

## The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R

Ahmed Imran KABIR<sup>1</sup>, Ridoan KARIM<sup>2</sup>, Shah NEWAZ<sup>3</sup>, Muhammad Istiaque HOSSAIN<sup>4</sup>

<sup>1</sup> Faculty of Business and Accountancy, University of Malaya, Malaysia.

<sup>2</sup> Faculty of Law, University of Malaya, Malaysia.

<sup>3</sup> Faculty of Business and Accountancy, University of Malaya, Malaysia.

<sup>4</sup> School of Business and Law, Central Queensland University, Australia.

iamahmedimran@gmail.com, ridoankarim@siswa.um.edu.my, saminewaz@gmail.com,  
iam.mistiaque@gmail.com

*Apparently, word clouds have grown as a clear and appealing illustration or visualization strategy in terms of text. Word clouds are used as a part of various settings as a way to give a diagram by cleansing text throughout those words that come up with most frequently. Generally, this is performed constantly as an unadulterated text outline. In any case, that there is a bigger capability to this basic yet intense visualization worldview in text analytics. In this work, we investigate the adequacy of word clouds for general text analysis errands and also analyze the tweets to find out the sentiment and also discuss the legal aspects of text mining. We used R software to pull twitter data which depends altogether on word cloud as a visualization technique and also with the help of positive and negative words to determine the user sentiment. We indicate how this approach can be viably used to explain text analysis tasks and assess it in a qualitative user research.*

**Keywords:** Big data, Text Analytics, Social Media Analytics, R, Sentiment analysis, Word Cloud, Twitter Analysis

### 1 Introduction

Microblogging or Social media sites have developed to wind up plainly a source of fluctuated sort of data. This is because of nature of social media on which individuals post constant messages about their opinion on an assortment of points, examine current issues, grumble, and express positive feeling for items they use in day by day life. Actually, organizations assembling or manufacturing such items have begun to survey these microblogs or social media to get a perception of general slant for sentiment for their items or product. Commonly these companies think about client responses and answer to clients on social media. One test is to develop technology to distinguish and outline a general sentiment.

While there has been a considerable amount of research on how assumptions are communicated in genres, for example, online reviews, blogs and news articles, how feelings or sentiments are communicated given the casual language and message-length requirements of microblogging or social networking

has been substantially less studied. Highlights, for example, programmed grammatical form labels and resources, for example, sentiment vocabularies have demonstrated valuable for sentiment examination or analysis in different areas, yet will they likewise demonstrate helpful for sentiment analysis in twitter? In this paper, we start to explore this question. Word clouds produced for a collection of text can fill in as a beginning stage for a more profound analysis [1-3]. For example, they help to decide whether a given text is applicable to a particular data require. One of their downsides is that they give a simply factual rundown of disengaged words without considering phonetic information about the words and their relations. Subsequently, word clouds are utilized rather statically as a way to outline message in many frameworks and they regularly give no or just restricted collaboration abilities. We think there is a bigger potential to this straightforward yet capable visualization worldview in numerous analyzing con-

texts. In this work, we, along these lines, investigate the potential outcomes by utilizing word clouds at the exceptionally focal point of text mining.

In this paper, we analyze one such prevalent microblog or social media called twitter and build R models for characterizing "tweets" into positive, negative and unbiased sentiment and also create word cloud to find out the most frequently used term. For twitter sentiment, we assemble models for twitter authentication, and then we will pull the data from twitter. Here we will use a political figure to analyze sentiment what type of words are being used by him in everyday life to figure out actually what is happening in his mind. By using the R models, we will basically create a graph of positive, negative and neutral words used by the twitter user.

To generate word cloud, we will first use R model to authenticate twitter. Then we will pull twitter data of a famous phone company. Then we will process the twitter data in a way that we can create a word cloud based on the dataset. The finalized word cloud will picture what the company is actually thinking. Meaning which words are being used frequently on this particular twitter account.

## 2 Managerial Contribution

In general, social media analytics are used for future forecasting for better business decision making and planning. In this study, we obtained and modified numerous source codes from different sources and run the algorithms by rendering a new technique to generate twitter sentiment and word cloud for further forecasting on political context. Although some of the results are already prevailing in scattered manner – we tried to put those in place – to illustrate the whole process of text mining for business analytics professionals or researchers.

## 3 Literature Review

Sentiment analysis is a developing area of Natural Language Processing with research extending from document level characterization [4] to taking in the extremity of words and

phrases [5, 6]. Given the character impediments on Tweets, characterizing the sentiment of twitter messages is most like sentence-level sentiment analysis [7, 8]. Be that as it may, the casual and specific dialect utilized as a part of Tweets, and in addition the very idea of the microblogging or social networking domain make twitter sentiment analysis is altogether a different assignment. How well the features and strategies utilized on more structured information will transfer to the microblogging or social media area. Just in the previous year there have been various papers taking a gander at twitter sentiment [9-12]. Different data scientists and researchers have started to investigate the utilization of grammatical features but the results are still mixed. Features regular to microblogging or social networking are likewise normal, yet there has been little examination concerning the handiness of existing sentiment resources created on non-microblogging information. Specialists have likewise started to examine different methods for naturally gathering training information. A few analysts depend on emojis for characterizing their training information [10]. Barbosa and Feng [13] misuse existing twitter sentiment sites for gathering training information or data. [12] likewise utilize hashtags for making training information, however they restrain their examinations to sentiment/non-sentiment grouping.

Being "Born outside the universe of PCs" [2], word clouds ended up plainly well known in the specific community oriented websites, for example, photo sharing website Flickr, advertising firm Technorati or social bookmarking website Delicious, that utilize tagging as an ordering strategy [14]. In the meantime, they have developed as a core system of data representation that is applied in a wide range of contexts.

One famous application range for tag clouds or word cloud is text outline [15-17]. Here, word cloud is utilized to give a natural and outwardly engaging diagram of a content by delineating the words that used regularly within it. Such a synopsis is useful to find out about the number and sort of topic introduce in an assortment of text. This statistical outline

is accomplished by decidedly corresponding the font size of the delineated words with the word recurrence. At the point when a word cloud representation is utilized along these lines, the words in the word cloud are words from a content. Therefore, the term word cloud is regularly favored over the term label cloud in these specific circumstances.

Although many assessments of text analyzing in social media written by the data scientists are highly persuasive, however, they often neglect the importance of text/data analyzing of social media as a tool to understand social and political sentiment of any given region. On the other hand, books, thesis and reports written by social and political scientists only focus purely on social and political assessment of social media and always ignore the scientific aspects of using text analysis to understand the social behavior of the people. As this research attempts to provide a brief scientific assessment of the social media of Bangladesh by using text analysis methods on the political aspects in Bangladesh and avoid unsubstantiated assertions, such study undoubtedly stands as a major reference work on the subject.

#### **4 Why Social Media Analytics or Text Mining in Essential**

In the beginning of social media or microblogging, public relation organizations would screen clients' posts on business websites trying to distinguish and oversee displeased clients. With the expansion of social media or microblogging sites and the quantity of use on them, this isn't sufficient. Consider the commonness of web-based social networking we can come up to these points: Social media or microblogging seems very simple but it is rich with opportunity. Number of social media users is 2.46bn worldwide in 2017 [18]. Also, Facebook, YouTube, and twitter are most-trafficked sites on the Internet.

In any case, even these insights fail to give a full record of the impact that social networking or microblogging sites are having. Users spend over 135 minutes per day online [19] through social media or microblogging sites. Facebook alone has an overall market infiltration rate 26.3% [20]; in North America it is

72.4% [20]. These rates are developing rapidly, with Facebook has 2.07bn users which was only 1.59bn in the end of 2015 [21].

YouTube's mining of its videos demonstrates 100 million individuals like, dislike, comment or share those videos every week [22]. Within two years this figure doubled. Facebook now incorporates social activities in its online promotions, for example by enabling user to investigate whether their partners have favored or chosen on items being advertised while they were watching YouTube videos. Essentially, hashtags on twitter have given clients another speedy approach to express their likes, dislikes or comments; and these offers opportunities for companies to study about their sentiments.

#### **5 Legal and Ethical Aspects of Social Media Mining**

In today's global marketplace, the social media data become one of the rich sources for companies to understand the market value and customer diversity [23]. In addition, political parties competing for state elections, may also find such text analyzing of social media very useful to understand the opinion of mass people and to make new political strategies for the elections. However, there are certain legal concerns in relation to the data mining from the social media.

There are some major legal issues facing by people when they plan to undertake the text data mining which are protected and governed by the database law, copyright law and contract law [24]. The copyright law gives protection automatically over the work which one not copied from elsewhere by recording and creating in a way it establishes a new inventory work [24]. The copyright protection somehow provides the owner the right whether to approve or decline on third party's' restricts acts which includes such as adaptation of the work, copying, re-dissemination for publishing it on the Web, or any substantial part of the original copyright work, as well as translation of the work into other languages.

The copyright law is more complicated on the practice. A simple explanation of facts that includes more data or information is not subject to the copyright issues, however, to identify on the classification short sentence or a fact that still remains unclear and controversial for the court to decide to resolve the matters. There is one recent European Court case confirmed that a sentence of 11 words was justified to be protected by the copyright even though this does not clearly imply that all similar sentences will be protected by copyright [25].

The copying act alleged to infringe the Intellectual Property Rights belongs to the original owners, where it will indicate to some legal ambiguity on the process of text data mining. Legal aspects are of course country dependent. However, it should be noted that, posting something publically and accessing to the public information on the Web is perfectly legal [26]. When we reading a webpage it is similar like we reading a billboard or a street sign, therefore assessing the information that shared or broadcast to the entire world consist to have no legal restrictions. However, there are some restriction where the law provided on the access to unethical webpage likes the pornographic material by minors thus the service provider has to check the consumer age and enforcement of the law so that the viewer is not doing anything illegal by merely viewing a public Web page.

Now the question remains open whether the text posted in social media stipulates the creation of any particular work or not. In most of the social media, people only tend to share the simple facts, which by its nature, can never considered as a copyright issues in legal arguments. Therefore, copyright issues relating to social media needs serious concern from Judges, Lawyers and Legal academicians.

It is vital to note that, it is not a problem on accessing by viewing the freely and publicly posted data or information that we needed via certain webpage and use that data of the information by saving it or share it to others and there is no difference if the data or information is accessed interactively with a browser by

downloading it automatically with others software. However, the condition will be different on the data or information that is not freely or publicly posted, but rather than it is being protected not for free viewing on the access on the webpage example like to have the password-protected account is required to access or to use the data or information on a Web page. In most of the scenario, the web page provider will impose on some conditions and restrictions for the user such as non-disclosure to third parties, payment of a fee, or any other Terms of Use or Terms of Service (ToS). This is more alike an agreement that exists between web page provider and user of the information or service. Some have justified that if ToS forbids some act it does not mean that it will be illegal. However, there are other laws binds with Tos, where it can be a legally make that agreement to be binding, and therefore if you fail to follow as provided by the ToS, then you will be violating the law that exists on the contractual arrangements, then give rise to contract law liabilities [27].

Even if it is lawful to gather information using social network site for the purposes of research, there may be ethical considerations that to make it inappropriate on the issue of privacy of personal data. There is information privacy policy which are protected by data protection laws that regulate the use of personal data information. However, a user who accesses to the personal data or information are protected not limited for one use of such information, except as specified by the ToS. However, depending on what is considered as ethical behavior, it may have different views about the gravity of breaching a ToS agreement, and the magnitude of violating a third party's right to privacy. It should typically consider such violation of privacy as unethical even it is not illegal.

Research on social network data has significant ethical consequences concerning the protection of the human subjects [28]. It has been argued and recommended that the research on social media is to be reviewed by an independent review panel to make sure that it is protected from any discrimination, abuse, risk,

privacy violations, and other potentially adverse factors. If there is any study conducted on online social network, it should be ensuring that no any personal identifiable information is to be publish however only to the information required which is about aggregate statistical analysis.

**6 The Process of Social Media Mining**

Social media or microblog mining includes a three-organize process: Gather, Analyze, and Visualize. “Gather” includes getting applicable web-based social media or microblog data

by observing different social media or microblog sources, documenting significant data and delivering relevant data. This procedure should be achievable by an institution itself or a data analyst. Not all information that are caught will be valuable. “Analyze” chooses important information for demonstrating, expels noise, low-quality information, and utilizes different propelled data analytics techniques to investigate the information restore and pick up bits of knowledge from it. “Visualize” manages showing discoveries from “Analyze” in every significant way

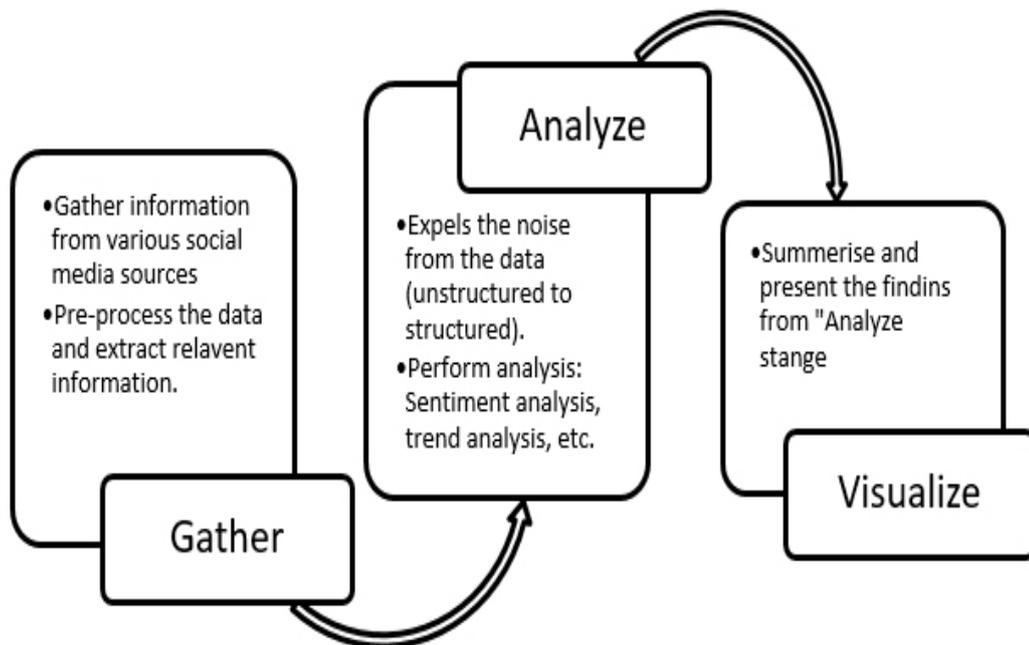


Fig. 1. 3 stages of social media mining or analytics.

For a business occupied with social network or microblog analytics, “Gather” enables it to recognize discussions on social media or microblog sites identified with its interest and activities. This is finished by gathering huge measures of applicable information crosswise over hundreds or thousands of social networking sources utilizing news feeds or APIs. “Gather” stage covers mainstream platforms, for example, Foursquare, LinkedIn, Facebook, Twitter, Instagram, YouTube, Tumblr, Pinterest, Google+, and so forth. To set up a data set for the “Analyze”, different pre-handling steps might be performed, including data modeling, record, linking from other

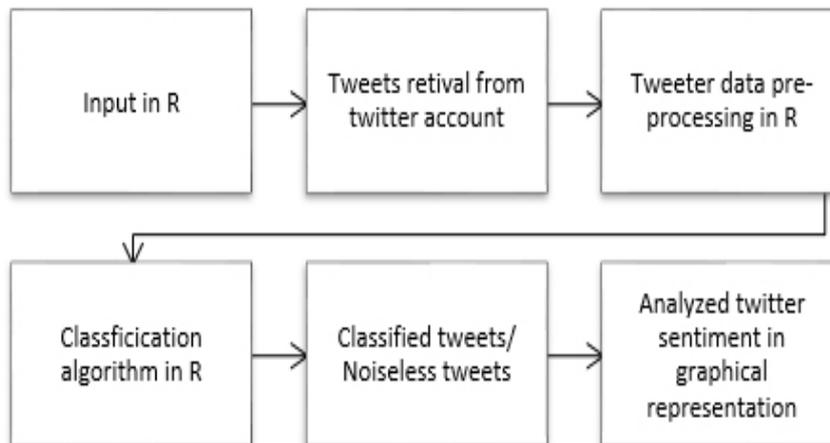
source, part of speech tagging or other syntactic and semantic techniques that help analysis. Once a firm has accumulated the comments and posts identified with its products and operations, it should next study their influence and create metrics necessary for decision making. This is the “Analyze” stage. Since the “Gather” stage accumulates information from numerous users and sources of social media or microblog, a sizeable bit might be noisy and may should be expelled before playing out any meaningful analysis. Straightforward, run based text classifiers or more modern classifiers prepared on marked noisy information might be utilized for this cleaning capacity. Assessing the importance of the cleaned data

can include various statistical methods and different methods got from text mining, machine learning, or and system inquiry. [29]. This stage gives data about users' sentiment. Numerous valuable measurements and patterns about users can be delivered in this stage, interests, covering their experiences concerns, etc.

### 7 Sentiment analysis using R

For twitter sentiment analysis, we used twitter account of a famous Bangladeshi businessman and politician. He serves as an Information and Communication Technology advisor to the Bangladesh Government and he is

also a member of Bangladesh Awami League. He is an influential person in Bangladesh. To analyze the twitter sentiment, we have to get twitter authentication from twitter developer website ([www.apps.twitter.com](http://www.apps.twitter.com)). Then, using R packages we pulled **700** tweets of Sajeeb Wazed and analyzed the positive, neutral and negative sentiments he used in his official twitter account. We collected the positive [30] and negative words [31] datasets from GitHub. According to process of social media mining, twitter sentiment also has gather, analyze and visualize stage.



**Fig. 2:** Stages of sentiment analysis in R

### 8 Model code on R for twitter sentiment analysis

In order to run the sentiment analysis with R, the source codes are obtained from two main sources R bloggers, GitHub [32, 33].

Both are focused on Twitter sentiment analysis using R. But to satisfy the analysis we modified and run the codes using R to get the sentiment analysis of the twitter user "sajeebwazed".

```

# debugging setting
debug <- TRUE

# install required packages
> if(!require('twitterR')) install.packages('twitterR')
> if(!require('ROAuth')) install.packages('ROAuth')
> if(!require('plyr')) install.packages('plyr')
> if(!require('dplyr')) install.packages('dplyr')
> if(!require('stringr')) install.packages('stringr')
> if(!require('ggplot2')) install.packages('ggplot2')
> if(!require('tm')) install.packages('tm')

# connect all libraries
> library(tm)
> library(twitterR)
> library(ROAuth)
> library(plyr)
> library(dplyr)
  
```

```

> library(stringr)
> library(ggplot2)

# Twitter developer account authentication
> api_key <- "XXXXXX"
> api_secret <- "XXXXXXXXXX"
> access_token <- "XXXXXXXXXXXX"
> access_token_secret <- "XXXXXXXXXXXX"
> setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)

# the function of tweets accessing and analyzing
> search <- ("sajeebwazed")

# access tweets and create cumulative file
> if(isTRUE(debug)) print("Fetching tweets...")
> list <- searchTwitter("sajeebwazed", n=700)
> if(isTRUE(debug)) print("Done.")

# creating dataframe
> if(isTRUE(debug)) print("Creating stack file...")
> df <- twListToDF(list)
> df <- df[, order(names(df))]
> df$created <- strftime(df$created, '%Y-%m-%d')
> if (file.exists(paste("sajeebwazed", '_stack.csv'))==FALSE) write.csv(df,
file=paste("sajeebwazed", '_stack.csv'), row.names=F)
> if(isTRUE(debug)) print("Done.")

# merge last access with cumulative file and remove duplicates
> if(isTRUE(debug)) print("Merging stack file...")
> stack <- read.csv(file=paste("sajeebwazed", '_stack.csv'))
> stack <- rbind(stack, df)
> stack <- subset(stack, !duplicated(stack$text))
> write.csv(stack, file=paste("sajeebwazed", '_stack.csv'), row.names=F)
> if(isTRUE(debug)) print("Done.")

# evaluation tweets function
> score.sentiment <- function(sentences, pos.words, neg.words, .progress='none')
+ {
+   require(plyr)
+   require(stringr)
+   scores <- laply(sentences, function(sentence, pos.words, neg.words){
+     sentence <- gsub('[:punct:]', "", sentence)
+     sentence <- gsub('[:cntrl:]', "", sentence)
+     sentence <- gsub('\\d+', "", sentence)
+     sentence <- tolower(sentence)
+     word.list <- str_split(sentence, '\\s+')
+     words <- unlist(word.list)
+     pos.matches <- match(words, pos.words)
+     neg.matches <- match(words, neg.words)
+     pos.matches <- !is.na(pos.matches)
+     neg.matches <- !is.na(neg.matches)
+     score <- sum(pos.matches) - sum(neg.matches)
+     return(score)
+   }, pos.words, neg.words, .progress=.progress)
+   scores.df <- data.frame(score=scores, text=sentences)
+   return(scores.df)
+ }

# creating scores file
> if(isTRUE(debug)) print("Creating scores file...")
> pos <- scan('positive-words.txt', what='character', comment.char=';') # folder with
positive dictionary
> neg <- scan('negative-words.txt', what='character', comment.char=';') # folder with
negative dictionary
> pos.words <- c(pos, 'upgrade')
> neg.words <- c(neg, 'wtf', 'wait', 'waiting', 'epicfail')
> Dataset <- stack

```

```

> Dataset$text <- as.factor(Dataset$text)
> scores <- score.sentiment(Dataset$text, pos.words, neg.words, .progress='text')
> write.csv(scores, file=paste("sajeebwazed", '_scores.csv'), row.names=TRUE) # save
evaluation results into the file
> if(isTRUE(debug)) print("Creating stack file...")

# total evaluation: positive / negative / neutral
> if(isTRUE(debug)) print("Performing Sentiment Analysis...")
> stat <- scores
> stat$created <- stack$created
> stat$created <- as.Date(stat$created)
> stat <- mutate(stat, tweet=ifelse(stat$score > 0, 'positive', ifelse(stat$score <
0, 'negative', 'neutral')))
> by.tweet <- group_by(stat, tweet, created)
> by.tweet <- summarise(by.tweet, number=n())
> write.csv(by.tweet, file=paste("sajeebwazed", '_opin.csv'), row.names=TRUE)
> if(isTRUE(debug)) print("Done.")

# create chart
> if(isTRUE(debug)) print("Creating chart/plot...")
> ggplot(by.tweet, aes(created, number)) + geom_line(aes(group=tweet, color=tweet),
size=2) +
+   geom_point(aes(group=tweet, color=tweet), size=4) +
+   theme(text = element_text(size=18), axis.text.x = element_text(angle=90,
vjust=1)) +
+   # stat_summary(fun.y = 'sum', fun.ymin='sum', fun.ymax='sum', colour = 'yellow',
size=2, geom = 'line') +
+   ggtitle("sajeebwazed")
> ggsave(file=paste("sajeebwazed", '_plot.jpeg'))
> if(isTRUE(debug)) print("Done.")

# END OF PROGRAM

```

## 9 Word cloud analysis using R

For generating word cloud using R, we used twitter account of famous Bangladeshi political and ruling party Awami League. The twitter account id of Bangladesh Awami League is “albd1971” which we used to get tweets. To analyze the twitter word cloud analysis to find out the frequently used words by a twitter user, we have to get twitter authentication from twitter like we have used in twitter sentiment analysis. Then using R packages, we pulled 200 tweets of Bangladesh Awami League party and we

got unstructured data of 200 tweets. Then we analyzed the data using R and create a word cloud visualization to figure out the importance of each words used in “albd1971” twitter account.

## 10 Model code on R for twitter word cloud analysis

Here, we cultivated and nurtured different source codes from GitHub, r-bloggers, and stackoverflow [34-36] in order to create the word cloud using R.

```

# installing packages
install.packages("tm")
install.packages("twitterR")
install.packages("ROAuth")
install.packages("wordcloud")

# importing installed libraries
library(tm)
library("twitterR")
library("ROAuth")
library(wordcloud)

# Twitter developer account authentication
> api_key <- "XXXXXXX"
> api_secret <- "XXXXXXXXXXXX"
> access_token <- "XXXXXXXXXXXX"

```

```

> access_token_secret <- "XXXXXXXXXXXX"
> setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)

# storing tweets
getTweets<-userTimeline("albd1971",n=200)

# displaying tweets
head(getTweets)

# transform tweets to data frame
df<-do.call("rbind",lapply(getTweets,as.data.frame))

# build a corpus
myCorpus<-Corpus(VectorSource(df$text))

# using tm to remove unwanted data
myCorpus <- Corpus(VectorSource(myCorpus))

myCorpus <- tm_map(myCorpus,removePunctuation)
myCorpus <- tm_map(myCorpus,removeNumbers)

myCorpus <- tm_map(myCorpus, PlainTextDocument)
myCorpus <- tm_map(myCorpus,tolower)

myStopwords <- c(stopwords('english'))
myCorpus <- tm_map(myCorpus,removeWords,myStopwords)

myCorpus <- Corpus(VectorSource(myCorpus))
myTdm <- TermDocumentMatrix( myCorpus,control=list( wordLengths = c(1,Inf) ) )

findFreqTerms(myTdm, lowfreq=20)

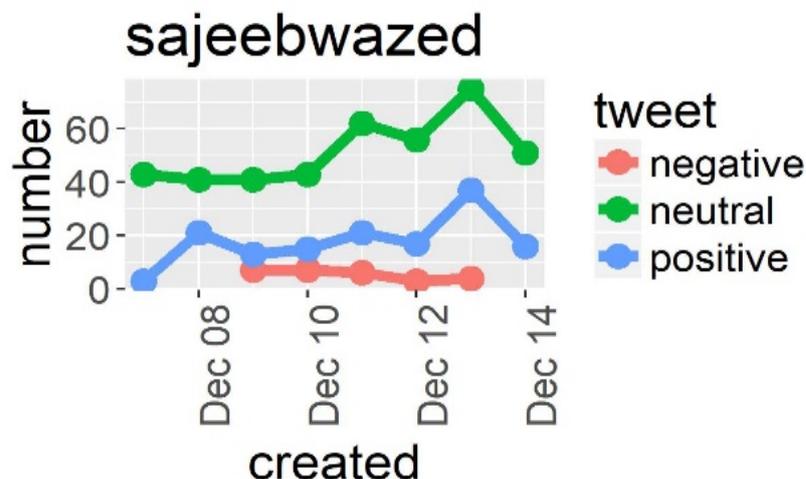
# Word Cloud
m<-as.matrix(myTdm)
wordFreq<-sort(rowSums(m),decreasing=TRUE)
set.seed(375)
grayLevels<-gray((wordFreq+10)/(max(wordFreq)+10))
wordcloud(words=names(wordFreq),freq=wordFreq,min.freq=3,random.order=F,col-
ors=grayLevels)

# END OF PROGRAM

```

**Output and explanation of Twitter sentiment analysis and twitter word cloud analysis**

After the twitter analysis on R, it generated the following chart which shows the positive, neutral and negative sentiments the user used in his twitter account.



**Fig. 3.** Sentiment analysis output for the twitter account “sajeebwazed”



562	12/8/2017	1	FALSE	9.38805E+17	FALSE	NA	NA	9.39E+17	Tsa151	8.32E+17	0	FALSE	AflahaMir	<a href="#">	@Tsa151	FALSE
563	12/8/2017	0	FALSE	9.38805E+17	FALSE	NA	NA	NA	NA	NA	0	FALSE	eeazim	<a href="#">	@UJ09A1	TRUE
564	12/8/2017	0	FALSE	9.38804E+17	FALSE	NA	NA	9.39E+17	Tsa151	8.32E+17	0	FALSE	AflahaMir	<a href="#">	@Tsa151	FALSE
565	12/8/2017	0	FALSE	9.38804E+17	FALSE	NA	NA	9.39E+17	AflahaMir	8.94E+17	0	FALSE	Tsa151	<a href="#">	@AflahaM	FALSE
566	12/8/2017	0	FALSE	9.38802E+17	FALSE	NA	NA	9.39E+17	Tsa151	8.32E+17	0	FALSE	AflahaMir	<a href="#">	@Tsa151	FALSE
567	12/8/2017	0	FALSE	9.38801E+17	FALSE	NA	NA	9.39E+17	Tsa151	8.32E+17	0	FALSE	AflahaMir	<a href="#">	@Tsa151	FALSE
568	12/7/2017	3	FALSE	9.38792E+17	FALSE	NA	NA	NA	NA	NA	0	FALSE	nazmulbd	<a href="#">	@UJ099C	TRUE
569	12/7/2017	0	FALSE	9.3877E+17	FALSE	NA	NA	9.39E+17	SyedTito7	8.64E+17	0	FALSE	SyedTito7	<a href="#">	@albd197	FALSE
570	12/7/2017	0	FALSE	9.3877E+17	FALSE	NA	NA	9.39E+17	SyedTito7	8.64E+17	0	FALSE	SyedTito7	<a href="#">	@albd197	TRUE
571	12/7/2017	19	FALSE	9.3877E+17	FALSE	NA	NA	NA	NA	NA	4	FALSE	prokashm	<a href="#">	@UJ09AC	FALSE
572	12/7/2017	0	FALSE	9.38759E+17	FALSE	NA	NA	9.39E+17	albd1971	2.29E+09	0	FALSE	SyedTito7	<a href="#">	@albd19	TRUE
573	12/7/2017	1	FALSE	9.38759E+17	FALSE	NA	NA	9.38E+17	cbaidulq	1.77E+08	0	FALSE	SyedTito7	<a href="#">	@cbaidu	TRUE
574	12/7/2017	0	FALSE	9.38758E+17	TRUE	NA	NA	NA	NA	NA	1	FALSE	SyedTito7	<a href="#">	RT	FALSE
575	12/7/2017	0	FALSE	9.38757E+17	FALSE	NA	NA	9.39E+17	SyedTito7	8.64E+17	0	FALSE	SyedTito7	<a href="#">	@Sohag5	TRUE
576	12/7/2017	0	FALSE	9.38757E+17	FALSE	NA	NA	9.38E+17	albd1971	2.29E+09	0	FALSE	afnanghai	<a href="#">	@albd197	FALSE
577	12/7/2017	0	FALSE	9.38757E+17	FALSE	NA	NA	9.38E+17	Schagsaif	7.73E+17	1	FALSE	SyedTito7	<a href="#">	@Schags	TRUE
578	12/7/2017	0	FALSE	9.38755E+17	TRUE	NA	NA	NA	NA	NA	1	FALSE	SyedTito7	<a href="#">	RT @Syed	FALSE
579	12/7/2017	2	FALSE	9.38752E+17	FALSE	NA	NA	9.39E+17	albek_bhi	8.11E+17	0	FALSE	lmtlaza67	<a href="#">	@albek_b	FALSE
580	12/7/2017	0	FALSE	9.38748E+17	FALSE	NA	NA	9.39E+17	SyedTito7	8.64E+17	0	FALSE	SyedTito7	<a href="#">	@albd197	TRUE
581	12/7/2017	0	FALSE	9.38748E+17	FALSE	NA	NA	9.37E+17	SyedTito7	8.64E+17	0	FALSE	SyedTito7	<a href="#">	@albd197	TRUE
582	12/7/2017	0	FALSE	9.38741E+17	FALSE	NA	NA	NA	NA	NA	0	FALSE	Tsa151	<a href="#">	@UJ0986	TRUE
583	12/7/2017	0	FALSE	9.38717E+17	FALSE	NA	NA	9.24E+17	albd1971	2.29E+09	0	FALSE	Najimud9	<a href="#">	@albd197	TRUE
584																

Figure 5: Collected stack file (unstructured data) from the twitter account “sajeebwazed”

Twitter word cloud analysis also has some limitations. In such data visualization like word cloud, the main downside is that we lose context. We cannot differentiate between positive and negative words. In word cloud analysis, we basically lose the ability to derive the meaning from it. For example, from the word

cloud we generated, there is one term “rohingyacrises”. This is a major issue for Bangladesh in recent times. It is a continuous issue that has exhausted Bangladesh-Myanmar relationship since 1970 [37]. So, in this context, it is difficult to learnt whether “rohingyacrisis” used in positive or negative reference.

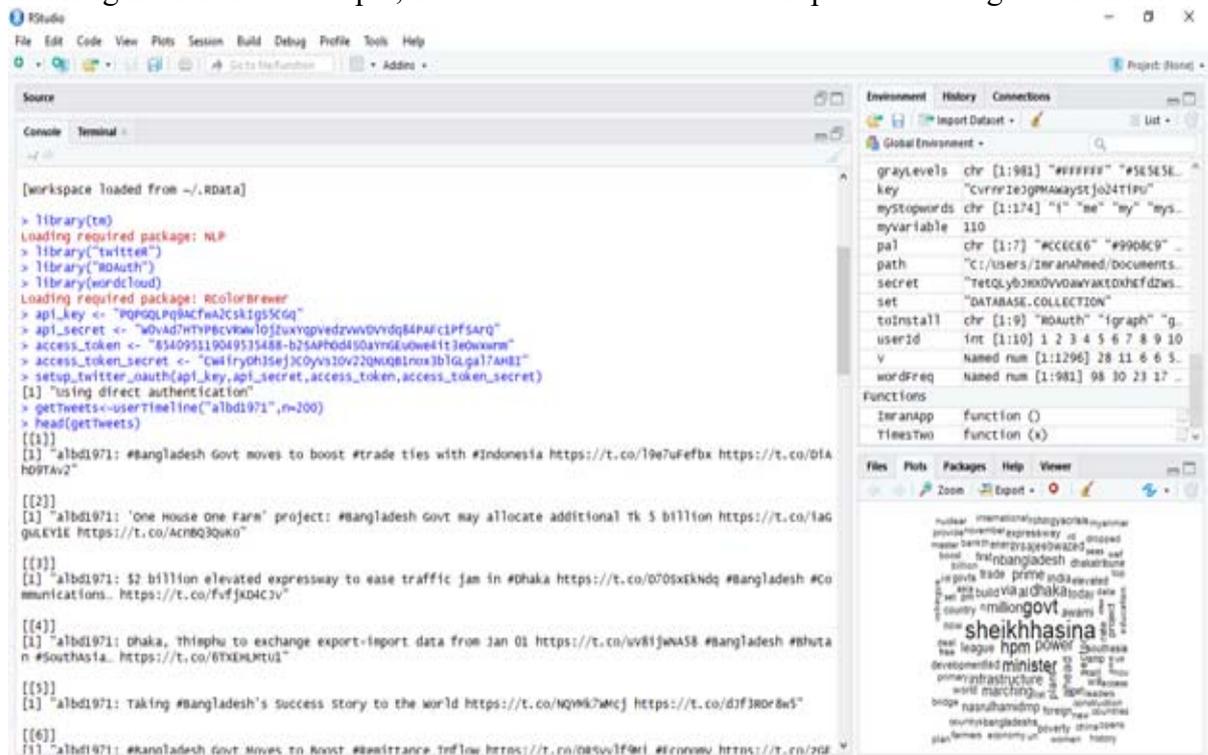


Figure 6: Word cloud generating from “albd1971” using twitter.

Generated word clouds need to be used carefully but they present a quick and easy way of data visualization.

## 12 Challenges and conclusions

The social networking scene is huge and evolving. Indeed, even as some online networking sites detonate into utilization, rapidly turning into everyday tools, new stages are going along with them continually. By and large, there are dozens of destinations with no less than one hundred thousand enlisted clients, and numerous more special guests—including sites a large portion of us have never known about. Indeed, even as organizations start to understand the danger in disregarding social networking content and, on the other hand, the open door it exhibits, their inquiries uncover how much stays obscure. Real time analysis through social media is a serious challenge. Social media data is so much noisy and unstructured and it's very challenging in the context of real time social media analysis from unstructured data and this analysis lags seriously when it comes to numeric analysis. Social media analysis (sentiment analysis, word cloud analysis) using text mining can't be done by computer effectively categorized because in social media the word "good" might mean good or it might mean bad, depending on perception, relationship and other variables. Despite all these challenges, new analytics tools with enough time can bring out meaningful comments and analysis across various social media sites.

### References:

- [1] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon, "Manyeyes: a site for visualization at internet scale," *IEEE transactions on visualization and computer graphics*, vol. 13, 2007.
- [2] F. B. Viégas and M. Wattenberg, "Timelines tag clouds and the case for vernacular visualization," *interactions*, vol. 15, pp. 49-52, 2008.
- [3] J. Sinclair and M. Cardew-Hall, "The folksonomy tag cloud: when is it useful?," *Journal of Information Science*, vol. 34, pp. 15-29, 2008.
- [4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, pp. 1-135, 2008.
- [5] A. Esuli and F. Sebastiani, "SentiWord-Net: a high-coverage lexical resource for opinion mining," *Evaluation*, pp. 1-26, 2007.
- [6] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997, pp. 174-181.
- [7] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004, p. 1367.
- [8] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 129-136.
- [9] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welp, "Predicting elections with twitter: What 140 characters reveal about political sentiment," *Icwsn*, vol. 10, pp. 178-185, 2010.
- [10] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *LREc*, 2010.
- [11] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in *International conference on discovery science*, 2010, pp. 1-15.
- [12] D. Davidov, O. Tsur, and A. Rapoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd international conference on computational linguistics: posters*, 2010, pp. 241-249.
- [13] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd international conference on computational linguistics: posters*, 2010, pp. 36-44.
- [14] G. Smith, *Tagging: people-powered metadata for the social web*: New Riders, 2007.
- [15] B. Y. Kuo, T. Hentrich, B. M. Good, and M. D. Wilkinson, "Tag clouds for

- summarizing web search results," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 1203-1204.
- [16] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, 2014, pp. 1833-1842.
- [17] J. Steele and N. Iliinsky, *Beautiful visualization: looking at data through the eyes of experts*: " O'Reilly Media, Inc.", 2010.
- [18] K. Gordon. Number of social media users worldwide from 2010 to 2021 (in billions) [Online]. Available: <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>
- [19] Unknown. Daily time spent on social networking by internet users worldwide from 2012 to 2017 (in minutes) [Online]. Available: <https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/>
- [20] Unknown. Percentage of global population using Facebook as of June 2017, by region [Online]. Available: <https://www.statista.com/statistics/241552/share-of-global-population-using-facebook-by-region/>
- [21] Unknown. Number of monthly active Facebook users worldwide as of 3rd quarter 2017 (in millions) [Online]. Available: <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>
- [22] K. Vance, W. Howe, and R. P. Delvalle, "Social internet sites as a source of public health information," *Dermatologic clinics*, vol. 27, pp. 133-136, 2009.
- [23] J. S. Brown and P. Duguid, *The Social Life of Information: Updated, with a New Preface*: Harvard Business Review Press, 2017.
- [24] P. Casanovas, L. De Koker, D. Mendelson, and D. Watts, "Regulation of Big Data: Perspectives on strategy, policy, law and privacy," *Health and Technology*, pp. 1-15, 2017.
- [25] R. Matulionyte, "Self-Publishing Boom and Its Legal Challenges," 2017.
- [26] D. O'Brien, J. Ullman, M. Altman, U. Gasser, M. Bar-Sinai, K. Nissim, *et al.*, "Integrating Approaches to Privacy Across the Research Lifecycle: When Is Information Purely Public?," 2015.
- [27] A. Vedder, C. Cuijpers, P. Vantsiouri, and M. Z. Ferrari, "The law as a 'catalyst and facilitator' for trust in e-health: challenges and opportunities," *Law, Innovation and Technology*, vol. 6, pp. 305-325, 2014.
- [28] S. P. Borgatti, M. G. Everett, and J. C. Johnson, *Analyzing social networks*: Sage, 2018.
- [29] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Communications of the ACM*, vol. 49, pp. 76-82, 2006.
- [30] J. Breen. positive-words.txt [Online]. Available: <https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/blob/master/data/opinion-lexicon-English/positive-words.txt>
- [31] J. Breen. negative-words.txt [Online]. Available: <https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/blob/master/data/opinion-lexicon-English/negative-words.txt>
- [32] M. Kmane. (2016, January 8). *R-Code-for-tweet-sentiment-analysis*. Available: <https://github.com/mayurkmane/R-Code-for-tweet-sentiment-analysis>
- [33] A. Core. (2014, January 8). *Twitter sentiment analysis with R*. Available: <https://www.r-bloggers.com/twitter-sentiment-analysis-with-r/>
- [34] E. Sevinç. (2012, January 8). *Simple word cloud*. Available: <https://gist.github.com/emres/3424557>
- [35] S. Overflow. (2015, January 8). *Making a simple wordcloud in R*. Available: <https://stackoverflow.com/questions/30386890/making-a-simple->

wordcloud-in-r?utm\_medium=organic&utm\_source=google\_rich\_qa&utm\_campaign=google\_rich\_qa

- [36] T. Kodali. (2015, January 10). *Building Wordclouds in R*. Available: <https://www.r-bloggers.com/building-wordclouds-in-r/>

- [37] S. N. Parnini, M. R. Othman, and A. S. Ghazali, "The Rohingya refugee crisis and Bangladesh-Myanmar relations," *Asian and Pacific Migration Journal*, vol. 22, pp. 133-146, 2013.



**Ahmed Imran Kabir** has graduated from the college of Business with a Master of Science in Business Analytics/ Data Science from Texas A&M University- Commerce, United States in 2017. He is now a doctoral candidate in the Department of Business and Accountancy, University of Malaya, Malaysia. His work focuses on healthcare analytics, business intelligence, text mining, and big data software.



**Ridoan Karim** has graduated from the Department of Law with a Master of Comparative Law from International Islamic University Malaysia (IIUM) in 2016. He is now a doctoral candidate in the Department of Law; University of Malaya, Malaysia. His work focuses of nuclear regulations, data protection law, block chain technology. He worked as an adjunct lecturer at East Delta University, Bangladesh in the Fall semester 2016.



**Shah Newaz** currently studies at the Faculty of Business and Accountancy, University of Malaya. Shah likes to do research in Manufacturing, Supply chain management and Business Administration. Their most recent publication is 'A Cross-Dock Model to Replace Traditional Supply Chain of Vegetable Industry through Eliminating Bottlenecks: Bangladesh Perspective'.



**Muhammad Istiaque Hossain** has graduated from the Department of Business and Law with a Master of Professional Accounting from Central Queensland University, Australia. His work focuses on financial analytics, data security in online banking and CSR activities and the financial performance of the companies.