Sentiment classification of Hotel reviews using Hybrid Method

Vinit Goel , Ankit Jain , Saksham Arora ,Shashank Shekhar ,Poonam Verma
Computer Science Department
Bharati Vidyapeeth's College of Engineering,
Guru Gobind Singh Indraprastha University
New Delhi, India
{vinitgoel150595, ankitjain19962013, saksham.190296, 01shashankshekhar}@gmail.com, poonam.verma@bharatividyapeeth.edu

***Abstract* – Sentiment analysis refers to a reckoning process that deals with treatment of opinions, emotions, and subjectivity . We contemplate problem of classifying a hotel review into sentimental categories like positive , neutral, and negative and thereby extracting sentiment of consumer and predicting behaviour on four traits : food, service, price, and locality that defines a hotel . Using Hotel reviews Data set from Trip Advisor and SentiWordnet dictionary (an opinion lexicon derived from WordNet database), our approach's sole focus is to gauge consumer sentiment and forecasting comportment using both lexicon features and machine learning algorithms . We conclude by devising a hybrid approach that collectively exhibit accuracy of machine learning techniques and speed of a lexicon approach .**

***Keywords* – Sentiment Analysis, SentiWordNet, WordNet.**

I. INTRODUCTION

Internet is the fastest growing technology since the inception of mankind. Today, it dominates every sphere of our lives. With this huge amount of user participation comes huge amount of unorganised data in the form of comments and opinions on public and private forums. This data coupled with that of rapidly growing on-line discussion groups and review sites covers a lot of information occupying the net. These opinions being unorganised and dirty cannot be handled and analysed using textbook methods while the sheer amount makes them impossible to be manually handled let alone being analysed. In an effort to better organize this information, researchers have been investigating this problem of automatic text categorization. An important part of Automatic text categorisation involves classifying the given information and converting it from a qualitative analysis to quantitative one. This process is called as sentiment analysis. This classification not only helps in managing the data but also forms an auxiliary or sometimes a crucial characteristic of the posted articles that is their sentiment, or overall opinion towards the subject matter — for example, whether a product review is positive or negative (quantitatively 0 or 1).The analysis is particularly useful for review sites, news apps and e-commerce sites where a major part of the functioning depends upon quantifying the user reviews and comments.

Recent years have seen a rapid growth in online discussion groups and review sites (e.g.www.tripadvisor.com) where a crucial characteristic of a customer’s review is their sentiment or overall opinion. There are two main approaches to Sentiment classification or opinion mining : machine learning based and lexicon based. Machine learning based approach deals with data collection, training data and classification while lexicon based method uses sentiment dictionary having sentiment words and correlate them with dataset to find the sentiment direction. In this paper we seek to turn words into quantified measurements using SentiWordNet[2, 5] Dictionary and analyse furthermore different aspects of the particular opinion i.e, if it has more good associations or bad associations. Also, the machine is trained for calculating ratings in other fields(location, price, food etc.) based on review's choice of words.

The objective of this paper is to find out the concepts of lexical based and machine learning approaches in the field of sentiment analysis and combining them to devise a new hybrid technique that collectively provides the positive aspects of both approaches . The paper is organised as follows : II section represents the related work in the field of sentiment classification done in the literature. III section provides a overview of sentiment classification, opinion lexicons, SentiWordNet [2, 5], Machine Learning and Hybrid analysis. IV section provides information about the Dataset used and the proposed methodology of our research work. V section tabularises the results of our research and VI section concludes our work with future improvements.

II. REVIEW OF LITERATURE

This section briefly overview previous work on sentiment-based text classification. A large amount of work in the field of sentiment analysis has been seen in the literature. There are various domains in which sentiment classification work has been observed such as :

1. Movie reviews [1, 3]
2. Hotel reviews [6, 21]
3. News Articles [7, 8]
4. Product reviews [9]

To gauge the sentiment direction of a sentence or a document , two approaches are commonly used : lexicon-based approach and machine learning-based approach.

The lexicon-based approaches are unsupervised learning to sentiment analysis because for classification of data it does not require any prior training [10]. A significant work has been seen in study of [11] ,which determines semantic orientation of reviews by using unsupervised classification. Average semantic orientation of Phrases in reviews that contain adjectives is calculated by using simple unsupervised learning algorithm. Reviews are classified as recommended (thumbs up) and not recommended (thumbs down) in [11]. Our proposed research under lexicon-based approach uses SentiWordNet [2, 5]. Similar work of sentiment classification using SentiWordNet[2, 5] is also presented in [4] on film reviews dataset , in which positive and negative term scores are counted to determine the polarity of the text and the results shows an improvement by building a dataset of pertinent characteristics using SentiWordNet [2, 5] as a source.

The machine learning-based approach in sentiment classification belongs to supervised leaning [10], which generally trains sentiment classifiers by applying some learning techniques. Another area of our proposed research revolves around machine learning techniques : Naive Bayes (NB), Support Vector Machines (SVM) and Random Forest (RF). Various work has been seen including machine learning-based work such as : Three different supervised machine learning algorithms (Naive Bayes, SVM and Maximum Entropy) are analysed and compared to determine the sentiment orientation in the context of movie reviews [12] and the results shows machine learning algorithms outperforms the human produced results as well as NB performing the worst and SVM tends to do the best. A number of probabilistic models are observed in [6] such as Laplace smoothing, semantic orientation, Naive Bayes and SVM to classify a review and results shows better results of Naive Bayes model than SVM with their dataset.

The Hybrid-based approach is a combination of both the lexicon and machine learning-based approaches. Some researches shows that hybrid-based approach performs well than both the previous approaches in the sentiment classification domain. Study of [13] shows that *p-Senti,* a combination of machine leaning and lexicon-based approach acquires high accuracy in sentiment polarity when compared with pure lexicon-based systems as well as have more clear results when compared with pure learning-based systems in the movie and software reviews domain. [14] represents an entity-level sentiment analysis method for twitter data. Their method firstly performs entity-level sentiment analysis using a lexicon approach ,secondly trains a classifier to assign polarities to the newly discovered tweets, thirdly training examples are given by a lexical approach , resulting in improvements and better results than the baselines.

III. SENTIMENT CLASSIFICATION

Sentiment refers to a thought, opinion, or idea based on a feeling about a situation, or a way of thinking about a particular subject. It is the overall direction in which an idea points to, in other words 'the general idea' or 'the central idea' of an entire string of messages written, and the process of employing tools for systematically studying and identifying the orientation of the particular subjective content is called as Sentiment Classification. It can also be said to be the analysis of contextual polarity of the subjective message. This is based on the fact that various parts of the opinions offered by their authors can be assumed to be such that they can be categorised based on a judgement factor and the degree to which a particular word present in the said text is positive or negative.

.

Sentiment Classification has found its uses in many fields especially in voice of customer [15] materials such as product reviews, customer service, natural language processing, social networking, chat bots etc. that is mostly the fields involving human-computer interaction. As these fields require knowing either the central idea behind the human-side of interaction or the knowledge of message polarity, sentiment classification provides an excellent tool for this purpose.

A. Opinion Lexicons (Lexicon Approach)

Opinions are the personal preferences expressed by the user in response to any product experience, situation or service. They can be biased or prejudiced too and are not be based on facts but rather on user experience and personal preferences. Thus it becomes difficult to assess opinions computatively. It may very well be just an individual's point of view and unique in every aspect.

While expressing an opinion of any particular kind, there exists a certain way of expressing it in the form of words carrying certain form of expression, feeling and weightage attached to it that determines the formality, effectiveness and seriousness of the message. This choice of words made by the author while expressing his opinion is called as lexicon, and procedural analysis of the lexicon in order to understand the sentiment behind the message is called as lexical analysis.

Many open-source resources also exist that contain word-sentiment relationships that help determining the polarity of the message.

B. SentiWordNet

Any kind of Lexicon-Based approach requires the study of lexemes(word phrases) to determine the sentiment value of the statement made. One such tool used for this purpose is SentiWordNet Dictionary [2, 5]. It associates sentiment information to each WordNet synset, where each WordNet synset is associated with sentiment scores describing how positive, negative and objective terms are in the synset are. The WordNet maps the relationships and associations attached and the glosses associated with the synsets are analysed. Due to the sheer amount of vocabulary used, the dataset needs to be structured such that it supports fast retrieval of words from the dictionary. The Glosses train the classifiers and the prediction from the classifiers determines the sentiment orientation of the statement.

It is designed so as to accurately predict the polarity as well as meaning of the phrase using grammar and syntactic structure of the sentence, and this final review of the text can be then used for multiple purposes. Here we use this review in association with the Machine Learning Process to improve upon the accuracy of the Sentiment Analysis.

C. Machine Learning Approach

Machine learning is the most adaptable approach for the sentiment orientation purpose due to its greater accuracy. Generally supervised learning techniques are deployed for sentiment analysis work. It is a four stage process [16] :

1. Data Collection - Data is acquired from different sources according to the field of application for analysis purpose.
2. Pre-processing - Data is prepared for feeding into classifier by the application of cleaning.
3. Training - This is the propellant for the classifier. Training data is fed to classifier with the objective of learning.
4. Classification - This is the core part of the whole approach. A machine learning technique is employed depending upon requirement and the classifier is set to be used for sentiment extraction work.

We employed three machine learning techniques in our proposed research work that are :

1. Naive Bayes (NB) - Classification technique based on *Baye's theorem* which predicts the probability for a given finite sequence of terms to belong to a particular class [17].
2. Support Vector Machine (SVM) - Classification technique that analyzes the data, figure out the hyper planes that classify the data into two classes with maximum margin [17].
3. Random Forest (RF) - Classification technique ensembles with a forest (collection of Decision tress) , in which classification is done by giving "votes" by each tree and the most votes classification is chosen by the forest [18].

D. Hybrid Approach

The lexicon based analysis involves using the syntactical presence of a word along with the grammatical structure of a sentence to analyse the meaning and eventually the sentiment of the word. So in lexicon based analysis the words and their individual polarities are the centre of attraction. While, the machine learning-based analysis requires training the classifier with labelled examples. This means that we need to gather a training dataset, extract the features/words from that dataset & then train the classifiers.

Clearly, machine learning-based approach requires a training dataset. If this training dataset provided to the machine learning technique is augmented by the correct outputs of the lexicon based approach, surely there will be an improvement in efficiency provided the sentiment value of the lexemes are already present. This approach of combining the output of the Lexicon based analysis and using them with machine learning-based analysis is called as Hybrid Approach to Sentiment Analysis. This approach improves the classification process because it takes the best of both approaches. Lexicon-based classifiers, for example, are very accurate in determining a text's polarity and strongly worded text is actually a good indicator of its polarity. Machine Learning on the other side is known to be highly domain adaptive and is be able to find deep correlations. Thus, a hybrid system encompasses the best of both worlds that is, lexicon based and machine learning based and being a more-than-single stage process, only the true positives move on to the next iterations thus ensuring a high success rate.

IV. METHODOLOGY

This section includes the information about the Dataset that we used in our research work on sentiment classification. System Architectures are also proposed in this section that we worked upon in our research .-

A. Dataset

We used the Trip Advisor Data Set for our research work in sentiment analysis. Parsed reviews used in our work are crawled from *http://www.tripadvisor.com* .The Meta Data includes : Author, Content, Date, Number of Reader, Number of Helpful Judgment and Overall Rating. It is freely available at the web source : *http://times.cs.uiuc.edu/~wang296/Data/* [19, 20] .

B. Proposed System Architecture

Our proposed research work revolves around three major research areas. Figure 1,2,3 collaboratively gives an architectural overview of our Sentiment classification system.

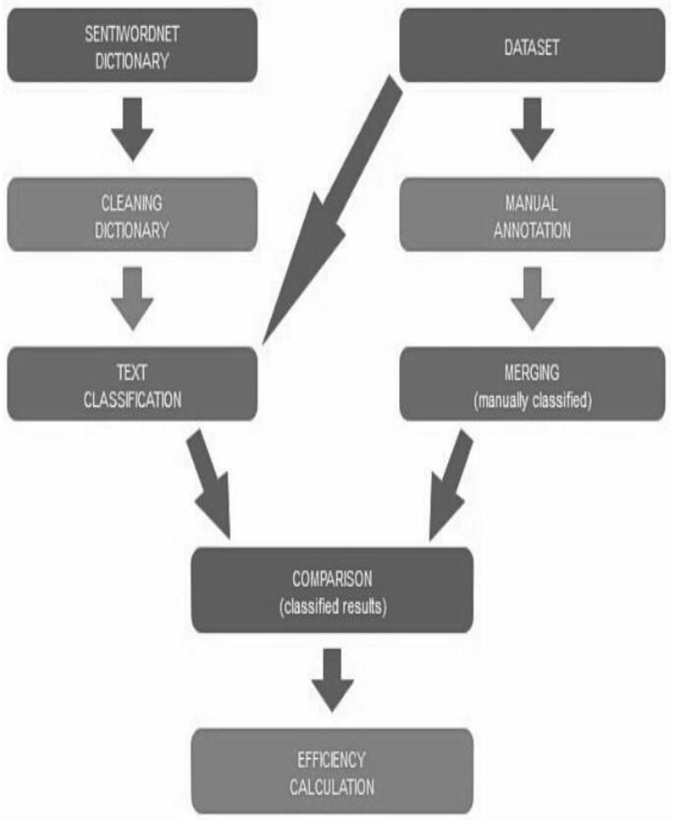

Fig. 1. Architectural Overview of Lexicon-based Classifier

First area of our research is Lexicon Based classification. Architectural overview of our Lexicon-based classifier is presented in Figure 1. This classifier performs Sentiment classification in three broad steps : (1) Manual annotation is applied on the Trip Advisor Dataset (2500+ hotel reviews) and the reviews are analyzed and classified by introducing a class tag under each review into five classes : Strongly Positive , Positive , Neutral , Negative and Strongly Negative . Four additional tags are also introduced under each review which are namely : Food, Service, Price and Locality which gives ratings on the scale of 1-5 (1 being the worst and 5 being the best) on the basis of review to each of the above mentioned tags. (2) A classifier is built with the vision of calculating efficiency using TRIE Data Structure (an information retrieval Data Structure) in Java that took the raw reviews as the inputs , used SentiWordNet [2, 5] (cleaned according to relevant requirements and suitable conditions) as the lexicon and classified the hotel reviews on basis of dictionary and scores in it. (3) The results obtained in first and second step are analyzed and compared and efficiency is calculated. There are also various sub steps included in each step.

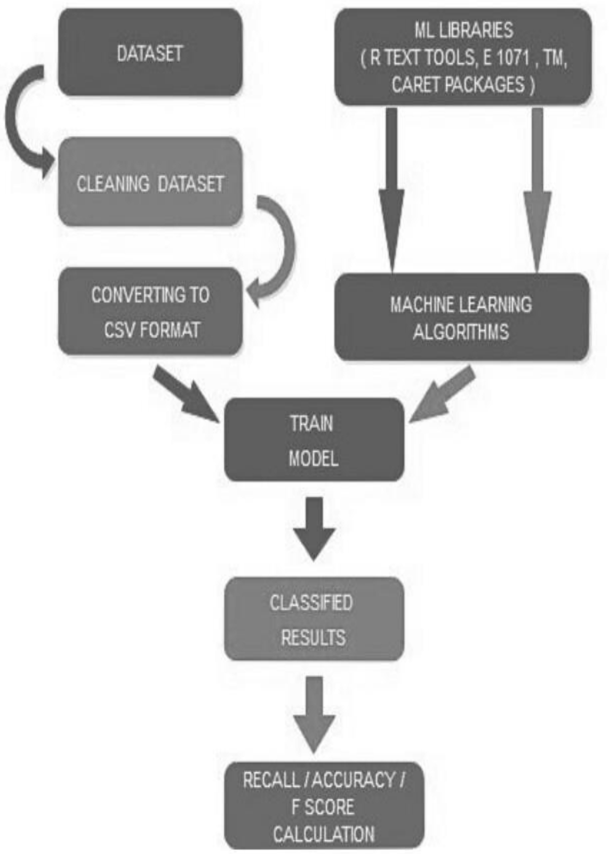
DATASET
ML LIBRARIES (R TEXT TOOLS, E 1071 , TM, CARET PACKAGES)
CLEANING DATASET
CONVERTING TO CSV FORMAT
MACHINE LEARNING ALGORITHMS
TRAIN MODEL
CLASSIFIED RESULTS
RECALL / ACCURACY / F SCORE CALCULATION

Fig. 2. Architectural Overview of Machine Learning-based Classifier

Second area of our research is Machine Learning-based classification. Architectural overview of our Machine learning-based classifier is presented in Figure 2. Working of this classifier is a five step process : (1) Dataset is cleaned using

Java and converted to CSV (comma separated values) format for training purpose. (2) Machine Learning Packages (e1071, R text tools, Text mining and Caret) are employed to deploy machine learning techniques (Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF)) on the CSV converted dataset for the training purpose. (3) In the training phase , a form of *Term Document matrix* is employed and the Dataset is partitioned into two parts to perform iterative *cross validations* to provide a strengthened measure that effectively expresses how our model will perform on an independent dataset, resulting in a trained model (classifier). (4) Classification is done by feeding raw inputs to the trained model and getting the results. (5) *Confusion matrices* (a table often used to describe performance of a classifier) are employed to give the recall accuracies for each of the machine learning techniques. There are also various sub steps included in each step.

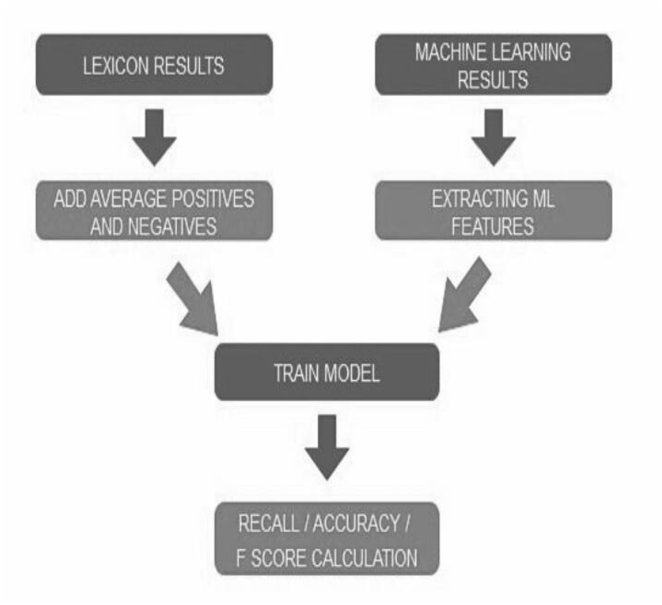
LEXICON RESULTS
MACHINE LEARNING RESULTS
ADD AVERAGE POSITIVES AND NEGATIVES
EXTRACTING ML FEATURES
TRAIN MODEL
RECALL / ACCURACY / F SCORE CALCULATION

Fig. 3. Architectural Overview of Hybrid-based Classifier

Third and main area of our research is a hybrid-based classification. Architectural overview of our Hybrid-based classifier is presented in Figure 3. Working of this classifier is also a five step process : (1) Average positives and negative scores are drawn out from the results of the Lexicon-based classifier. (2) Machine Learning features (food, service, price and locality) are extracted from the results of the Machine Learning-based classifier. (3) The results obtained in the first and second step are combined to train the model with the application of *Term Document matrix* and *Cross Validations*. (4) Classification is done by feeding raw inputs to the trained model and getting the results. (5) *Confusion matrices* are employed to give the recall accuracies for each of the techniques. There are also various sub steps included in each step.

V. RESULTS

This section represents the results obtained from the three approaches used for the sentiment classification work.

Lexicon-based approach uses the SentiWordNet [2, 5] lexicon in the task of Sentiment analysis of Hotel Reviews. The efficiency of this classification is found to be **65%** (which is calculated by ratio (Total number of hits/Total number of reviews)).

Machine Learning and Hybrid-based approach results are tabularized in the Table 1 and plotted graphically in Figure 4.

TABLE I. RECALL ACCURACIES

Approach	**Technique**		
	Naive Bayes	***SVM***	***Random Forest***
Machine Learning	0.5790	0.6954	0.7279
Hybrid	0.6122	0.6974	0.7755

Machine Learning approach's results shows that Random Forest technique performs best while Naive Bayes performs the worst in determining the sentiment direction of the text. Also, results depicts that Hybrid-based approach outperforms the Machine Learning-based results for all the three techniques in the sentiment classification task.

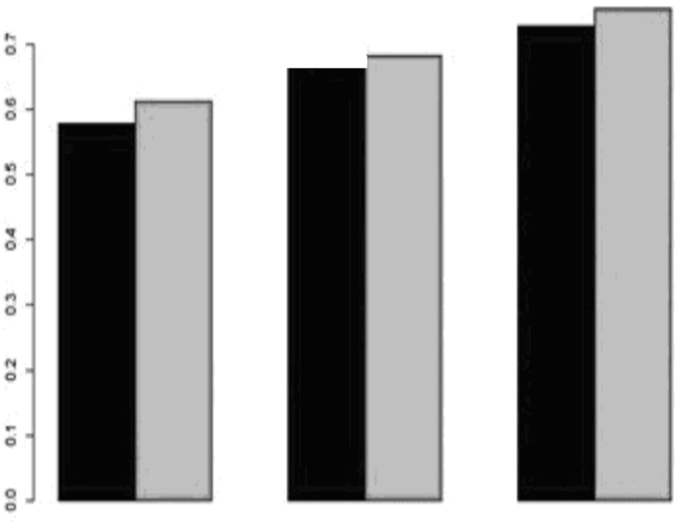

Fig. 4. Comparison of Machine learning approach results (black) and Hybrid approach results (white)

An additional research is also carried out by removing neutral reviews and remaining reviews are classified by Machine Learning and Hybrid-based approach using Naive Bayes algorithm. Results of this research are tabularised in Table II and plotted graphically in Figure 5.

TABLE II. RECALL ACCURACIES
(NAIVE BAYES REVISITED WITHOUT NEUTRAL CLASS)

Technique	**Approach**	
	Machine Learning	***Hybrid***
Naive Bayes	0.7015	0.7958

Results shows an improvement in Hybrid-based approach over Machine Learning-based approach when Naive Bayes is revisited without Neutral class. Also, accuracy of Hybrid-based approach is increased to **79.58%** from **61.22%** when compared with accuracy of original classification (including neutral class reviews also).

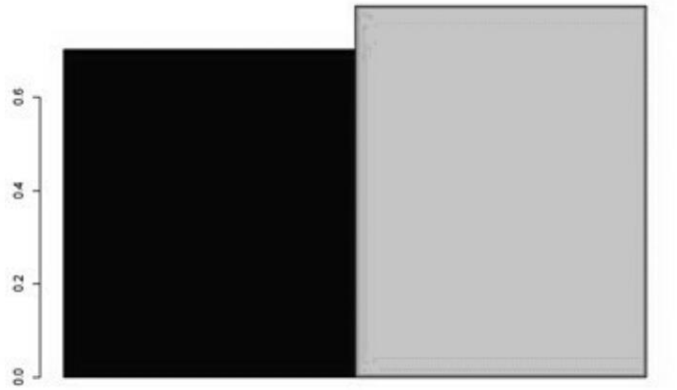

Fig. 5. Comparison of Machine learning approach results (black) and Hybrid approach results (white)
(Naive Bayes revisited without neutral class)

VI. CONCLUSION AND FUTURE SCOPE

This research worked on both opinion lexicons (used SentiWordNet opinion Lexicon) and machine learning approaches (Naive Bayes , SVM, Random forest techniques) and devised a new hybrid technique combining the features of two, which shows improvements over all existing techniques. A further research was done using Naive Bayes technique by removing neutral reviews, and the accuracy was found to be increased in comparison to the research having all positive, negative and neutral reviews.

Future aspects of our research will evolve around making or modifying our own dictionary resource that can be advantageous in overcoming some of the limitations seen in the context and also improving the efficiency by devising an algorithm to increase scores of hotel based feature words in the SentiWordNet dictionary. The efficiency of various techniques with neutral reviews can be worked upon and improved . Web application for classification of hotel reviews will serve as a good way to commercialize and practically use this research in the current world. We shall work upon building a fully fledged application that will provide an intuitive interface for the user to interact with and present the results of analysis in an interactive manner with some options to alter and focus on the more relevant part of the results.

ACKNOWLEDGEMENT

We want to thank Andrea Esuli and Fabrizio Sebastiani (Italian Institute of Information Science and Technology) for the availability of the SentiWordNet lexical resource used in this research work.

REFERENCES

[1] Pouransari, Hadi , Saman Ghili. “Deep learning for sentiment analysis of movie reviews.” (2014).

[2] Esuli A, Sebastiani F. , "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining", Proceedings from International Conference on Language Resources and Evaluation (LREC) , Genoa, 2006

[3] K. Yessenov , S. Misailovi´c : "Sentiment Analysis of Movie Review Comments ",Spring, 2009

[4] Ohana, B. & Tierney, B. , "Sentiment Classification of reviews using SentiWordNet" , 9th.IT&T conference , Dublin,2009

[5] Baccianella S., Esuli A., & Sebastiani F., "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining ",Italy

[6] Vikram Elango and Govindrajan Narayanan. 2014. Sentiment Analysis for Hotel Reviews, http://cs229.stanford.edu/projects2014.html

[7] Kiran Shriniwas Doddi, Dr. Y. V.Haribhakta, Dr. Parag Kulkarni “Sentiment Classification of News Articles”,(IJCSIT) International Journal of Computer Science And Information Technologies, Vol.5 (3), 2014, pp 4621- 4623

[8] Padmaja,” Comparing and Evaluating the Sentiment on Newspaper Articles” Department Of CSE, Hyderabad,Science and information conference 2014

[9] Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande, 2015, "Amazon Review Classification and Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 6

[10] S.M.Vohra,2 PROF.J.B.Teraiya,”A Comparative Study OF Sentiment Analysis Techniques”. Journal Of Information , Knowledge And Research In Computer Engineering .

[11] P. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”, Proceedings of the Association for Computational Linguistics (ACL), 2002, pp. 417–424.

[12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, EMNLP ’02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[13] A. Mudinas, D. Zhang, M. Levene, “Combining lexicon and learning based approaches for concept- level sentiment analysis”, Proceedings of

the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY, USA, Article 5, pp. 1-8, 2012.

[14] 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.

[15] Morrison, Scott (2008-01-28). "So Many, Many Words". The Wall Street Journal. Retrieved 2010-04-14.

[16] Harsh Thakkar and Dhiren Patel Approaches for Sentiment Analysis on Twitter: A state-of-art study accepted at the International Network for Social Network Analysis conference (INSNA), Xi‘an, China, July 2013.

[17] Vimalkumar B. Vaghela and Bhumika M. Jadav “Analysis of Various Sentiment Classification Techniques”, International Journal of Computer Applications (IJCA), 20156, pp: 23-27.

[18 Web resource: https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/

[19] Hongning Wang, Yue Lu & ChengXiang Zhai.: "Latent Aspect Rating Analysis without Aspect Keyword Supervision." ,The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2011), P618-626, 2011.

[20] Hongning Wang, Yue Lu and Chengxiang Zhai :" Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach.",The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2010), p783-792, 2010.

[21] Wojoud Al-Abdullatif, Yasser Kotb, “Using Online Hotel Customer Reviews to improve booking process” International Journal of Computer Applications (0975 – 8887) Volume 97– No.16, July 2014.