Sentiment Analysis of Movie Reviews Using Heterogeneous Features

Madhuri Dixit[1], Ashok Kumar Verma[2]
[1]Dept of CSE
[2]Associate Professor, Dept of CSE
[1, 2] Gyan Ganga Institute of Technology and Sciences
Jabalpur, Madhya Pradesh, India.

Abstract- *Data classification is highly significant in data mining which leads to a number of studies in machine learning with preprocessing and algorithmic technique. Class imbalance is a problem in data classification wherein a class of data will outnumber another data class. Sentiment Analysis is an evaluation of written and spoken language which determines a person's expressions, sentiments, emotions and attitudes and is commonly used as dataset in machine learning.*

Twitter is an emerging platform to express the opinion on various issues. Plenty of approaches like machine learning, information retrieval and NLP have been exercised to figure out the sentiment of the tweets. We have used movie reviews as our data set for training as well as testing and merged the naive bayes and adjective analysis for finding the polarity of the ambiguous tweets. Experimental outputs reveal that the overall accuracy of the process is improved using this model. In this work we have focused on two areas like: Feature Selection and entropy method, and second using machine learning techniques. We use "Twitter" movie review dataset. We also use accuracy comparison framework for comparing algorithms based on execution time.

Keywords- Sentiment analysis, Twitter, Adjective analysis, Naive bayes, Entropy method.

I. INTRODUCTION

Today's information-based society, also sometimes referred to as the digital age, is distinguished by the rapidly increasing amount of informational data. The younger generation in particular, covers a major percentage of online users, generating a vast amount of informal subjective content. This web based data is useful to extract meaningful information to study and add value for multiple application domains [1]. The data produced is majorly unstructured data (opinionated text) which is processed in sentiment analysis and then labeled into various categories, namely positive and negative. Sentimental analysis undergoes various challenges, some of which outlined and categorized by Ohana [2] are as follows:

- *Implicit Sentiment*: There are examples of such sentences which implicitly own a strong sentiment without containing any significant word to showcase the sentiment. E.g. one has to on quite some medicines in order to make a documentary like this.
- *Domain Dependency Sentiment*: Polarity and subjectivity both are expressed via structure and vocabulary which are hence influenced by domain. Many words denote differing polarity in various domains, as shown. E.g. The TV show was inspired from an American novel. I was inspired from the book.
- *Let Down Expectations*: It so happens that sometimes the writer showcases a positive context of the content and then in the end he disapproves it. E.g. A fun read book which is full of inspiring and motivating traits for adolescents/college students but all in vain due to lack of creativity and imagination.
- *Pragmatics:* The elements of communication of the user need to be recognized.
  E.g. It was good to see after so long time. It rain destroyed my mood.
- *World Knowledge*: Sometimes one has to have the knowledge and whereabouts of an entity mentioned in the statement in order to recognize the sentiment. E.g. She is as kind as a witch. So a person needs to know about "witch" in order to realize the accurate sentiment in the sentence.
- *Subjectivity Detection*: It is necessary to differentiate between the sentences which are rich in sentiment and the sentences which are just neutral in meaning. E.g. I love cars. I hate the movie "love in car‟.
- *Entity identification*: When multiple entities exist in a particular sentence it becomes necessary to recognize and differentiate which sentiment is used for which entity. E.g. Roger is better than Nadal. In this statement, a positive sentiment is for Roger whereas negative for Nadal.

- *Negation Handling*: Handling a negation in a sentence sometimes becomes very difficult.

In this thesis we have proposed a negation handling technique instead of automatically determining the scope of negations. This technique is can be broadly classified for detection and handing of syntactic and morphologic negation. The methodology has been used for a dataset consisting of movie reviews [3, 4]. The process of sentiment analysis can be depicted through the following flowchart (Fig. 1.1):

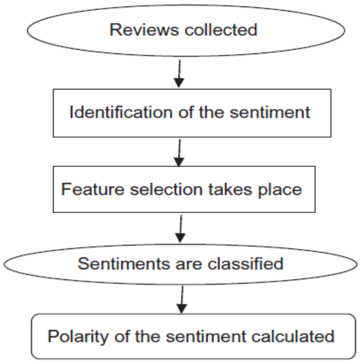

Reviews collected
Identification of the sentiment
Feature selection takes place
Sentiments are classified
Polarity of the sentiment calculated

Figure 1.1: Tweets review process using sentiment analysis.

Sentiments perform very important role for predicting future sales performance, a mix of good and even bad reviews will create a positive effect on the sales performance and sales prediction. In this thesis, the different types of issues are encountered like modelling the online reviews and tweets, sales prediction, and deriving the actionable knowledge. The sentiments, movie past sales performance these factors are important for predicting sales performance of movies. A most current sentimental approach is that classifies the reviews into positive and negative which not gives complete understanding of sentiments. So that we done mining of sentiments based on Probabilistic Latent Semantic Analysis (PLSA) is S-PLSA model which is totally different from traditional PLSA model. These models consist of the sentiments from reviews and tweets as the joint result of hidden factors and handle multifaceted nature of sentiments. S-PLSA model considered sentiments as a substitute of topics. So instead of using bag of words only sentimental words are considered. In this model only appraisal words from reviews and tweets are exploited for composing the feature and used for gathering the hidden sentiment factors.

1.1 Opinion Mining Overview: Textual information available on the web can be broadly categorized as fact or opinion. A sentence which contains fact is known as objective sentence whereas the sentence which contains opinion is known as subjective sentence [3]. With rapid expansion of ecommerce culture, there has been a substantial increase in the number of opinions that are available online in the form of reviews, blogs etc. Online merchants enable customers these days to review the products they have purchased. People who are performing online shopping mostly rely on these reviews, while they are also used by businesses and organizations to provide better product or service. This not only enhances shopping experience of existing customers, but also helps potential customers to form an opinion about that product or service. This sort of sentiment evaluation is gaining much more importance due to increasing trend of online shopping. However, due to the huge amount of available reviews, it is difficult for the new customer to analyze them manually and make an informed decision. Therefore, identifying and analyzing these reviews has become one of the major problems. This process of extracting of people’s opinion, experience and emotions from reviews, blogs and other sources is known as opinion mining [3]. An opinion is a quintuple of five things: entity name, aspect of entity, the orientation of opinion, name of opinion holder and the time at which the opinion is expressed. Researchers in past have used different approaches to extract opinions. Broadly opinion mining can be extracted in two ways: machine learning based approach and lexicon-based approach [4].

1.2 Challenges in opinion Mining:

The major challenges in Opinion Mining are listed below:

- Difficulties in effectively measuring similarity among short texts leading to ambiguity issues.
- Text retrieval time is slower.
- Accuracy issues in may produce faulty results while detecting polarity.
- Ambiguous words may affect further Classification and clustering strategies.
- Sentiment is difficult to comprehend because of poor abbreviations, spelling mistakes and grammatical errors.

If the challenges of sentiment analysis are to be well thought of, the foremost thing that is to be kept in mind is the fact that human beings favor giving multifaceted judgments, where the lexical substance can itself be deceiving. Dealing with cynicism, mockery, and repercussions is a big issue in the field of sentiment analysis. While dealing with the opinions, variations in matter or reversals within the wordings is also to be considered. The categorization factor is also of importance in terms of analyzing, for example, we can grade the consumers or the text itself, or the sentences along with paragraphs, or the preset adjective expressions, or even single words or a single comment. Short phrases, for example, can serve as building blocks of sentiment analysis. Phrases like "highest prices" and "lowest quality" brings out the actual essence of the sentiment present in a text, so a method should be devised before advancing toward sorting. If we consider that there is some connection between equal polarity words and reviews, a set of keywords might be favorable enough to identify polarity. Other than human-generated keyword lists, there is existence of information-driven methods as well which yield better lists than those of humans, but unigram means can give up to 80% accuracy while listing keywords. One of the popular ways to address sentiments is by analyzing smiley's or emoticons in tweets or texts (Figure 1.2). Smileys have the advantage that they have very short texts associated with them, hence the overhead is low and sentiments can be expressed very precisely with these emoticons.

Sentiment research-related datasets are very much domain sensitive and at the same time it is complicated to accumulate or construct them.

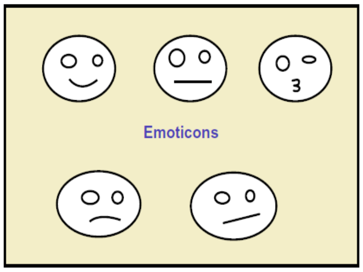


Figure 1.2: Emoticons.

Pang and Lee provides reviewed movie datasets, scaled sentiment datasets as well as subjectivity datasets. If datasets for sentiment analysis is to be created, both self-annotated and hand-annotated data have to be used. While the former has built in labeling by the creator, the latter has annotated data autonomous of the creator resulting in the process to be more laborious and varies on reliability. It has been observed that to extract the sentiment of a phrase, the best possible technique is to assign a real number measure and categorize as affirmative or negative sentiment of that phrase. From here rose the concept of polarity, which is a dual value, either representing positive or negative sentiment of the phrase, in which Pos and Neg categories are used to categorize the relevant words [5]. Another way of extracting sentiments from words is by using the Wordnet, where similar meaning words are grouped in synsets and the relationship between words were found out. Even, polarity identification was combined with WordNet [6], where a set of similar adjectives of identified orientation is started with. To find out the proximity of unknown adjectives, synonymy or antonymy is used to group them. Labeling is done based on the familiarity to positive or negative words. Ultimately, latest labeled words are added to the set. Then to uncover the polarity of a sentence, judgment-based sentences which contain the predefined set of adjectives are extorted. At the end, the sentences are evaluated based on the number of positive or negative word counts. This experiment reveals high accuracy results, and it is done quickly, with no training data being necessary.

The data we using in this experiment are movie reviews. We have collected about 17000 movie reviews from twitter. The movie reviews contains reviews of different movies. Reviews can categorize in three ways:

1. Positive reviews: messages in which people liked the movie.
2. Negative reviews: messages in which people not liked the movie.
3. Neutral reviews: messages in which people doesn't have any emotion or based on mere fact.

We show comparison between the different machine learning classifiers and find out which will give best results among these.

II. LITERATURE REVIEW

2.1 Machine Learning

Machine Learning is a field of artificial intelligence. In 1959 Arthur Samuel defined this as a "Field of study that gives computers the ability to learn without being explicitly programmed" [7]. Machine Learning is a study of algorithms which can learn from given data and which can make predictions on new data. Such algorithms can be made from a model which takes training data as input and will make a prediction on new input data based on knowledge gained from training data rather than following some static instructions. It can also add newly predicted data to its training data-set in order to improve efficiency. It can be categorized into three categories as supervised learning, unsupervised learning and Reinforcement learning.

2.2 Naïve Bayes Classifier

Naive Bayes classifier uses a bayes theorem as name suggests. Here Naive stands for the assumption that features used in classification are strongly independent of each other i.e. value of one feature does not depend on the value of other feature [8]. It is one of the standard method in case of text (or document) classification such as positive or negative, spam or not-spam etc. It uses word count as features for classification. The probability of each feature will be calculated for all the classes and text will be classified into a class for which it got a maximum probability. As it works on probability of features so it is a probabilistic classifier. It works very well under supervised learning technique. One main advantage of naive Bayes is it needs fewer amounts of labeled data for training and classification.

2.3 Random Forest Classifier

Random Forest Classifier uses ensemble method i.e. more than one decision trees. These decision trees choose random samples and features from given training data for decision making. These chosen data can overlap with each other. Also, size of chosen data may not be same for all trees. New data predictions will be made based on the majority vote from predictions of all trees. The advantage of Random Forest classifier is, it does not expect linear features and moreover, as trees are being used it can handle a large number of training data easily [9].

2.4 Support Vector Machine

Support Vector Machine is a supervised machine learning algorithms. Labeled data will be passed to it for training purpose. Training data will be having only two classes (say either positive or negative). These data will be analyzed to build a model that can be used to predict results for new data. Hence, SVM does not use probability for classification and also classification is taking place only in one of two classes so it is a non-probabilistic binary linear classifier [10, 11].

2.5 Multinomial Naive Bayes

According to Rajni Singh and Rajdeep Kaur [12], they proposed combined dictionary based on social media keywords and online review and also find hidden relationship pattern from these keyword. In Dataset review text and class label (overall user sentiment) are only two attributes. They set simple rules for scaling the user review. For dataset, a user rating greater than 6 is considered as positive, between 4 to 6 considered as neutral and less than 4 considered as negative then they applied Preprocessing, in which word parsing and tokenization, Removal of stop words, stemming (used Snowball). Here, naïve bayes multinomial classifier has been used. Combined word of twitter dataset and online review dataset forms a dictionary. As after classifying each word probability as positive, negative and neutral. Compare the probability for each word and categorize each word into three different dictionaries based on highest polarity of each word.

Some of the example features are:

i) Terms and their frequency
ii) Part of speech
iii) Sentiment words and phrases
iv) Rules of opinions
v) Sentiment shifters
vi) Syntactic dependency.

2.6 Corpus Based Approach

Corpus Based Approach is a technique of classification of Sentiment Analysis. It is mainly well suited for twitter data. We shall go with the lexicon based approach for our sentiment analysis on the twitter data as Lexicon based approach gives the best accuracy result with twitter data. A bag of positive, negative and neutral words is introduced and classification of the tweet is based on it. New abbreviations and commonly used words which aren't in the bag are added separately to each of the bag to increase the accuracy of the result. For emoticons we have introduced emojis dictionary which decodes every emoticon available into simple words which can be easily classified into the 3 bag of words available.

A) *Subjectivity/Objectivity-* To perform sentiment analysis we first need to identify the subjective and objective text. Only subjective text holds the sentiments. Objective text contains only factual information.

Example-

1.) Subjective**:** Titanic is a superb movie. (This sentence has a sentiment (superb), thus it is subjective)
2.) Objective**:** James Cameron is the director of titanic. (This sentence has no sentiment, it is a fact, and thus it is objective)[13].

B) *Polarity-* Further subjective text can be classified into 3 categories based on the sentiments conveyed in the text.

1.) Positive: *I love new Samsung galaxy mobile.*
2.) Negative: *The picture quality of camera was awful.*
3.) Neutral: *I* usually get hungry by noon.

This sentence has user's views, feelings hence it is subjective but as it does not have any positive or negative polarity so it is neutral. This positive, negative and neutral nature of text is termed as polarity of text. There is a lot of debate whether to take two or three classes but it is found that by considering neutral class accuracy gets increased. There are two ways for it: either classify text into two classes positive/negative and neutral and then further handling positive/negative or classify text into three classes in first step only [14].

C) *Sentiment level-* sentiment analysis can be performed at various levels –

- Document Level- In it the whole document is given a single polarity positive, negative or objective [15].
- Sentence Level – In it document is classified at sentence level. Each sentence is analyzed separately and classified as negative, positive or objective. Thus overall document has a number of sentences where each sentence has its own polarity.
- Phrase Level- It involves much deeper analysis of text and deals with identification of the phrases or aspects in a sentence and analyzing the phrases and classifies them as positive, negative or objective. It is also called aspect based analysis.

Liza et al [16] they suggest three phases of text mining i.e. pre-processing, processing and validation. After applying primary pre-processing, it performs weighting schemes and use Naïve Bayes as a classification algorithm. Then after in validation phase uses 10-fold cross validation testing. Yunchao et al [17] present both unigram and bigram as feature extraction and cluster the texts using K-Means clustering. And after Naïve Bayes classification algorithm is applied. Rishabh et al [18] proposed cluster-than-predict approach, first cluster the tweets using K-Means clustering and then perform classification using CART (Classification and Regression Trees) to improve the accuracy.

Gang Li et al [19] use TF-IDF scheme as feature extraction. Then for improve the result and detect neutral polarity they suggest voting mechanism and distance measure approach. After that apply K-Means clustering to find the review i.e. positive, negative or neutral. Hima et al [20] a novel fuzzy clustering model and compare it with K-Means and Expectation Maximization algorithms. And the result is practicable for high quality twitter sentiment analysis. Nagamma et al [21] applied TF-IDF for feature extraction and after Fuzzy C-Means clustering is used to improve the result of classification and after apply Support Vector Machine and Naïve Bayes. Then predict the revenue from the reviews about movie. Yunchao et al [22] address to estimate the sentiment of unlabeled data; they use a two-step-merge method. They use clustering for sparsity problem and NB classifier for categories the text. It gives the better result than bag of words method.

N. Mitta l& B. Agarwal [23] proposed “A Hybrid Approach for Twitter Sentiment Analysis” Here; the author had proposed that the system has three stage of sentiment extraction. The polarity of tweets calculated with predefine list of words. In this paper hybrid approach has used. The author conclude hybrid approach improve accuracy. In this paper, the author Alexande Pak and Patrick Paroubek [24] author had applied the data mining process the author had applied data mining process which is applied for tweets as well as for obtain better accuracy they are using data mining algorithm called Knearest neighbour (IBK) out forms. For classification of tweets in positive, negative and neutral they are used three classifiers which are Random Forest, BaysNet, and Naïve Bayesian. Because they are used Knearest neighbour they not need to use ensemble of classifier for sentiment prediction of tweets.

Karthika, S. Priyadharshini and Assistant Professor [25] in 2017 proposed “Survey on Location based sentiment analysis of Twitter data” in this system the author has worked on big data concept. The big data is used for accumulating, reserving, and examining large volume of a data and provides decision making. Twitter as dataset applied. It is a survey paper. In this paper the text of tweet along with emojis are analyzed.

III. PROPOSED METHOD

OTP Comments, reviews and opinion of the people play an important role to determine whether a given population is satisfied with the product, services. It helps in predicting the sentiment of a wide variety of people on a particular event of interest like the review of a movie, their opinion on various topics roaming around the world. These data are essential for sentiment analysis. In order to discover the overall sentiment of population, retrieval of data from sources like Twitter, Facebook, Blogs are essential.

The limitations of available systems are not sufficient to deal with the complex structure of the big data. In this section, we present some of the limitations that are present in the existing system.

1) Classifiers such as SVM and Naïve Bayes used in previous system do not give much accuracy.
2) Inadequate reviews that leads to wastage of time and money. This issue is overcome in present system by the process of pre-processing. The exact reviews are provided to the users in the form of graphs based on polarity result which is easily understood by the user and it saves time as well as money of users.
3) In existing system feature scores are not present, so it becomes difficult to decide which phrase is either completely positive or completely negative for example phrase such as

“NOT VERY BAD”. This problem is solved by providing labels to each features, we have used five feature labels such as strong positive, strong negative, and neutral features by which the exact result based on polarity is obtained.

The proposed method helps to eliminate all the drawbacks mentioned above.

3.1 Proposed Algorithm:

Input: dataset.

Output: classified output.

1. Take a data set as input.
2. If that set has more features then apply the feature selection technique as pre-processing technique.
3. Apply parallelism from step 4 to step 6.
4. Evaluate the entropy value and information gain ratio.
5. Construct the models separately using proposed algorithm based on entropies.
6. Find the accuracy and execution time of each model and store the value in array.
7. Find a model that has maximum Accuracy.
8. If two have maximum accuracy then
9. Find a minimum execution time of the model that has maximum accuracy.
10. Classify by that model which has minimum execution time.
11. Else classification done by the model which has maximum accuracy.
12. End

1. Input Data: The data we using in this experiment are movie reviews. We have collected about 17000 movie reviews from various social sites. The movie reviews contains reviews of different movies. After releasing of any new movie the reviews of that movie are added to the dataset. Reviews can categorize in three ways:

a). Positive reviews: messages in which people liked the movie.

b). Negative reviews: messages in which people not liked the movie.

c). Neutral reviews: messages in which people don’t have any emotion or based on mere fact.

2. Pre-Processing: We then remove the stop words from the collected corpus to make the content free from commas, full stops etc. Stop words are the words like “a”,”is”,”the”, “etc” etc; these words has nothing to do with the emotion, so has to be discarded from the message. Now next step is to train the data using supervised classifier.

We have found that to get desired results from the classifier we have to make sure that the tweets can be processed properly. As tweets can be in user language, so we have to clean every data which are irrelevant to the data. The following things which can be irrelevant to the data are:-

- URL’s: URL’s in the message will not make any sense as it simply distracts the result of classifier.
- Username: Removal of username can be necessary for cleaning purposes as it can effect falsely to our results.
- Repeated characters: If the character is repeated more than two times then it can be comprise new word but the meaning is same, so we have to eliminate that word and make the word genuine. For example good can be written as goooood.
- Repeated words: If the message contains word which has been appeared more than two times continuously then it has to be change into two times. For example great great great great movie can be covert to great great movie.

3. *Feature Extraction*

In the process of feature extraction, movie features are extracted from every sentence. For finding the polarity of text document, it is necessary to understand the sentiment score with its usage as well as their relationship with all the nearby words. Following are some features that affect the polarity of the document.

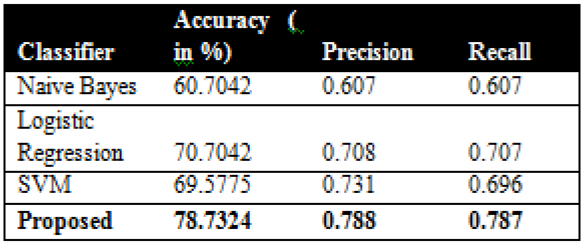
For primary feature extraction, we have used N-gram tokenizer which tokenizes the input tweet into word *n*-grams such as unigram, bigram etc. Frequency of an n-gram feature in a tweet is considered as the feature value. This is a collection of positive, negative and neutral words along with their broad part-of-speech categories. For our defined feature, the feature value is calculated based on how many polarity words of particular type are contained in a tweet. For example, if there are three sentiment classes such as positive, negative and neutral, we consider three SentiWordnet features i.e. how many positive words found in the SentiWordnet are also found in the tweet, how many negative words found in the SentiWordnet are also found in the tweet and how many neutral words found in the SentiWordnet are also found in the tweet. If there are *m* number of n-gram features, our feature set contains a total of *m+3* features where 3 is for the SentiWordnet features. We transform each tweet into vector presentation of length *m+3* and the vector is labeled with the class of the training tweet under consideration.

IV. RESULTS & EVALUATION

Multithreading Evaluations of various algorithms according to different parameters are displayed below:

The classification performance can be evaluated in three terms: accuracy, recall and precision as defined below. Accuracy explains correctly classified instances. Precision and Recall are in weighted average for positive and negative terms.

Table 4.1: Performance evaluation.



Classifier	Accuracy (in %)	Precision	Recall
Naive Bayes	60.7042	0.607	0.607
Logistic Regression	70.7042	0.708	0.707
SVM	69.5775	0.731	0.696
Proposed	**78.7324**	**0.788**	**0.787**

Chart below represent comparison of accuracy between different algorithms.

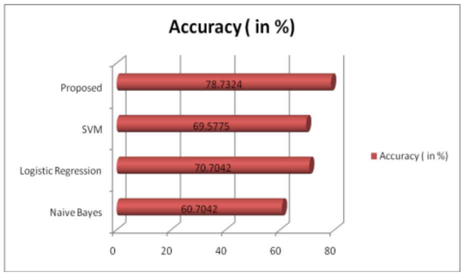


Figure 4.1: Evaluation of accuracy.

Chart below represent comparison of precision between different algorithms.

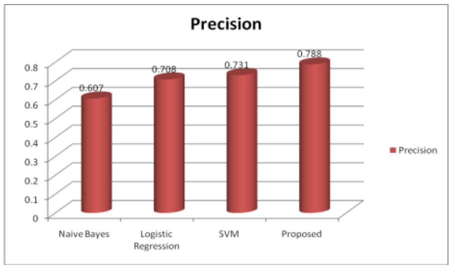


Figure 4.2: Evaluation of precision.

Chart below represent comparison of recall between different algorithms.

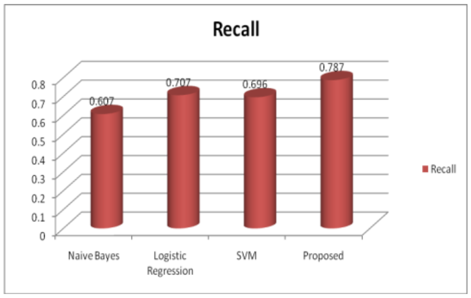


Figure 4.3: Evaluation of recall.

V. CONCLUSION AND FUTURE WORK.

The goal of our work is to overcome the problem of negation handling and feature extraction faced during sentiment analysis. In this thesis, we have proposed a system, which treats both syntactic as well as morphologic negations. Sentiment classification approaches suffer when negation is involved in the datasets. The system has been evaluated on the basis of annotated datasets (positive (1) and negative (0)). The syntactic negation was handled by exploiting grammatical relations among words. Stanford parser was used to identify the scope of negation. On the other hand, in order to handle morphological negations we analyzed the structure of words and its prefixes. Overcoming the limitations of existing window based negation handling, combination of both the techniques led to a better analysis of negative sentences. Our proposed methodology gives a higher accuracy, precision value and recall.

REFERENCES

[1] Agarwal, B. Xie., I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of twitter data”, in Proc. ACL 2011 Workshop on Languages in Social Media, 2011, pp. 30–38.

[2] B. Ohana and B. Tierney, “Sentiment classification of reviews using SentiWordNet”, IT&T Conference, 2009.

[3] Liu, Bing, and Lei Zhang. "A survey of opinion mining and sentiment analysis." In Mining text data, pp. 415-463. Springer US, 2012.

[4] Serrano-Guerrero, Jesus, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera Viedma. "Sentiment analysis: A review and comparative analysis of web services." Information Sciences 311 (2015): 18-38.

[5] V. Hatzivassiloglou, K.R.McKeown, Predicting the semantic orientation of adjectives, in: Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, ACL, New Brunswick, NJ, 1997, pp. 174–181.

[6] M. Hu, B. Liu, in: Mining and summarizing customer reviews, Paper Presented at the Proceedings of the Tenth

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[7] P. Simon, Too big to ignore: the business case for big data. Hoboken, New Jersey: John Wiley & Sons, Inc, 2013.

[8] I. Rish, “An empirical study of the naive bayes classifier,” in IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22. IBM New York, 2001, pp. 41–46.

[9] G. Malakar, “What is random forest algorithm?” https://www.youtube.com/watch?v=LIPtRVDmj1M, 2014.

[10] T. S. Korting, “How svm (support vector machine) algorithm works?” https://www.youtube.com/watch?v=1NxnPkZM9bc, 2014.

[11] Ashish Shukla and Rahul Misra, “Sentiment Classification and Analysis Using Modified K-Means and Naïve Bayes Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering 5(8), August- 2015, pp. 80-85.

[12] Singh, Rajni, and Rajdeep Kaur. "Sentiment Analysis on Social Media and Online Review." International Journal of Computer Applications 121, no. 20 (2015).

[13] Arora, Piyush. “Sentiment Analysis for Hindi Language.” Diss. International Institute of Information Technology Hyderabad, 2013.

[14] Pang B, Lee L., “Opinion mining and sentiment analysis” Found Trends Inform Retriev: 1–135, 2008.

[15] B. Liu, Sentiment analysis and opinion mining, Syn. Lect. Human Lang. Tech. (2012).

[16] Liza Mikarsa, Sherly Novianti Thahir, "A Text Mining Application of Emotion Classifications of Twitter’s user using Naïve Bayes Method", IEEE, 2015

[17] Yunchao He, Chin-Sheng Yang et al, "Sentiment Classification of Short Texts based on Semantic Clustering", IEEE,2015.

[18] Rishabh Soni, K. James Mathai, "Effective Sentiment Analysis of a Launched Product using Clustering and Decision Tree", International Journal of Innovative Research in Computer and Communication Engineering, vol.4, pp.884-891, January 2016.

[19] Gang Li, Fei Liu, “Sentiment Analysis based on Clustering: A Framework in Improving Accuracy and Recognizing Neutral Opinions”, Springer, September 2013.

[20] Hima Suresh, Dr.Gladston Raj. S, "An Unsupervised Fuzzy Clustering Method for Twitter Sentiment Analysis", IEEE, 2016.

[21] Nagamma P, Pruthvi H.R et al, "An Improved Sentiment Analysis of Online Movie Reviews based on Clustering for Box-Office Prediction", IEEE, 2015.

[22] Yunchao He, Chin-Sheng Yang, Liang-Chih Yu, K. Robert Lai, Weiyi Liu, “Sentiment Classification of Short Texts based on Semantic Clustering”, IEEE, 2015.

[23] Namita Mittal, Basant Agarwal, Saurabh Agarwal, Shubham Agarwal, Pramod Gupta, “A Hybrid Apprach for Twitter Sentiment Analysis,” Proceedings of ICON-2013: 10 th International Conference on Natural Language Processing, Noida, India, 2013, pp: 116-120.

[24] Alexander Pak, Patrick Paroubek, ”Twitter as a Corpus for Sentiment Analysis and Opinion Mining”,University de Paris-Sud, Laboratoire LIMSI-CNRS, Batiment 508, F-91405 Orsay Cedex, France2016.

[25] Karthika, 2 S. Priyadharshini, Assistant Professor, 2PG Scholar,”Survey on Location based sentiment analysis of Twitter data” 2017 IJEDR | Volume 5, Issue 1 | ISSN:2321-9939.