Improving Sentiment Classification using Supervised Learning

Prajakta Gosavi[1] Vaishali Shirsath[2]
[1]Department of Computer Engineering [2]Department of Information Technology
[1]Shree L.R. Tiwari College of Engineering, Mira Road, Thane, India. [2]Vidyavardhini College of Engineering and Technology, Vasai, India.

Abstract— Sentiment classification is one of the techniques of sentiment analysis. It refers to the computational techniques for classifying whether the sentiments of text are positive or negative. SentiWordNet is a popularly used lexicon for extraction of sentiment words from documents. Many existing opinion mining or sentiment classification models works on positive and negative polarities while ignoring the objective words. If More than 90 percent of the words in SentiWordNet are objective words, Sentiment classification may be affected due to this major portion of objective words. This article proposes a MovieSentiNet framework, which will reassign a proper sentiment value & tendency to objective words aiming at improving sentiment classification. The sentiment values to objective words are assigned based on three steps. (a) Extracting sentiments on the words from SentiWordNet, (b) Calculating relevance of objective words based on frequency of objective words in positively & negatively tagged documents. (c) Then performing sentiment classification using SVM. The experimental results on a 1000 positively & 1000 negatively tagged documents demonstrate the effectiveness of proposed approach. Original objective words that are reassigned with sentimental orientation have contributed in the improvement of sentiment classification.

Key words: Sentiment, Supervised Learning

I. INTRODUCTION

The Web holds valuable, vast, and unstructured information about public opinion .With the explosion of Web 2.0 there has been the enormous increase in user-generated content, in the form of blogs, forums, micro blogging services, and, in general, social media.

"What other people think" has always been an important piece of information for most of us during the decision-making process. Earlier the main sources for this information were friends and specialized magazines or websites but now the web has changed from read only to read-write. Because of this change, number of enthusiastic users are interacting & sharing information through social networks, online communities, blogs, wikis, and other collaborative media. This information is unstructured as it's produced for human consumption it is not processable by machines Capturing public opinion about social events, political movements, company strategies & market campaigns are increasing demand today & resulting emerging field is known as sentiment analysis or opinion mining[1]. Sentiment Analysis is the task of calculating the intensity and the polarity of sentiments expressed by large number of user who post their views online. Data could be posts on social networking sites like tweets posted on twitter, or they may be reviews written on review such as amazon.com.This data is massive and so not possible for any individual to troublesome for an organization to read and process individual posts. In other words we can say, Sentiment analysis and opinion mining mainly focuses on opinions which express or imply positive or negative sentiments. Polarity is the stuff through which sentiment is measured Sentiment is usually considered to have the "poles" positive and negative. Sentiment analysis is really considered useful for telling us what is "good" and "bad" in our information stream. Two main types of textual information are available. Facts & opinion. Anything that is universally accepted is called as a fact. Any numbers of people have same answer for this object. For example: 10 X10 = 100 Every individual may have different sentiment towards an object. But anything for which multiple persons may or may not same views is sentiment. Sentiment is expressed as a result of persons experience with the object or may be experience of trusted person.

For example:

1) India played well
2) Yeh! India won
3) What an innings by India ☺
4) Indian fielding was poor.

Many sentiments are expressed by the users all over the world. Processing these sentiments using computational techniques to figure out the sentiments of the masses is termed as sentiment analysis.

In several ways sentiment analysis can be useful. For example, in marketing it helps in judging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features. Another example includes identifying unpopular features of BMWs. Automatic analysis of thousands of comments in BMW car forum. Identify features and sentiment. Identify if a new computer is popular. Automatic analysis of all blogs Compare to results for other computers. Identify impact of TV advertising campaign.

Sentiment analysis has to face several challenges. The first is an opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinions in a same way. The traditional text processing relies on the fact that minute differences between two pieces of text don't change the meaning very much. In Sentiment analysis, however, "the movie was great" is very different from "the movie was not great".

People can be contradictory in their statements, which is easy for humans to understand, but more difficult for a computer to parse. Sentiment analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. Sentiment classification is one of the main fields of research in Sentiment analysis. It is refers to the computational techniques for classifying whether the sentiments of text are positive or negative. Sentiment classification deals with classifying entire documents according to the opinions towards certain objects. Sentiment

analysis techniques can be broadly classified as supervised learning and unsupervised learning techniques.

A term with suitable sentiment tag is essential for sentiment classification. Most existing sentiment classification techniques or opinion mining models focus on classifying whether the sentiment of text are positive or negative while completely ignoring the objective words. For a given dataset or document containing reviews or opinion about particular domain the main task is to identify the subjectivity and objectivity of that document. Sentiment classification works on subjective sentences or words, objective sentences or words are of no use for sentiment classification. These objective words might have some effect on sentiment classification because they can be affected by their co-occurring sentiment words. Our work is to reevaluate the sentiment value and tendency to objective words which will contribute to the improvement of sentiment classification.

The rest of this paper is organized as follows. Section II related work in opinion mining. Section III elaborates on popular lexicon SentiWordNet. Section IV presents traditional sentiment classification using SentiWordNet. section V presents Sentiment classification using revised SentiWordNet , Section VI presents experimental results and analysis, while Section IV concludes this paper.

II. RELATED WORK

This literature survey is done to study the opinion mining problem in-depth and to familiarize with other works done on the subject. The aim of this study is to get access to the latest works in opinion mining as they frame new ideas and further develop the practice. Supervised machine learning techniques for Sentiment Classification was developed by Pang, Lee and Vaithyanathan [2]. They are also the pioneers for extracting, transforming and tagging 2000 reviews of the popular movie review dataset. S. Yueheng, W. Linmei, and D. Zheng [3] describes an approach for automatic sentiment analysis using generating positive and negative sentiment words & determining the sentiment orientation of ambiguous words according to their contexts, and later on setting up proper weight factors to different part-of-speeches, and then expanding the initial sentiment words by an iterative process. Bruce and Wiebe [4] describe the technique of manual tagging of sentences as subjective or objective based on the opinion of different judges. Judges tag individual sentences and final classification is based on agreement among them. SenticNet [5] exploits AI and Semantic Web techniques for mining opinions from natural language text at semantic level. It creates a collection of common sense concepts with strong positive or negative polarities while discarding the neutral polarities. In this, each commonsense concept is associated with only one value per concept. SentiWordNet [6] describes, a lexical resource in which each synsets is associated with numerical values, describing how objective, positive, and negative the terms present in the synset are. The method used to develop SentiWordNet is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification. SentiWordNet 3.0[7], for supporting sentiment classification and opinion mining applications SentiWordNet is a lexical resource explicitly used. Improved version of SentiWordNet 1.0 is evolved. Instead of manually formulating glosses they used Semi-supervised learning process & random-walk process to get the accuracy improvement of about 20 percent with respect to older version of SentiWordNet. In [8] For business-related data source SentiWordNet was applied for opinion classification. They have developed a set of tools to interpret and classify opinions, it concentrated on the use of a lexical resource and a text analysis tool to create features that enable us to carry out classification of short and long text reviews in the business domain. They applied the techniques to coarse as well as fine grained classification using Support Vector Machines statistical models. Bruce Ohana and Brendan Tierney [9] evaluated the function of sentimental scores in SentiWordNet for the automatic sentiment classification of film reviews. An approach using Sentiment value in SentiWordNet was slightly outperforms the approach using only frequency of sentimental words. It comprised counting positive and negative term scores to determine sentiment orientation, and an improvement is presented by building a data set of relevant features using SentiWordNet as source, and applied to a machine learning classifier. The results indicated SentiWordNet could be used as an important resource for sentiment classification tasks. Many sentiment-based classification tasks extract sentimental words directly from SentiWordNet to avoid using a manual sentiment lexicon. Chihli Hung [10] applied SentiWordNet for tagging sentimental orientations and classifying documents into five qualitative categories. Each word in SentiWordNet has its specific sentimental influence. Document quality classification approach, which extracts sentiment value from SentiWordNet and accumulates the different sentimental influence of each word based on a document level. It achieves a better classification performance than the approach in which SentiWordNet is not used. 93.75% of the synonymous sets in SentiWordNet are ignored as they have a stronger objective tendency than positive and negative [11].

III. SENTIWORDNET

SentiWordNet is a lexicon which provides an extension for WordNet; all synsets can be associated with a value concerning the negative, positive or objective connotation SentiWordNet3.0 is the improved version of SentiWordNet 1.0 and publicly freely available for research purpose with a web interface. This extension labels each synset with a value for each category between 0.0 and 1.0. The sum of the three values is always 1.0.The web interface allows the user to search for any synset belonging to WordNet with its associated SentiWordNet scores. Additionally the user is able to see a visualization of that scores.

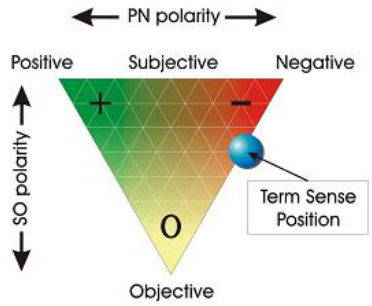

Fig. 1: Visualization of synsets in SentiWordNet 3.

Figure 1 each category is linked to a colour, which is red for negativity, blue for objectivity and green for positivity. The advantage of using synsets instead of terms is to offer different sentiment scores for each sense of one word, because the connotations can differ in one word depending on the sense. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms that is synsets, each expressing a distinct concept. Existing system which uses SentiWordNet for sentiment classification tasks simply ignores objective words which might have some effects on sentiment classification. MovieSentiNet is a hybrid approach using supervised learning & unsupervised learning. MovieSentiNet will initially use unsupervised approach by extracting positive and negative polarities from SentiWordNet. The remaining objective words will be processed using supervised approach. The objectivity of these words will be reevaluated & reassigned if needed. By reevaluating sentiment value and tendency for objective words which are considered useless in traditional sentiment classification. As an example, consider two sentences where each word contains three sentiment values as indicated in brackets—that is, positive, objective, and negative— the values are derived from SentiWordNet. Sentence 1: I (p:0, o:1, n:0) will (p:0, o:1, n:0) meet (p:0, o:1, n:0) you (p:0, o:1, n:0) later(p:0, o:1, n:0). Sentence 2: It (p: 0, o: 1, n: 0) is (p: 0, o: 1, n: 0) my (p: 0, o: 1, n: 0) pleasure (p: 0.875, o: 0.125, n: 0) to(p: 0, o: 1, n: 0) meet(p: 0, o: 1, n: 0) you(p: 0, o: 1, n: 0). Sentence 1 doesn't contain any sentiment words so, obviously it's an objective sentence. In Sentence 2 the summation of positive values is greater than that of negative values so, it considered as a positive sentence. Initially the words in sentence 1 have no clear sentiments, so they are not useful for sentiment classification. However, the word "meet" from sentence 1 appear in positive sentence i.e. sentence 2.So we can say that this word is having a positive tendency. With this proposed approach we able to know that sentiment of text how much positive or how much negative then it will definitely contribute to improvement in sentiment classification.

IV. TRADITIONAL SENTIMENT CLASSIFICATION USING SENTIWORDNET

The traditional sentiment classification approach, which classifies documents based on SentiWordNet, is consisting of two modules (a) Document preprocessing (c) Sentiment Classification using support vector machine.

A. Document preprocessing:

Extracted reviews or documents will be used to perform document level sentiment classification. Firstly full text review is divided into sentences. Most of the English sentences include words like "a, an, "of, the", "I", "it", "you", and. Such words do not carry particular meaning. Information extraction fromnatural language can be done effectively and clearly by avoiding those words which occurs very often. Text file is created consists of stop words to remove such words from sentences and further replaces it with white spaces. The Process of assigning different parts of speech tags such as noun, preposition, verb, adjective and adverb to a given text are known as Part-Of-Speech tagging. It is a special application of natural language processing. The part-of-speech is a category used in linguistics that is defined by a syntactic or morphological behavior of a word. The traditional English language grammar classifies partsof-speech in the following categories: verb, noun, adjective, adverb, pronoun, preposition, conjunction and interjection. The reason why POS tagging is so important to information extraction is the fact that each category plays a specific role within a sentence. The OpenNLP parser is used to generate the POS tagging of each word, present in the sentence It is very essential as it helps in finding general language patterns. Opinion Mining requires adjectives and adverbs to find the sentiment polarity. These can be done by using the part-of-speech tagger.

B. Sentiment Classification:

This module includes two steps extracting sentiments on the words from SentiWordNet & then performs sentiment classification using SVM.

A word in SentiWordNet contains different sentiments i.e. positive, negative, and objective. After removal of stop-word, only nonstop words exist in their base forms in each sentence. Then a word document frequency count matrix will be constructed. Words in this matrix having positive or negative orientation in SentiWordNet will be assigned their respective polarities from SentiWordNet itself.

For word i, posWi specifies the sentiment value in positive orientation, negWi specifies the sentiment value in negative orientation, and objWi specifies the sentiment value in objective orientation. Thus, the sum of posWi, negWi, and objWi for word i equals 1. A word whose sentiment value is the greatest in positive, negative, or objective orientation is defined as a positive, negative, or objective word, respectively. A document will be classified as positive by SVM if total weight of positive words in document will be more than total weight of negative words. Similarly A document will be classified as negative if total weight of negative words in document will be more than total weight of positive words.

Weight can be calculated by multiplying the term frequency i.e the occurrence of particular word in that document with the positive sentiment value or negative sentiment value from SentiWordNet which is shown in the following equations.

$$Wi = TFi \times posWi \quad (1)$$

$$Wi = TFi \times negWi \times (-1) \quad (2)$$

If the word doesn't contain any sentiments then its weight calculated as 0 i.e. objective words. If a document

was originally tagged as positive and also classified as positive then it will contributed to true Positive in confusion matrix if a document was originally tagged as negative and also classified as negative then it will contribute to True Negative in confusion matrix. If a document was originally tagged as positive but classified as negative then it contribute to False Negative in confusion matrix. If a document was originally tagged as negative but classified as positive then it will contribute to False Negative in confusion matrix. Based on the confusion matrix accuracy of traditional sentiment classification has calculated.

V. SENTIMENT CLASSIFICATION USING REVISED SENTIWORDNET (MOVIESENTINET)

Existing system which uses SentiWordNet for sentiment classification tasks simply ignores objective words which might have some effects on sentiment classification. as SentiWordNet comprising of 93.75% of objective words which are considered as of no use for sentiment classification. Our approach of MovieSentiNet as figure 2 shows, is a hybrid approach using supervised learning & unsupervised learning. MovieSentiNet will initially use unsupervised approach by extracting positive and negative polarities from SentiWordNet. The remaining objective words will be processed using supervised approach. MovieSentiNet consisting of three modules. (a)Document Preprocessing (b)Modification of objective words in SentiWordNet (c) Sentiment Classification. The first and last modules are same as described in the traditional sentiment classification using SentiWordNet. Here, we are focusing on second module i.e. Modification of objective words in SentiWordNet. This module will consist of three steps:

1) Extraction of sentiments on the word.
2) Calculating the relevance of an objective word.
3) Sentiment classification with revised sentiment values.

A. Extraction of sentiments on the word

For a given document a word document frequency count matrix has been constructed. Words in this matrix having positive or negative orientation in SentiWordNet were assigned with their respective polarities from SentiWordNet itself. Remaining words of the word document matrix are objective words as defined in SentiWordNet. In further steps an attempt has made to reassign the sentiment value to these objective words if needed.

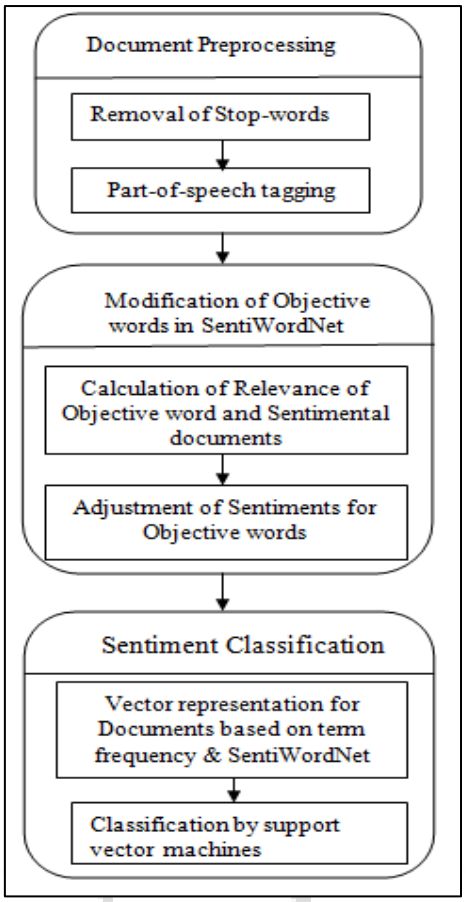

Fig. 2: Proposed Architecture For MovieSentiNet

B. Calculating the relevance of an objective word

In this step we calculated the relevance of an objective word considering its associated documents. The basic concept is that a positive or a negative document has some sentimental influence on its associated objective words. This influence on objective words will be calculated by considering the frequency distribution of that object positively tagged documents and negatively tagged documents. If the frequency of an objective word is greater in positively tagged documents then it has assigned with the proportionate positive value. If the frequency of an objective word is g negatively tagged documents then it has assigned with the proportionate negative value. Objective word has assigned with positive value otherwise negative value will be assigned. If the frequency of originally objective word is equal in both positively tagged documents and negatively tagged documents then polarity of objective word will remain unchanged. Simply, posDoc indicates a positive document, negDoc indicates a negative document, and probability (Pr) indicates the relationship between a word and its associated sentence. The sentiment value of an original objective word i is modified. An original objective word is remaining as it is if its appearance in both positive and negative orientation is equal. For example, an objective word in positive, negative, and objective sentences five, ten and five times, respectively. This objective word is reassigned as a negative word with a negative value of 0.5 because $\Pr(\text{negSen}, \text{Wordi}) = 10/20$ is greater than $\Pr(\text{posSen}, \text{Wordi}) = 5/20$ (as Equation 3 shows) for a revised positive word when this objective word appears in a positive sentence more often than in a negative sentence.

$$\text{posWi} = \Pr(\text{posSen}, \text{Wordi}) = \frac{\text{Psi}}{\text{Fri}}$$
$$\text{negWi} = 0$$
$$\text{objWi} = 1 - \text{posWi} \quad (3)$$

Where Pr (posSen, Wordi) > Pr (negSen, Wordi indicates the frequency of word i sentence, and fri indicates the frequency of word a dataset. (as Equation 4 shows) for a revised neg when this objective word appears in a negative sentence more often than in a positive sentence.

$$\text{negWi} = \Pr(\text{negSen}, \text{Wordi}) = \frac{\text{Nsi}}{\text{Fri}}$$
$$\text{posWi} = 0$$
$$\text{objWi} = 1 - \text{negWi} \quad (4)$$

Where $\Pr(\text{negSen}, \text{Wordi}) > \Pr(\text{posSen}, \text{Wordi})$,Nsi indicates the frequency of word i sentence, and fri indicates the frequency of word a dataset.

As we have assigned the sentiment values to objective words with the formulas described above, but if a sentence or a word contains small sentiment value then it wouldn't impact on achieving improvement in sentiment classification influence on its associated objective word. Thus, we have to set the threshold for a Modification to an objective word only if its sentiment values are greater than the set threshold.

VI. EXPERIMENTAL RESULTS

We selected a movie review dataset from Internet Movie Database (IMDb), which includes 27,886 movie review articles. Among this review dataset, 1,000 articles have a preassigned positive sentiment tag, 1,000 articles have a preassigned negative sentiment tag, and others have no preassigned sentiment tag. From Pang lee's movie review database already tagged positive and negative articles are used for constructing sentiment classification model. we extract sentiments on the word and sentence levels in 1000 positively tagged & 1000 negatively tagged documents and reassign a proper sentiment value and orientation to objective words in the IMDb dataset. For calculating the accuracy we have used the traditional accuracy criterion as shown in (5)

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (5)$$

For better understanding, we present evaluation criterion for classification in a confusion matrix as shown in Table 1.

		Predicted sentiment orientation	
		Positive review	Negative review
Actual Sentiment Orientation	Positive review	**True positive**	**False negative**
	Negative review	**False positive**	**True negative**

Table 1: (Accuracy evaluation criterion)

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the four different possible outcomes from tagged documents.

True positive means that a document is tagged as a positive and also classified as a positive. True negative means that a document is tagged as a negative and classified as a negative. Both true positive and true negative are correct classifications. False positive means document is tagged as a positive but classified as a negative. False negative means that a document is tagged as a negative but classified as a positive.

We executed several experiments to determine at which point we are getting highest accuracy. Finally we set threshold to 0.5 for document and got the significant results. For experimental result we have used first subset of 200 positively tagged documents and 200 negatively tagged documents and extracted its sentiment values from a very well known SentiWordNet which is an unsupervised approach. SentiWordNet usually contains several senses for a particular word and each sense has its own sentiment value in three sentiment orientations i.e. positive, negative & objective. Table 2 shows the experimental result between traditional sentiment classification using SentiWordNet & sentiment classification using revised SentiWordNet i.e. proposed approach of MovieSentiNet. Accuracy of the positively tagged documents and negatively tagged documents is calculated with confusion matrix table as shown in table 1.

Subset no	Original Method (%)	Refined Method (%)
01	65.00%	88.59%
02	70.50%	82.50%
03	70.64%	79.50%
04	70.65%	81.00%
05	70.87%	82.50%

Table 2: (Experimental results comparing traditional & proposed approaches on 5 subsets of tagged document)

For the table 2 we have plotted the graph which clearly shows that when we reassigned sentiment value to objective word it has contributed to the improvement in sentiment classification which is shown in figure 3.

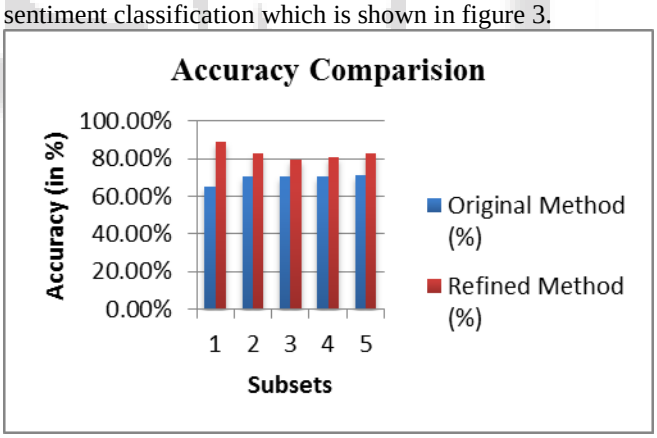

Fig. 3: Graphical comparison between traditional & proposed approach

Complete dataset	Original Method (%)	Refined Method (%)
01	62.9%	90.75%

Table 3: (Experimental results comparing traditional & proposed approach on complete dataset)

For experimental result now we have used entire dataset i.e. 1000 positively tagged and 1000 negatively tagged documents. Table 3 shows the experimental results comparing traditional & proposed method on a complete dataset of tagged documents. Figure 4 shows the graphical comparison of traditional and proposed method. Revised SentiWordNet outperforms the non-revised i.e traditional sentiment classification in all experiments.

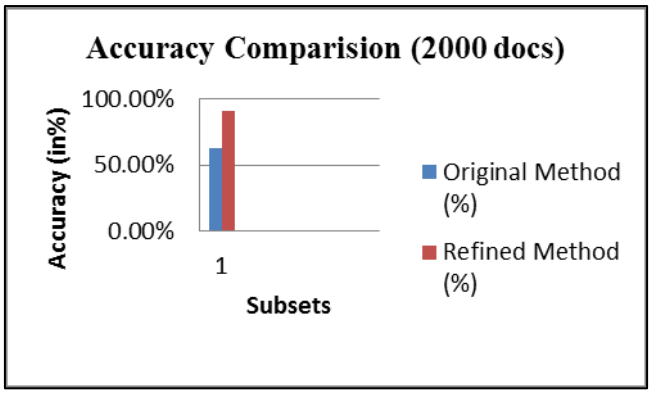

Fig. 4: Graphical comparison between traditional & proposed approach on complete dataset

VII. CONCLUSION

In this article, we have proposed framework for MovieSentiNet in which reassignment of objectives words with proper sentiment orientation has contributed to the improvement in sentiment classification results. As SentiWordNet consisting of more than 90 % of objective words which are considered as no use. By identifying objective words and their frequency distribution in positive and negative tagged documents, sentiment orientation has been assigned to those words & then sentiment classification is performed using SVM.Based on the confusion matrix and accuracy evaluation criterion final accuracy has calculated for both traditional and proposed approach. Experimental results shown that the revised SentiWordNet outperforms the non revised SentiWordNet approaches in all experiments.

REFERENCES

[1] Erik Cambria, Bjorn Schuller, Yunqing Xia, Catherine Havasi, "New Avenues in Opinion Mining and Sentiment Analysis", IEEE Intelligent Systems, vol.28, no. 2, pp. 15-21, March-April 2013.
[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, vol. 2, nos. 1–2, pp. 1–135, 2008.
[3] Sun Yueheng; Wang Linmei; Deng Zheng, "Automatic Sentiment Analysis for Web User Reviews", Information Science and Engineering (ICISE), 1st International Conference on , vol., no., pp.806,809, 26-28 Dec. 2009.
[4] Rebecca F. Bruce and Jaynce M. Wiebe , " Recognizing subjectivity: a case study in manual tagging", Natural Language Engineering,vol. 5, pp 187-205,1999.
[5] E. Cambria et al., "SenticNet: A Publicly Available Semantic Resource for Opinion Mining", Proc. AAAI Commonsense Knowledge Symp., Assoc. for the Advancement of Artificial Intelligence (AAAI), pp. 14−18,2010.
[6] A. Esuli, and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining", Proc. 5th Int'l Conf. Language Resources and Evaluation, ELRA, 2006.
[7] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0:An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," Proc. Int'l Conf. Language Resources and Evaluation, ELRA, pp. 2200−2204,2010.
[8] H.Saggion,A.Funk,"Interpreting SentiWordNet for opinion classification", Proceeding Of the seventh conference on International language resources & evaluation LREC 10,PP.1129-1133,2010
[9] B.Ohana and B. Tierney, "Sentiment Classification of Reviews Using SentiWordNet," Proc.9th IT&T Conf., Dublin Inst. Of Technology, 2009
[10] C. Hung, C.-F. Tsai, and H. Huang, "Extracting Word-of-Mouth Sentiments via SentiWordNet for Document Quality Classification, "Recent Patents on Computer Science", vol. 5, no. 2,pp. 145−152,2012.
[11] Chihli Hung; Hao-Kai Lin, "Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification," Intelligent Systems, IEEE , vol.28, no.2, pp.47,54,March-Apri l2013.

