Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 152 (2019) 341–348

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Pervasive Computing Advances and Applications – PerCAA 2019

# The Impact of Features Extraction on the Sentiment Analysis

Ravinder Ahuja[a], Aakarsha Chug[a], Shruti Kohli[a], Shaurya Gupta[a], and Pratyush Ahuja[a]

[a]*Jaypee Institute of Information Technology, Noida 201301, India*

**Abstract**

In today's world, everyone is expressive in one way or other. Many social websites and android applications whether being Facebook, WhatsApp or Twitter, in this highly advance and the modernized world is flooded with views and data. One of the most global and popular platforms is Twitter. This is seen as the main source of sentiments where almost every enthusiastic or social person tends to express his or her views in form of comments. These comments not only express the people but also give the understanding of their mood. Text present on these medias are unstructured in nature, so to process them firstly we need to pre-process, six pre-processing techniques are used and then features are extracted from the pre-processed data. There are so many feature extraction techniques such as Bag of Words, TF-IDF, word embedding, NLP(Natural Language Processing) based features like word count, noun count etc. In this paper we analysed the impact of two features TF-IDF word level and, N-Gram on SS-Tweet dataset of sentiment analysis. We found that by using TF-IDF word level (Term Frequency-Inverse Document Frequency) performance of sentiment analysis is 3-4% higher than using N-gram features, analysis is done using six classification algorithms(Decision Tree, Support vector Machine, K-Nearest Neighbour, Random Forest, Logistic Regression, Naive Bayes) and considering F-Score , Accuracy, Precision, and Recall performance parameters.



*Keywords:* Accuracy; Classification; N-gram; TF-IDF; Sentiment; Twitter

## 1 Introduction

Due to increase in contents over social media such as Twitter, Facebook, and Trip advisor, expressing opinion about products, services, or any government policy among others. Twitter having 336 million[1] active users monthly is now a main source of feedback for government, private organization, and other service providers. On Twitter around 500 millions tweets are produced per day[2], generating huge amount of unstructured text data. Text classification is the process of processing the text data generated on social media for various applications such as email categorization, web search, topic modeling, and information retrieval. Sentiment analysis (opinion Mining) is used to retrieve the insight information from the tweets posted by users. Twitter sentiment is used to classify the tweets into neutral, positive, or negative. Many researchers have presented classification methods in sentiment analysis [19, 20].

* Corresponding author. Tel.: +91-9582073264; fax: +91120-2400986.
*E-mail address:* ahujaravinder022@gmail.com


10.1016/j.procs.2019.05.008

First step in sentiment classification is to preprocess the text, this process will make the unstructured data present on the web containing noise in such a form that can be used for classification. Preprocessing involves tasks such as tokenization, stop word removal, lower case conversion, stemming, removing numbers etc. Next stage is to features extraction. There are different types of text features such as count vectors, bag of words, TF-IDF, word embeddings, NLP based. Next stage is to select the features, generally mutual information, information gain, chi-square, Ginni index is used. Final stage is to apply machine learning algorithms such as support vector machines, decision trees, naïve bayes, artificial neural network etc. for classification.

Some researchers have analyzed the impact of pre processing techniques on sentiment analysis from Twitter data. The paper will address the impact of different features (TF-IDF and Bag of Words) on the performance of sentiment analysis. six pre-processing techniques are applied and further two types of features(TF-IDF and BOW) are extracted from the text after preprocessing than six classification techniques have been applied to identify which features is better.

The remaining paper is structured as: Section 2 contains existing word done in sentiment analysis, proposed methods is given in section 3 describes, section 4 contains classification algorithms used, performance parameters are presented in section 5, Section 6 contains results followed with conclusion and future work in Section 7.

## 2. Literature Survey

The paper [19], impact of pre-processing is analyzed. Tweets considered are full of symbols, unidentified words, abbreviation. URLs, punctuation, user mentions, stop words were removed, and they find out the importance of slang words and spelling correction and SVM classifier is used in their experiment. In this paper [2] Vector representations have been utilized for Natural Language processing tasks. Authors here targeted on utilizing the efficiency of word vector representations for providing the solution to sentiment analysis problem. Three tasks, of retrieval of sentiment words, the polarity of sentiment words identification, and forecasting text sentiment, have been given the primary importance. They scrutinized the potency of vector representations over unique text data and checked the quality of vectors depending upon different domains. The representations have been also used to calculate various vector-based features to provide and check effectiveness. They state that they have achieved F1_score and the accuracy to be 85.77% and 86.35% respectively for text sentiment analysis for APP reviews. In paper [3], impact of preprocessing was investigated on movie reviews dataset. They have considered removal of stop words, removing negations, removal of non English letters, stemming, and prefix ‘NOT_’, considering SVM classifier. In paper [4], authors have considered four datasets HCR, Sentiment140 (only 3000 Tweets), Sanders, and Stanford 1k and considered Bag of words features, lexicon based features and Part of Speech based features. They have applied three machine learning techniques namely SVM, Logistic Regression, Naïve Bayes, and ensemble of these. In this paper [5] old approaches for sentiment analysis of short text lack the dependence of emotion words and modifiers and simply collect the sentiment of the sentence to seek the sentiment of short text, they managed to mitigate the difficulties through sentiment structure and the sentiment computation norms. In the paper, the proposed approach shows how the dependency parsing deduces the sentiment structure with the relational migration and adjusted distance, which provides good contribution to knowing the sentiment of short text. The sentiment of short text is gathered as per the distinct influence of mappings between the modifier and the emotion word. Their experiment results ensure the effectiveness of the approach they actually proposed for mitigating the problems through sentiment structure. In paper [6] authors considered text (feedback of e-learning) in Greek and extracted part of speech features and text based features are extracted and evaluated the impact of these features on the performance of sentiment classification. In paper [7] Joseph D. Prusa applied ten different feature selection techniques and four classifiers. They find out that using feature selection technique will improve the performance of sentiment analysis. In paper [21], authors have applied three levels of feature extraction techniques. They have used SVM, J48, and Naïve Bayes classifiers. In paper [22], authors have surveyed different classification techniques with different feature selection technique without any implementation. In paper [23], authors have considered dataset of Twitter (total 1000 comments) and applied various machine learning approach and ensemble approach (majority voting) to classify the comments. They have used twitter specific features as an input to classifier for classification. In paper [24], authors have access tweets about Samsung galaxy phone by using API and classified the tweets into positive, negative, and neutral categories. In paper [25], authors have induced decision rule based on Samsung

galaxy G5 product reviews on the website of Samsung using LEM2 rough set algorithm. This will help business analyst to understand product in different dimensions with different attributes and association between them.

## 3. Proposed Approach

As shown in figure 1, we have taken firstly SS-Tweet dataset than we applied six pre processing techniques on the dataset and extracted features using N-grams and TF-IDF techniques. In the next step we applied seven classification techniques and evaluated four parameters.

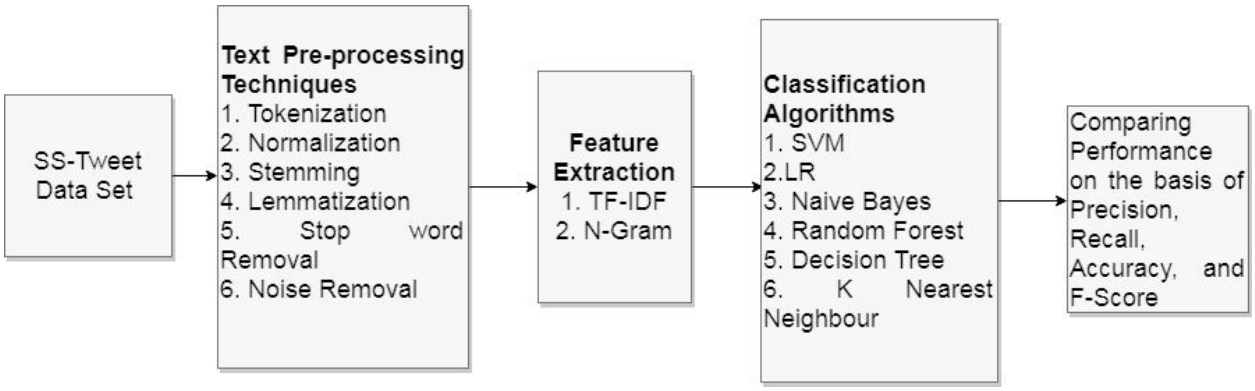


Figure 1: Proposed Methodology

### *3.1 Dataset – SS Tweet*

SS -TWEET stands for Sentiment Strength Twitter Dataset. This dataset is annotated manually. It contains total 4242 tweets, 1037 are negative tweets, 1953 are neutral tweets, and 1252 are positive tweets.

### *3.2 Pre-processing Techniques*

#### *3.2.1 Tokenization*

This step breaks the large paragraphs called chunks of text is broken into tokens which are actually sentences. These sentences can further be broken into words. For example, consider the sentence, before tokenization the it is? PhD is a tuff job to do and after tokenization it comes: {‘?', ‘ PhD',' is', ‘tuff’, 'job', 'to',' do'}

#### *3.2.2 Normalization*

There are many tasks performed simultaneously to achieve normalization. It includes the conversion of all text to either upper or lower case, eliminating punctuations and conversion of numbers to their equivalent words. This increases the uniformity of preprocessing on each text.

#### *3.2.3 Stemming*

The stemming process is used to change different tenses of words to its base form this process is thus helpful to remove unwanted computation of words. For example: fishing, fish, fisher to fish, Argue, arguing, argues to argue

#### *3.2.4 Lemmatization*

Lemmatization is the process of merging two or more words into single word. This analyzes the word morphology and eliminates the ending of the word like shocked to shock, caught to catch etc.

*3.2.5 Removing Stop Words*

Stop words refer to most common words in the English language which doesn't have any contribution towards sentiment analysis. Some of the stop words are "are", "of", "the", "at" etc. So these need to be eliminated.

*3.2.6 Noise removal*

The datasets taken comes in raw form. We have applied manual cleaning of raw data along with the use of regular expression in NLP used to eliminate noises. The noise removal is done very carefully as it sometimes eliminates a few numbers of rows of the dataset which leads to decreased accuracy. The regular expression used on datasets cleaning was able to remove unnecessary white spaces and bring data in proper columns.

**3.3 Feature Extraction**

*3.3.1 TF-IDF*

The term frequency-inverse document frequency (also called TF-IDF), is a well-recognized method to evaluate the importance of a word in a document. Term Frequency of a particular term (t) is calculated as number of times a term occurs in a document to the total number of words in the document. IDF (Inverse Document Frequency) is used to calculate the importance of a term. There are some terms like "is", "an", "and" etc. which occurs frequently but don't have importance. IDF is calculated as IDF (t) = log(N/DF), where N is the number of documents and DF is the number of document containing term t. TF-IDF is a better way to convert the textual representation of information into a Vector Space Model (VSM). Suppose there is a document which contains 200 words and out of these 200 words mouse appears 10 times than term frequency will be 10/250=0.04 and suppose there are 50000 documents and out of these only 500 documents contains mouse. Than IDF (mouse) = 50000/500=100, and TF-IDF (mouse) will be 0.04*100= 4.

*3.3.2 N-Gram*

N-Gram will form the features of text for supervised machine learning algorithms. These are sequence of n tokens from the given text. Value of n can be 1, 2, 3, and so on. If we consider the value of n to be 1 it is called unigram, for n=2, bigram and for n=3 trigram and so on. If we consider a sentence **"Jaypee is better Institute".** If we consider N=2 than it will produce "Jaypee is", "is better", "better Institute".

**4. Classification Algorithms**

*4.1 Logistic Regression*

This is a popular classification algorithm which belongs to class of Generalized Linear Models. The probabilities describing outcome of a trial is modeled using logistic regression [9]. This algorithm is also called Maximum Entropy.

*4.2 Naive Bayes*

This is powerful algorithm for classification used for classifying data on basis of probabilities. With millions of records also this algorithms works awesomely. It simply works on Bayes theorem and uses various probabilities to classify data. In Naïve Bayes class with maximum probability is considered to be as the predicted class. Naïve Bayes is also known as Maximum a Posterior Naïve Bayes has various advantages and disadvantages across

different domains. It is a fast and highly scalable algorithm. It is used on both Multiclass and Binary Classification. It can also be used on small datasets and thus also gives good results [10].

### *4.3 Support Vector Machine*

This is an efficient algorithm for regression as well classification purpose. It draws a hyperplane to separate classes. This algorithm works extremely well with regression, the effect of SVM increases as we increase dimensional space. SVM also perform well when the dimension number is larger than the sample number [11]. There exists a drawback too it does not perform well on huge datasets. SVM extensively uses cross-validation to increase its computational efficiency.

### *4.4 Decision Tree*

This algorithm can be used for both regression and classification. The core idea is to divide the dataset into smaller subsets and at the same time tree associated is incrementally created. This can handle both categorical as well numerical data. We can use Gini index as well information gain parameter to decide which attribute will be used for further division of dataset. If we use Gini index than decision tree is called CART (classification and regression tree) and if we use information gain than it is called ID3. This algorithm can be easily used for any type of application [12].

### *4.5 K-Nearest Neighbour (KNN)*

This algorithm is simple and has applications mainly in pattern recognition; intrusion detection and many more are also there. In this distance between data point of which we want to identify class is calculated using Euclidean distance (other measures like Manhattan distance etc.) is calculated with the existing data points and the k nearest neighbour (value of k is initially decided can be 3, 4 etc.) will vote for the class of new data point. Majority voting will decide the class [13].

### *4.6 Random Forest*

It is an ensemble of decision tree algorithms which can be used for both classification and regression. In this algorithm generally, more trees correspond to better performance and efficiency. In a given training set, extract a sample set of data points by using bootstrap method. After this construct a decision tree based on the output of previous step. Apply previous two steps and we will get number of trees (in our case 100). Every tree constructed will vote for data point. Calculate the majority voting of all the decision tree classifier [18].

## 5. Performance Parameters

### *5.1 Accuracy:*

This is the ratio of true positives plus true negative to the true positives plus true negatives plus false positive plus false negative as shown in equation 1. It calculates how much percentage of cases is correctly classified

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{\text{True Positive}+\text{True Negative}+\text{False Positive}+\text{False Negative}} \quad \ldots\ldots (1)$$

### *5.2 Precision:*

Ratio of predicted positive observations to the total number of positive observation is known as precision. It is computed as follows:

$$Precision = \frac{True\ Positive}{True\ Positive+False\ Positive} \quad \ldots\ldots (2)$$

*5.3 Recall:*

Ratio of correctly predicted positive observations to all observations in actual class yes is known as recall. It is computed as follows:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad (3)$$

*5.4 F-score:*

Weighted average of recall and precision is called f-score. More important parameter than accuracy when having an uneven class distribution in data. It is calculated as follows:

$$F_{score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

## 6. Results

In this paper, we considered two features TF-IDF (word level) and N-Grams (value of n=2) on the Twitter sentiment analysis dataset (SS-Tweet). Table 1 shows the output(four performance parameters i.e accuracy, precision, recall, and f-score) of six classification techniques (Random Forest, Decision Tree, Naive Bayes, SVM, Logistic Regression, and KNN) using TF-IDF feature. Table 2 shows the output (four performance parameters i.e accuracy, precision, recall, and f-score) of six classification techniques (Random Forest, Decision Tree, Naive Bayes, SVM, Logistic Regression, and KNN) using N-gram features. As it can be seen from both the tables logistic regression is performing better in both the cases and our job is find out which features is performing better as compared to other, this has been shown in figure 2.

Table 1: SS-Tweet results through word level

| SS-Tweet Dataset - (WORDLEVEL TF-IDF) | | | | |
|---|---|---|---|---|
| **ML Algorithms** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F-Score (%)** |
| KNN | 46 | 32 | 33 | 21 |
| Decision Tree | 46 | 43 | 42 | 42 |
| SVM | 46 | 15 | 33 | 21 |
| Logistic Regression | 57 | 57 | 50 | 50 |
| Naïve Bayes | 53 | 56 | 44 | 42 |
| Random Forest | 51 | 47 | 44 | 44 |

TF-IDF at word level feature is performing around 3-4% better as shown in figure 2 because it considers each and every word equally important.

Table 2: SS-Tweet results through N-gram Feature

| SS-Tweet Dataset - (N-GRAMS VECTORS) | | | | |
|---|---|---|---|---|
| **ML Algorithms** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **Score (%)** |
| KNN | 46 | 41 | 34 | 26 |
| Decision Tree | 48 | 44 | 42 | 42 |
| SVM | 46 | 15 | 33 | 22 |
| Logistic Regression | 49 | 51 | 39 | 54 |
| Naïve Bayes | 50 | 52 | 41 | 38 |
| Random Forest | 51 | 49 | 43 | 42 |

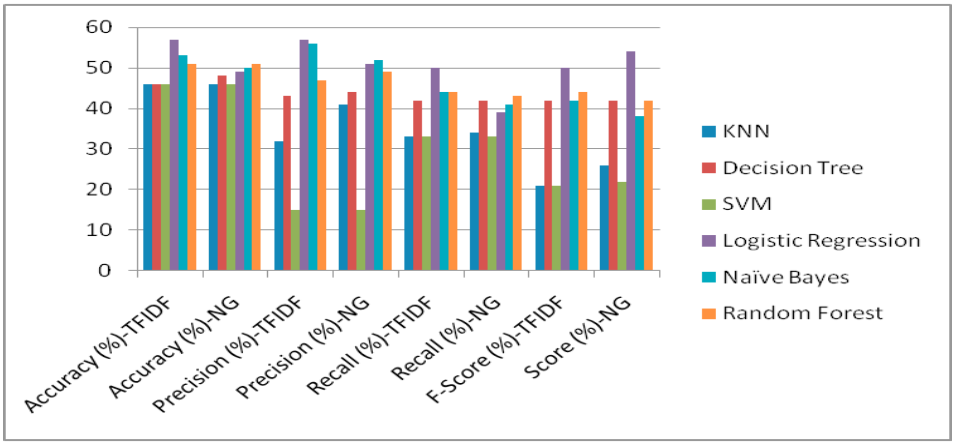


Figure 2: Comparison of TF-IDF and N-gram approach on SS-Twitter Data set

**NG - N-grams**

## 7. Conclusion

In this paper, we have thus applied 6 different algorithms of classification on the SS-Tweet dataset considering two features (TF-IDF and N-Grams). Thus after doing sentiment analysis of these tweets, we founded that, TF-IDF features are giving better results (3-4%) as compared to N-Gram features. Thus we can conclude that if we are going to use machine learning algorithm for the text classification than TF-IDF is the best choice of features as compared to N-Gram. By overall comparison of machine learning algorithms, we found out that logistic regression gave best predictions of sentiments by giving maximum output for all four comparison parameters namely – accuracy, recall, precision, and f-score and for both feature extraction methods namely – N-Gram and word-level TF-IDF. Thus Logistic regression is the best algorithm for sentiment analysis and both feature extraction techniques are good enough. In future, comparison of other features like word polarity score features, word embeddings, twitter specific features etc.

## References

[1] Y. Woldemariam, "Sentiment analysis in a cross-media analysis framework," *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, Hangzhou, 2016, pp.1-5.

[2] X. Fan, X. Li, F. Du, X. Li, and M. Wei, "Apply word vectors for sentiment analysis of APP reviews," *2016 3rd International Conference on Systems and Informatics(ICSAI)*, Shanghai,2016,pp.1062-1066.

[3] Shi, Y., Xi, Y., Wolcott, P., Tian, Y., Li, J., Berg, D., Chen, Z., Herrera-Viedma, E., Kou, G., Lee, H., Peng, Y., Yu, L. (eds.): Proceedings of the First International Conference on Information Technology and Quantitative Management, ITQM 2013, Dushu Lake Hotel, Sushou, China, 16–18 May 2013, Procedia Computer Science, vol. 17. Elsevier (2013)

[4] Fouad M.M., Gharib T.F., Mashat A.S. (2018) Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble. In: Hassanien A., Tolba M., Elhoseny M., Mostafa M. (eds) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018). AMLTA 2018. Advances in Intelligent Systems and Computing, vol 723. Springer, Cham

[5] J. Li and L. Qiu, "A Sentiment Analysis Method of Short Texts in Microblog," *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, Guangzhou,2017,pp.776-779.

[6] Spatiotis N., Paraskevas M., Perikos I., Mporas I. (2017) Examining the Impact of Feature Selection on Sentiment Analysis for the Greek Language. In: Karpov A., Potapova R., Mporas I. (eds) Speech and Computer. SPECOM 2017. Lecture Notes in Computer Science, vol 10458. Springer, Cham

[7] Prusa, Joseph D., Taghi M. Khoshgoftaar, and David J. Dittman. "Impact of Feature Selection Techniques for Tweet Sentiment Classification." In *FLAIRS Conference*, pp. 299-304. 2015.

[8] Y. Ji, S. Yu, and Y. Zhang, "A novel Naive Bayes model: Packaged Hidden Naive Bayes," *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, Chongqing, 2011, pp. 484-487.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

[10] M. Rathi, A. Malik, D. Varshney, R. Sharma, and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, 2018, pp. 1-3

[11] İ. İşeri, Ö. F. Atasoy and H. Alçiçek, "Sentiment classification of social media data for telecommunication companies in Turkey," *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, 2017, pp. 1015-1019

[12] YAcharya, U. R., Molinari, F., Sree, S. V., Chattopadhyay, S., Ng, K. H., and Suri, J. S., Automated diagnosis of epileptic EEG using entropies. Biomed. Signal Process. Control 7(4):401–408, 2012.

[13] Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbour (knn) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications*, *3*(5), 605-610.

[14] S. Wu and H. Nagahashi, "Parameterized AdaBoost: Introducing a Parameter to Speed Up the Training of Real AdaBoost," in *IEEE Signal Processing Letters*, vol. 21, no.6,pp.687-691,June2014.doi: 10.1109/LSP.2014.2313570

[17] *R. M. Esteves, T. Hacker, and C. Rong, "Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets," 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, 2013, pp. 17-24.*

[18] Breiman, L., Random forests. Mach. Learn. 45(1):5–32, 2001.

[19] Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." In *Proceedings of the workshop on languages in social media*, pp. 30-38. Association for Computational Linguistics, 2011

[20] Mohammad, Saif M., Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. "Sentiment, emotion, purpose, and style in electoral tweets." *Information Processing & Management* 51, no. 4 (2015): 480-499.

[21] Angulakshmi, G., and Dr R. Manicka Chezian. "Three level feature extraction for sentiment classification." *International Journal of Innovative Research in Computer and Communication Engineering* 2, no. 8 (2014): 5501-5507.

[22] Kamale, Mr Amit S., Pradip K. Deshmukh, and Prakash B. Dhainje. "A Survey on Classification Techniques for Feature-Sentiment Analysis." *International Journal on Recent and Innovation Trends in Computing and Communication* 3, no. 7 (2015): 4823-4829.

[23] Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, Prof. Dr. D. R. Ingle “Sentiment Analysis in Twitter” International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 01 | Jan-2018 pages 880-886.

[24] Das, Tushar Kanti, D. P. Acharjya, and M. R. Patra. "Opinion mining about a product by analyzing public tweets in Twitter." In *Computer Communication and Informatics (ICCCI), 2014 International Conference on*, pp. 1-4. IEEE, 2014.