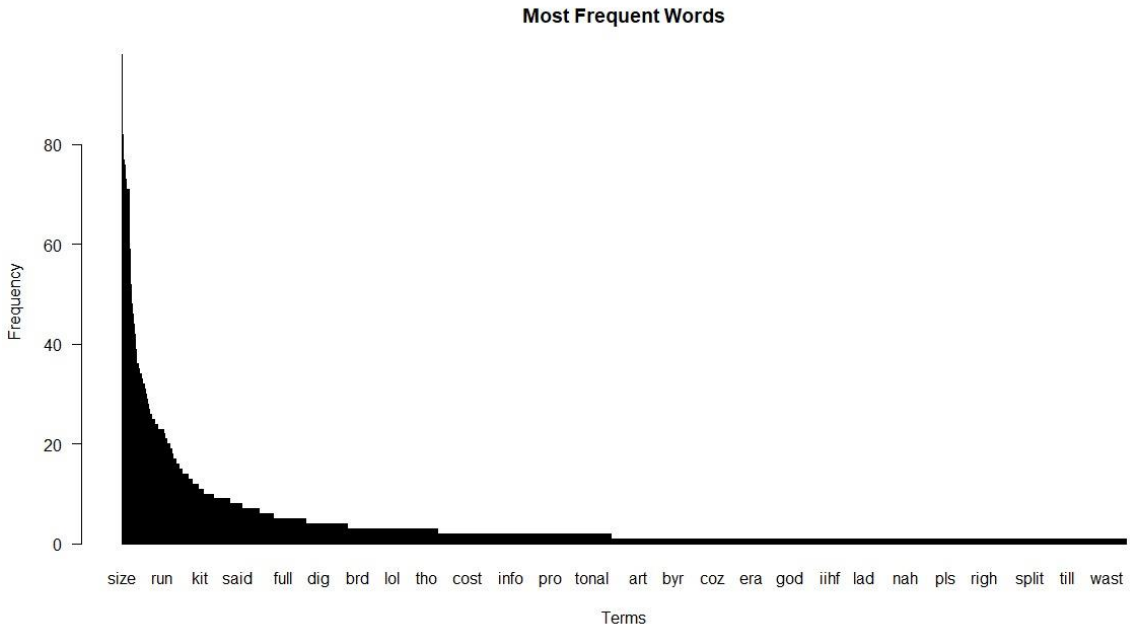

Fig. 6 Bar plot with most frequent words

Figures 3-6 above shows the most popular words obtained from the tweets scrapped. The majority of the words are related to the products that Nike sells. These include air for air, new, size,color etc. They are also words which are associated with social issues like women (see appendix 1). However the word 'Mbhokodo' is not associated with Nike and appeared as people use viral hashtags to get attention to their posts.

4.2 Sentiment Analysis

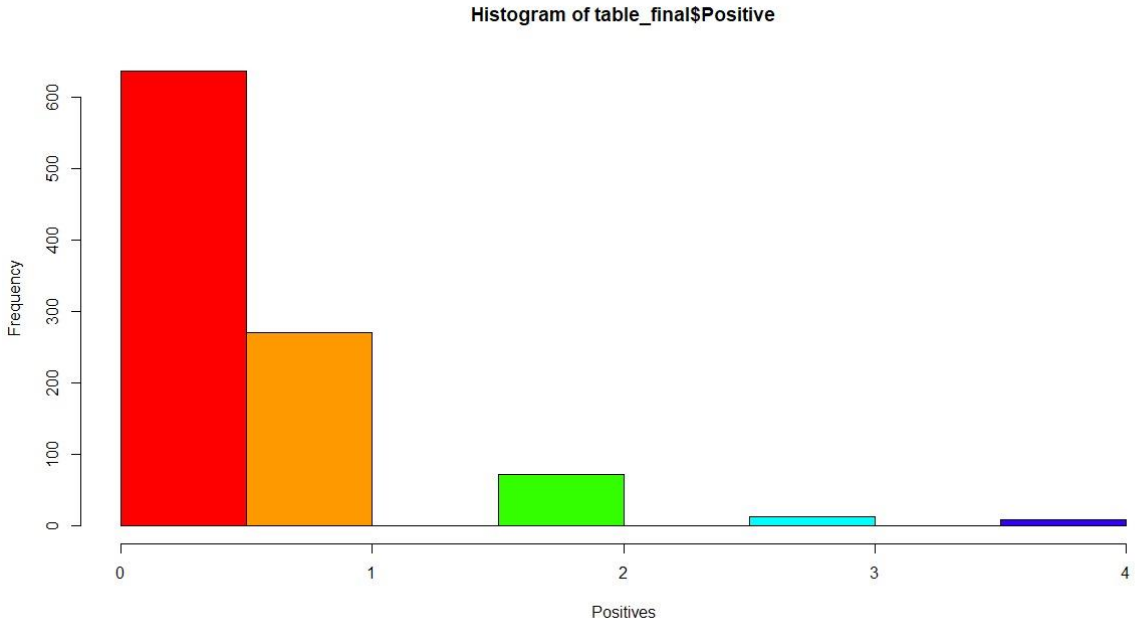

Fig. 7 Positive sentiment of Tweets

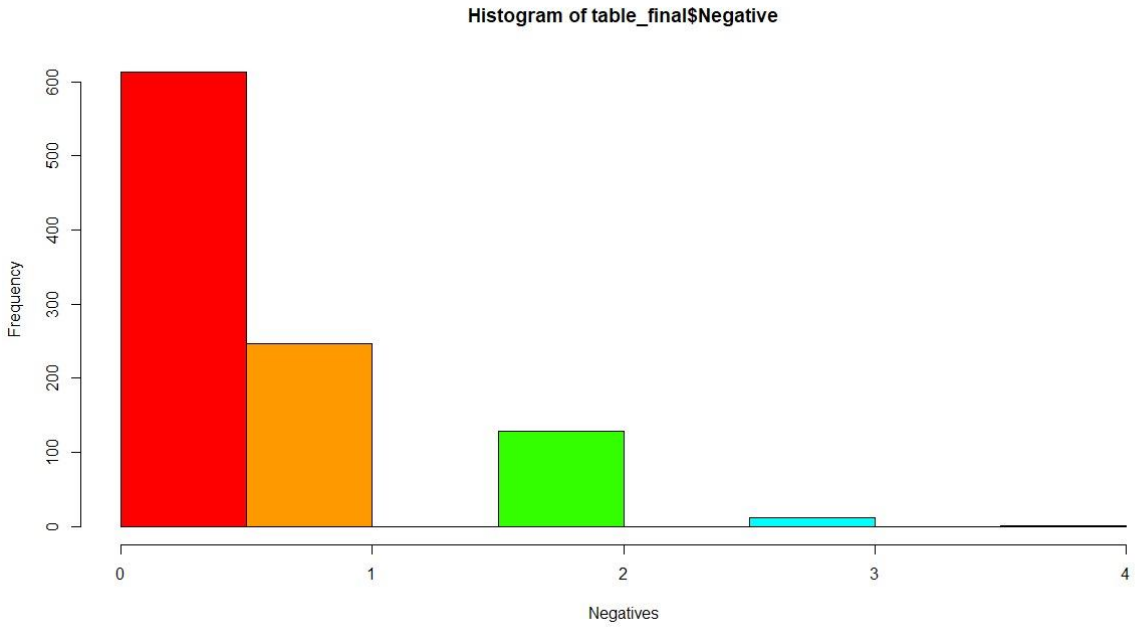

Fig. 8 Negative sentiments of Tweets

Figure 7 and 8 show that there is not much positive or negative on the extremes and that most of it is close to neutral sentiment.

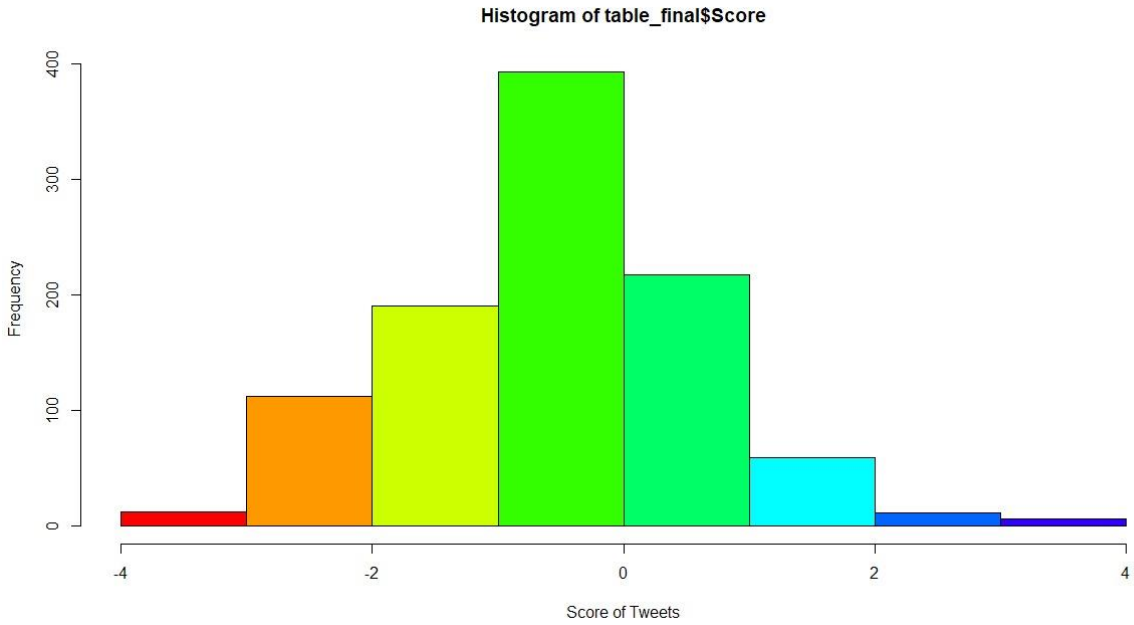

Fig. 9 Overall sentiment of Tweets

The overall sentiment is neutral with a slight skew to the negative side.

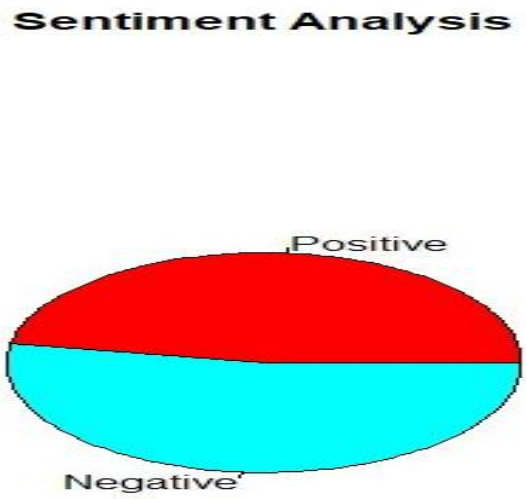

Fig. 10 Sentiment Analysis of Tweets

Figure 10 above shows us that the sentiment of the tweets regarding Nike are more negative than positive.

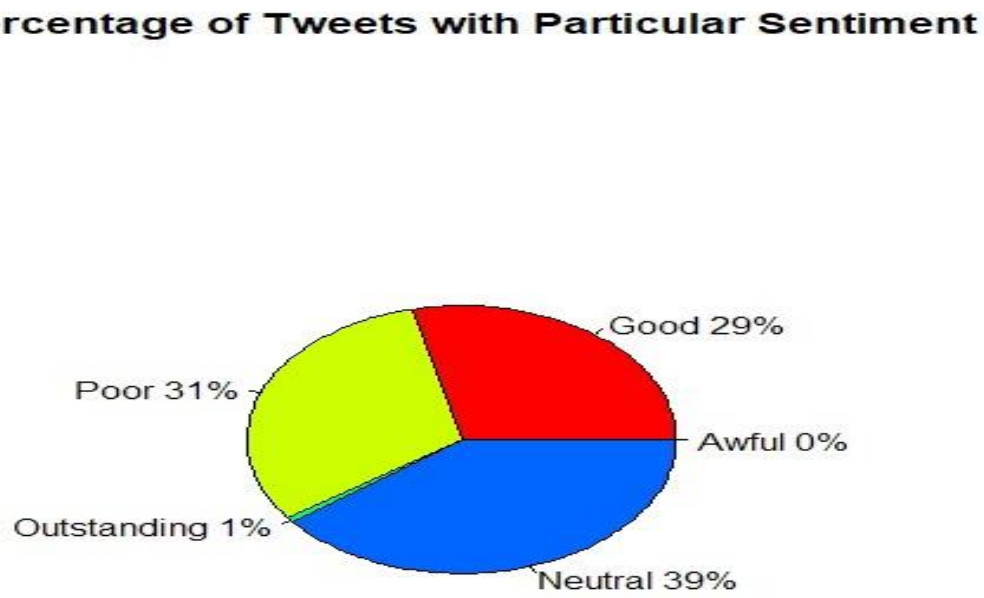

Fig. 11 Sentiment of Customer Tweets

Figure 11 above shows that the majority of the sentiment is neutral which is at 39%. The sentiment which is good or positive is 29%, while the poor sentiment is 31%.

The results in both figure 9 and 10 is understandable considering that Nike as a brand involves itself with social issues. An example would be the racism against African-Americans involving Colin Kaeperrnick (BBC Sport, 2019). It also sponsors a lot of athletes which have an affect on sentiment and perception towards Nike due to their activism and behaviour.

5. Conclusions and Implications

In conclusion, the use of social media mining and text mining can help companies improve their marketing strategies. The large amount of data provided by social media platforms allow for important insights to be gained. Nike should use its social channels to increase engagement with their brand and use the data produced to get insights.

Nike is a brand which has always involved itself in social issues due to its values. It also understands that taking certain social stances will lead to backlash and therefore negative sentiment. However they understand that taking these stances will have long term positive effects as a lot of the younger customer or potential customers are likely to share the same views (BBC News, 2019). Therefore they are more likely to gain more customers than they lose.

6. References

Amaresan, S. (2019). What Is Social Listening & Why Is It Important?. [online] Blog.hubspot.com. Available at: https://blog.hubspot.com/service/social-listening [Accessed 17 May 2019].

Barker, S. (2019). How Social Media is Influencing purchase decisions. [online] Social Media Week. Available at: https://socialmediaweek.org/blog/2017/05/social-media-influencing-purchase-decisions/ [Accessed 17 May 2019].

BBC Sport. (2019). Colin Kaepernick: Nike suffers #justburnit backlash over advertising campaign. [online] Available at: https://www.bbc.co.uk/sport/american-football/45407340 [Accessed 10 Jun. 2019].

BBC News. (2019). Nike sales defy Kaepernick backlash. [online] Available at: https://www.bbc.co.uk/news/business-45472399 [Accessed 10 Jun. 2019].

Ebner, T. (2019). What is a Social Media Firestorm? A clear Checklist and Definition - social media #facts. [online] social media #facts. Available at: http://www.socialmediafacts.net/firestorms/firestorm-definition [Accessed 17 May 2019].

Fray, B., Keller, E., Larkin, R. and Pauwels, K. (2019). Deriving Value From Conversations About Your Brand. [online] MIT Sloan Management Review. Available at: https://sloanreview.mit.edu/article/deriving-value-from-conversations-about-your-brand/ [Accessed 17 May 2019].

Kolkur, S., Dantal, G. and Mahe, R. (2019). [online] Inpressco.com. Available at: http://inpressco.com/wp-content/uploads/2015/03/Paper32768-770.pdf [Accessed 17 May 2019].

Macale, S. (2019). Twitter users are more likely to impact your brand than any other social network. [online] The Next Web. Available at: https://thenextweb.com/twitter/2011/08/18/twitter-users-are-more-likely-to-impact-your-brand-than-any-other-social-network/ [Accessed 17 May 2019].

MonkeyLearn. (2019). Sentiment Analysis: Nearly Everything You Need to Know | MonkeyLearn. [online] Available at: https://monkeylearn.com/sentiment-analysis/ [Accessed 17 May 2019].

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. and Mozetič, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. PLOS ONE, 10(9), p.e0138441.

Sarawgi, K. and Pathak, V. (2019). [online] Pdfs.semanticscholar.org. Available at: https://pdfs.semanticscholar.org/b8d6/dee8923bb1344bb098979c3648676d879326.pdf [Accessed 17 May 2019].

Scott, D. (2015). The new rules of marketing and PR. 5th ed. Hoboken, N.J.: Wiley.

Shobana, G., Vigneshwara, B. and Maniraj Sai, A. (2019). [online] Ijrte.org. Available at: https://www.ijrte.org/wp-content/uploads/papers/v7i4s/E1989017519.pdf [Accessed 17 May 2019].

Socialbakers.com. (2019). Nike Statistics on Twitter followers. [online] Available at: https://www.socialbakers.com/statistics/twitter/profiles/detail/415859364-nike [Accessed 9 Jun. 2019].

W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.

www.statista.com. (2019). Topic: Social Media Statistics. [online] Available at: https://www.statista.com/topics/1164/social-networks/ [Accessed 17 May 2019].

7. Appendices

7.1 Appendix 1. Women

7.2 Appendix 2. R Syntax ……

```
rm(list=ls()) # clears all objects in "global environment"
cat("\014") # clears the console area
getwd()
##########################################################################
###

### install packages and load Libraries
pkgs <- c("twitteR", "ROAuth", "httr", "tidyverse", "quanteda", "tm","wordcloud",

"openNLP","openNLPdata","tidytext","dplyr","ggplot2","reshape","plotrix","stringr",
"plyr")
library(pacman)
pacman::p_load(char=pkgs,install=TRUE,character.only=TRUE)
## Installing OpenNLP package from local folder as it is not available for R 3.5.2
# install.packages("openNLPmodels.en_1.5-1.tar.gz", repos = NULL, type = "source")
library(openNLPmodels.en)
##########################################################################
###

### twitterConnection

consumer_key <- 'yBZ9gWqoSGzC7dwhTuPquPtNi'
consumer_secret <-
'1V9v6DJcauhDm32xQ4dZD6e0Y6CHVWO7SN48kO5bnmXnm6a9VC'
access_token <- '2806191819-Ej4Kpj1evGWD4IugbG3B7acqx9xcprHS8nxfvaf'
access_secret <- 'ikVFCYuMdj2W3d2yJOKQzMNp9FSxjXcMIB2PgfjkqP9SK'

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

##########################################################################
###

### searchTweet
## Below code is commented as the dataframe is made static post one extraction
and used as is. Dataset is shared.

# fn_twitter <- searchTwitter("@Nike",n=2500,lang="en")
```

```
# fn_twitter_df <- twListToDF(fn_twitter) # Convert to data frame
# fn_twitter_df_1000 <- head(fn_twitter_df, n = 1000)
# write.csv(fn_twitter_df_1000, 'Raw_Data.csv')

##########################################################################
###

### import_Select
nike <- read.csv("Raw_Data.csv", stringsAsFactors = FALSE) #read the CSV file

nike_eText <- as.vector(nike[,2]) # get only the relevant text i.e second column
nike_eText <- sapply(nike_eText,function(row) iconv(row, "latin1", "ASCII", sub="")) #
convert the corpora to vector

##########################################################################
###

### preprocessing
##Data preprocessing: POSIX characters Handling
nike_eText <- gsub("<.*>","",nike_eText) #1. get rid hex characters
nike_eText <- gsub("https.*","",nike_eText) #2. get rid of urls
nike_eText <- gsub("[\n\r]","",nike_eText) #3. get rid of new line and carriage return
nike_eText <- gsub("@[a-z,A-Z]*","",nike_eText) #4. get rid twitter names
nike_eText <- gsub("(f|ht)tp(s?)://(.*)[.][a-z]+", "", nike_eText) # get rid of other odd
chars

##########################################################################
###

##consolidate the original text file the text into the single corpora
text <- paste(nike_eText,collapse = "")
## get the data into the Volatile corpus
corpus_Nike <- tm::VCorpus(VectorSource(nike_eText))

##########################################################################
###

### textMining
```

```
corpus_L <- tm::tm_map(corpus_Nike, tm::content_transformer(tolower)) #1. convert the text to the lower case
corpus_Punc <- tm::tm_map(corpus_L, removePunctuation) #2. remove punc with tm package
corpus_Numbers <- tm::tm_map(corpus_Punc, tm::removeNumbers) #3. remving numbers from the text
stop <- c(stopwords('en'),'will','nike','shoe') # Creating a vector of common English stopwords and project related stop words
corpus_Stopwords <- tm::tm_map(corpus_Numbers, tm::removeWords, stop) #4 removing the stop words , command order is a key here.
corpus_RemoveSpace <- tm::tm_map(corpus_Stopwords, tm::stripWhitespace) #5 remove the space between words

################################################################################

### analysis
## starting the analyses the corpus
dtm <- tm::DocumentTermMatrix(corpus_RemoveSpace) #6 DTM counts the frequency of each word (min 3 characters)
tm::inspect(dtm) #examine the frequency of the words used in the document and disperseness

thewords_used <- dtm$dimnames$Terms #7 dtm: terms are here already in the system as dtm is used
wordsIndoc <- thewords_used[dtm$j] #8 two documents index numerator

selectDoc <- (dtm$i) #9 determine whether the word belongs to doc 1 here or 2
wordsInSelectedDoc <- wordsIndoc[selectDoc] #10 name and enumerate the words from the 1 st doc
# print(wordsInSelectedDoc) # print the results

freqConcepts <- dtm$v[selectDoc]  #11 determine the frequency of words in doc 1
lookInside <- data.frame(Term = wordsInSelectedDoc,freqConcepts = freqConcepts) #12 133 put the previous result into the table view
# print(lookInside)

################################################################################
```

```
### Stemmisation
corpus.stem <- tm::tm_map(corpus_RemoveSpace, tm::stemDocument, lang =
"English") #13 get the stems of the words
# print(corpus.stem[[24]]$content) #Testing for random tweet24

dtm <- tm::DocumentTermMatrix(corpus.stem) # get the frequency of themmed doc
twit 24 as an example

#  selectDoc <- (dtm$i)
# print(data.frame(Term = dtm$dimnames$Terms[dtm$j[selectDoc]],Freq =
dtm$v[selectDoc]) ) #print the results as mentioned in 12
# ### dimentionality has been reduced from 23 words to 20

freq <- colSums(as.matrix(dtm)) # get the freq of words for the entire doc i.e. all
tweets
# print(freq)

ordered <- freq[order(freq,decreasing = TRUE)] # get the freq of 10 most popular
words
print(head(ordered,n = 10))

### plot the results  xlab-ais names...  las-words in x axis, csx.names-fot size, las 2
vertical y axis, labeled x axis
# wordcloud(names(freq),freq,min.freq = 1) #word cloud with min freq of 1 word
used
set.seed(423)
suppressWarnings(wordcloud(names(freq), freq, min.freq = 3,
        max.words = 200,random.order = FALSE, rot.per = 0.35,
        colors = brewer.pal(8, "Dark2")))

# Plot word frequencies
barplot(ordered,cex.names = 1.0, las = 1,
      main = "Most Frequent Words",
      ylab = "Frequency",xlab = "Terms")

## Explore frequent terms and their associations
# Printing terms appearing at least 30 times in the selected tweets
print(tm::findFreqTerms(dtm,lowfreq =30))
```

```
##########################################################################
###

###posneg: Finding out the Positive and Negative words in the tweets
## Impoting positive and negative word sets by Prof. Minqing Hu and Bing Liu
pos.words <- scan('positive-words.txt', what='character', comment.char=';')
neg.words <- scan('negative-words.txt', what='character', comment.char=';')

#Adding project specific positive words to the word sets by Prof. Minqing Hu and
Bing Liu
pos.words<-c(pos.words, 'good', 'best', 'love', 'loved', 'thnx', 'Grt',
          'gr8', 'thank','thanks', 'trendy', 'awesome', 'nice',
'light','lightweight','nyc1','wonderful','comfortable','comfy','cool')
neg.words <- c(neg.words, 'shit', 'shitty', 'heavy', 'damn', 'no', 'not','bleh','boo')

##########################################################################
###

##sentimentAnalysis
scSentiment <- function(sentences, pos.words, neg.words, .progress='none') #tweets
prameterisd as a sentence
 {
 list<-lapply(sentences, function(sentence, pos.words, neg.words)
 {#Regular expressions to ensure that the received corpora is clear
  ##useful if Sentiment Analysis is run separately.
  sentence <- gsub('[[:punct:]]',' ',sentence)
  sentence <- gsub('[[:cntrl:]]','',sentence)
  sentence <- gsub('\\d+','',sentence)
  sentence <- gsub('\n','',sentence)

  sentence <- tolower(sentence) #pre processing for safety
  #bringing all tweets to a single list
  list_words <- str_split(sentence, '\\s+') # generating a word 'list' from the tweet
  #changing to vector
  unlist_word <- unlist(list_words) # unlisting the list for match action
  #Matching words with the positive and negative lists, which generates a binary
result
  pmatch <- match(unlist_word, pos.words)
  nmatch <- match(unlist_word, neg.words)
```

```
  # getting rid of the 0s (or non-matches in both variables)
  pmatch <- !is.na(pmatch)
  nmatch <- !is.na(nmatch)
  # Count of postive and negative words in a tweet
  pp<-sum(pmatch)
  nn <- sum(nmatch)
  score <- sum(pmatch) - sum(nmatch) # Score of a tweet = (No. of Positive Words -
No. of Negative Words)
  #Storing all three params in a list and returning from function
  list1<-c(score, pp, nn)
  return (list1)
 }, pos.words, neg.words)
 # Attaching the list elements to separate variables, and making separate dataframes
of these
 score_new<-lapply(list, `[[`, 1)
 pp1=score=lapply(list, `[[`, 2)
 nn1=score=lapply(list, `[[`, 3)
 #Generating separate DFs for the threeech sentiment
 scores.df <- data.frame(score=score_new, text=sentences)
 positive.df <- data.frame(Positive=pp1, text=sentences)
 negative.df <- data.frame(Negative=nn1, text=sentences)

 # Returning all dataframes from the function call
 list_df<-list(scores.df, positive.df, negative.df)
 return(list_df)
}

result <- scSentiment(nike_eText, pos.words, neg.words)

##########################################################################
###

#Creating three different data frames for Score, Positive and Negative
#Removing text column from data frame
test1<-result[[1]]
test1$text<-NULL
test2<-result[[2]]
test2$text<-NULL
test3<-result[[3]]
```

```
test3$text<-NULL

# Taking the sentiment scores in variable sc
ss1<-test1[1,]
ss2<-test2[1,]
ss3<-test3[1,]   #q1---ss1.....qq1--ssc qq2--ssp...qq3--ssn
ssc<-melt(ss1, var='Score')
ssp<-melt(ss2, var='Positive')
ssn<-melt(ss3, var='Negative')
ssc['Score'] <- NULL # the score
ssp['Positive'] <- NULL # the postive sentment
ssn['Negative'] <- NULL # the negative sentiment
# For Visualisation taking it into a data frame (with the scores)
table1 <- data.frame(Text=result[[1]]$text, Score=ssc)
table2 <- data.frame(Text=result[[2]]$text, Score=ssp)
table3 <- data.frame(Text=result[[3]]$text, Score=ssn)

#Merging all the three tabes into 1
combined<-data.frame(Text=table1$Text, Score=table1$value,
Positive=table2$value, Negative=table3$value)

#Histogram showing the Positive and Negative words in the Tweets, and the overall
score of Tweets on the Pos-Neg Scale
hist(combined$Positive, col=rainbow(10), xlab = 'Positives')
hist(combined$Negative, col=rainbow(10), xlab = 'Negatives')
hist(combined$Score, col=rainbow(10), xlab = 'Score of Tweets')

#Pie Chart of Positives-Negatives in the Tweets
pie <- c(sum(combined$Positive), sum(combined$Negative))
label <- c("Positive", "Negative")
suppressWarnings(pie(pie, labels = label, col=rainbow(length(label)),explode=0.00,
main="Sentiment Analysis"))

##########################################################################
###

#Positive Percentage

#Taking the sentiment out in separate variable
```

```
posSc<-combined$Positive #with +ive
negSc<-combined$Negative #with -ive

# Calc +ive %age
combined$PosPercent <- posSc/ (posSc+negSc)

# Removing Non-Numbers
pp <- combined$PosPercent
pp[is.nan(pp)] <- 0
combined$PosPercent <- pp

#Negative Percentage

# Calc -ive %age
combined$NegPercent <- negSc/ (posSc+negSc)

# Removing Non-Numbers
nn <- combined$NegPercent
nn[is.nan(nn)] <- 0
combined$NegPercent <- nn

##Finding out the scores for each level of Sentiment

#Good

Sc <- combined$Score

#Output of following is FALSE or TRUE
good <- sapply(Sc, function(Sc) Sc <= 3 && Sc > 0)
#Converts to actual value
# Sc[good]
list_good <- Sc[good]
value_good <- length(list_good)

#Very good

vgood <- sapply(Sc, function(Sc) Sc > 3)
#Converts to actual value
# Sc[vgood]
```

```
list_vgood <- Sc[vgood]
value_vgood <- length(list_vgood)

#Bad : Unsatisfactory

#Output of following is FALSE or TRUE
bad <- sapply(Sc, function(Sc) Sc >= -3 && Sc < 0)
#Converts to actual value
# Sc[bad]
list_bad <- Sc[bad]
value_bad <- length(list_bad)

#Very bad : Poor

#Output of following is FALSE or TRUE
vbad <- sapply(Sc, function(Sc) Sc < -3)
#Converts to actual value
# Sc[vbad]
list_vbad <- Sc[vbad]
value_vbad <- length(list_vbad)

#Neutral
neutral <- sapply(Sc, function(Sc) Sc == 0)
list_neutral <- Sc[neutral]
value_neutral <- length(list_neutral)

slices1 <- c(value_good, value_bad , value_vgood , value_neutral , value_vbad )
lbls1 <- c("Good", "Poor", "Outstanding", "Neutral", "Awful")
pct <- round(slices1/sum(slices1)*100) #Percentage
lbls1 <- paste(lbls1, pct) # display pecent
lbls1 <- paste(lbls1,"%",sep="") # display pecent
pie(slices1,labels = lbls1, col=rainbow(length(lbls1)),
   main="Percentage of Tweets with Particular Sentiment")
```