

Sentiment Analysis of Tweets using Naïve Bayes Algorithm through R Programming

Annie Syrien¹, M. Hanumanthappa², B. Sundaravadivazhagan³

¹ Department of Computer Science and Applications, Bangalore University, Bengaluru, India

² Department of Computer Science and Applications, Bangalore University, Bengaluru, India

³ Department of Information Technology, Al-Musanna College of Technology, Sultanate OMAN

*Corresponding Author: syrien01@gmail.com

Available online at: www.ijcseonline.org

Accepted: 12/Nov/2018, Published: 30/Nov/2018

Abstract — The enlargement in development of web 2.0 and web enabled devices smacked huge user generated data hence attracted many researchers in the past years in the field of social media mining. The focal point in mining social media is for obtaining the important decision making opinions, attitudes, sentiments, and emotions. This paper uses naïve bayes algorithm to classify the sentiments and polarity of the tweets of Bengaluru traffic in detail with the help of opinion lexicon through R studio. The tweets on Bengaluru traffic are first accessed from twitter through streaming API, then preprocessed and functions containing naïve bayes classifier is used to classify the tweets into emotions and polarity, through classify emotions and classify polarity functions. Classify emotions functions makes use of naïve bayes algorithm for classifying the emotions into seven categories such as anger, disgust, fear, joy, sadness, surprise, and best fit. Classify polarity function receives two arguments, cleaned tweets and naïve bayes algorithm for classifying the polarity into positive sentiment and negative sentiment. The results are represented through plots in R studio.

Keywords— Sentiment analysis, Naïve bayes, R programming, Data mining and Polarity detection.

I. INTRODUCTION

Social media is vastly maneuvered in the era of internet [1]. Data generated from these social networking sites are large [2] and can be used for marketing, prediction, or sentiment analysis [3]. Social media mining is attracting larger business segment and researchers due to its rapid growth of data size and people interest [4]. Opinion mining and sentiment analysis mainly focuses on what people feel and think about products, persons, services, and business entities [5] [6]. These opinions contributes important role in decision making process in any organization or business segment.

Social media mining is obtaining functional information from the user missive. Social media mining is gravitated research topic which attracted many researchers because of the elevation of number of users in social media sites, whereas sentiments of the users are simply speculated from their shares, likes, and clicks. Individual's sentiments like positive or negative helps in making major decisions with regard to product and services by empowering the customer to share their view point [7].

Micro blogging sites has become conveniently way of expressing opinions of the users [8]. Twitter is one of the famous micro blogging website [9] [10] [11] where users express their opinions easily. Twitter daily generates 200 million tweets on various topics [12]. Sentiment analysis on twitter data provides real time monitoring of users feelings [13]. The objective of this research work is to extract tweets online from twitter streaming API, preprocess and analyze their sentiments through naïve bayes classifier in R studio. The comprehended results will be represented via plots in R studio.

The paper is organized as follows. Section II related works are discussed. Section III gives the methodology of the work constituted. In section IV conclusion and recommendations is discussed.

A. Social Media

Social media is an internet based communication aid which enables people to share facts, news, data, and information. The term social media can be better understood by defining social and media. Social implies relating to the society, connecting with the people and disbursing time to build and

enhance the relationships, media implies aid for information transmission like newspaper, TV, radio, Internet so on., here emphasis is internet as social media is an electronic podium for fraternizing people.

It is an effective technique to spread information and express sentiments. Examples of social media sites are Twitter, Facebook, YouTube, LinkedIn, Digg etc. Recent advancements of cellular phones [14] and internet development lead into the situations of tranquil access to social networking sites anytime [15], hence the sharing of views from micro aspects to macro aspects increased gradually, people developed phubbing leading mass data production online every second, whereas these data can be excavated to produce useful information.

B. Data Mining

Data mining is a Step in KDD process for analyzing data and producing specific patterns. Data mining is one of the major fields, which is attracting many researchers due its large applications and challenges. The research on data mining has produced successfully innumerable tools, algorithms for solve real-world problems. Data mining on social networking sites has been most attracted research topics [16].

Data mining research has successfully produced numerous methods, tools, and algorithms for handling large amounts of data to solve real-world problems. The Figure 1 shows the different steps involved in data mining process.

Data mining is excavating the large set of raw data in pursuance of useful knowledge, which has different stages such as

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Patter evaluation
- Knowledge presentation

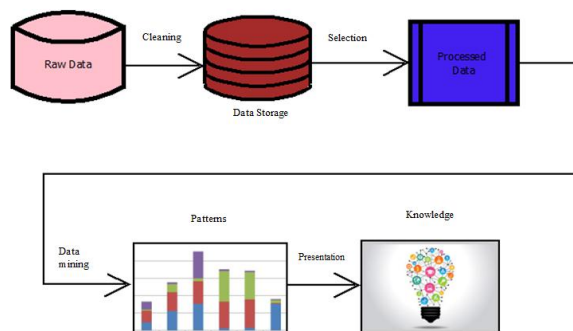


Figure 1: Data Mining Process

C. R Programming

R was created by Ross Ihaka and Robert Gentleman. R is powerful, effective statistical tool. It is open-source and implemented in the language S. R has adopted Object oriented principle polymorphism. R provides flexibility to work with programming languages like Java, Python, C/C++ and databases such as ODBC, MySql and Packages like SAS. R contains 7000+ packages which can be imported.

II. RELATED WORK

Sentiment analysis in social media has enchanted many researchers' from past few years [17]. Sentiment analysis extracts opinions from the user tweets [18]. There are several researchers pivoted their attentions on social media mining through experimenting various tools [19]. David Zimbra et al [20] presented the research work in brand related twitter sentiment analysis through feature engineering and dynamic architecture for Artificial neural network, which addressed the challenges related to twitter language and sentiment expressions which was very essential to identify the sentiments of the tweets. They referred to the star bucks brand related tweets for processing, they conducted two sets of experiment, five-class and three-class tweet sentiment classification. They used rule based approach, which yielded accuracies above 80% in both five-class and three-class tweet mild sentiment classification compared to Chinese algorithm from the different

Huma Parveen et al [21] worked on Sentiment Analysis on Twitter Data set using Naïve Bayes Algorithm, they have used Hadoop framework for movie data set preprocessing from twitter. Reviews, comments, and feedback were the attributes focused upon. Their results were presented as positive, negative, and neutral sentiments about the data set. Their methodology involved the following tasks.

The data was collected through twitter API and stored in HDFS (Hadoop Distributed File System), preprocessed, SentiWordNet dictionary was downloaded, both dataset and SentiWordNet was fed to mapper, naïve bayes algorithm was split into two phases, map and reduce phase.

The map phase contained two major tasks.

First, produce a hash map for replenishment of polarity of the tweets.

Secondly, exercise the overall polarity of the tweets through naïve bayes algorithm.

The reduce phase detected the polarity on positive, negative, extreme positive, extreme negative, and neutral and output was displayed on HDFS.

Mondher Bouazizi et al [22] contributed significant work on Sentiment Analysis: from Binary to Multi-Class Classification, A pattern-based approach for multi-class sentiment analysis in twitter. The research work focused on classifying the sentiments from binary (positive and negative) to seven-class classification (happiness, sadness, anger, love, hate, sarcasm, and neutral). Random Forest classifier is used for the proposed work. They generated results of multi-class are compared with binary classification and ternary classification. F-measure is used to compare the accuracy between binary, ternary and multi-class classifications. Finally confusion matrix is used to represent the multi-class classification values.

Neetu M S et al [23] presented the work on Sentiment Analysis in Twitter using Machine Learning Techniques, which comprised of naïve bayes, SVM, maximum entropy, and ensemble classifier for sentiment classification in Matlab. The performance of these classifiers were shown in the table, naïve bayes had better precision compared to other three classifiers, but slightly lesser accuracy and recall.

III. METHODOLOGY

The statistical package R Studio is used for data collection and proposed work. The data from twitter is collected through twitter streaming API by establishing secure authorization. R Studio served as fast, powerful, and effective tool for data collection and text preprocessing. Data preprocessing involved the major task of removing noise such as detaching user mentions, special characters, hashtags, and URL's. Once the data is cleaned simple naïve bayes algorithm with the help of sentiment package in R is applied to evaluate the sentiments of the tweets into emotions and polarity, through classify emotions and classify polarity functions, where classify emotions function is shown in the figure 4. Classify emotions functions makes use of naïve bayes algorithm for classifying the emotions into seven categories such as anger, disgust, fear, joy, sadness, surprise, and best fit. Classify polarity function receives two arguments, cleaned tweets and naïve bayes algorithm for classifying the polarity into positive sentiment and negative sentiment. The results are represented through plots in R studio as shown in figure 7. The figure 5 represents lexicon with each word having a different adjective. Figure 6 shows the implementation of classify emotion and classify polarity functions in the R studio. The flow of the work is represented in the figure 2. Naïve bayes, Maximum entropy, SVM, and ensemble classifier are the other data mining techniques available for classification of tweets in sentiment analysis, though naïve bayes is simple classifier it gives better precision and accuracy [24][25] compared to other classifiers and vastly used in the sentiment analysis.

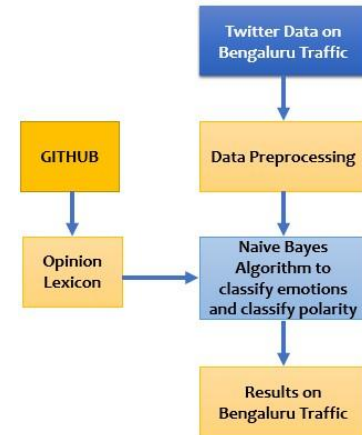


Figure 2: Sentiment Analysis of Bengaluru Traffic

A. Data Set

The data set used is obtained from twitter streaming API on bengaluru traffic, 2000 tweets were requested but streaming API returned 725 tweets, the number of tweets the API returns based on the tweets tweeted on the twitter, for some experiments the streaming API has also returned nearly 2000 tweets. Every user tweets contains 140 characters of length, which comprises of text, hashtag, a shortened URL, username along with video or image as shown in the figure 3. Generally tweet contains the metadata about the tweet, when, who sent so on. However, the twitter has doubled the size of user tweet from 140 to 280 recently.



Figure 3: A sample tweet

Profile Picture: Every twitter user may or may not have a profile picture, this appears first on the tweet.

User name: Username starts with @ symbol. Users are identified with username. In the above Figure 1 the user name is @Moto_IND

Hash tag: Hash tags starts with # sign and used to organize the updates for twitter search engines.

Retweet: Retweet is used to mention that user is posting someone else post. It is abbreviated as RT, the format is RT @username, and here the username is twitter name of the person who's retweeted.

Tweet: Tweet is twitter update, user's space to share his thoughts or messages and so on

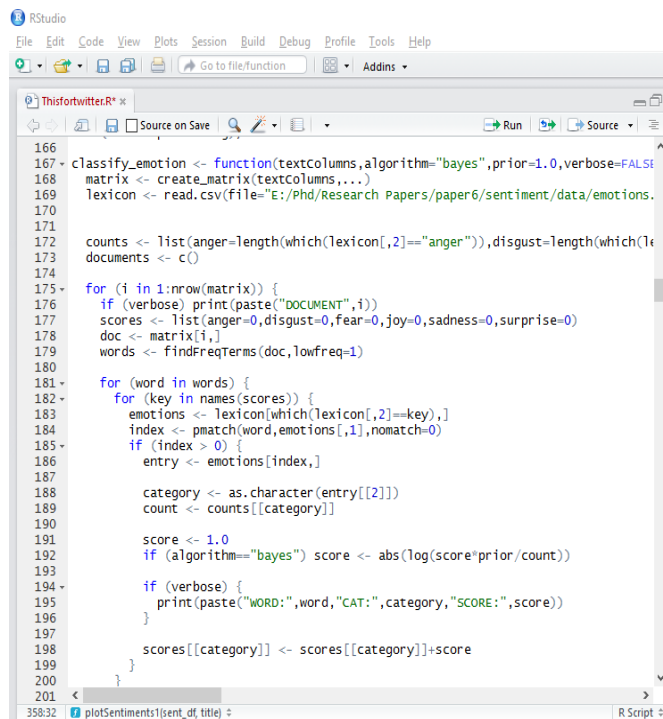
The Table 1 refers to the sample tweets downloaded from online through Twitter streaming API.

TABLE1: USER TWEETS

@sampad @Satyajee Full disclosure required. If Delhi kills you by pollution Bangalore kills you by traffic.
@nagraik_arzoo I arrived 15 days back and I already see the difference in traffic between Delhi and Bangalore.
@HithendrarR Absurd demand! Trust Bangalore Traffic police will not buckle to the pressure but book violators
#Bangalore traffic is so bad even google maps is giving up!
@simpleharish once I reached from Delhi to Bangalore on a flight by 730 pm or so and reached home at 11 midnight

B. Naïve bayes

Naïve bayes is a probabilistic algorithms with probability theory and bayes' theorem to predict the type of sample like customer review, sentiment of the product so on. Probabilistic indicating for a given sample the calculation takes place, category with the highest one is given as output based on bayes theorem, which describes the probability on prior knowledge of conditions that is related to the feature. Positive, negative words, and tweets from the twitter classes are computed through probability based on bayes theorem. Despite naïve bayes being the simplest probabilistic classifier algorithm it generates astonishing results.

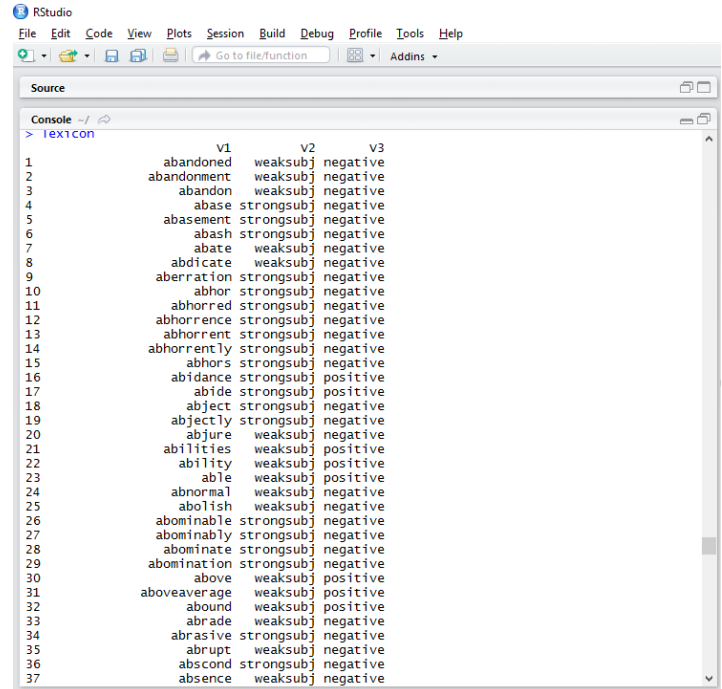


```

166
167 classify_emotion <- function(textColumns,algorithm="bayes",prior=1.0,verbose=FALSE
168 matrix <- create_matrix(textColumns,...)
169 lexicon <- read.csv(file="E:/Phd/Research Papers/paper6/sentiment/data/emotions.
170
171
172 counts <- list(anger=length(which(lexicon[,2]=="anger")),disgust=length(which(1
173 documents <- c()
174
175 for (i in 1:nrow(matrix)) {
176   if (verbose) print(paste("DOCUMENT",i))
177   scores <- list(anger=0,disgust=0,fear=0,joy=0,sadness=0,surprise=0)
178   doc <- matrix[i,]
179   words <- findFreqTerms(doc,lowFreq=1)
180
181   for (word in words) {
182     for (key in names(scores)) {
183       emotions <- lexicon[which(lexicon[,2]==key),]
184       index <- pmatch(word,emotions[,1],nomatch=0)
185       if (index > 0) {
186         entry <- emotions[index,]
187
188         category <- as.character(entry[[2]])
189         count <- counts[[category]]
190
191         score <- 1.0
192         if (algorithm=="bayes") score <- abs(log(score*prior/count))
193
194         if (verbose) {
195           print(paste("WORD:",word,"CAT:",category,"SCORE:",score))
196         }
197
198         scores[[category]] <- scores[[category]]+score
199
200       }
201     }

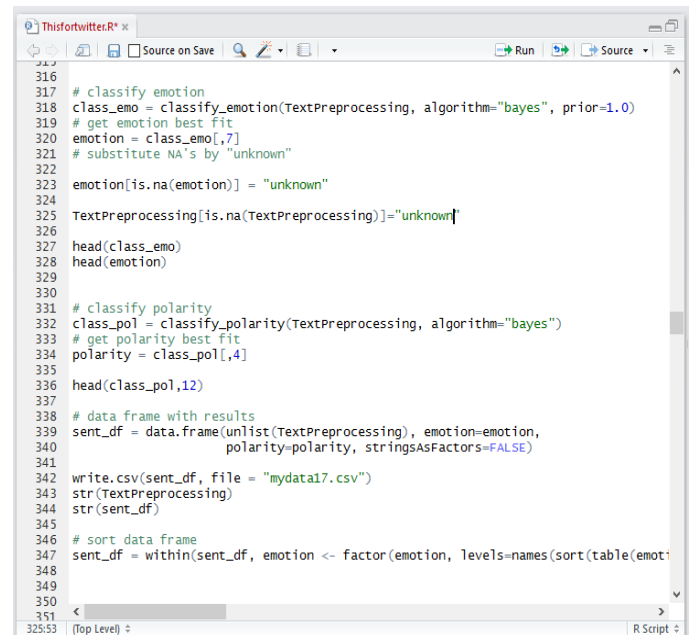
```

Figure 4: Classify_emotion function



	v1	v2	v3
1	abandoned	weaksbj	negative
2	abandonment	weaksbj	negative
3	abandon	weaksbj	negative
4	abase	strongsubj	negative
5	abatement	strongsubj	negative
6	abash	strongsubj	negative
7	abate	weaksbj	negative
8	abdicate	weaksbj	negative
9	aberration	strongsubj	negative
10	abhor	strongsubj	negative
11	abhorred	strongsubj	negative
12	abhorrence	strongsubj	negative
13	abhorrent	strongsubj	negative
14	abhorrently	strongsubj	negative
15	abhors	strongsubj	negative
16	abidance	strongsubj	positive
17	abide	strongsubj	positive
18	abject	strongsubj	negative
19	abjectly	strongsubj	negative
20	abjure	weaksbj	negative
21	abilities	weaksbj	positive
22	ability	weaksbj	positive
23	able	weaksbj	positive
24	abnormal	weaksbj	negative
25	abolish	weaksbj	negative
26	abominable	strongsubj	negative
27	abominably	strongsubj	negative
28	abominate	strongsubj	negative
29	abomination	strongsubj	negative
30	above	weaksbj	positive
31	aboveaverage	weaksbj	positive
32	abound	weaksbj	positive
33	abrade	weaksbj	negative
34	abrasive	strongsubj	negative
35	abrupt	weaksbj	negative
36	abscond	strongsubj	negative
37	absence	weaksbj	negative

Figure 5: lexicon with every word categorized as positive and negative



```

316
317 # classify emotion
318 class_emo = classify_emotion(TextPreprocessing, algorithm="bayes", prior=1.0)
319 # get emotion best fit
320 emotion = class_emo[,7]
321 # substitute NA's by "unknown"
322
323 emotion[is.na(emotion)] = "unknown"
324
325 TextPreprocessing[is.na(TextPreprocessing)]="unknown"
326
327 head(class_emo)
328 head(emotion)
329
330
331 # classify polarity
332 class_pol = classify_polarity(TextPreprocessing, algorithm="bayes")
333 # get polarity best fit
334 polarity = class_pol[,4]
335
336 head(class_pol,12)
337
338 # data frame with results
339 sent_df = data.frame(unlist(TextPreprocessing), emotion=emotion,
340 polarity=polarity, stringsAsFactors=FALSE)
341
342 write.csv(sent_df, file = "mydata17.csv")
343 str(TextPreprocessing)
344 str(sent_df)
345
346 # sort data frame
347 sent_df = within(sent_df, emotion <- factor(emotion, levels=names(sort(table(emot
348
349
350
351
352
353

```

Figure 6: Functions implementing classify_emotion and classify_polarity

IV. RESULTS AND DISCUSSION

The figure 7 shows the results of emotion categories, the number of tweets obtained matched with the opinion lexicon through classify emotion function using naive bayes and

found that many expressed different emotions, which is categorised such as anger, disgust, fear, joy, sadness, and surprise. The graph shows the emotion joy as elevated emotion, but when compared to other similar emotions such as anger, disgust, fear, sadness, surprise, the expression joy remains lesser count indicating people are not happy with the traffic situation in the Bengaluru.. The Figure 8 illustrates how tweets are polarized into positive, negative, pos/negative and best fit, where as positive polarity turns to be more compared to negative due to the more positive keywords even though emotions expressed are negative.

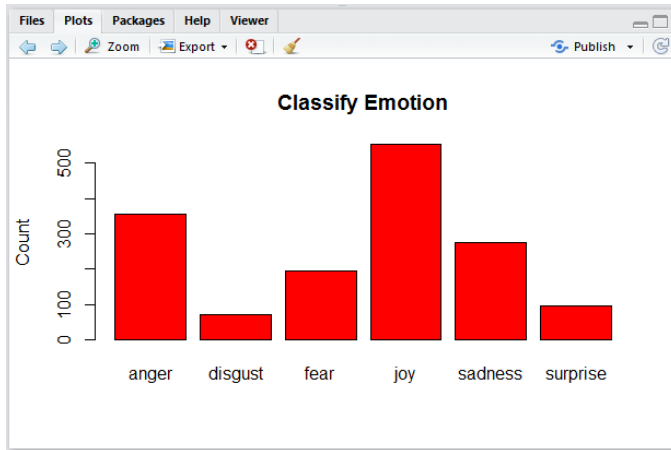


Figure 7: Plot representing the sentiments of tweets

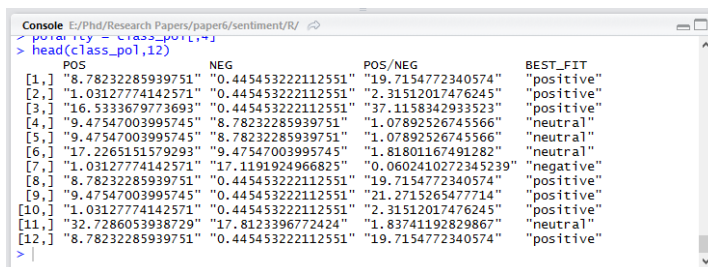


Figure 8: Output of Class Polarity function

V. CONCLUSION

The classification of tweets into positive, negative and various emotions was carried out through R Studio which is an open source and free IDE. Sentiments of tweets is estimated through simple naïve bayes algorithm, the data set on Bengaluru city traffic is used. R is a constructive tool for data analysis. Even though there are great evolutions in machine learning algorithms in the last few years, it is proven that naïve bayes is not only simple, but fast, reliable and accurate.

REFERENCES

- [1] Peerapon Vateekul and Thanabhat Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," in Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on, 2016, pp. 1-6.
- [2] Zahra and Jalali, Mehrdad Rezaei, "Sentiment Analysis on Twitter using McDiarmid Tree Algorithm," in 7th International Conference on Computer and Knowledge Engineering (ICCKE 2017), October 26-27 2017, Ferdowsi University of Mashhad, 2017, pp. 33-36.
- [3] Monu Kumar and Anju Bala, "Analyzing Twitter sentiments through big data," in Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, 2016, pp. 2628-2631.
- [4] Zhou Jin, Yujia Yang, Xianyu Bao, and Biqing Huang, "Combining user-based and global lexicon features for sentiment analysis in twitter," in Neural Networks (IJCNN), 2016 International Joint Conference on, 2016, pp. 4525-4532.
- [5] Manju Venugopalan and Deepa Gupta, "Exploring sentiment analysis on twitter data," in Contemporary Computing (IC3), 2015 Eighth International Conference on, 2015, pp. 241-247.
- [6] Paramita Ray and Amlan Chakrabarti, "Twitter sentiment analysis for product review using lexicon method," in Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on, 2017, pp. 211-216.
- [7] Zhao Jianqiang and Cao Xueliang, "Combining Semantic and Prior Polarity for Boosting Twitter Sentiment Analysis," in Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on, 2015, pp. 832-837.
- [8] Umme Aymun Siddiqua, Tanveer Ahsan, and Abu Nowshed Chy, "Combining a rule-based classifier with weakly supervised learning for twitter sentiment analysis," in Innovations in Science, Engineering and Technology (ICISSET), International Conference on, 2016, pp. 1-4.
- [9] Hima Suresh and others, "An unsupervised fuzzy clustering method for twitter sentiment analysis," in Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on, 2016, pp. 80-85.
- [10] Aliza Sarlan, Chayanit Nadam, and Shuib Basri, "Twitter sentiment analysis," in Information Technology and Multimedia (ICIMU), 2014 International Conference on, 2014, pp. 212-216.
- [11] Suvarna D. Tembhurnikar and Nitin N. Patil, "Topic detection using BNGram method and sentiment analysis on twitter dataset," in Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference on, 2015, pp. 1-6.
- [12] Rif'at Ahdi Ramadhani, Fatma Indriani, and Dodon T. Nugrahadi, "Comparison of Naive Bayes smoothing methods for Twitter sentiment analysis," Comparison of Naive Bayes Smoothing Methods for Twitter Sentiment Analysis, 2016.
- [13] Zhao Jianqiang, "Combing Semantic and Prior Polarity Features for Boosting Twitter Sentiment Analysis Using Ensemble Learning," in Data Science in Cyberspace (DSC), IEEE International Conference on, 2016, pp. 709-714.
- [14] Sonia Anastasia and Indra Budi, "Twitter sentiment analysis of online transportation service providers," in Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on, 2016, pp. 359-365.
- [15] Abhijit Janardan Patankar, Kshama V. Kulhalli, and Kotrappa Sirbi, "Emotweet: Sentiment Analysis tool for twitter," in Advances in Electronics, Communication and Computer Technology (ICAECCT), 2016 IEEE International Conference on, 2016, pp. 157-159.
- [16] Fotis Aisopos, Dimitrios Tzannetos, John Violos, and Theodora Varvarigou, "Using n-gram graphs for sentiment analysis: an extended study on Twitter," in Big Data Computing Service and Applications

- (BigDataService), 2016 IEEE Second International Conference on, 2016, pp. 44-51.
- [17] Chintan Dedhia and Jyoti Ramteke, "Ensemble model for Twitter sentiment analysis," in Inventive Systems and Control (ICISC), 2017 International Conference on, 2017, pp. 1-5.
- [18] Mohit Mertiya and Ashima Singh, "Combining naive bayes and adjective analysis for sentiment detection on Twitter," in Inventive Computation Technologies (ICICT), International Conference on, vol. 2, 2016, pp. 1-6.
- [19] Anurag P. Jain and Vijay D. Katkar, "Sentiments analysis of Twitter data using data mining," in Information Processing (ICIP), 2015 International Conference on, 2015, pp. 807-810.
- [20] David Zimbra, Manoochehr Ghiassi, and Sean Lee, "Brand-related Twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks," in System Sciences (HICSS), 2016 49th Hawaii International Conference on, 2016, pp. 1930-1938.
- [21] Huma Parveen and Shikha Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," in Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on, 2016, pp. 416-419.
- [22] Mondher Bouazizi and Tomoaki Ohtsuki, "Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis," in Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on, 2015, pp. 1594-1597.
- [23] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, 2013, pp. 1-5.
- [24] Omar Abdelwahab, Mohamed Bahgat, Christopher J. Lowrance, and Adel Elmaghraby, "Effect of training set size on SVM and Naive Bayes for Twitter sentiment analysis," in Signal Processing and Information Technology (ISSPIT), 2015 IEEE International Symposium on, 2015, pp 46-51.
- [25] Angelpreethi, P. kiruthika , S. BrittoRameshKumar, "A Methodological Framework for Opinion Mining" in International Journal of Computer Sciences and Engineering, 2018, Vol.6(2)