AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

Text Mining for unstructured data on Twitter hashtags using R.

Loveleen Gaur, Gurinder Singh, Saurabh Minotra, Aditya Varshney

Amity University, Noida, UP, India.

Address For Correspondence:
Loveleen Gaur, Amity University, Noida, UP, India.
E-mail: lgaur@amity.edu,

ARTICLE INFO

Article history:
Received 18 February 2017
Accepted 15 May 2017
Available online 18 May 2017

Keywords:
#Demonetisation, R, Sentiment Analysis,Sentiment Polarity, Emotion Analysis

ABSTRACT

Background: On 8th November 2016, Hon'ble PM of India Shri Narendra Modi announced that Rs. 500 and Rs 1000 will not be considered as legal tender. This announcement created lots of hue and cry among Indian citizens. The social media was full of mixed emotions and opinions **Objective:** This paper has been developed with the following objectives: To find out the sentiment polarity of demonetization (#demonetisation) on fellow citizens. To conduct the emotions analysis (Anger, anticipation, fear, sadness, surprise) using R by fetching #demonetisation tweets through twitter. **Results:** To analyze the actual sentiments and emotion, sentiment analysis is a form of social listening, it means monitoring the posts and discussions on the social media platform, then figuring out how the consumers react to a particular brand or event. **Conclusion:** The paper analyzed twitter hashtags related to demonetisation in India.

INTRODUCTION

Microblogging sites have turn out to be a basis of diverse information, individuals in today's world post their opinions and discuss upon the present-day topics/issues. Also, people complain or express their sentiments about the recently launched products which they use in their day-to-day life. As per the Twitter official website, as of 2016, monthly active users on twitter were more than 313 million**.** *As per The New York Times, in 2016, Twitter had turned out to be the leading source of breaking news with 40 million tweets, on the day of US presidential election.*

Sentiment analysis (at times known as emotion AI or opinion mining) discusses about the usage of natural language processing (NLP), computational linguistics, text analysis, and biometrics to methodically identify, abstract, compute, and study affective states and subjective data. It is broadly functional in the voice of the customer materials such as reviews and survey responses, online and social media, other materials for applications that range from marketing to customer service.

Here, authors have extracted the Tweets from Twitter and use it as an unstructured data for the analysis. It gives the user the power to create and share ideas and information instantly, without barriers. To categorize and search tweets people use the hashtag symbol (#), it is prefixed to the appropriate keyword or phrase. The top industry verticals namely E-commerce, Finance and aviation sector are widely utilizing the sentiment analysis and yielding positive results for their businesses. The companies receive the real time feedback which is unaltered, as in case of the traditional feedback approach which can be manipulated by the user filling questionnaire or the user can be biased while answering. The analysis helps examining reputation of the brand and identifies fluctuations in the market. The idea behind this paper is to analyze the sentiments post demonetization on Indian citizens. According to the news broadcast by NDTV (Abhinav Bhatt, Nov 2016), demonetization was announced by Prime Minister of India Shri Narendra Modi on 8 November. The currency notes unexpected withdrawal from circulation, had caused chaos, with marketplaces, petrol pumps and other

To Cite This Article: Loveleen Gaur, Gurinder Singh, Saurabh Minotra, Aditya Varshney., Text Mining for unstructured data on Twitter hashtags using R.. ***Aust. J. Basic & Appl. Sci.,*** *11(8): 49-55, 2017*

retailers, but it was a step aimed at fighting corruption, According to The Economics Times and Live Mint, here are a set of numbers to understand how the Narendra Modi governments to demonetize high- value currency notes is panning out/impacting the Indian citizens :

Table 1: Demonetization impact by number

S.No.	Figure	Comments
1	Rs. 14.2 Trillion	As per RBI data (Live Mint, Dec 2016), worth of currency notes (Rs. 500 & Rs. 1000) in circulation as on 31st March, 2016.
2	Rs 12.4 Trillion	As per RBI data (Live Mint, Dec 2016), the amount deposited till 10 December post demonetisation
3	2.11 %	According to Central Statistics Office (Live Mint, Dec 2016), food inflation reduced from 3.32 to 2.11%
4	3.36%	Indian retail inflation softened from 4.2%

The paper is arranged as follows. The first section discusses the introduction of the sentiment analysis and the demonetization. The second section briefly discusses earlier researches/literature review. Section 3, briefly discusses about the dataset that we have used for this paper and data preprocessing techniques adopted. Section 4 discusses the sentiment analysis technique and the data interpretation. The section 5 includes the conclusion of the paper.

Literature Review:

Sentiment analysis has been used as a Natural Language Processing. The Preliminary research was done at a document level classification task, they have found the relation between subjectivity detection and polarity classification (Turney, 2002; Pang and Lee, 2004), later studies done at a sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and further work was done at the phrase level (Wilson *et al*., 2005; Agarwal *et al*.,2009). Some of the early results on sentiment analysis of Twitter data are by Go *et al*. (2009), (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010). Go *et al*. (2009), have used distant learning to acquire sentiment data. The tweets ending with ":)" ":-)" were considered as positive and emoticons ":(" ":-(" were taken as negative tweets. They build models using MaxEnt, Support Vector Machines (SVM) and Naive Bayes, and they report SVM outperforms other classifiers. (Alexander Pak *et al*.), they have done the research using the automatic collection of corpus. Sentiment categorization is essentially a classification problem, where features that contain opinions or sentiment information should be identified before the classification. (Pang and Lee), in their research recommended feature selection to remove the objective sentences by fetching and mining the subjective sentences. They proposed a text-categorization technique, which is capable of identifying the subjective content. Gann *et al*., in the research the author worked upon the 6,799 tokens based on Twitter data, a sentiment score was given to the token, viz. TSI(Total Sentiment Index), introducing it as a positive token or a negative token. Specifically, a TSI for a certain token is computed as: $TSI=(p-tp / tn\times np) / (p-tp / tn\times np)$; where *p* is the number of times a token appears in positive tweets and *n* is the number of times a token appears in negative tweets. Tp/tn is the ratio of total number of positive tweets over total number of negative tweets. The research done by (Sasikumar Natarajan, Dec 2016) related to sentiment analysis on demonetization in India, give the hypothesis test results that people of India are having a Negative sentiment on the decision taken by the Government.

Methodology:

Problem Conceptualization:

As per Joel Rebello & Gayatri Nayak (Economic Times, Nov 29, 2016), Indian economy is largely dependent on cash, the banking system for money transactions is used by only half of the population, demonetisation has hit hard on the trade and consumption. Due to the disruption in the flow of money, there has been a drop in spending which has put down the growth during the current fiscal year. As per the economist, Amartya Sen (Hindustan Times, December 31, 2016), demonetization would have adverse effects.

There were mix reviews by the people, strategists, economists, political parties and various other sources on the demonetization. In order to do the reality check of this political issue the questionnaire won't serve the purpose, so the sentiment analysis on tweets posted on Twitter would be able to capture the real-time emotions. As compared to other medium, social media is more reliable for checking the pulse of fellow citizens, as it offers an opportunity for the real-time analysis of the expressed mood. There are various emotions in the tweets joy, love, anticipation, fear, anger, sadness, disgust and surprise.

Samples and Measures:

The paper collected Twitter data which used in a wide range of applications, Twitter data represents the largest archive of human behavior in existence, and what we can learn from it is virtually unlimited. With an objective to do the sentiment analysis, the data is collected from the Twitter using the API. A data set is known

as corpus; it is a comprehensive dataset which ensures that the accuracy of algorithm meets the standards we expect.

As per the document by Jeff Gentry, the package twitteR has the function **searchTwitter**, it is used to collect the data. Depending upon the arguments passed to the function, it will search the Twitter.

Sampling Technique:

The complete data set is fetched from the twitter on the basis of the hashtags demonetization and black money and is stored in the corpus, then filtering and cleansing is done on the corpus. In cleansing and filtering, redundant data, white spaces, punctuations, numbers, stop words are removed, Stemming is done on the data set.

Table 2: Sampling Data

Max Sample Size (Max argument value passed)	10000 tweets
Max Tweets Collected (Actual Sample Size)	8238 tweets
Data Collection Tool	RStudio
Search String:	tweet <- searchTwitter("#Demonetisation or #Blackmoney", n=10000, since="2016-11-08"). Here, in the SearchTwitter function, we have passed the hashtag demonetization as an argument for the maximum tweets limit 10000 since the date 8 November, 2016. The maximum value is passed as an input and no other arguments values are passed as we want fetch the complete data set.

There are other arguments as well, which can be used for different scenarios: searchTwitter (searchString, n=10, since= NULL, geocode=NULL, lang=NULL, until=NULL, maxID=NULL, locale=NULL ,resultType= "popular", retryOnRateLimit=120, ...) Rtweets(n=25, lang=NULL, since=NULL, ...).

Table 3: Arguments in the Search Twitter Function

S.No.	Argument	Comments
1	geocode	If not null, the tweets within the radius of the latitude and longitude are returned. Unit can be mile or kilometers Example: geocode='37.781157,-122.39720,1mi'
2	resultType	"Recent" – It returns the most recent results "Popular"- it returns the most popular results Default is Mixed- it includes both the real time and popular results.
3	retryOnRate Limit	If non-zero value is passed, it will block the retry up to x times. It might take a longer to run, but the task will eventually complete if the retry count is high enough.

Process Adopted:

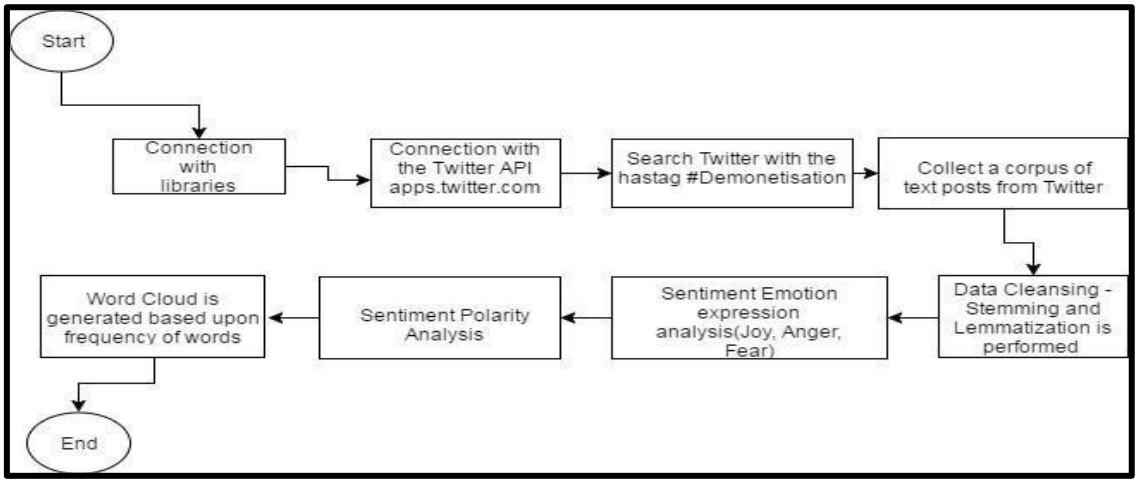

Fig. 1: The Process adopted in the sentiment analysis

The above is the flow diagram of the process adopted in the sentiment analysis, the below section follows the detailed information about the same.

Sentiment Analysis:

Fetch tweets from Twitter:

In order to fetch the tweets, one need to connect to libraries and make a connection with the Twitter API. Once the connection is made successfully, the data is searched and fetched from the Twitter using the keyword "demonetization". The below are the steps involved:

Step 1: Connection with libraries

Step 2: Connection with Twitter API[1] - First the app is created on Twitter apps.twitter.com/ and API keys and access tokens are generated. Now, in order to access the Twitter in R, the authentication is done using api_key, api_secret, access_token, access_token_secret

Step 3: Search the Twitter with the hashtag- We search the Twitter using the function **searchTwitter** with the hastag - "#Demonetisation OR #Blackmoney". The tweets list is converted to the data frame using the twListToDF package.

Cleanse the data:

Once the data is fetched from twitter we cleanse the data.

Step 4: Collect a Corpus – A corpus consists of a databank of natural texts, compiled from writing and/or a transcription of recorded speech. We extracted the tweets and collected, it will be the "corpus" (body) of texts we are mining, a vector source is created from the corpus; mycorpus <- Corpus(VectorSource(tweets.text)). We collected a corpus of text which have both the positive and negative sentiments. No human effort is needed for classifying the texts. The size of the collected corpus can be arbitrarily large.

Step 5: Processing and cleansing of tweets – the text is converted to lower case, white spaces, stop words, punctuations and numbers are removed. The stemming is performed on the data for linguistic normalisation. Words need to be stemmed to retrieve their radicals. The function stemDocument stem words in a text document using Porter's stemming algorithm (Ingo Feinerer, Text Mining package, 2017).

Sentiment analysis and Polarity of tweets:

Emotion recognition is a special case of sentiment analysis. The output of sentiment analysis is produced in terms of either polarity (e.g., positive or negative) or in the form of rating (e.g., from 1 to 5). Emotions are a more detailed level of analysis in which the result are depicted in more expressive and fine-grained level of analysis. Sentiment analysis deals with only text, while emotions can be expressed by text,

images, audio, video, facial signs etc.(Gaston) (Veera 2016). We use the Syuzhet package in R to do the analysis the package has four sentiment dictionaries, it was developed in the NLP group at Stanford and it provides a robust technique to retrieve the data and is considered to be the affluent sentiment extraction tool. In our research we are using the get_nrc_sentiment, it implements Saif Mohammad's NRC Emotion lexicon. According to Mohammad, "the NRC emotion lexicon is a list of words and their relations with eight emotions (joy, love, fear, anger, sadness, disgust, surprise & anticipation) & two sentiments (negative and positive)". In the example below, we store the sentences in the object, in the next step the object is passed as an argument to the get_nrc_sentiment function. The emotions values data is further stored in separate columns and it can be retrieved individually or in sets. Here we classify the item(s) that the NRC lexicon identifies as joyful; *"joy_items <- which (var_nrc$joy > 0)"*. The score of all the emotions can be viewed using the pander library "*pander::pandoc.table(nrc_data[, 1:8], split.table = Inf)"*.The score of the emotions is stored in a file and graph is plotted from the value. The positive and negative valence is tested; *pander::pandoc.table(var_nrc[, 1:8], split.table = Inf)*

A sentiment token is a word or an expression that expresses sentiment, a word token comprises of a positive or negative word and its part-of-speech tag. Formula for t's sentiment score(SS) is

$$SS(t)=\sum^{i=1}5\ i\times\ \gamma\ 5,i\times Occurrencei(t)$$

$$\sum^{i=1}5\ \ \gamma\ 5,i\times Occurrencei(t)$$

Where t is the token, Occurrence i (*t*): In i-star rating reviews where i varying from 1 to 5, t's number of occurrence. The data is discrete as the reviews are varied with different star ratings. In our dataset, major amount is of the 5-star reviews, hence ratio, $\gamma\ 5,i$ is

$$\gamma 5,i=|5-star|$$

$$|i\text{-}star|$$

Where i varies from 1..5, the number of 5-star reviews is taken up in the numerator and the denominator is the number of *i*-star reviews. Therefore, if the dataset were balanced, γ 5,*i* would be set to 1 for every *i*. Consequently, every sentiment score should fall into the interval of [1,5]. The median of sentiment score exceeding 3 is considered to be the positive word token and the score less than 3 is taken as a negative work token. Below table has sentiment score, which is returned as an output of the get_nrc_sentiment function and it is stored in csv file. From the table below we find that, out of actual sample size of 8238 tweets of data set, after cleansing and passing the dataset to the get_nrc_sentiment function, the function returns 2376 tweets as an output because emotions were captured in these tweets. Below score indicates that the function is smart enough to find the emotions

Table 4: Sentiment Score

Emotions	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Negative	Positive
Score	204	193	114	230	145	199	103	279	438	471

For the above set of sentiment score returned from the package. Below is the graph plotted using the function ggplot.

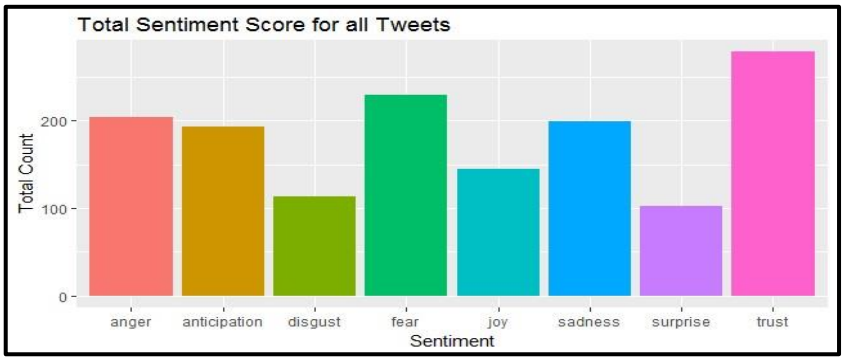

Fig. 2: Output of emotion analysis on Rstudio

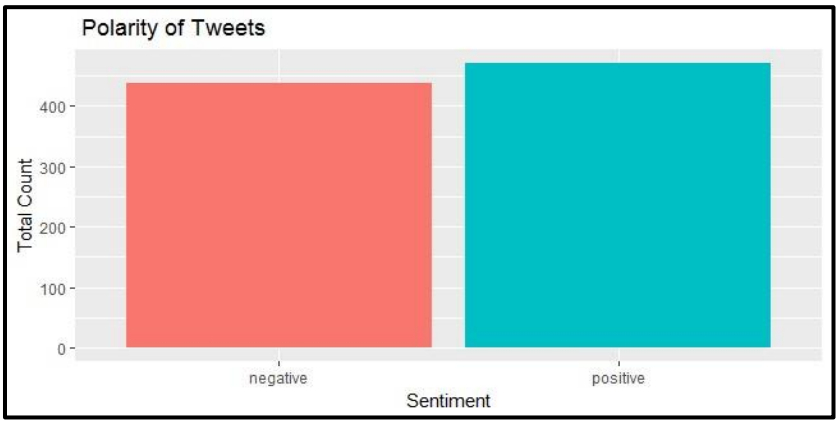

Fig. 3: Output of Sentiment Polarity on RStudio

From the Figure 2, 3, 4: we can see that maximum number of people (279 count) are having trust on the decision taken on the demonetization, which is around 12 % of the total output of 2376 tweets. Around 10% of the people who tweeted were in fear, around 8%, 9% were in anger and sadness respectively, around 8% in anticipation, 6% of the people who tweeted were in joy, 5% in disgust and 4% were surprised. The sentiment polarity: 471 of the tweets were positive which is around 20% and 438 were negative tweets which is around 18%.

Generate the Word Cloud:

Word cloud is the visual representation of the text data. To generate it, the cleansed data is passed as an input to create the word cloud. The words which are repeated more than three times are included in the word cloud. As per the document by Ian Fellows, 2015, the package wordcloud is used to plot a cloud of words.

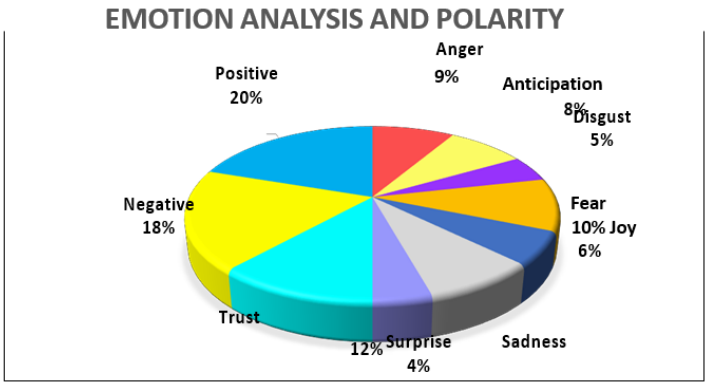

Fig. 4: Percentage analysis Emotion and Polarity

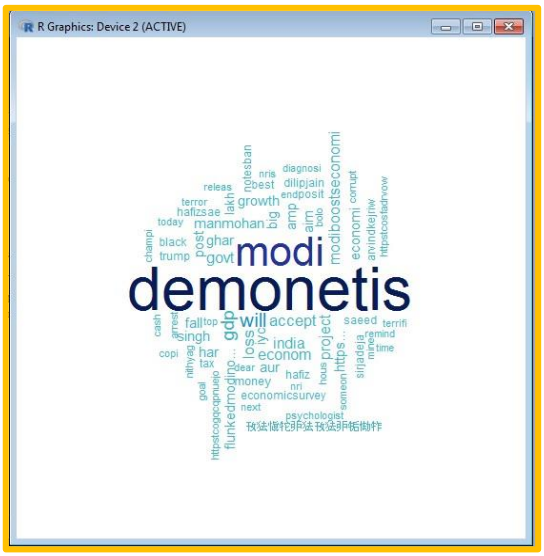

Fig. 4: Word Cloud

wordcloud(words,freq,scale =c(4,.5), min.freq=3, max.words = Inf, random.order =TRUE, random.color = FALSE, rot.per = .1, colors="black", ordered.colors=FALSE, use.r.layout=FALSE, fixed.asp=TRUE, ...)

min.freq: the words having the frequency less than the threshold will not be plotted. max.words: the words having the frequency less than the threshold will be plotted. Least frequent terms dropped random.order: The word cloud is generated in the random order. In case of false, it will be plotted in decreasing frequency.

Conclusion:

In the paper, authors have done the sentiment analysis on tweets posted on Twitter to capture the emotions of the Indian citizens on the demonetization decision taken by the Government of India. Despite of the lot of chaos, difficulties in trade faced they businesses, various reports from the economists that the decision would have the side effects etc. we find that the citizens of India are having a Positive sentiment towards this major decision taken by the Government of India.

REFERENCES

Twitter Usage / Company Facts, 2017. https://about.twitter.com/company

Isaac, Mike; Ember, Sydney, 2016. "For Election Day Influence, Twitter Ruled Social Media". The New York Times. https://www.nytimes.com/2016/11/09/technology/for-election-day-chatter-twitter- ruled-social-media.html

Abhinav Bhatt, 2016. PM Modi Announces Notes Ban In Anti-Corruption Move, Millions Face Cash Crunch. NDTV, http://www.ndtv.com/india-news/pm-modi-speaks-to-nation-tonight-at-8-pm-1622948

2017. Demonetisation: RBI's own figures indicate return of 15 lakh crore of banned notes, The Economic Timeshttp://economictimes.indiatimes.com/news/economy/finance/demonetisation-rbis- own-figures-indicate-return-of-15-lakh-crore-of-banned-notes/articleshow/56536621.cms

2016. Demonetisation's impact by the numbers, Live Mint. http://www.livemint.com/Politics/9XNrKKKjknca6owQm8qm7K/Demonetisation-impact-by-numbers-as-of-14-December.html

2016. Demonetisation will have adverse effects, says economist Amartya Sen, Hindustan Times.

http://www.hindustantimes.com/india-news/demonetization-will-have-adverse-effects-says-economist-amartya-sen/story-RZKZ6rhVjoLMWS7nY7CI2H.html

Joel Rebello and Gayatri Nayak, ET Bureau, 2016. Demonetisation and its side-effects, Economic Times, http://economictimes.indiatimes.com/articleshow/55678393.cms?utm_source=content ofinterest&utm_medium=text&utm_campaign=cppst

Hu, M., B. Liu, 2004. Mining and summarizing customer reviews In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168–177.. ACM, New York, NY, USA.

Pang, B., L. Lee, 2008. Opinion mining and sentiment analysis. Found Trends Inf Retr, 2(1-2): 1-135.

Alexander Pak, Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508, F-91405 Orsay Cedex, France

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, Sentiment Analysis of Twitter Data. Department of Computer Science, Columbia University New York

Suresh Kumar, R Tutorial, 2016. http://www.bigdatanews.com/profiles/blogs/learn-everything-about-sentiment- analysis-using-r

Sasikumar Natarajan, 2016. Sentiment Analysis on Demonetization in India, http://techienasa.blogspot.in/2016/12/sentiment-analysis-on-demonetization-in.html

Jeff Gentry, R Based Twitter Client, 2016. https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

Ingo Feinerer [aut, cre], Kurt Hornik [aut], Text Mining Package https://cran.r-project.org/web/packages/tm/tm.pdf

Gaston, Sentiment Analysis with "sentiment" https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment

Veera Raghava Reddy, 2016. Sentiment Analysis Using R Language, http://www.evoketechnologies.com/blog/sentiment-analysis-r-language/

Matthew Jockers, Introduction to the Syuzhet Package, 2016 https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

Saif Mohammad, NRC Word-Emotion Association Lexicon http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Xing Fang and Justin Zhan, Sentiment analysis using product review data https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2