ISSN 2278-3091
Volume 8, No.4, July – August 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse25842019.pdf
https://doi.org/10.30534/ijatcse/2019/25842019

Human Annotation and Interpretation of Public Sentiments about Jio Coin marked in Social Networks using Machine Learning Algorithms

Deepa Mary Mathews[1], Sajimon Abraham[2]
[1]Research Scholar, School of Computer Sciences, Mahatma Gandhi University, India
deepamarymathews@gmail.com
[2]School of Management and Business Studies, Mahatma Gandhi University, India
sajimabraham@rediffmail.com

ABSTRACT

The basic errand in Sentiment Analysis is to categorize the orientation of a given review and subsequently identifying whether the sentiment implied is positive, negative or fair. In this article the authors present the following lines of experimentation and outcomes. One is related to human annotation of Tweets and assessment of their quality and dataset properties. Another is about training sentiment classifiers, their performance and comparisons. The authors' presents a comprehensive assessment about various supervised machine learning techniques to interpret the public sentiments about 'Jio Coin' marked in social networks. Various evaluation measures like Precision, Recall, F-Score, Matthews Correlation Coefficient, Jaccard Index and Kappa statistics depicts the efficiency rate of the models in the datasets. The learning time and the predicting time taken by various classifiers depicted in the article helps to choose the classifier that suits well if time is a constraint.

Key words: Machine Learning, Opinion Mining, Sentiment Analysis, Vectorization

1. INTRODUCTION

The acceleration of online social networks and media over the last decade has reformed the approach of the individuals' interaction and the industry dealings. Millions of newcomers entered the digital world due to the demonetization and the attempted changeover to a cashless fund. Extracting and analyzing the online shared information is gaining prominence as the current trends and viewpoints are updated directly on such platforms. With the tremendous accessibility of archives that expresses suppositions on various issues, the challenge arises to analyze it properly and churn out meaningful information from it. Sentiment analysis will enable us to figure out and employ this data successfully and help to augment the decision making process.

1.1 Sentiment Analysis

Sentiment Analysis is broadly used to tract user attitudes and views. To split up the users' viewpoints and to identify the vital patterns among this data, opinion mining can be utilized [1]. One of the basic tasks in opining mining is categorizing the orientation of a particular review at document, sentence, or feature level to know whether the communicated sentiment is positive, negative or fair. Machine Learning methods these days turn into an inexorably critical facet in various emerging areas which aid in taking decisions, analysis and automation. Among the two methodologies (Supervised and Unsupervised) utilized for Sentiment Analysis, in this article the authors used supervised machine learning way to do classification. Provided with the labeled data, the model classifies the content by means of any machine learning algorithm.

Though the crypto currency fascination is mesmerizing the whole humankind, particularly Bitcoin, chances of possible misuses of such virtual currencies have to be overlooked. So the authors in this article analyze the standpoints of the people to know whether the world is supporting or opposing the entry of Jio Coin. In this article the authors present the following lines of experimentation and outcome. One is related to human annotation of Tweets and assessment of their quality and dataset properties. Another is about training sentiment classifiers, their performance and comparisons.

The article is framed as follows. The subsequent section explains the works associated with the study followed by the section 3 which explains the various methodologies used to implement the work. The Implementation section portrayed the framework of the proposed work and the actual experimentation values. In the Results and Discussion section, the outcomes are investigated which is trailed by the Conclusion section which concludes the work along with a description of future works.

2. RELATED WORKS

Document based classification has been done for news comments using different supervised machine learning approaches [3].The approaches for sentiment analysis are explained in the survey papers [4,5,6]. Various supervised machine learning methodologies are compared and evaluated by [7]. The proposed system for sentiment mining visualizes in real world data set and resulted in to an experiment that distinct the positive and negative sentiments [8]. The authors made a comparison between Naive Bayes Classifier (NB) and Support Vector Machines (SVM) on online reviews related to travel destinations [9,10]. A query expansion ranking method is proposed by authors which is based on query expansion term weighting methods [11]. The algorithms like Max Entropy, NB, Decision Trees and SVM are used by [12,13]. The authors introduced a dynamic learning to a structure that includes ensemble methods for opinion mining [14]. The authors' leverages J48, NB, OneR and BF Tree models for the optimization of opinion mining [15]. The authors [16] explored the impact of pre-processing methods for classifying sentiments of twitter datasets. The classifiers like NB and Neural Network are integrated for classifying movie reviews [17]. Sentiment Classification is done using NB and SVM classifiers and vectorization of features is done using Count_Vectorizer and TFIDF_Vectorizer[18]. The authors done a comparison between four data mining toolkits for classification purposes, nine different datasets were used to judge the four toolkits tested using six classification algorithms namely; NB, C4.5, SVM, KNN, OneR, and Zero Rule and found that no tool is better than the other [24]. The authors developed a method for predicting the helpfulness of online product reviews using sentiment and emotions [25]. The authors present a review on prediction approaches for epidemic disease outbreaks dependent via web-based social media information [26].

3. METHODOLOGIES

The procedure for learning and evaluating the user viewpoints are described in this section. We portrayed different methodologies used for experimentation to retrieve the tweets, to pre-process the sentiments and the various classification algorithms used to analyze the sentiments.

3.1 Data Collection

The data for experimentation can be collected either online or offline. Using Python programming language which uses the Twitter API credentials, the tweets based on "Jio Coin" are acquired. The offline labeled datasets downloaded from various repositories and used for experimentation are IMDB Movie Reviews, Yelp Reviews and Amazon Reviews. The figure 1 depicts the overview of the process flow. The statistics of the datasets used is depicted in the Table 1.

3.2 Noise Removal

Social media reviews don't tag on any grammar, mainly short messages, and spell faults are common, and have many irrelevant words. So a pre-processing phase is required. The data from reviews are tokenized into convenient units to build a representation of the data [20]. This phase removes the data which is not required for analysis. Natural Language Processing (NLP) is particularly utilized in Sentiment Analysis as it tries to bridge the gap between human and machine, by hauling out the valuable data from characteristic dialect messages [18]. The pre-processing steps executed to create a Bag-of-Words picture of the data are as:

Table 1: Statistics of the Datasets

Dataset	No: of Samples	No: of Features	% of Positive Reviews	% of Negative Reviews
IMDB	25000	70752	50.00%	50.00%
Yelp	10000	26834	83.24%	16.76%
Amazon	996	1650	50.10%	49.90%
Jio Coin	246	771	34.96%	28.00%

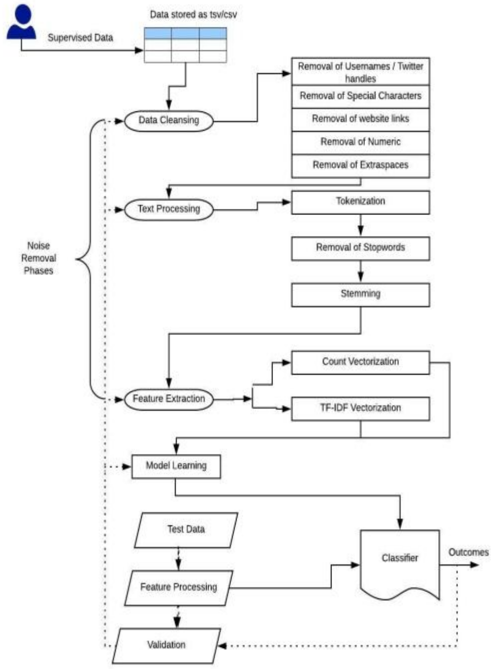

Figure 1: Schematic Representation of the Proposed Process Flow

- Cleansing - Removal of website links, Removal of Twitter handles, Remove tweets with Not Available text, Removal of special characters, Removal of digits
- Text processing - Tokenization, Stemming, Stopword Removal
- Feature Extraction and Selection

The Shingle technique is applied to the datasets where it considers successive terms and treated them as a single object. The Bag of Words (BOW) model of the review datasets is actually the arrangement of 1-shingles. In this article, authors' considered 1-Shingles (unigrams) for the experimentation. Feature extraction and selection is a vital pre-processing step to machine learning problems. The data after feature extraction is in the form of vectors. The model which represents these review documents as vectors of identifiers are called Vector Space Model. The significance of the terms in the dataset can be measured using TFIDF measure which is the most widely used effective weighting functions. The Count_Vectorizer as well as the TFIDF_Vectorizer are utilized in this article for vectorization of the corpus. The most commonly occurring pre-processed tokens are identified based on the occurrence of the tokens. The counting function for the Term_Frequency(termfreq) can be given as:

$$termfreq(term,document) = \Sigma_{t\epsilon document}\, fr(t,term)$$

where the *fr(t,term)* is defined as:

$$fr(t, term) = \begin{cases} 1 \text{ ; if } (t = term) \\ 0 \text{ ; otherwise} \end{cases}$$

Inverse_Document_Frequency (InvDF) depicts how recurrent the term amid all other review documents. Low InvDF value is given to the most prevalent term since such terms are less significant in the classification [20]. The TFIDF measure used by the vectorizer can be calculated as

$$\text{TFIDF} = \text{termfreq}_{(term,document)} * \text{InvDF (term)}$$

Feature Selection trims down the dimensionality in capacious data and spots out the fussy features that lessen computational overhead and augment the accuracy of the classifier.

```
Algorithm 1: SentiAnal
Retrieve and Save the tweets about Jio Coin into a .csv file;
Store the offline datasets in csv format;
Let C be Corpus of the reviews Dataset D
enum noisy = {URLs, special char, hashtags, extraspaces};
enum Vectorizer = {CountVectorizer, TFIDFVectorizer};
enum SplitRatio = {50:50, 60:40, 70:30, 80:20, 90:10};
enum Classifier = {BNB,MNB,SVM,LR,kNN,RF,DT,GB,xGB,MLP};
tempRevList ← D;
foreach D in C do
    RevList ← 0;
    foreach row in tempRevList do
        if tempRevList.contains noisy then
            tempRevList.remove(noisy);
            RevList ← tempRevList;
        end
        RevList ← RevList.apply(nltk.word tokenize);
        RevList ← RevList.remove(stopwords);
        RevList ← RevList.apply(nltk.P orterStemmer());
        RevList ← RevList.remove(Sentiment ==’neutral’);
    end
    foreach value in Vectorizer do
        FList ← build.BOW(RevList);
    end
    foreach value in SplitRatio do
        Train a Classifier using the FList;
        Generate the Classification Report;
    end
    Set the SplitRatio with max(Acc);
    foreach value in Classifier do
        Train the Classifier;
        Fit and Test the Classifier using test data;
        Calculate the time to Train and Predict;
        Generate Confusion Matrix;
        Calculate the values of Prec, R, F-Score, Acc;
        Calculate the values of JI, MCC, Kappa;
    end
end
```

3.3 Classification Algorithms

Classification algorithms that use supervised approach like Bernoulli Naive Bayes (BNB), Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), Logistic Regression (LR), k-Nearest Neighbor (kNN), Decision Tree (DT), Multi Layer Perceptron (MLP) and various ensemble methods like Random Forest (RF), Gradient Boosting (GB) and eXtreme Gradient Boosting (xGB) are considered in this article for classification. Naive Bayes is used for twofold and multifold classification problems [20]. The BNB classifier is mostly employed when the nonexistence of a term matters. The MNB classifier is applied when several occurrences of the term signify in classification. The SVM algorithm is based on ruling a parting between hyper planes distinct by the classes of data [9]. The LR is a classification algorithm for analyzing a dataset in which there are one or more independent variables that determine an outcome [21]. The k-NN classification algorithm considers the resemblance between the k nearest neighbors. In this algorithm, an item is classified by majority voting of its neighbors [20]. The DT, where the data is continuously split according to parameters like Gini index or Entropy to create a model that predicts by learning simple decision rules inferred from the data features [13]. Random Forests brings out multi-altitude DTs. The linkage amid trees is abridged by arbitrarily picking trees and so the prediction accuracy boosts and thereby increases the effectiveness [14,21].

4. IMPLEMENTATION

Labeled datasets like IMDB movie reviews, Amazon product reviews and Yelp labeled reviews are used to train the classifiers. The labeling or the sentiment distribution in the Jio Coin dataset is calculated based on the user ratings. The user ratings with value greater than 3 is considered as positive, less than 3 is taken as negative and 3 is taken as neutral. All the user reviews dataset are pre-processed and the missed values are removed, if any. The algorithm 1 (SentiAnal) explains the implementation process used for extracting tweets and the steps for building the model. Classifiers are learned or trained on a finite training set. The classifiers that taken, envisages the case in point of a testing set as either positive or negative that may lead to four outcomes – TP (True_Positive), TN (True_Negative), FP (False_Positive) and FN (False_Negative). These values can be visualized using Confusion Matrix. Various evaluation measures derived from the Confusion Matrix used in experimentation are

- Precision (Prec) = TP / (TP + FP)
- Recall (R) = TP / Pos
- F-Score = 2.Prec.R / (Prec + R)
- Accuracy (Acc) = (TP+TN) / (TP+TN+FP+FN)
- Matthews_Correlation_Coefficient=(TPxTN-FPxFN)/ sqrt((TP+FP)(TP+FN)(TN+FP)(TN+FN))
- Jaccard Index, JI = TP / (TP+FP+FN)

The trained classifier has to be tested on new test set experimentally. The values of various evaluation metrics like Precision, Recall, F-Score, Accuracy, Jaccard Index, MCC, Kappa statistic were taken as the criterion function for assessing the classifier performance experimentally. The value of MCC is in between -1 and +1 where +1 symbolizes an ideal, 0 an average random prediction and -1 a contrary prediction. The Kappa statistic which measures the closeness of the instances after classification with the data labeled as ground truth, controls the accuracy of a random classifier as measured by the expected accuracy. Landis and Koch interprets the Kappa statistic as 0-0.20 as minor, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost ideal.

4.1. Experimental Evaluation

The variants of Naive Bayes classification algorithms have been used for the experiment 4.1.1. The experiment is conducted by varying the size of the Training:Testing split to find out which split size shows better accuracy values. The experiment 4.1.2 and 4.1.3 computes the values of various evaluation metrics of different models used for classifying the reviews on different datasets using Count Vectorizer and TFIDF Vectorizer respectively. Experiment 4.1.4 is conducted for finding out which classifier comparatively takes lesser time to learn the model and the time to predict.

A. Training: Testing Split Ratio

The ratio of the labeled dataset taken for the test case is determined by considering the accuracy values. The experiment is conducted on IMDB Movie review dataset and on Yelp reviews dataset. The experiments depicts that the more the trained data, more accurate the classification. The Training:Testing split ratio 90:10 shows better accuracy values and so this split size is considered for the remaining experiments.

B. Scores of Various Evaluation Metrics using Count Vectorizer

The experimentation results depicted that the LR classifier is having the high accuracy value in IMDB and Yelp datasets and in the other two datasets also it gives better score. Even though the LR classifier is having high precision value, the F-Score value and accuracy value is higher for GB classifier in the Jio Coin dataset. The kNN classifier is having the lowest values for all the metrics. It is found that the Jaccard Index is same as the accuracy values of the classifiers. The MCC score is high for the LR classifier (0.7644) in IMDB dataset comparatively with other classifiers. The experiments show that an average random classification is done by the classifiers for predicting the class of the sentiment.

C. Scores of Various Evaluation Metrics using TFIDF Vectorizer

The experimentation results depicted that the LR classifier is having high accuracy value in the datasets under consideration except in Yelp dataset. Most of the classifiers shows high accuracy score if using TFIDF Vectorizer. The experimentation value shows that classifiers like MNB, SVM, LR and MLP (IMDB) Kappa statistics are substantial. In this experiment too, the MCC score is high for the LR classifier (0.7850) but with an increase compared to the count vectorizer values.

D. Learning Time and Predicting time of various Classifiers

The two phases of classification are learning the model and then predicting the new cases. It is found that the learning time and the predicting time for the MLP and DT classifiers are shown high in my core i5 Intel processor. Even though MLP is taking comparatively significant time to learn, it takes very less time to predict. The experimentation values show that xGB classifier takes lesser time compared to the GB classifier. Variants of Naive Bayes classifier are more efficient when considering time as a factor to select the classifier for classification.

5. RESULTS AND DISCUSSION

The weighted Precision, Recall and F-Score values calculated for top 4 classifiers using Count Vectorization and TFIDF Vectorization is diagrammatically represented on figure 2 and 3 respectively. The figures in 4 and 5 depicted the accuracy values of top four classifiers using Count_Vectorizer and TFIDF_Vectorizer respectively. The figure 6 depicted that the most of the classifiers outperformed if using TFIDF Vectorization. MNB classifier is more efficient when considering time as a factor to select the classifier for classification as per table 5 and figure 7.

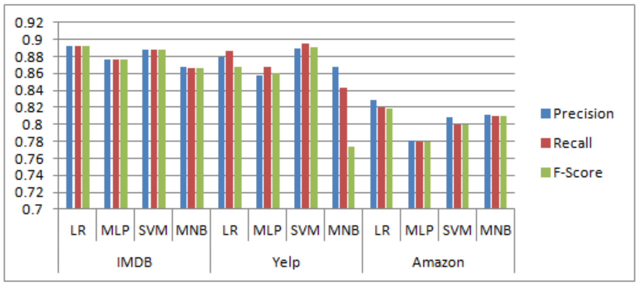

Figure 2: Precision, Recall and F-Score values of Classifiers using Count Vectorization

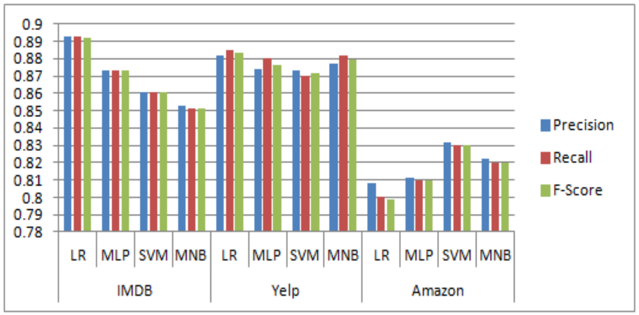

Figure 3: Precision, Recall and F-Score values of Classifiers using TFIDF Vectorization

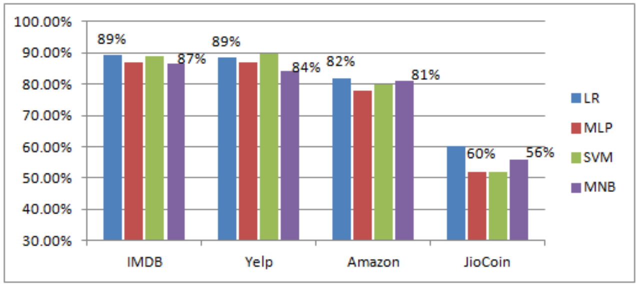

Figure 4: Accuracy of top performing Classifiers using ount_Vectorizer

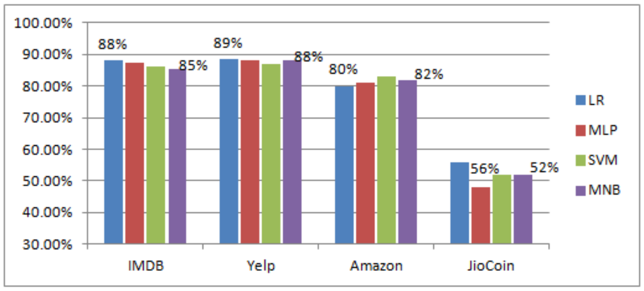

Figure 5: Accuracy of top performing Classifiers using TDIDF_Vectorizer

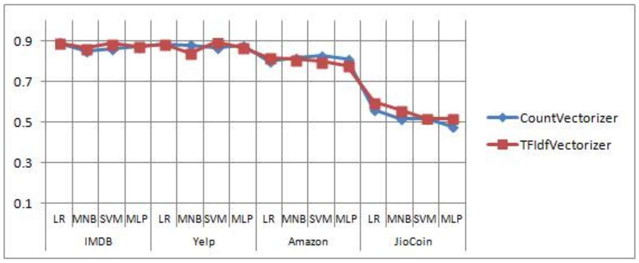

Figure 6: Accuracy Scores using Count_Vectorizer and TFIDF_Vectorizer

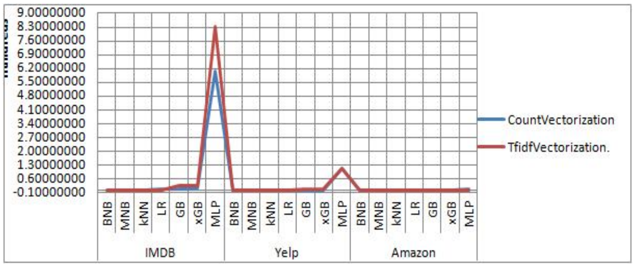

Figure 7: Learning Time and Predicting Time

6. CONCLUSION

Choice of classifiers will be determined based on resources, accuracy demand, time constraints and so forth. As per the experimental values, while in view of opinion mining, LR classifier obviously has an upper hand with high accuracy and recall values. We can infer that if accuracy is at our most astounding need then we should opt a classifier model like LR or MLP that consumes high learning time however has best accuracy. If process power and memory is an issue then the NB classifier ought to be chosen due to its low memory & processing power necessities. MNB classifier is more efficient when considering time as a factor to select the classifier for classification.

The authors used varied supervised machine learning algorithms for Sentiment Analysis and are found that supporters as well the critics on the news thread about Jio Coin are almost on same strength. The major constraint in the study about Jio Coin is that the training set contained reviews with Hindi words written in English language (Eg: hogi, sach) and these non English words are considered as neutral in this study that may misdirect and influence the execution of the classifier leading to moderate classifiers.

REFERENCES

1. S. Chen, B. Mulgrew, and P. M. Grant. **A clustering technique for digital communications channel equalization using radial basis function network**s, *IEEE Trans. on Neural Networks*, Vol. 4, pp. 570-578, July 1993.
https://doi.org/10.1109/72.238312
2. Liu, Bing. **Sentiment analysis and opinion mining.,** *Synthesis lectures on human language technologies* 5.1: 1-167,.2012
https://doi.org/10.2200/S00416ED1V01Y201204HLT016
3. Bird, Steven, Ewan Klein, and Edward Loper. **Natural language processing with Python: analyzing text with the natural language toolkit.** *O'Reilly Media, Inc.*, 2009.
4. Zhao, Yan, Suyu Dong, and Leixiao Li. **Sentiment analysis on news comments based on supervised learning method.**, *International Journal of Multimedia and Ubiquitous Engineering,* Vol.9, No.7 pp.333-346, 2014
https://doi.org/10.14257/ijmue.2014.9.7.28
5. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. **Thumbs up?: sentiment classification using machine learning techniques,** *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*- Volume 10. Association for Computational Linguistics, 2002.
https://doi.org/10.3115/1118693.1118704
6. Liu, Bing, and Lei Zhang. **A survey of opinion mining and sentiment analysis,** *Mining text data. Springer US*, 415-463, 2012.
https://doi.org/10.1007/978-1-4614-3223-4_13
7. Mohammad, Saif M., **Sentiment analysis: Detecting valence, emotions, and other affectual states from text.,** *Emotion measurement,* 201-237, 2016.
https://doi.org/10.1016/B978-0-08-100508-8.00009-6
8. Padmaja, S., and S. Sameen Fatima. **Opinion mining and sentiment analysis-an assessment of peoples' belief: A survey.** *International Journal of Ad hoc Sensor & Ubiquitous Computing,* 4.1 : 21, 2013
https://doi.org/10.5121/ijasuc.2013.4102

9. Jin, Wei, Hung Hay Ho, and Rohini K. Srihari. ”**OpinionMiner: a novel machine learning system for web opinion mining and extraction** *Proc of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM*, 2009. https://doi.org/10.1145/1557019.1557148
10. Ye, Qiang, Ziqiong Zhang, and Rob Law. **Sentiment classification of online reviews to travel destinations by supervised machine learning approaches** *Expert systems with applications*, 36.3: 6527-6535, 2009. https://doi.org/10.1016/j.eswa.2008.07.035
11. Khan, Khairullah, Baharum B. Baharudin, and Aurangzeb Khan. **Mining opinion targets from text documents: A review**” *Journal of Emerging Technologies in Web Intelligence* 5.4 : 343-353, 2013 https://doi.org/10.4304/jetwi.5.4.343-353
12. Parlar, Tuba, Selma Ay¸se O¨ zel, and Fei Song. **QER: a new feature selection method for sentiment analysis**. *Human-Centric Computing and Information Sciences* 8.1: 10, 2018. https://doi.org/10.1186/s13673-018-0135-8
13. Pang, Bo, and Lillian Lee. **A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.** *Proc of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics*, 2004. https://doi.org/10.3115/1218955.1218990
14. Agarwal, Basant, and Namita Mittal. **Prominent feature extraction for review analysis: an empirical study.,** *Journal of Experimental & Theoretical Artificial Intelligence* 28.3: 485-498, 2016 https://doi.org/10.1007/978-3-319-25343-5
15. Aldo˘gan, Deniz, and Yusuf Yaslan. **A comparison study on active learning integrated ensemble approaches in sentiment analysis** *Computers & Electrical Engineering* 57: 311-323, 2017 https://doi.org/10.1016/j.compeleceng.2016.11.015
16. Singh, Jaspreet, Gurvinder Singh, and Rajinder Singh. **Optimization of sentiment analysis using machine learning classifiers.** *Human-centric Computing and Information Sciences* 7.1 : 32, 2017 https://doi.org/10.1186/s13673-017-0116-3
17. Saif, Hassan, et al. **On stopwords, filtering and data sparsity for sentiment analysis of twitter.,** 810-817,2014
18. Dhande, Lina L., and Girish K. Patnaik. **”Analyzing sentiment of movie review data using Naive Bayes neural classifier.”** *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* 3.4 (2014): 313-320.
19. Wagner, Wiebke. ”Steven bird, Ewan Klein and Edward Loper: **Natural language processing with python, analyzing text with the natural language toolkit.”** *Language Resources and Evaluation* 44.4: 421-424, 2010 https://doi.org/10.1007/s10579-010-9124-x
20. Palmer, David D.”**Tokenisation and sentence segmentation.”** *Handbook of natural language processing* : 11-35, 2000.
21. Dey, Lopamudra, et al. ”**Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier**.” *arXiv preprint arXiv:1610.09982* (2016). https://doi.org/10.5815/ijieeb.2016.04.07
22. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. **”Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors**). ” *The annals of statistics* 28.2 (2000): 337-407. https://doi.org/10.1214/aos/1016120463
23. Mathews, Deepa Mary, and Sajimon Abraham. **Analytic thinking of patients’ viewpoints pertain to spa treatment.** *Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on. IEEE,* 2017. https://doi.org/10.1109/NETACT.2017.8076772
24. Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa, E. M. **“A comparison study between data mining tools over some classification methods.”** *International Journal of Advanced Computer Science and Applications***,** *8(2), 18-26***,** 2011
25. Oueslati, Oumayma, Ahmed Ibrahim S. Khalil, and Habib Ounelli. **"Sentiment Analysis for Helpful Reviews Prediction."** *International Journal of Advanced Trends in Computer Science and Engineering, 7.3***,** *2018***.** https://doi.org/10.30534/ijatcse/2018/02732018
26. S. Ravi Kumar, Dr. M. Vamsi Krishna, Dr. Anurag, “**A survey on prediction approaches for epidemic disease outbreaks based on social media data”** *International Journal of Advanced Trends in Computer Science and Engineering.,8.3,2019* https://doi.org/10.30534/ijatcse/2019/86832019