Analysis of Various Sentiment Classification Techniques

Vimalkumar B. Vaghela
Assistant Professor,
L. D. College of Engineering
Ahmedabad, India

Bhumika M. Jadav
M.E. Scholar,
L. D. College of Engineering
Ahmedabad, India

ABSTRACT
Sentiment analysis is an ongoing research area in the field of text mining. People post their review in form of unstructured data so opinion extraction provides overall opinion of reviews so it does best job for customer, people, organization etc. The main aim of this paper is to find out approaches that generate output with good accuracy. This paper presents recent updates on papers related to classification of sentiment analysis of implemented various approaches and algorithms. The main contribution of this paper is to give idea about that careful feature selection and existing classification approaches can give better accuracy.

Keywords
Sentiment analysis, Text mining, Classification, Feature selection

1. INTRODUCTION
Sentiment analysis also refers to opinion mining which is a natural language processing problem. Natural language processing is related to area of human computer interaction. The task of identifying opinion of review is called sentiment analysis. Opinion may be positive, negative or neutral polarity. Sentiment analysis is classified into three different level namely they are document level, sentence level and entity-aspect level. Overall opinion is to be considered in document level whether opinion of particular sentence is to be considered in sentence level and Focus is directly on opinion itself is called entity and aspect level [1].

Machine learning technique is applied on Movie review dataset and proved that machine learning technique performs well than human generated result [2]. Text databases are increasing day by day due to large collection of information in from of electronic document so information retrieval is the process through which information is retrieved from large collection of textual database. Support vector machine, Maximum Entropy(MaxEnt) and naïve bayes classifiers are the most widely used algorithm in sentiment analysis. There are some issues in sentiment analysis, among them the major issue is classification accuracy so classification accuracy can be increased by choosing good preprocessing, feature selection and classification techniques. The main aim of this paper is to analyze existing method and find techniques that perform well in sentiment classification.

Sentiment analysis process is as showed in figure 1. Customers post their review in comment, forum or blog. These reviews are in form of unstructured data so first unstructured dataset is converted into structured form then extracts features from structured review using feature selection method then classification technique is applied on extracted features to classify them into its sentiment polarity that is namely either positive or negative. Some researchers also have classified review as neutral category.

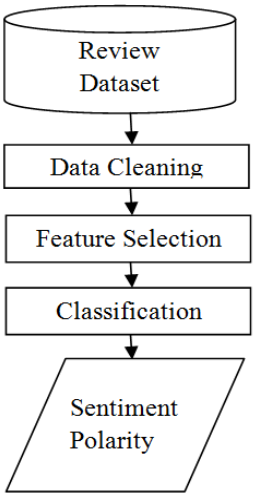

Figure 1: Sentiment Analysis Process Flow

2. RELATED WORK
Sentiment analysis is performed as showed in figure 1. These sections contain data cleaning, features, feature selection and classification steps.

2.1 Data cleaning
Stop word removal and special character removal method has been performed in [5]. Unwanted punctuation, new lines, ASCII code and stop word are removed and stemming is also performed as a part of preprocessing technique [6]. Stop word, stemming and tagging has been performed [7] [14]. Tokenization [9] divides given text into token. Conjunction rule, Negation rule, Part of Speech tagging by POS tagger and baseline approach has been implemented in [10]. NLTK is used in many existing papers with python to preprocess the dataset.
Following various methods are used in sentiment analysis as a part of preprocessing.

- Convert upper to lower case letter, remove Punctuation which are unwanted, remove New Line, remove special character, remove ASCII code, remove extra white spaces etc.
- Stemming: M.F. porter stemmer is most widely used algorithm which stems the word.
- Negation rule: this method removes negation word which reverses meaning of word in review.
- Conjunction rule: This method extracts meaning from review using grammatical rule.

2.2 Features

Unigram, bigram and N-gram features have been used to represent features in [1][2][9][10]. Bag of features framework is implemented in current paper. $d=(n_1(d),n_2(d),n_3(d),\ldots n_m(d))$ where document is defined as d, features as f, $n_i(d)$ is number of time f_i occurs in d [2]. N-gram feature model from shareware Rubryx tool is used as classifier to classify review [3]. Unigram with bag of words produce best feature for feature set than other n-gram method. The main aim of doing tagging is that adjectives are the most indicator of opinion of review so it helps to perform identification for sentiment classification. Unigram with stemming without stop word and unigram without stemming with stop word gives better result among all other result [6]. Selecting single feature from review is termed as unigram feature. Selecting pair of continue double as feature is termed as bigram features. Selecting n continue features from reviews is termed as n-gram feature. Bag of words model and feature hashing is used to select feature [15]. Bag of word is represented as $tweet_i = (a_{i1}, a_{i2},\ldots,a_{im})$, where a_{ij} is the frequency of term t_j in the $tweet_i$. Feature hashing is used to reduce the number of features provided as input to algorithm so original high-dimensional space is reduced to lower-dimensional space by mapping features to hash keys. Multiple features are mapped to the similar hash key [15].

Part of speech tagging is the process of marking up word in text to its corresponding tag. It is used to tag the word using various taggers like Penn Treebank POS tagger, Stanford tagger NLTK (natural language tool kit) tagger in the field of sentiment analysis. There are mainly nine parts of speech in English that are noun, verb, article, adjective, pronoun, adverb, conjunction and interjection. There are around 50 to 150 tags if we consider all tags [18].

2.3 Feature Selection

Various feature selection methods like TF-IDF(Term Frequency–Inverse Document Frequency), IG(Information Gain), MI(Mutual Information), Feature Vector, Unigram, Bigram and N- gram methods. CountVectorizer and TF-IDF are two feature selection techniques which are discussed in [5]. TF-IDF score is to be taken into consideration to balance most weighted and less weighted word [5]. Chi square method gives good result for both positive and negative class. Mutual information, Chi-square, TF-IDF and Information Gain techniques are used to select feature from high dimensional data [6]. Minimum support threshold is used to extract aspect [7]. TF-IDF feature selection is used [9]. Feature Vector is constructed from feature set [10]. Unigram feature extraction technique has been used to extract feature and feature vector list is produced [11]. Opinion words are extracted using Wilson lexicon list [12]. Unigram, Bigram, Unigram with bigram and Unigram with Pos tagging technique are used as part of feature selection to extract features and emotions are taken as noisy label to improve the accuracy level [13]. Most widely used feature selection methods have been defined below.

2.3.1 Count Vector

It is defined by number of occurrences of feature in review.

2.3.2 TF-IDF

It is defined by multiplying value of frequency of word in review (TF) and frequency of word in whole corpus (IDF).

$$\text{TF-IDF}_i = t_{i,j} * \log\left(\frac{N}{df_i}\right) \quad (1)$$

TF-IDF_i is the weight of a term i. $t_{i,j}$ is the frequency of term i in sample j. N is the total number of samples in the corpus. df_i is the number of samples containing term i.

2.3.3 Information Gain

Information gain is the most widely used attribute selection measures in the area of sentiment analysis. It determines the relevant features to predict review by studying the presence or absence of feature in document.

$$IG(f,c) = -\sum_{c,c} P(c)\log P(c)+\sum_{f,f} P(f)\sum_{c,c} P(c|f)\log P(c|f) \quad (2)$$

P(c|f) is the joint probability, class is c and feature is f and P(c) denotes the marginal probability.

2.3.4 Mutual Information

MI is the process of selecting features that are not uniformly distributed across the sentiment classes because they are informative of their classes and we can see that MI gives more importance to only few terms.

$$MI(f,c) = \sum_{c\epsilon C}\sum_{f} P(f,c)\log\frac{P(f,c)}{P(f)P(c)} \quad (3)$$

Where P(f,c) represents joint probability distribution function, P(f) and p(c) represent marginal probability distribution of f and c. c is positive and negative classes.

2.3.5 Chi-square

Chi-square measures observed count and expected count and analyzed how much deviation occurs between them.

$$\lambda^2(f,c) = \frac{N(WZ-YX)^2}{(W+Y)(X+Z)(W+X)(Y+Z)} \quad (4)$$

W, X, Y, Z represents the frequencies, represent the presence or absence of feature in the sample. W is the count of samples in which feature f and c occurred together. N=W+X+Y+Z. f represents the feature and c represents the class.

2.4 Classification approaches

GIS based approach [14] is proposed in which longitude and latitude is considered in classification process. The main aim of this visual classification is just to provide view for easy understanding. ESRI ArcGIS Viewer and ArcGIS API have been used to implement their proposed word of semantic orientation. Framework has been proposed to see easily understand. Various evaluation parameters like Precision, Recall, F-measure and accuracy are used to know the result.

Mainly approaches are classified into two categories namely lexicon based approach and machine learning based approach. Lexicon based approach is further divided into two category namely dictionary based and corpus based approach. In dictionary based approach, sentiment is identified using synonym and antonym from lexical dictionary like WordNet. In corpus based approach, it identifies opinion words by considering word list. Corpus based approach further more classified as statistical and semantic approach. In statistical approach, co-occurrences of words are calculated to identify sentiment. In semantic approach, terms are represented in semantic space to discover relation between terms [16].

Machine learning is further divided into two category namely supervised and unsupervised learning. Supervised classification algorithms are probabilistic classifier, linear classifier, decision tree and rule based classifier. Supervised learning technique is based on labeled dataset which is provided as input to train the model and this model is applied to test data to generate output. Sentiment classification in machine learning consists of two steps. First one is to extract feature and store in feature vector and second one is to train feature vector by using classification algorithms.

Table 1. Comparison between lexicon and machine learning[16]

Criteria	Lexicon-based	Machine learning
Domain	Independent	Dependent
Classification approach	Unsupervised	Supervised
Require prior training	No	Yes
Adaptive learning	No	Yes
Time of result generation	Fast	Slow
Maintenance	Need maintenance of corpus	Do not require maintenance
Accuracy	low	high then lexicon

2.4.1 Machine learning approach

SVM and Naïve bayes classifiers are used to classify sentiment and stated that SVM performs better than Naïve bayes[5][10]. SVM is used to classify text as either positive or negative [6][11].

Naïve bayes algorithm is used to classify sentiment and this sentiment orientation performs well with more accuracy [7]. Ensemble technique increases 3-5 % performance than machine learning techniques. Random Forest, Decision Tree, Extremely Randomized Trees and Ada Boost are ensemble to improve performance [8]. LIBSVM is used which is a library of SVM. Linear SVM performs better than other four kernels[10]. Bag of word model with support vector machine algorithm is used in existing paper which produces good result [9]. SVM, Naïve and Maximum Entropy classifiers are used to classify twitter dataset [13]. Multinomial Naive Bayes, Logistic Regression, Random Forests and Support vector machine (LIBSVM) algorithms are ensemble to improve the performance [15]. Most widely approaches have been discussed below.

2.4.1.1 SVM

Support vector machine examines the data, identify hyper plane that classify data in to two classes with maximum margin. SVM also supports classification and regression in statistical learning. A separating hyper plane is written as:

$$W*X + b = 0 \quad (5)$$

Where $W = \{w_1, w_2, w_3, \dots, w_n\}$. w_n is defined as weight vector of n attributes. b is defined as bias. Distance from separating hyper plane to any point on H1 is 1/|W| and same to any point on H2 is 1/|W|. So maximum margin is 2/|W|. if hyper plane value > 0 then +ve category, if hyper plane value < 0 then –ve category, if hyper plane value = 0 then all points are perpendicular to W. if value of margin is large then large penalty is assigned to errors/margin errors. If value of margin is small then some points become margin error and orientation of hyper plane is changed.

$$W = \sum_j \alpha_j c_j d_j \ , \alpha_j \geq 0 \quad (6)$$

Let c(1,-1) is class (positive, negative) for document d.

2.4.1.2 Naïve Bayes

It is used to predict the probability for a given tuple to belong to a particular class. It is used because of its easiness in both during training and classifying steps. Pre-processed data is given as input to train input set by classifier using naïve bayes and that trained model is applied on test data to generate either positive or negative sentiment. The bayes theorem is as follows.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (7)$$

H-Hypothesis, X-Tuples, P(H|X) represents Posterior probability of H conditioned on X i.e. the Probability that a Hypothesis holds true given the value of X , P(H) represents Prior probability of H i.e the Probability that H holds true irrespective of the tuple values, P(X|H) represents posterior probability of X conditioned on H i.e. the Probability that X will have certain values for a given Hypothesis, P(X) represents Prior probability of X i.e the Probability that X will have certain values.

2.4.1.3 Maximum Entropy

Maximum entropy maximizes the entropy that is defined in the conditional probability distribution where c is the class, d is the tweet. It processes same as described in naïve bayes algorithm.

$$P_{ME}(c|d) = \frac{\exp[\sum \lambda_i f_i(c,d)]}{\sum_{c'} \exp[\sum \lambda_i f_i(c,d)]} \quad (8)$$

2.4.2 Lexicon based approach

Semantic orientation of phrases is determined as positive if it is more related to "best" and is considered to negative if it is more related to "poor". So SO(Semantic Orientation) values are dependent on calculation result which is calculated by taking average of SO value of all phrases which has been extracted. Classification accuracy is improved by including semantics, word sense disambiguation and WordNet which help to find Synset of words [8]. Three Fuzzy set approaches has been proposed that are namely positive fuzzy set, negative fuzzy set and neutral fuzzy set [9]. Some paper describes domain specific lexicon and adjective/adverb score is calculated using WordNet database dictionary [10]. Synonym of feature is identified using WordNet and calculated adjective scores. Feature is compared with every feature of feature vector list and if features are same then they are replaced with its synonym word [11]. Semantic orientation is determined by assigning +1 to positive word,+0.5 to weak positive word, -1 to negative word, -0.5 to weak negative word and 0 to neutral word as semantic orientation score. Sentence Sentiment Scoring Function (SSSF) calculates the score of orientation of sentiment for each entity e_i in s. Entity Sentiment Aggregation Function (ESAF) calculates the total sentiment scores for an entity e_i [12]. Emotion classifier, Bag of word classifier and SentiWordNet classifier have been proposed. Emotions of tweets generate best performance so it is used to improve accuracy in first emotion classifier. First emotion classifier is performed if generated result is neutral then Bag of words model is performed and if it still generates neutral result then SentiWordNet classifier is performed [14].

2.4.2.1 WordNet

It is English database dictionary where every term is associated with each other via link. Mostly WordNet is used to check similarity with words and to calculate sentiment score. It links to sets of syntactic category which are verb, adjective, adverb and noun. They are linked with semantic relations those are termed as synonym, antonym, hyponymy, meronymy, troponymy, Entailment etc [4]. WordNet[11] has improved the performance in existing research paper. First they find out synonym of feature from feature vector list and compare it with each feature of feature list. For example glad is near to "happy" word then glad is replaced by beautiful word so synonym of word is also expressed after final result. Adjectives are scored with help of WordNet so it helps in classification of polarity.

WordNet has 166000 above sense pairs and word form. Form is represented by string of ASCII characters and a sense is represented by set of synonyms [17].

2.4.2.2 SentiWordNet

It is lexical resource derived from WorNet. It contains score for positive, negative and objective terms. Score is already defined for every words and range of that word is showed in table 3. A strong word contains higher score and a weak word contains lower score so sentiment classification is done based on score.

Table 2. SentiWordNet Database Content Structure

Field	Detail
POS	It is called syntactic category N=Noun A=Adjective V=Verb R=Adverb
Offset	It is numerical ID which can uniquely identify POS tagging.
Positive score	Numerical value between 0 to 1
Negative score	Numerical value between 0 to 1
Synset Terms	All terms of Synset

Scoring rule is as follows.

$$Pos(s) + Neg(s) + Obj(s) = 1 \quad (9)$$

$$Obj(s) = 1 – (Pos(s) + Neg(s)) \quad (10)$$

Positive and negative score are already given so objective score can be calculated using given scores. Scores information with its POS tag with group is contained in text file.

3 PERFORMANCE EVALUATION ANALYSIS

Analyzed papers includes various dataset those are namely Cornell movie review dataset, Twitter dataset, Customer dataset (amazon.com, epinions.com, cnet.com), Sanders twitter dataset, Stanford twitter dataset, Obama-McCain Debate twitter dataset and Health care reform twitter dataset.

Among them movie review mining is more challenging reviews than other dataset review because real life word and ironic terms are mixed in movie review. For example unpredictable terms indicate negative opinion but it gives positive opinion for movie review [3].

The performance of sentiment analysis is calculated by using help of confusion matrix which is generated when algorithm is implemented on dataset. Various performance measures are used that are Precision, Recall, F-measure and Accuracy.

Table 3. Confusion Matrix

	Correct Labels	
	Positive	**Negative**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Accuracy: It is one performance evaluation parameter and it is calculated by number of correctly predicted reviews divide by total number of reviews present in the corpus. The formula is given as below.

$$\text{Accuracy} = \frac{Tp+Tn}{Tp+Tn+Fn+Fp} \quad (11)$$

Table 4. Comparison of Evaluated Result

Reference	Technique	Dataset	Accuracy
[2]	SVM MaxEnt NB	Movie Review	82.9 % 81 % 81.5%
[3]	Supervised Unsupervised	Movie Review	83.54% 77%
[5]	SVM NB	Movie Review	94% 89.5%
[6]	NB	Customer Review	92.37%
[7]	SVM	Movie Review	83%
[9]	Fuzzy classifier	Movie Review	good
[10]	SVM	Customer Review	78%
[11]	WordNet SVM MaxEnt NB	Twitter	89.9% 85.5% 83.8% 88.2%
[14]	Unsupervised	Twitter	80% up
[15]	Ensemble Classifier (RF, NB, LR, SVM)	Sanders Twitter Stanford Twitter OMD Twitter HC Twitter	84.89% 87.20% 76.81% 78.35%
[13]	NB MaxEnt SVM	Twitter	82.7% 83% 82.2%

Figure 2 shows graphical view of the comparison of existed methods accuracy which is already shown in the table 4. SVM performs very well among the given classifier for the given dataset.

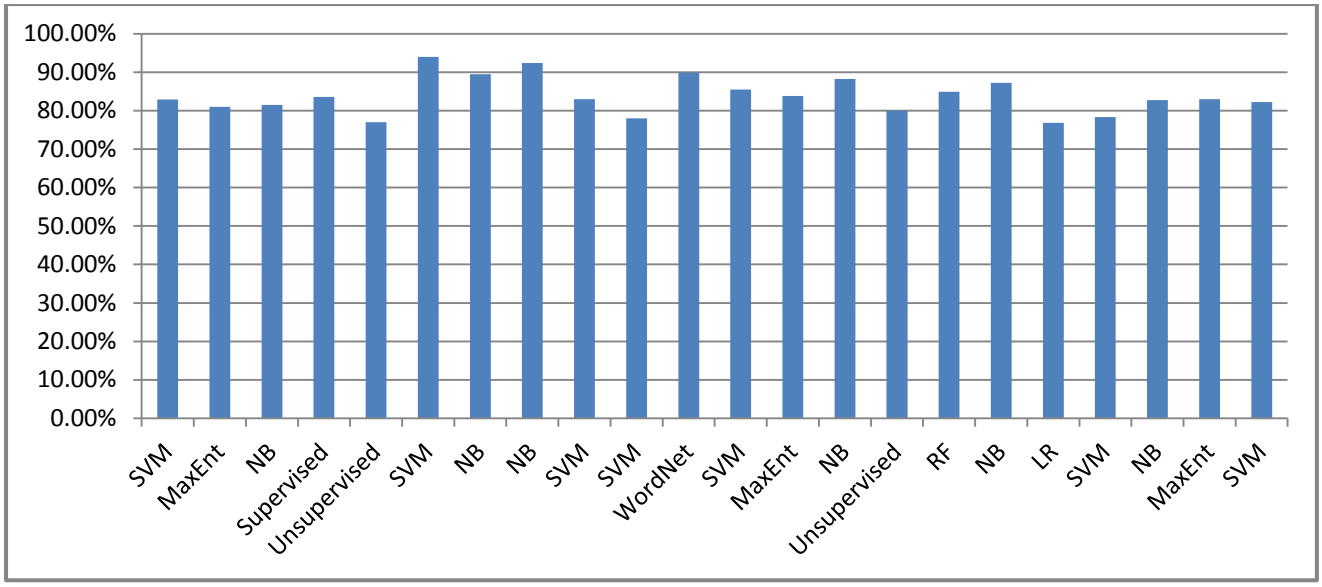

Figure 2: Accuracy Comparison Of Existing Techniques

4. DISCUSSION & CHALLENGES

Unigram improves performance than other features but still accuracy is issue because classifier performs average classification on dataset and accuracy is comparatively low than topic based categorization [2]. POS tagging is suggested to increase accuracy. Proper selection method is not used and some meanings conveyed are not captured [3]. SVM performs better than Naïve Bayes. By using KNN, maximum entropy classifier and stochastic gradient classifier can improve accuracy than present SVM [5]. Unigram performs very well. Various feature selection technique is applied but ensemble of feature selection can further improve accuracy and unigram with bag of word gives best accuracy [6]. POS tagging identifies tagging of word and produces improved result [9]. Ensemble of algorithms improves performance [8]. Accuracy is improved by considering emotions as noisy label in twitter dataset and use of WordNet dictionary generates better result than SVM, Maximum entropy and Naïve Bayes. Accuracy can be still improved by doing careful feature selection and proper classification technique [11]. Bag of word produce good accuracy compare to feature hashing but take more computational effort than feature hashing. Feature hashing takes less computational effort compare to bag of word but it is less accurate then bag of word. So both approaches have their pros and cons.

Accuracy is major issue in sentiment analysis. Another is sentiment lexicon that is to identify opinion from phrases and idioms. Some sentiment may have opposite opinion. Some sentences like interrogative and conditional have opinion and some time has not any opinion. Some sentences are sarcastic so opinion is changed. Many sentences without sentiment words also imply opinions. So these types of challenges can be solved using innovative approaches [1].

5. CONCLUSION

Referred papers have generated review as either positive or negative using classification techniques for sentiment analysis which produces result with vary accuracy. So the use of careful feature selection, POS tagging using Stanford tagger, SentiWordNet dictionary and proper classification algorithm has generated improved result so accuracy can be improved by using such combination of technique. Support vector machine is most widely used classification algorithm for sentiment analysis so it can generate better result. SVM have many non linear kernel functions which are Radial basic function, Polynomial Function and sigmoid kernel. RBF kernel function of SVM has hyper parameter which are gamma γ and margin constant C so these hyper parameter can enhance the performance of sentiment analysis by modifying different value of (C,γ) and chooses best pair of (C,γ) which gives better accuracy. So performance can be increased by modifying these hyper parameter values and also can find good value for these parameters for particular dataset.

6. REFERENCES

[1] Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp.1-167.

[2] Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[3] Chaovalit, P. and Zhou, L., 2005, January. Movie review mining: A comparison between supervised and unsupervised classification approaches. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on (pp. 112c-112c). IEEE.

[4] Miller, G.A., 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11), pp.39-41.

[5] Tripathy, A., Agrawal, A. and Rath, S.K., 2015. Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science,57, pp.821-829.

[6] Shahana, P.H. and Omman, B., 2015. Evaluation of Features on Sentimental Analysis. Procedia Computer Science, 46, pp.1585-1592.

[7] Jeyapriya, A. and Selvi, K., 2015, February. Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on (pp. 548-552). IEEE.

[8] Kanakaraj, M. and Guddeti, R.M.R., 2015, February. Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In Semantic Computing (ICSC), 2015 IEEE International Conference on (pp. 169-170). IEEE.

[9] Mouthami, K., Devi, K.N. and Bhaskaran, V.M., 2013, February. Sentiment analysis and classification based on textual reviews. In Information Communication and Embedded Systems (ICICES), 2013 International Conference on (pp. 271-276). IEEE.

[10] Bhadane, C., Dalal, H. and Doshi, H., 2015. Sentiment analysis: Measuring opinions. Procedia Computer Science, 45, pp.808-814.

[11] Gautam, G. and Yadav, D., 2014, August. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Contemporary Computing (IC3), 2014 Seventh International Conference on (pp. 437-442). IEEE.

[12] Zhou, X., Tao, X., Yong, J. and Yang, Z., 2013, June. Sentiment analysis on tweets for social events. In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on (pp. 557-562). IEEE.

[13] Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1, p.12.

[14] Khan, F.H., Qamar, U. and Javed, M.Y., 2014, November. Sentiview: A visual sentiment analysis framework. In Information Society (i-Society), 2014 International Conference on (pp. 291-296). IEEE.

[15] da Silva, N.F., Hruschka, E.R. and Hruschka, E.R., 2014. Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 66, pp.170-179.

[16] Michael W. Berry, Soft Computing in Data Science, First International Conference, Scds 2015, Putrajaya, Malaysia, September 2-3, 2015, Proceedings (Communications in Computer and Information Science)

[17] Patel, S.N. and Choksi, M.J.B., 2015. A Survey of Sentiment Classification Techniques. Journal for Research| Volume, 1(01).

[18] https://en.wikipedia.org/wiki/Part-of-speech_tagging

7. AUTHOR PROFILE

Prof. (Dr.) Vimalkumar B Vaghela, holds Ph.D. in Computer Science & Engineering. This author is Young Scientist awarded from Who's Science & Engineering 2010-2011 & also his biography is included in American Biographical Institute in 2011. His publication is also available in ieeeexplorer and also in spocus online database. He is currently working as Assistant Professor in Computer Engineering Department at L.D. College of Engineering, Ahmedabad, Gujarat, India. He received the B.E. degree in Computer Engineering from C. U. Shah College of Engineering and Technology, in 2002 & M.E. degree in Computer Engineering from Dharmsinh Desai Univrsity, in 2009. He has published book titled "Ensemble Classifier in Data Mining" in LABERT Academic Publisher, Germany, 2012 & "Operating System" in Dreamtech-India 2015. His research areas are Relational Data Mining, Ensemble Classifier, Pattern Mining. Author has published / presented more than 23 international papers and 5 national papers.

Bhumika M. Jadav, is currently pursuing her M.E. in Computer Science & Engineering Department at L.D. College of Engineering, Ahmedabad. She has received the degree of B.E. from G.E.C. Rajkot in 2010.