Sentiment Analysis of Texual Review

Hiteshkumar N Vegda[1], Prof.Tejal R Patel[2], Prof.Bhargesh B Patel[3]

[1]*Information Technology, G.H.Patel College of Engineering and Technology, Vallabh Vidhyanagar, Gujarat, India, hvegda@gmail.com*
[2]*Information Technology, G.H.Patel College of Engineering and Technology, Vallabh Vidhyanagar, Gujarat, India, tejalrpatel@gcet.ac.in*
[3]*Information Technology, G.H.Patel College of Engineering and Technology, Vallabh Vidhyanagar, Gujarat, India, bhargeshpatel@gcet.ac.in*

Abstract-Our day-to-day life has always been influenced by what people think. Ideas and opinions of others have always affected our own opinions. The explosion of Web has led to increased activity in Podcasting, Blogging, Tagging, Contributing to RSS, Social Bookmarking, and Social Networking. As a result there has been an eruption of interest in people to mine these vast resources of data for opinions. Sentiment Analysis or Opinion Mining is the computational treatment of opinions, sentiments and subjectivity of text. I proposed new algorithm for sentiment analysis using Bags of word model, where I distinct any review as positive, negative or neutral. And finally calculates the performance parameters like accuracy, precision and recall. After classifying the dataset I have calculated the performance parameter and the highest and lowest accuracy are 79.87% and 62.76%. And while increasing the dataset size it gives the constant accuracy.

Keywords- Machine learning, Opinion mining, Sentiment analysis (SA), Bags of word (BOW).

I. INTRODUCTION

1.1 WHAT IS SENTIMENT ANALYSIS?

Sentiment Analysis [1] is a Natural Language Processing and Information Extraction task that aims to obtain writer's feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document. In recent years, the exponential increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment Analysis today. The Web is a huge repository of structured and unstructured data. The analysis of this data to extract latent public opinion and sentiment is a challenging task.

1.2 SENTIMENT ANALYSIS PROCESS

The opinion mining process is explained in the figure1. The raw data is collected from various social media, Review Sites, Blogs,we can also write a crawler to extract the data from it. The Data set available online for research work.

After data collection we preprocess the data to have a structured set of reviews, Pre-processing the data is the process of cleaning and preparing the text for classification. The whole process involves several steps: online text cleaning, white space removal, expanding abbreviation, stemming, stop words removal, negation handling and finally feature selection. so that we can apply classification techniques to classify the opinion either as positive, negative or neutral.

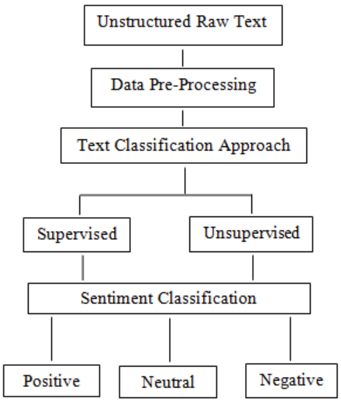

Figure1- Sentiment Analysis Process

The supervised classification, Feature Extraction, It is the process where properties are extracted from the data, because the whole input data is too large to use in classification. The different feature extraction methods are N-gram model, Tf-idf (term frequency-inverse document frequency) Measure, Part-of-speech Tagger (POS). The Classification algorithm are nearest neighbour, naive Bayes, maximum entropy and Support vector machine (SVM) are applied. And in Unsupervised classification Point mutual information (PMI) is used.

II. APPLICATIONS OF SENTIMENT ANALYSIS

1. Applications to Review-related Websites Reviews and feedbacks [5] on almost everything, ranging from product reviews, to feedbacks on political issues are abundantly available over the Internet.Much like a search engine, a sentiment engine can be built to utilize this information.
2. Applications in Business Intelligence For many businesses [2], online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace. This statement highlights the importance of sentiment analysis in businesses. The most obvious usage of Sentiment Analysis in business intelligence lies in understanding the user reviews to improve their products, and in turn, their reputation.
3. Opinion mining for Ad Placement In online systems that display ads as sidebars [2], it is helpful to detect WebPages that contain sensitive content inappropriate for ads placement; for more sophisticated systems, it could be useful to bring up product ads when relevant positive sentiments are detected, and perhaps more importantly, nix the ads when relevant negative statements are discovered.
4. Opinion mining for trend prediction Organization could perform trend prediction in sales using Opinion mining by tracking public viewpoints. In stock market we can analysis the sentiment related to detect whether the stock price will be higher or lower and help the investor to take decision related to buying or selling the stock [2].
5. Opinion mining for political domain Opinions matter a great deal in politics [2]. Sentiment analysis has specifically been proposed as a key enabling technology, allowing the automatic analysis of the opinions that people submit about pending policy or government-regulation, understanding what the voters is thinking, predicting the outcome of elections etc.

III. CHALLENGES FOR SENTIMENT ANALYSIS

Sentiment Analysis approaches aim to extract positive and negative sentiment bearing words from a text and classify the text as positive, negative or else objective if it cannot find any sentiment bearing words. In this respect, it can be thought of as a text categorization task. In text classification there are many classes corresponding to different topics whereas in Sentiment Analysis we have only 3 broad classes. Thus it seems Sentiment Analysis is easier than text classification which is not quite the case. The general challenges can be summarized as:

1. Implicit Sentiment and Sarcasm
   A sentence may have an implicit sentiment even without the presence of any sentiment bearing words [3].
2. Domain Dependency
   There are many words whose polarity changes from domain to domain [3].
3. Thwarted Expectations
   Sometimes the author deliberately sets up context only to refute it at the end [3].
4. Pragmatics
   It is important to detect the pragmatics of user opinion which may change the sentiment thoroughly [2]. World Knowledge Often world knowledge needs to be incorporated in the system for detecting sentiments.
5. Subjectivity Detection
   This is to differentiate between opinionated and non-opinionated text. This is used to enhance the performance of the system by including a subjectivity detection module to filter out objective facts. But this is often difficult to do [2].
6. Entity Identification
   A text or sentence may have multiple entities. It is extremely important to find out the entity towards which the opinion is directed.

IV. SOURCES OF DATA FOR SENTIMENT ANALYSIS

A. Review Sites

B. Blogs
C. Micro-blogging
D. Forums
E. Social Networks
- Twitter
- Facebook

V. TOOLS AVAILABLES FOR SENTIMENT ANALYSIS

A variety of open-source text-analytics tools like natural language processing for information extraction and classification can be applied for sentiment analysis. The tools listed below can work on textual sources only [1].

LingPipe, OpenNLP, Stanford Parser and Part-of-Speech (POS) Tagger, NLTK, Opinion Finder, Tawlk/osae, GATE, Textir, NLP Tool suite, Review Seer tool, Web Fountain, Red Opal, Opinion observer.

Along with these automated tools, there are various online tools like Twitrratr, Twendz,Social mention, and Sentimetrics are available to track the opinions in the web.

VI. TECHNIQUES FOR SENTIMENT ANALYSIS

All In order to give more imminent into the problem of opinion mining, in the following sections we discuss the domain overview and various types of opinion mining. The opinion mining is frequently associated with the topic information retrieval. The information retrieval algorithm works on factual data but the opinion mining works on subjective data. The task of opinion mining is to find the opinion of an object whether it is positive or negative and what features does it depict, and what features are appreciated, which are not etc. The notion of an opinion mining is given by Hu and Liu. They put most impact on their work and said that the basic components of an opinion are:

Opinion holder: it is the person that gives a specific opinion on an object.

Object: it is entity on which an opinion is expressed by user.

Opinion: it is a view, sentiment, or appraisal of an object done by user.

A. Task of Opinion mining at Document level
B. Task of opinion mining at Sentence level
C. Task of opinion mining at feature level
D. Opinion mining in compound sentence

In this sub-section the following methodology we use to determine the opinion in compound sentence of a movie review domain [4]:

- Sentence classification
- Splitting of the document into sentences
- Determining whether the sentence is opinionated
- Determining whether the opinionated sentence is compound sentence
- Opinion Mining in Compound sentence

VII. TECHNIQUES FOR SENTIMENT CLASSIFICATION

The literature survey done indicates two types of techniques include machine learning and semantic orientation.

A. Machine Learning

Several Machine Learning methods have been studied [12]. Prominent methods are: Naive Bayes Classification, Maximum Entropy Classification, and Support Vector Machines. In his work, Pang Lee et al., compared the performance of Naïve Bayes, Maximum Entropy and Support Vector Machines in SA on different features like considering only unigrams, bigrams, combination of both, incorporating parts of speech and position information, taking only adjectives.

B. Sementic Orientation

Problem of Opinion mining can be categorized as sentiment classification and feature based opinion mining. Problem of Opinion mining can be categorized as sentiment classification and feature based opinion mining [12].

Classification of Approaches of Sentiment Orientation

- Corpus Based Approach
- Dictionary Based Approach

C. Features for Opinion Mining

Feature engineering is an extremely basic and essential task for Opinion Mining. Converting a piece of text to a feature vector is the basic step in any data driven approach to Opinion. Some commonly used features in Opinion Mining and their critiques have been discussed in the following sections.

- Term Presence vs. Term Frequency
- Term Position
- N-gram Features
- Tf-idf Measure
- Parts of Speech
- Adjectives only

VIII. EVALUATION OF SENTIMENT CLASSIFICATION

In general, the performance of sentiment classification is evaluated by using four indexes. They are Accuracy, Precision, Recall and F1-score. The common way for computing these indexes is based on the confusion matrix as shown below:

	Predicted positives	Predicted negatives
Actual positive Instances	Number of True Positive instances (TP)	Number of False Negative instances (FN)
Actual negative Instances	Number of False Positive instances (FP)	Number of True Negative instances (TN)

Table1- Confusion Matrix [1]

These indexes can be defined by the following equations [1]:

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call}$$

Accuracy is the portion of all true predicted instances against all predicted instances. An accuracy of 100% means that the predicted instances are exactly the same as the actual instances. Precision is the portion of true positive predicted instances against all positive predicted instances. Recall is the portion of true positive predicted instances against all actual
positive instances. F1 is a harmonic average of precision and recall.

IX. PROPOSED ALGORITHM

I proposed the algorithm for sentiment analysis. It is divided in three major steps. The three steps are explained below.

1. Collect text data from web
2. Bag of words model

```
A. Class Labeling (positive/Negative/Neutral)
   - Enter Positive/Negative/Neutral text
     Syntax: [Text] [Sentiment]
     E.g. [I love my car] [positive]
B. Extraction of words
   I. combines text into one file
        - extract words with length >3(remove stop words)
      - combine all words into one file
      - get word features
   II.Get frequency distribution of words
      - Count each word occurrence with its sentiment
      - Make frequency distribution of words
      - Using frequency distribution create training set
C.Use above training set as test data set for classification
   - Train naïve bayes classification
   - Using frequency distribution, train classifier
   - Test classifier for accuracy using test data
```

3. After the classifier has been trained and tested, the user gives in an input text. And finally classified it as positive, negative or neutral.

BAG OF WORDS MODEL

The bag-of-words (BoW)[1] model is a representation method used in natural language processing and information retrieval where, a text is represented as an unordered collection of words, disregarding grammar and even word order. The bag-of-words model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier.

As mentioned above, the purpose of using bag-of-words in our implementation is for two reasons. First being, to extract words easily from a text/document/sentence. And second, to get the weightage of the words/features in the document, which in turn would help us to identify the text category-aiding in classification. This also enhances the performance of the classifier when combined with the naïve bayes classification, as now the text is first filtered for word/feature weightage using BoW and then trains classifier using these words.

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. Recently, the bag-of-words model has also been used for computer vision.

The bag-of-words model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier.

An early reference to "bag of words" in a linguistic context can be found in Zellig Harris's 1954 article on Distributional Structure.

The following models a text document using bag-of-words.Here are two simple text documents:

“John likes to watch movies. Mary likes too”.
“John also likes to watch football games.”
Based on these two text documents, a dictionary is constructed as:

"John"	:1
"Likes"	:2
"To"	:3
"Watch"	:4
"Movies"	:5
"Also"	:6

"Football"	:7
"Games"	:8
"Mary"	:9
"Too"	:10

Which has 10 distinct words. And using the indexes of the dictionary, each document is represented by a 10-entry vector:

[1,2,1,1,1,0,0,0,1,1]
[1,1,1,1,0,1,1,1,0,0]

Where each entry of the vectors refers to count of the corresponding entry in the dictionary (this is also the histogram representation). This vector representation does not preserve the order of the words in the original sentences. This kind of representation has several successful applications, for example email filtering.

The two main modules that we are dealing with are-

- Word/feature extraction
- Classification

Process 1

The basic and the first step for any text classification. This can be either static data-fed in manually by the user or dynamic-collected from the web in real time. For simplicity and explanation, we will be considering the static data. The dynamic data poses the problem of dealing with information coming in continuously within very short periods of time and in very high quantity. Handling and classifying the information instantaneously is being researched.

There are three kinds of data needed. First the input data for training the classifier. Second, input data for testing the classifier. And the final is the user input data which is to be classified.

The text is manually classified with corresponding sentiment as positive, negative or neutral. This can also be a file containing documents of text labeled under each sentiment.

Process 2

Once the positive, negative and neutral text has been entered, the next is the extraction of words. The document is filtered to remove all the stopwords such as is, and, the, etc. The BoW is used here to extract words/features. The words extracted are weighed and their order in the text does not matter at all. Once the feature set is got, the frequency distribution function is used to create the training set.

Process 3

The most essential part of the sentiment classification tool- the classifier. After the removal of stop words, application of BoW and the creation of the training set, the classifier is the last step. We get the Naïve Bayes classifier. This classifier is trained by giving it the training set that was created. Once the classifier has been trained, it is tested using the test data set and once the needed accuracy is reached, it can be implemented via an interface.

X. IMPLEMENTATION AND ANALYSIS

1. IMPLEMENTATION DETAILS

I have tested the algorithm with different dataset and calculated the accuracy, precision and recall. The detailed is shown below.

No.	Dataset	Accuracy	Precision	Recall
1	Twitter	79.87%	78.13%	44.86%
2	Mobile Review(Nokia 6610)	76.87%	58.13%	74.86%
3	Phone Review	62.76%	51.08%	63.04%
4	Camera	74.58%	70.25%	41.24%
5	TV Review	71.42%	64.24%	76.24%
6	Laptop Review	77.24%	75.12%	42.35%

Table2 - Accuracy, Precision & recall of proposed algorithms

Plotted the graph for accuracy, Precision and recall.

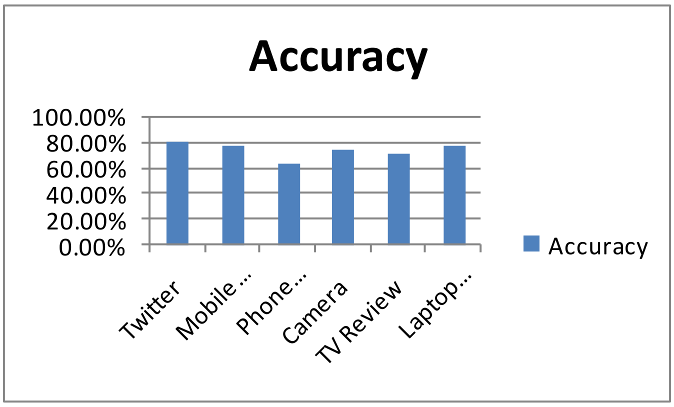

Figure2- Graph for Accuracy

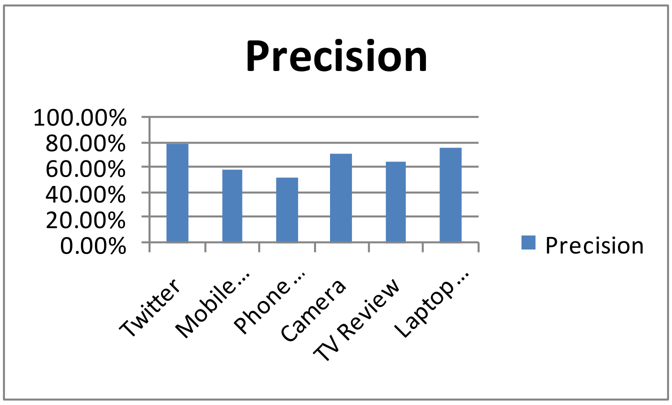

Figure3- Graph forPrecision

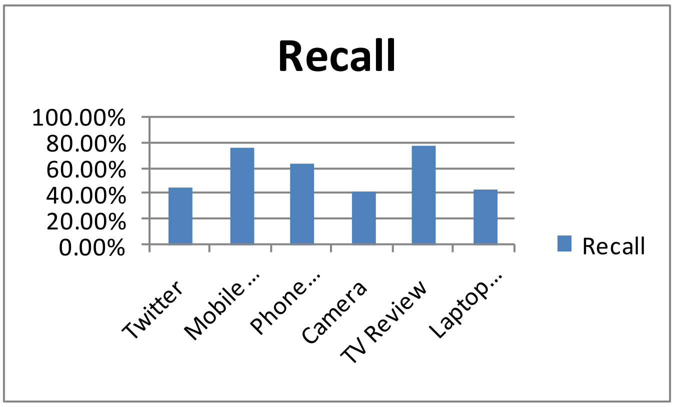

Figure4- Graph for Recall

2. ANALYSIS

Comparing theProposed algorithm with existing algorithms. The table below gives the accuracy of existing algorithm for sentiment analysis.

Algorithm	Yea r	Dataset	Accuracy
Pang et al.	200 2	Reviews	81.5%
Dave et al.	200	Reviews	87.0%(81.9

	3		-87.0)
Chen et al.	2006	Reviews	77.5%
gindl and Liegl	2008	Reviews	66.0%
Annett and Kondrak	2008	Reviews	77.5%
Go et al.	2009	Microblogs	82.7%
Bifet and Frank	2010	Microblogs	82.5%
Zhang et al.	2011	Reviews	84.5%

Table3- Accuracy of Existing algorithms [1]

The graph shows the comparison of proposed algorithm and existing sentiment analysis algorithm. The highest and lowest accuracy of the existing algorithm is 87.0 % and 66.0 % while the proposed algorithms have highest and lowest accuracy 79.87% and 62.76%. The proposed algorithm get the more accuracy for some dataset but it didn't achieved the highest accuracy of the existing algorithm.

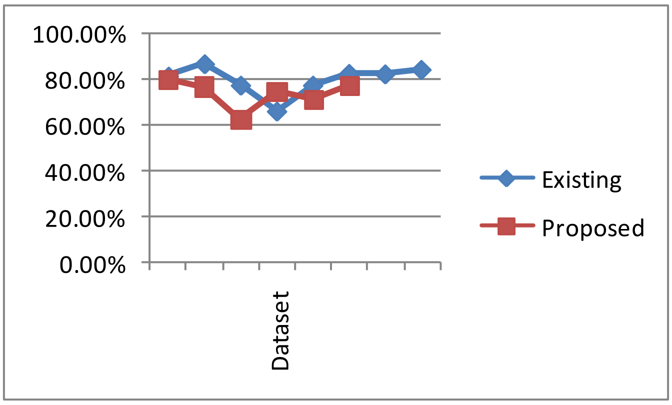

Figure5- Graph of comparisons

The existing algorithm gives the constant accuracy as the dataset size increased but the time taken by algorithm is also increased.

Dataset	Accuracy	Approx.Time taken(Minutes)
800	81.16	8
1500	79.20	18
2500	79.87	30

Table4- Accuracy vs. Time

The graph below is for accuracy verses time as the dataset size is increased the accuracy is decreases but still it is not more decreased remains constant and the time taken for execution is increased.

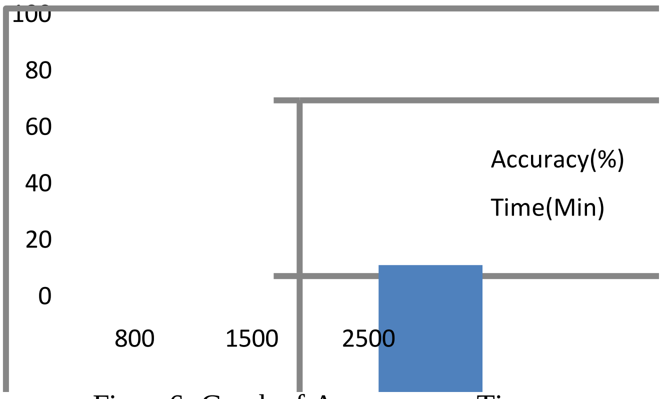

Figure6- Graph of Accuracy vs. Time

XI. CONCLUSION AND FUTURE WORK

This report discusses the various approaches to Sentiment Analysis. It provides a detailed view of the different applications and challenges of Sentiment Analysis that makes it a difficult task.We have proposed algorithm for sentiment analysis that will classify the text reviews in three labels Negative, Positive or Neutral.The performance of sentiment classification is evaluated by using indexes. Theyare Accuracy, Precision, Recall after evaluating performancewe had compared it with existing methods.The algorithm gets the more accuracy for some dataset than the accuracy but not achieved the more accuracy than the existing algorithms, the algorithm performs well when the size of data increases it gives the accuracy constant, decreases but not with more difference and time taken for the large dataset is also increased.

The algorithm didn't achieves the more accuracy than existing one every times so to increase the accuracy some useful steps may takenas the future work, also the time taken is more so to minimize the time consumed is also the challenge.

REFERENCES

[1] S Padmaja and Prof. S Sameen Fatima, "Opinion mining and sentiment analysis – an assessment of peoples' belief: a survey", International journal of ad hoc, sensor & ubiquitous computing (IJASUC) vol.4, no.1, Feb 2013.

[2] Mrs. Vijyalaxmi m, Mrs. Shalu Chopra, Mrs. Sangeeta Oswal, Mrs. Deepshikha chaturvedi, "The How, When and Why of Sentiment Analysis", International journal of computer technology& application, vol 4 (4), 660-665, 2013.

[3] Akshi kumar, Aeeja mary sebastian, "Sentiment Analysis: a perspective on its past present and future", I.J Intelligent systems and applications, ecs,page 1-14,2012.

[4] Nidhi Nishra, C.K.Jha, "Classification of Opinion Mining Techniques", International Journal of Computer Applications, Volume 56– no.13, Oct 2012.

[5] V. s. Jagtap, Karishma Pawar, "Analysis of Different Approaches to Sentence-level Sentiment Classification", International Journal of Scientific Engineering and Technology, Volume 2, Issue 3, April 2012.

[6] Pravesh kumar singh, Mohd Shahid husain, "Analytical study of Feature Extraction Techniques in Opinion Mining", International Conference on Advance in Computing and Information Technology, page 85-94, 2013.

[7] S.l. Ting, W.h. Ip, Albert H.c. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification?", International Journal of Software Engineering and Its Applications vol. 5, no. 3, July, 2011.

[8] Emma Haddi, Xiaohui Liu, Yong Shi, "The Role of Text Pre-Processing Sentiment Analysis", Information Technology and Quantitative Management, Procedia Computer Science 26 -32, Elsevier 2013.

[9] P. Waila, Marisha, V.k. Singh, M.k. Singh, "Evaluating Machine Learning and Unsupervised Semantic Orientation Approaches for Sentiment Analysis of Textual Reviews", IEEE International Conference on Computational Intelligence and Computing Research, 2012.

[10] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal of Advances In Information Technology, VOL. 1, February 2010.

[11] Huifeng Tang, Songbo Tan , Xueqi Cheng, "A Survey on Sentiment Detection of Reviews ", Expert Systems with Applications, page 10760 - 10773, Elsevier 2009.

[12] Nilesh m. Shelke, shriniwas Deshpande, Vilas Thakre, "Survey of Techniques for Opinion Mining" , international Journal of Computer Applications (0975 – 8887) Volume 57– No.13, November 2012.

[13] Mr. Saifee Vohra, Prof. Jay Teraiya, "Applications and Challenges for Sentiment Analysis: A Survey", International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 2, February- 2013.