Hybrid Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets

G. Vaitheeswaran[1], L. Arockiam[2]

[1]*Research Scholar*, [2]*Associate Professor*

[1,2] *Department of Computer Science,*
[1,2] *St. Joseph's College (Autonomous),*
[1,2] *Tiruchirappalli, India*

***Abstract*— The opinion of others toward a product or event plays an important role in the decision-making process. In recent times, the explosion of social media over the web has a rich impact on an individual's and the organization's decision-making process about certain content. Twitter which is a leading micro-blogging website allows the people to post their opinions, state of mind, or status toward products or any events such as politics and sports. Sentiment analysis performs an essential role to extract the content of the tweets. This research work consists of two components: a lexicon builder and a sentiment classifier. The proposed hybrid based Senti_*i*Lexi_*i*Classi approach is the combination of lexicon and machine learning based approaches. In sentiment analysis, the context-dependent word is a major problem. To solve this problem, the existing contextual valence shifter (CVS) method is improved and used to assign polarity values using the proposed Context Polarity Measure (CPM). To find the polarity of context-dependent text, the n-grams model is used. The tweets are classified more accurately and produces better results.**

***Keywords*— sentiment analysis, context dependent word, hybrid based approach, retweet count, support vector machines.**

I. INTRODUCTION

Everyday people share their state-of-mind on the internet via social media such as Facebook, Twitter, Linkedin, YouTube, etc. Among them Twitter is a powerful microblogging website. The tweets posted on twitter are knowledge rich documents. The twitter contains a rich source of information for processing sentiment analysis. The ability to extract such hidden knowledge of those documents is referred as 'sentiment analysis'. Sentiment analysis is the computational study of the 'state-of-mind' towards a product / item / entity. It is also known as opinion mining.

One of the major issues and challenging task in the process of sentiment analysis is to measure the sentiment depth of a context word in a sentence. In order to overcome this issue, a new approach is proposed by combining the idea of lexicon based and machine learning based approaches named Senti_*i*Lexi_*i*Classi approach. The context dependent words are analyzed using the proposed Senti_*i*Lexi approach. Later the labelled datasets are analyzed using the Senti_*i*Classi approach.

The consecutive sections are organized as follows: The context dependent word is explained in the background study. In the related works section, a review of some related works on sentiment analysis using hybrid approach has been deliberated briefly. The main objective of this proposed work has been provided in the section 4. The methodology diagram and proposed approach for sentiment analysis have been analyzed in the sections 5 and 6 respectively. To what extent the work has been done is summarized in the last section of this article.

II. BACKGROUND STUDY

A. *Context Dependent Word*

Most opinion mining systems rely on the extraction of sentiment words to detect opinions. These words, refer as polarized words, and deliver valuable information about the semantic orientation (positive or negative) of a text. However, the appearance of context in these words may modify their valence in many ways. Although being of importance, this issue has been investigated recently and is now increasing the attention of the researchers.

Polanyi et al. [1] had proposed the existence of contextual valence shifters, which are contextual phenomena altering the prior polarity of a term. Afterwards, some of these phenomena (such as negative or conditional syntactic structures) were dealt with on a case by case basis [2] [3] [4]. Later [5] [6] aimed at the best modelling, the expression of opinions before embedding context dependent word in a classification system. The main purposes of these studies are to determine a list of contextual valence shifters that impact the polarity of a term as well as to define the nature of this impact.

However, these lists are often manually built from linguistic intuitions and not learned from language data. Works depend on a corpus of texts to develop resources that reflect the best actual roles played by the linguistic context for opinion mining are limited. [7] had suggested a technique to automatically select valence shifting features in order to improve a sentiment classification system based on a machine-learning approach. All these studies agree that the contextual valence shifters can have diverse impacts on polarized words.

To understand in simple terms, the following are the examples for the context dependent and context free word:

Context-free:

– e.g., "an interesting film" (*Positive*)
– e.g., "a terrible situation" (*Negative*)

Context-dependent:

– e.g., "high quality", "low cost" (*Positive*)
– e.g., "high cost", "low quality" (*Negative*)

B. Tokenization into N-grams:

It is a process of creating bag-of-words from the text. Tokenization of social-media data is considerably more difficult than the tokenization of general text. The social media data contain many emoticons, URL links, special characters and abbreviations which cannot be easily separated as whole entities. It is a general practice to combine accompanying words into phrases or n-grams, which can be unigrams, bigrams, trigrams, etc. Unigrams are single words, bigrams are collections of two neighbouring words and trigrams are collections of three neighbouring words. In simple term, n-grams define a subsequence of n items from a given sequence. N-grams method can decrease bias, but it may increase statistical sparseness. It has been shown that the use of n-grams can improve the quality of text classification [8] [9]; however, there is no unique solution for the size of n-grams.

III. RELATED WORKS

Lei Zhang et al. [10], proposed the lexicon-based approach which had given high precision but low recall. The Pearson’s chi-square test was used to improve recall and also to identify the opinionated word that are not in the lexicon. To assign polarities of the identified tweets, the SVM based classifier was accomplished. Instead of being labeled manually, the examples were given by the lexicon-based approach. The test data were used as the result of chi-square’s resulted data. The LMS method produced better accuracy of 85.4% than the standard model.

Taken into consideration, the unigrams and bigrams features together will have produced better accuracy. The original lexicon was built with the unigrams pattern. In some cases, to find the polarity information of the sentiment word became difficult, because the original lexicon was built with the unigrams pattern. To strengthen the meaning of the sentiment, it can be built with the bigram patterns that has an adverb or a negative word.

Hanhoon Kang et al. [11], had suggested the better accuracy can be achieved in the lexicon based approach by encompassing the unigrams pattern along with the bigrams pattern (such as negative words and intensive adverbs) in the lexicon.

Hashtags, URLs, texts were used as the features to evaluate the performance of the support vector machines classifier [12].

Twitter is a micro-blogging website, and the tweets are limited to 140 characters. Geeta et al. [13] had suggested, there is no need to use external dictionaries or any other lexicons of polarized words to find the sentiment polarity in tweets. Automatically a training dataset was induced by referring to the sentiment present in the tweets containing emoticons. It is able to map all common expressions with new words, slangs, and errors.

Alistair Kennedy et al. [14] had made a study about contextual valence shifters (CVS) characteristics and its occurrence in the text documents. It proposed two algorithms CVS and Term Counting and their results are compared. From the comparison the higher accuracy is achieved by the CVS. Vo Ngoc Phu et al. [15] improved CVS algorithm and proposed a methodology for sentiment analysis, which is a combination of CVS and Term Counting. By combining both processes, researchers have noticed an increase in the accuracy rate.

The author [16] suggested a methodology, to process the sentiment analysis effectively by using the big data analysis method. The author discussed the tools for sentiment analysis such as Hbase, Mahout, and Hadoop framework.

Hasan Saif et al. [17] proposed a lexicon-based approach for sentiment analysis on Twitter. Recently, it has attracted much attention, due to its wide applications in both commercial and public sectors. In this article, the author the recommended SentiCircles approach which is different from typical lexicon based approaches. The co-occurrence patterns of words in different contexts, in tweets are to capture their semantics, update their reassigned strength and polarity in sentiment lexicons accordingly.

IV. OBJECTIVE

From the above related works, it is revealed that, most of the works had been carried out using the emoticon and context dependent words separately to find the polarity of tweets. Very few amount of research work had been carried out in the context dependent word for short text messages. This motivated to build a new hybrid model based on combining the lexicon and machine learning methods to enhance the accuracy.

The objectives of this article are as follows:

- To provide a novel hybrid-based approach using emoticon and contextual word identification.
- To classify the sentiment words of tweets and also to establish a classifier model to produce better accuracy than the existing method.

V. METHODOLOGY DIAGRAM

The proposed approach is explained briefly along with a methodology diagram. Figure 1 shows an architectural overview of the proposed hybrid (Lexicon and Machine Learning) - based approach.

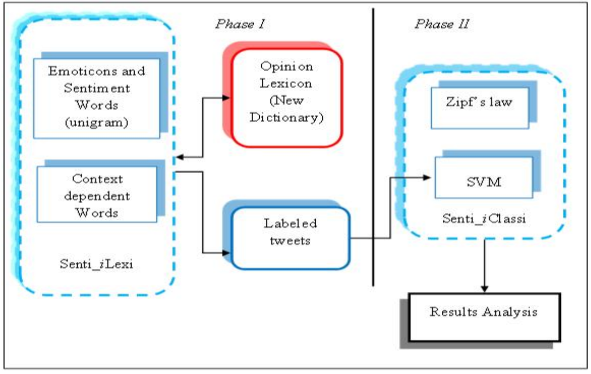

Fig 1 Methodology Diagram for Hybrid Approach

Phase I:

In the phase-I the raw tweets are preprocessed and stored into the text file. The preprocess steps include the following mechanisms:

- remove the URLs,
- remove hashtags,
- remove stopwords [list of stopwords created for this work],
- punctuation marks are not to be removed; since the emoticons contain punctuation symbols and special characters,
- convert the upper case letters to lower case,
- apply stemming and lemmatization; Porter stemmer is used for this approach,
- apply POS (Part-Of-Speech) tagging to tag each words.

The preprocessed tweets are tokenized using the POS tagger. The treebank POS tagger is used for this approach. To find the polarity for the context dependent word each and every word must be tagged with the part-of-speech. The POS tagged word helps to identify the adjective and nouns of the N-grams. Based on the token of the words the polarity of each word is calculated using the Context Polarity Measure (CPM) method.

Phase II:

Through the results of the existing article's contribution, it is easily understood that the machine learning based classifier produced better accuracy than the lexicon based classifier.

VI. PROPOSED APPROACH

Senti_*i*Lexi_*i*Classi Approach

After cleaning tweets, the sentence segmentation is performed. This segmentation separates the collected tweets into individual sentences. Afterward, the words are tokenized and tagged using Treebank POS tagger in each sentence. Figure 2 represents the procedure of the hybrid model for sentiment classification. The proposed hybrid method contains two functions. They are as follows:

(i) Senti_*i*Lexi() – to label the collected dataset as positive, negative, and neutral.

(ii) Senti_*i*Classi() – to build a machine learning-based classifier.

```
Procedure: Senti_iLexi_iClassi  /*Lexicon and
Machine Learning*/
Input    : Tweets
Output   : Accuracy
    1. Begin
    2. Preprocessed tweets
    3. Senti_Lexi()
    4. Print positive, negative and neutral tweets
    5. Senti_iClassi()
    6. Print precision, recall, F-score and Accuracy
    7. End
```

Fig 2 Procedure for Hybrid Approach

A. Senti_iLexi

To assign polarity values for the context-dependent text, the existing Contextual Valence Shifter (CVS) [15] method is enhanced and a new approach namely Context Polarity Measure (CPM) has been developed. The tweets are very short messages and limited to 140 characters. Even though the presence of n-grams (N>4) are very little possible, to find the context dependent word the n-grams technique is applied. The procedure for labeling the preprocessed datasets is given in the figure 3.

```
Procedure: Senti_iLexi
Input    : Preprocessed Tweets
Output: labeled tweets
    1. Begin
    2. Each tweet
    3. Apply n-grams
       Context_Polarity_Measure()

       Calculate positive, negative and neutral
    4. End
    5. Print total positive, negative and neutral tweets
    6. End
```

Fig 3 Procedure for Senti_*i*Lexi

Contextual Polarity Measure (CPM)

To identify the context-dependent word sentiment, the steps are proposed in the Context Polarity Measure (CPM) approach. For example the processing of connector step is explained briefly below.

– eg. *"The movie was awesome, but the plot was unexpected :)"*.

After preprocessing step, the stopwords are removed. Applying the usual n-grams (n=1) technique, the extracted sentiment words are "awesome, unexpected". The polarity value for the above sentence is "0."

In detail, the polarity for awesome is +1 and unexpected is -1. By adding the values of the sentiments in this sentence, gives the neutral polarity. But the actual polarity value of the sentence is negative. This is the identified context-dependent word problem.

To overcome this context-dependent problem, the following CPM approach is used. Let's consider the same example mentioned above. In the above mentioned sentence, the connector "but" is present, which performs the 'connectors processing' of step 2 in the CPM approach.

CPM is calculated based on CVS (Contextual Valence Shifters) method.

According to Processing connectors

Awesome (but) => +1 => 0 (adjust, since the presence of but)

Unexpected => -1 (the polarity value of negative word)

Sum of the valence => -1.

This indicates the statement is negative.

Figure 4 demonstrates the pseudo code for calculating the contextual polarity measure.

```
Approach: Context Polarity Measure(); Find the polarity
          for the context dependent words and
          emoticons.
Input: preprocessed tweets
Output: positive, negative and neutral tweets.
Begin
1. For each sentence in step 1, calculate valence;
    1.1if the sentences contain connector then go to 2;
    1.2 if the sentence contains ‘either .. or..’ then go to 3;
    1.3 if the sentence contains ‘neither .. or .. ‘ then go to 4;
    1.4 if the sentence contains comparative or superlative then go to 5;
    1.5 if the sentence contains intensifier or diminisher, then go to 6;
    1.6 if the sentence contains negative forms (not, never,
        nothing, no one, no more, none, none of) then go
        to 7.
    1.7 if the sentence contains emoticons then go to step
        8
    1.8 perform the following:
        Calculate valence of pharse, sentence and idiom;
        then, remove them out of sentence;
    For each individual word, calculate valence of each word;
    Sum all valence;
    Return valence of the sentence.
2. Processing connectors.
3. Processing ‘either..or..’.
4. Processing ‘neither ..or..’.
5. Processing comparative, superlative.
6. Processing intensifier, diminisher.
7. Processing negative forms.
8. Processing emoticons.
End
9. Sum all valence of each word.
10. Return valence of the tweet.
11. End
```

Fig 4 Procedure for Context Polarity Measure

B. Senti_iClassi

The labelled dataset obtained from the Senti_*i*Lexi is used to find the accuracy of the proposed model. The cross validation method is used for training and testing the model. The value for cross validation is set to 10. The results of the proposed model are analyzed using the four parameters namely accuracy, F-measure, precision, and recall, are discussed in the results and discussions section. Figure 5 depicts the process of the SVM classifier to find the accuracy.

```
Approach: Senti_iClassi. To enhance the accuracy
for the labeled tweets using SVM.
Input: Labeled tweets
Output:
        o  Find the best learning model.
        o  Enhance the accuracy of sentiment
           classification
Method:
    1. Preprocessed datasets
    2. Apply Zipfs’ law to select the relevant
       features
    3. Crossvalidation method (10 fold method)
    4. Apply SVM using SMO model
    5. Evaluate the model
    6. Results
```

Fig 5 Procedure for Senti_*i*Classi

Feature Selection

Feature selection is the process of selecting a subset of relevant features which have the highest predictive power. To enhance the accuracy, the n-grams are added as a new feature. The selected features for learning the model are as follows:

- *Unigrams*: The presence of positive, negative and neutral words in the tweets.
- *Emoticons*: Presence of positive, negative and neutral emoticons in the tweets.
- *Bigrams*: The negation words are tokenized along with the opinionated word and it is considered as the features.
- *Stemming*: The elongated words are stemmed and tokenized. The stemmed words are considered as the features in order to increase the accuracy.
- *Retweet count*: The positive, negative and neutral tweets posted by people are shared by other users, to show their state of mind towards the target. This shows the target contains strong polarity. To reduce the features space, the unigram and bigram features are extracted along with the presence of retweet count. This produces more accuracy.
- *n-grams*: Usually the context dependent text is found in the phrases or idioms of the sentences. Since the phrases are the mixed terms of more words, the n-grams technique is applied to validate the context dependent polarity. The presence and absence of n-grams are indicated using 1 and 0 respectively.

Frequency based Feature selection

In the existing work, the term-counting method was used to select the topmost frequency for the context-dependent word. Instead of the term-counting method, the Zipfs’ law is used to rank the top occurring words. The tweets are very short to form n-grams. In the condition of n-grams, if n>1, the words are tokenized and form into a single word. The n-grams are found very less in the obtained processed tweets.

Zipf's law is a law about the frequency distribution of words in a language (or in a collection that is large enough so that it is representative of the language). Let r be the rank of a word, Prob(r) be the probability of a word at rank r. The concept of the law is to find the rank of the word based on the frequency. The equation 1 represents the definition of the Zipfs' law.

By definition

$$Prob(r) = \frac{Freq(r)}{N} \quad \text{..........Eq (1)}$$

Where,
Freq(r) = the number of times the word at rank r appears in the collection.
N = total number of words in the collection (not number of unique words).

Support Vector Machines (SVM) Classifier

The SVM is a binary classifier that is the class labels can only take two values: +1 or -1.

In this case, the labelled data sets contain three classes namely positive, negative and neutral and these come under the 'three-class' problem. To build binary classifiers for multi-class SVM the class labels are distinguished as follows:

(i) between one of the labels and the rest (*one-versus-all*) or
(ii) between every pair of classes (*one-versus-one*).

The one-versus-one based classifier is used to learn the model. To evaluate the model, the binary Sequential Minimal Optimization (SMO) is applied by using the following one-versus-one classifier:

- positive, neutral
- positive, negative
- neutral, negative

The 10-fold cross validation is performed. The dataset is broken into 10 parts, keep 1 part aside for testing and training on 9 of them. The process is repeated for 10 times with each of the 10 parts used exactly once for testing. SVM classifier is trained and tested by using different features such as unigram, bigram, n-grams, emoticons, retweet count and stemming.

The kernel function used for the evaluation is, Polynomial kernel of degree d. The equation 2 states the function of polynomial kernel.

$$k(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d \quad \text{..... Eq (2)}$$

VII. RESULTS AND DISCUSSIONS

The computational time taken to find the polarity for the context dependent text is relatively high when compared with the unigram and bigram model, since tweets are very short text messages and contains less number of n-grams. This resulted in the lack of building n-grams model. Tokenizing each word with the POS tagger takes more time to produce the required output. An increase in correctly classified tweet has been found after considering the connectors and nouns of the words. This method gives more accuracy than the unigram model.

Figure 6 represents the correctly classified tweets using the proposed Senti_iLexi approach. X-axis represents the positive, negative and neutral label. Y-axis represents the count of the classified tweets. The positive, negative and neutral counts obtained are 555, 613 and 827 respectively. This is relatively high when compared with the existing lexicon based approach.

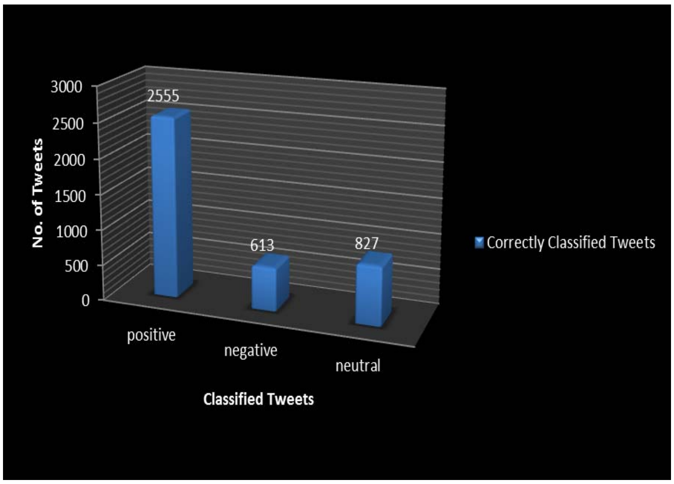

Fig 6 Correctly Classified Tweets

Figure 7 represents the positive, negative and neutral percentages of the classified tweets from the raw tweets. The positive tweets with 63.95%, negative tweets with 15.34% and neutral tweets with 20.70% are obtained. Using the context polarity measure, the positive and neutral tweets result in high accuracy. This result shows the importance of context-dependent word in the sentiment analysis.

Figure 8 represents the accuracy of the context-dependent word using the SVM classifier. X-axis represents the common measures. Y-axis represents the percentage of common measures. The obtained values for precision, recall, f-score and accuracy are 83.70%, 82.27%, 82.98% and 84.30% respectively.

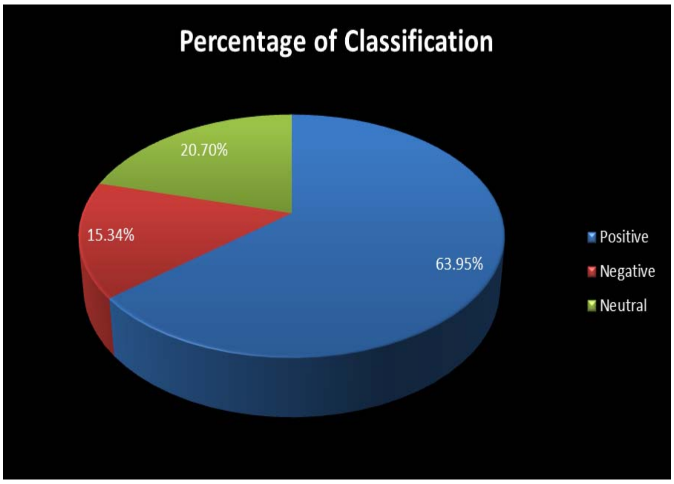

Fig 7 Percentage of Positive, Negative and Neutral for Senti_*i*Lexi

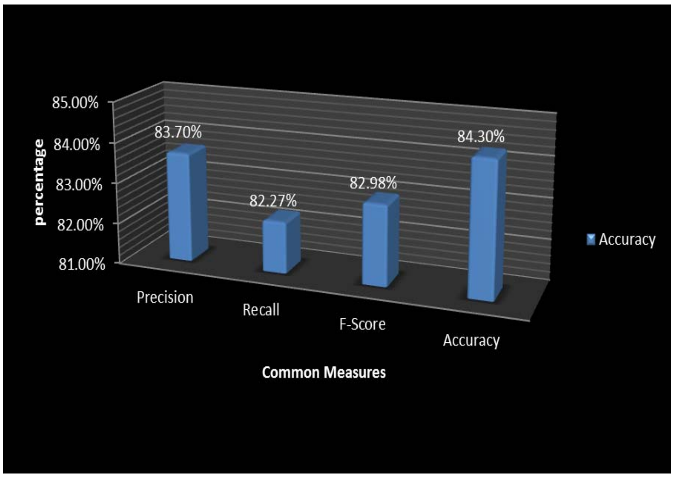

Fig 8 Accuracy of Classified Tweets for Senti_*i*Lexi_*i*Classi Approach

VIII. CONCLUSION

The methodological diagram discussed in this article delivers a big picture for processing hybrid approach of sentiment analysis on short text data. This article has presented a novel hybrid-based approach for analyzing sentiments. A new approach *Senti_iLexi_iClassi* has been proposed to provide better accuracy. Adding n-grams as a feature creates the probability sparseness. The time taken for calculating the accuracy in Senti_iLexi_iClassi is higher than the existing lexicon based approach.

REFERENCES

[1] L. Polanyi and A. Zaenen. “Contextual valence shifters”, In Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text, 2004, pp. 106–111.

[2] S. Das and M. Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. In Proceedings of the 8th Asia Pacific Finance Association Annual Conference, 2001, pp. 37–56.

[3] J.C. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou. “Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews”, Advances in Knowledge Organization, 2004, pp. 49–54.

[4] Councill I. G., McDonald R., and Velikovich L. “What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis”, In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP '10, Association for Computational Linguistics, 2010, pp. 51-59.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. “Lexicon-based methods for sentiment analysis”. Computational Linguistics, Vol 37, Issues 2, 2011, pp. 267–307.

[6] S. Morsy and A. Rafea. “Improving documentlevel sentiment classification using contextual valence shifters”, Natural Language Processing and Information Systems, 2012, pp. 253–258.

[7] S. Li, S.Y.M. Lee, Y. Chen, C.R. Huang, and G. Zhou. “Sentiment classification and polarity shifting”, In Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010, pp. 635–643.

[8] Raskutti B., Ferr´a H. L., and Kowalczyk A. “Second Order Features for Maximizing Text Classification Performance”, In Proceedings of the 12th European Conference on Machine Learning, EMCL '01, pp. 419-430, London, Springer-Verlag, 2001.

[9] Diederich J., Kindermann J., Leopold E., and Paass G, “Authorship Attribution with Support Vector Machines”. In Applied Intelligence, Hingham, MA, USA. Kluwer Academic Publishers, Vol 19, 2003, pp. 109-123.

[10] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu. “Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis”, Technical report, HP Laboratories, 2011.

[11] Hanhoon Kang, Seong Joon Yoo, and Dongil Han. “Senti-Lexicon and Improved Naïve Bayes Algorithms for Sentiment Analysis of Restaurant Reviews”, doi:10.1016/j.eswa.2011.11.107, 2012, ISSN: 0957-4174.

[12] Prerna Chikersal, Soujanya Poria and Erik Cambria. “SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning”, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 647-651.

[13] Geetika Gautam and Divakar yadav. “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis”, Contemporary Computing (IC3), 2014 Seventh International Conference, IEEE, 2014, pp. 437 – 442.

[14] Alistair Kennedy and Diana Inkpen. “Sentiment Classification of Movie Reviews Using Contextual Valence Shifter”, Computational Intelligence, Vol 22, Issue 2, 2006, pp. 110–125.

[15] Vo Ngoc Phu and Phan Thi Tuoi. "Sentiment Classification using Enhanced Contextual Valence Shifters", International Conference on Asian Language Processing, IEEE, 2014, pp. 224 - 229.

[16] Kalyankumar B. Waddar, and K. Srinivasa. “Opinion Mining in Product Review System using Big Data Technology Hadoop”, International Journal of Advanced Computational Engineering and Networking, Vol 2, Issue 9, 2014.

[17] Hassan Saif, Yulan He, Miriam Fernandez and Harith Alani. “Contextual Semantics for Sentiment Analysis of Twitter”, Information Processing & Management, Vol 52, Issue 1, 2016, pp. 5–19.