IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 7, Issue 1, January 2020

Use of Tidytext Lexicons Approaches for Sentiment Analysis

Niharica Choubey[1], Vilender Kumar[2], Sanjay Kumar Gupta[3]

M.Phil., Student of School of Studies in Computer Science and Applications,
Jiwaji University Gwalior (M.P.), India[1]
Associate Professor, Gitarattan International Business School, New Delhi 110085[2]
Professor, School of Studies in Computer science & Applications,
Jiwaji University Gwalior (M.P.), India[3]

Abstract: Use of Sentiment analysis is increasing day by day know the sentiments of public on social media. It analyzes sentiments, emotions and opinion of people which are shared on social sites. Twitter is a powerful social networking tool used frequently to online sharing of views. In this paper, tidytext mining technique is used for sentiment classification of twitter data. Sentiment analysis is performed on tweets of two popular Indian leaders' for prediction of 2019 election which are extracting from twitter posts from 2014 to 2018, and for this R language is utilized that provides Tidy text package. Tidytext package contains three lexicons NRC, BING, and AFINN to analyze emotions and comparison of three lexicons.

Keywords: Twitter, sentiment analysis, Tidytext lexicon approach: NRC, BING, AFINN, and wordcloud, Tidy text mining

I. INTRODUCTION

In some past years, social media is a vast platform for sharing thoughts, views, counsel, and opinion about different topics. So, social media platform contains a huge amount of raw data. These data are very useful for sentiment analysis by which we can know about human thinking on several topics. In sentiment analysis, we analyze about sentiments, emotions, and views of people which are shared by using of social media platform. Many social media platforms are as facebook, twitter, Whatsapp, instagram etc. Sentiment analysis is one of the fastest growing research areas in the field of computer science, making it challenge to keep track of all the activities in the area [2]. Sentiment analysis tries to explain the situation of a speaker, writer, essayist, or additional subjects in terms of theme via extreme emotional or turned on responses to associate degree archive, speech communication, or juncture. The situation might be a judgment or opinion, filled with feeling [1].

In this paper, we have collected tweets 2014 to 2018 from social media platform twitter regarding two prestigious Indian leaders Sh. Narendra Modi who is the prime minister of India belongs to Bhartiya Janta Party (BJP) and second, Sh. Rahul Gandhi who belongs to Congress party which is anti party of BJP. Lexicon based approaches are utilized for sentiment classification and referred to as unsupervised technique. Unsupervised approaches do not utilize the training and testing process on dataset, and it contains assortment of words, that make manually. Polarity of dataset may be found by lexicon approach in several sorts. Firstly, we will see, how to get a polarity by sentence; for example:-“this phone battery is extremely good”. This is often a positive sentence, and therefore, sentiment score is +1, if we write a sentence is “Sh. Narendra Modi isn't an acceptable person for prime minister”, then it's a negative sentence and sentiment score is -1. Now, to get a polarity by words, during the sentences will be broken into words. For example; “this phone battery is extremely good”, then sentence has to be broken into words like ‘this’, ’phone’, ’battery’, ’is’, ’very’, ’good’. After that, lexicon compare own words with dataset words and classify the polarity. In the above words we notice that the positive word is ‘good’, and therefore the sentiment score is +1 for good word, if negative word found then the sentiment score is -1 [3]. The aim of Sentiment analysis is to identify & analyze people opinion on social media posts in term of emotions as positive, negative, neutral, sad, angry, and many more. This will enhance the interpretation capabilities from huge unstructured data of twitter by treating text as data frames of individual words and integrate it with processing for analysis to provide a sound way to understand the attitudes and opinions of people expressed in tweets. The R language provides tidytext package that contains three lexicons NRC, AFINN and BING. Further comparison of these lexicons are performed on collected dataset, and also provide the best lexicon for sentiment analysis to analyze the tweets of 2014 to 2018 related to prediction of 2019 election.

The outline of this paper is as follows. We start with introduction of sentiment analysis in section one. Section two contains the literature review of outstanding work done by human and researchers for sentiment analysis across the globe. In section three, we present the methodologies which are used for sentiment analysis. Results are mentioned in section four. In Section five contains discussion and sixth describes conclusion and future work.

II. RELATED WORK

In past years, several researches have done on sentiment analysis by researchers. Some researcher's work on sentiment analysis is explained in this section. [1] This research paper is a study on sentiment analysis classification techniques of twitter data. Researcher describes about different sentiment classification techniques. First describes about machine learning classifiers, classifiers names are naive bayes, support vector machine and maximum entropy. The next approach is define in this paper is document level sentiment analysis approach, by this approach, detect the sentiments polarity of documents using unsupervised document based sentiment analysis system. This system classifies emotions in positive and negative. By this system, classify the sentiment polarity of any document based on the majority of opinion vocabularies which appear in documents. Sentence –level sentiment analysis is another approach on which researcher have done own study. The sentence level sentiment analysis focused on classifying sentences into different categories regarding sentences whether the sentence is positive, negative, and neutral. Twitter sentiment analysis consider as an example of sentence level sentiment analysis. For analysis of twitter sentiment using different techniques which are supervised machine learning approaches, ensemble approaches, lexicon approaches and hybrid approaches. [2]This research paper defines two classifier techniques linear classifier and probabilistic classifier. Tweets collected from twitter.com in this research paper. For linear classification the researcher use support vector machine with linear kernel and probabilistic classifier use two classifier naive bayes and logistic regression. Both types of classifier apply on dataset and represents comparative model of linear, probabilistic and discriminative classifier. [4] This research paper study on sentiment analysis of twitter data regarding Donald trump that means researcher analyze of people emotions which are revealed about Donald trump during many debates. For this work, researcher collects data from twitter OAuth API. The researcher use “sentiment” package for emotions classification and this package are given by R language which is used for programming by researcher. Researcher describes many emotions like positive, negative, anger, disgust, neutrality and many more and also gets the percentage of each emotion regarding Donald trump. [5] This paper describes the prediction about Indian election that is which party won the election in 2016. Tweets are collects in Hindi language from Hindi twitter archive. Tweets are collects regarding five Indian political parties: BJP, Congress, AAP, BSP, and NCP total tweets collects are 42,345. The researcher obtains polarity of tweets by three algorithms dictionary based approach, support vector machine, and naive bayes. These algorithms apply on labeled dataset and compares of three algorithms. This paper aims to given prediction about elections that which party won the election in 2016 and also compare three algorithms which are used for prediction. The researcher found out the accuracy of these algorithms by which researcher get the best algorithm for prediction or classification.

III. PROPOSED METHODOLOGY

This section is partitioned into four parts: 1. Dataset collection, 2. Data pre-processing, 3. Tidytext Lexicons approach, 4. Results and Visual image.

Part 1: Dataset Collection: - As we all know that if we've no food then we cannot cook anything like if we have no dataset then we cannot analysis, so the first thing to do is get some data from twitter[6]. Data extraction from social media depends on application programming interface (API) which is provided by social media platform itself [4]. Different social media platform offers API for extracting the data. We have collected data from English twitter. R language provides twitteR package, it's developed by Jeff Gentry, and it is the better and fastest way to get data from twitter [6]. We have collected tweets regarding three Indian politicians sh. Narendra Modi, Sh. Rahul Gandhi, Sh. Akhilesh Yadav. 2,500 tweets are collected of each politician but for getting a popularity of sh. Narendra Modi and Sh. Rahul Gandhi, we use two separate dataset of both politicians. For comparison of three lexicons we use combined dataset which is a collection of three dataset, 7,500. TwitteR package contains functions: setup_twitter_oauth ( ) function. It provides interface to the twitter web application. This function helps to launch the connection between R compiler and twitter. Some basic steps are as follows:

Step1:- First we create an account on twitter.

Step2:- Fill the form which is given on twitter account and get permission to extract the tweets from twitter.
Step3:- Then, produce app on developer.twitter.com. We have used three apps for this work. After producing app, twitter platform provides some keys: apikey, api secret key, access token, and access token secret key. These keys are important for obtaining tweets.

DOI 10.17148/IARJSET.2020.7110

Step4:- Load the library twitteR on R compiler and writes the keys with setup_twitter_oauth ( ) functions, the keys provided by twitter platform. Once running the code on R compiler, it'll provide the path for tweets.

Step5:- Write a function is SearchTwitter ( ) which is given by twitteR package to seek out tweets regarding specific person. Example: - searchTwitter ("narendra modi", n = 500, lang= "eng"). Three parameters are utilized in this function name, "n" suggests that number of tweets, "lang" suggests that language of that tweets are to be shown.

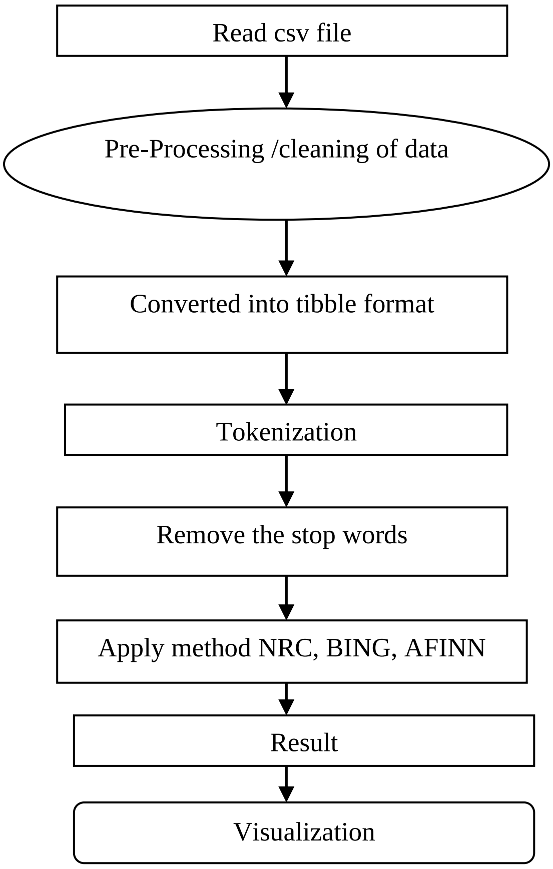
Step6:- Tweets are shown in list format on R compiler. Then, list will be converted in to a data frame using twListTODF ( ) function for further use.

Part 2: Data pre-processing:- It is extremely vital for data pre-processing. During this half, cleanup of information is performed. Once tweets are extracted from twitter in part 1, then those tweets are shown in unreliable format. It contains hyperlinks, numbers, @ sign, punctuation marks etc. We tend to take away these garbage collections from tweets using R function Gsub ( ) that removes hyperlinks, digits, @ signs, punctuation marks from data. In data pre-processing, data are converted from upper case to lower case also.
Example: - before cleaning

RT @SurjyaDas6: @KarunaduUpdates @INCKarnataka @AgentSaffron @CPBlr @BlrCityPolice Good step forward to eradicate fake story paddlers.

After cleaning good step forward to eradicate fake story paddlers

Part 3: Tidytext lexicon approaches: - In this part, we apply three lexicons on pre processed dataset. Sentiment classification depends on insight that the polarity of a part of text can be obtained from the ground polarity of the words which compose it in lexicon based methods [7]. Lexicon based approaches contains bag of words unit in one place like wordbook or dictionary. So, some researchers also known as wordbook or dictionary based approach. In this approaches, dictionary words are compared with dataset words to calculate positive, negative, neutral polarity. R language provides a Tidytext package that contains three lexicons NRC, BING and AFINN for this work. Tidy data sets allow manipulation with a standard set of "tidy" tools, including popular packages like dplyr define by Wickham and Francois 2016, tidyr developed by Wickham 2016, ggplot2 specify by Wickham 2009, and broom by Robinson 2017 [8].When we use tidy text package some steps are taken below:


Read csv file
Pre-Processing /cleaning of data
Converted into tibble format
Tokenization
Remove the stop words
Apply method NRC, BING, AFINN
Result
Visualization

DOI 10.17148/IARJSET.2020.7110

➢ Firstly convert dataset into tibble format using dplyr package. Tibble format is updated version of data frame by which data are shown in table format.

➢ Sentences are split in to tokens; token also called as word which means sentences broken in to words, and this process is called tokenization in tidy text mining.

➢ After splitting sentences, we remove stop words from tokens. Tidytext structure contains own stop words. Stop words are those words that are utilized in making sentence like ‘the’, ‘of’,’ ‘about’ and many more.

➢ Now, use three lexicons which are contain in tidytext package. Three lexicon names are NRC, AFINN, and BING. NRC lexicon developed by Saif Mohammad and Peter Turney, that specify10 type’s of emotions such as positive, negative, fear, anger, sadness, disgust, trust, anticipation, joy, surprise. BING lexicon developed by Bing Liu and collaborators, that turn out two emotions positive and negative, and AFINN lexicon developed by Finn Arup Nielsen, specify two emotions like positive and negative but it produce the results in 5 to -5 numerical digit range if produce 5,3,2,1 i.e. called positive emotion and if produce result -1,-2,-3,-4,-5 then call negative emotion [8].

➢ Results are shown in Visual image using ggplot2 package which is taken from R language.

Part 4: Result And Visual image: - In our analysis, first we analyze the tweets regarding Sh. Narendra Modi, and Sh. Rahul Gandhi by which we have to discover the popularity of both politicians in 2019 elections. We tend to use three lexicons NRC, BING, and AFINN.

Table of percentage of NRC emotions table 1:-

Sentiment name	Sh. Narendra Modi	Sh. Rahul Gandhi
Positive	23%	17%
Negative	15%	16%
Trust	14%	15%
Anger	9%	11%
Anticipation	8%	11%
Fear	8%	8%
Joy	8%	7%
Sadness	6%	7%
Surprise	4%	6%
Disgust	4%	3%
Neutral	60%	61%

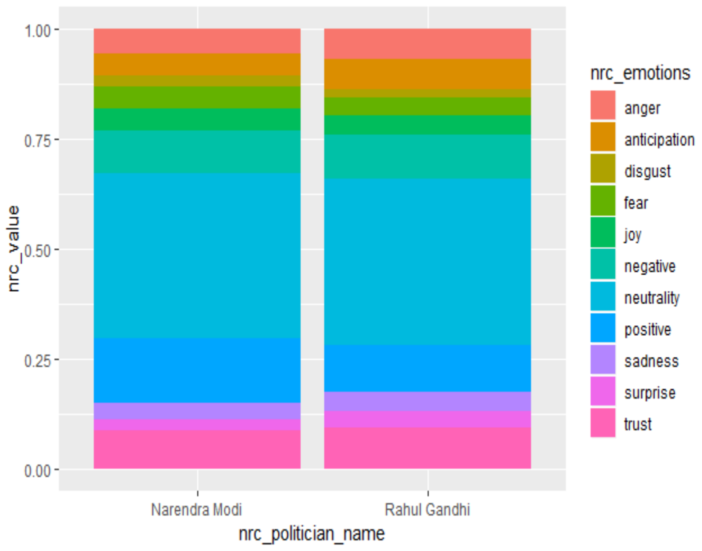


Figure 1. NRC classification of tweets

DOI 10.17148/IARJSET.2020.7110

BING emotions table 2:

Sentiment name	Sh. Narendra Modi	Sh. Rahul Gandhi
Positive	45%	33%
Negative	55%	67%
Neutral	92%	91%

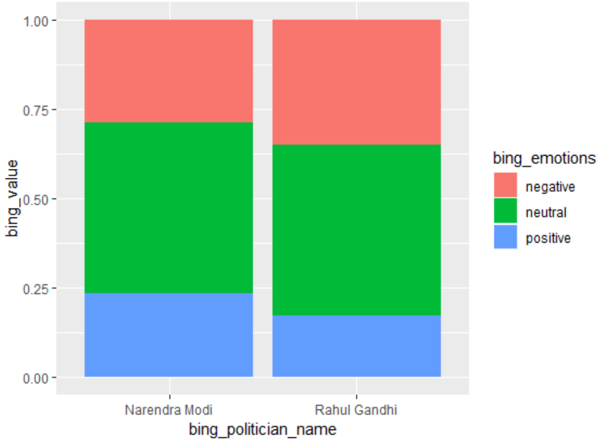


Figure 2. BING classification of tweets

AFINN emotions table 3:-

Sentiment name	Sh. Narendra Modi	Sh. Rahul Gandhi
Positive	50%	37%
Negative	50%	63%
Neutral	93%	91%

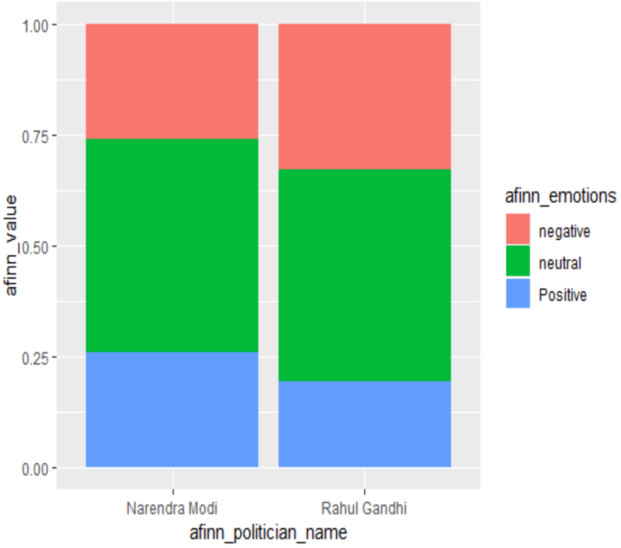


Figure 3. AFINN classification of tweets

As, we are able to see that Sh. Narendra Modi percentage is higher than Sh. Rahul Gandhi in positivity and lower in negativity. For this analysis we use 2,500 dataset of each politician. When we perform average accuracy of three

DOI 10.17148/IARJSET.2020.7110

lexicons then we use combined dataset which is 7,500.Once we perform three lexicons on combined dataset that is 7,500 so, average percentage table are going to be:-

Lexicon name	Accuracy
NRC	35%
BING	8%
AFINN	7%

This table shows NRC lexicon get highest average accuracy than BING, and AFINN. So, we are able to say Sh. Narendra Modi is popular politician in 2019 election. The aim of the visual image portion is to interpret the results of sentiment analysis diagrammatically [4]. For making bar charts, we use ggplot2 package which is provided by R language. In the visual part, we present the sentiments of both politicians in the visual images. Figure 1, show percentage of ten different emotions by NRC lexicon regarding both politicians. As you can see that, positive, and joy emotion of Sh. Narendra Modi is more rather than Sh. Rahul Gandhi whereas negative, anger, sad emotion of Sh. Rahul Gandhi is more over Sh. Narendra Modi according to NRC lexicon. Figure 2 and 3; turn out two emotions positive, negative by BING and AFINN lexicons. In both lexicons positive emotion of Sh. Narendra Modi is more over Sh. Rahul Gandhi whereas negative emotion of Sh. Rahul Gandhi is extremely higher than Sh. Narendra Modi. During this half, we also produce a wordcloud of lexicon words. Wordcloud also known a tag cloud and tag could be a single word [9]. It is a representation of text data or words which is selected from data using lexicons. So, different lexicons choose different words. Here, Wordcloud package is used from R language for creating a word cloud.

Sh. Narenadra Modi

Sh. Rahul Gandhi

disgust
fear
anticipation
joy
anger
hatred
idiot
vote
negative
sun
policy
trust
positive
surprise
sadness
disgust
fear
anticipation
joy
anger
congress
negative
raid
trust
positive
surprise
sadness

Wordcloud 1. NRC emotions

Sh. Narendra Modi

Sh. Rahul Gandhi

negative
comprehensive
love
positive
negative
love
protect
positive

Wordcloud 2. BING emotions

Sh. Narendra Modi

Sh. Rahul Gandhi

positive
love
admit
protect
negative
positive
comprehens
negative

Wordcloud 3. AFINN emotions

IV. DISCUSSION

In this study, we have collected 2,500 tweets regarding three Indian leaders Sh. Narendra Modi, Sh. Rahul Gandhi, and Sh. Akhilesh Yadav. But, we find the popularity of only two politicians Sh. Narendra Modi, and Sh. Rahul Gandhi in 2019 elections. To find out the popularity, we have used tidytext lexicons approaches NRC, BING, AFINN. We are able to see that in section III, Sh. Narendra Modi positive emotions in three lexicons is higher and negative emotions is less than Sh. Rahul Gandhi. After that the results of section III, we can say that Narendra Modi is famous Indian leader in 2019 election. We have also discovered a lot of information related to three lexicons. These three lexicons don’t perform neutrality of words. Neutral words are those words that do not exist in positive, negative and all other emotions. That means, those words means nothing and polarity score is 0. Therefore, we find the neutrality of tweets is manual. To find out the neutrality, we performed three lexicons on 2,500 dataset of Sh. Narendra Modi and Sh. Rahul Gandhi. First we calculate the total words of both dataset. The total calculated words are 44808 of Sh. Narendra modi dataset and 46469 of Sh. Rahul Gandhi dataset. After that we perform the cleaning process on the words then the total words are 24,262 of Sh. Narendra modi dataset and 24288 of Sh. Rahul Gandhi dataset. Now we apply NRC, BING, and AFINN lexicons on cleaned words for classification. So, total classify words are 9725 of Sh. Narendra modi dataset and 9574 of Sh. Rahul Gandhi dataset by NRC lexicon,1869 and 2301 by BING lexicon, and 1810 and 2218 by AFINN lexicon out off 24262 and 24288 words. The NRC lexicon not found 14,537 words from Sh. Narendra modi dataset which are 24,262 and 14,714 words from Sh. Rahul Gandhi dataset which are 24,288. Like NRC lexicon, BING and AFINN lexicons are also not able to found the words from both dataset; the words are not found by BING lexicon are 22,393 and 21,987, while words are not found by AFINN lexicon are 22,452 and 22070 from 24,262 and 24288 words. So, we finally calculate the neutrality of both dataset by performing a percentage calculation on total find words and words which are not found by three lexicons. The neutrality produce by NRC lexicon is 60% of Sh. Narendra modi dataset and 61% of Sh. Rahul Gandhi dataset. Like NRC, the BING lexicon produce neutrality is 92% and 91%, and AFINN lexicon is 93% and 91%. Thus, we can say that the upper neutrality produce by AFINN and BING lexicons which are not classify the mostly words. The results of neutrality is shown above, we can say that which words are neutral in NRC lexicons there is not necessary those words are also neutral words in BING and AFINN lexicon. As we can see in above calculation the NRC lexicon classifies more words and less neutrality comparison to BING and AFINN. After that, we also find the average accuracy of three lexicons. So, we use combine dataset then the total collection of tweets is 7,500. In Previous section, we have seen that NRC lexicon has 35%, BING lexicon 8% and AFINN lexicon 7% accuracy. So, NRC lexicon has higher accuracy instead of BING and AFINN. Therefore, we can say that the results of sentiment analysis on tweets are accurate which is produced by the NRC lexicon.

V. CONCLUSION AND FUTURE WORK

In this research paper, we aim to analyse people sentiments regarding Indian politicians that mean what people think about Sh. Narendra modi and Sh. Rahul Gandhi. Therefore, we collect tweets from twitter social media platform on both politicians from 2014 to 2018 so that we nalyse about people thinking and which leader is most likeable by pubic. By which we predict which politician government won the 2019 elections. So, we calculate percentage of positive sentiments of both politicians by which we get a popular leader of 2019 elections. For analysis of sentiments, we use Tidytext lexicons approaches. Here from this work we get two conclusions from this work first; NRC lexicon is best approach in lexicon based approaches. So, NRC lexicon results will be termed as accurate for sentiment analysis. Secondly, Narendra Modi was popular leader in 2019 elections by NRC lexicon and we have seen in 2019 elections that Sh. Narendra modi government was won the election in 2019. So, our analysis about people emotions is more accurate according to twitter data. In future work, researchers can collect more tweets for more analysis and also increase the accuracy of methods. R language provides additional lexicons or dictionaries and packages that are used for further work. Lexicon based approaches don’t perform training and testing on dataset therefore, the accuracy isn’t high. If we are able to increase accuracy then machine learning algorithms should be used.

REFERENCES

[1]. Abdullah Alsaeedi, Mohammad Zubair Khan “A Study on Sentiment Analysis Techniques of Twitter Data” Department of Computer Science, College of Computer Science and Engineering Taibah University Madinah, KSA.2019

[2]. Arvind Singh Raghuwanshi M.Tech, Satish Kumar Pawar Asst. prof. “Polarity Classification of Twitter Data using Sentiment Analysis” Computer Science and Engineering Department Samrat Ashok Technological Institute Vidisha(M.P.),India.

[3].Prabu Palanisamy, VineetYadav and Harsha Elchuri “Serendio:Simple and Practical lexicon based approach to Sentiment Analysis”Serendio Software Pvt Ltd Guindy, Chennai 600032, India

[4]. Malak Abdullah, mirsad hadzikadic “Sentiment analysis of twitter data: Emotions revelead regarding Donald trump during the 2015-2016 primary debates” college of computing and informatics university of north Carolina at charlotte, north Carolina,U.S.

[5]. Parul Sharma, Teng-Sheng Moh “ Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter” Department of Computer Science San Jose State University San Jose, CA, USA

[6]. https://sites.google.com/site/miningtwitter/basics/getting-data

[7]. Cataldo Musto, Giovanni Semeraro, Marco Polignano “A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts” Department of Computer Science University of Bari Aldo Moro, Italy

[8]. https://www.tidytextmining.com/, Text Mining with R book -A Tidy Approach Julia Silge and David Robinson.

[9]. https://www.r-graph-gallery.com/