ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

# Incremental Learning on Sentiment Analysis Using Weakly Supervised Learning Techniques

P.Bharathi, Student,Department of CSE, Kongu Engineering College, Perundurai, Erode - 638052, Tamil Nadu, India
PCD. Kalaivaani, Faculty, Department of CSE, Kongu Engineering College, Perundurai, Erode – 638052, Tamil Nadu,India

***Abstract – Due to the advanced technologies of Web 2.0, people are participating in and exchanging opinions through social media sites such as Web forums and Weblogs etc., Classification and Analysis of such opinions and sentiment information is potentially important for both service and product providers, users because this analysis is used for making valuable decisions. Sentiment is expressed differently in different domains. Applying a sentiment classifiers trained on source domain does not produce good performance on target domain because words that occur in the train domain might not appear in the test domain. We propose a hybrid model to detect sentiment and topics from text by using weakly supervised learning technique. First we create sentiment sensitive thesaurus using both labeled and unlabeled data from multiple domains. The created thesaurus is then used to classify sentiments from text. This model is highly portable to various domains. This is verified by experimental results from four different domains where the hybrid model even outperforms existing semi-supervised approaches.***

***Index*** **– Sentiment Analysis, Opinion Mining, Joint Sentiment topic (JST) model, Sentiment Classification, Naïve Bayes algorithm**

## I. INTRODUCTION

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. Sentiment classification techniques can help researchers to study sentiment information on the Internet by identifying text containing opinions and emotions. This is useful to determine whether a text contains positive or negative sentiments. Machine learning technique and semantic- orientation approach are used in previous research. It does not require prior training].Most of the current opinion mining research has focused on business and e-commerce applications, such as product reviews and movie ratings. Little research has tried to understand opinions in the social and geopolitical context [7]. Sentiment classification is a very domain specific problem. Training a classifier using the data from one domain may fail when testing against data from another. As a result, real application systems usually require some labeled data from multiple domains, guaranteeing an acceptable performance for different domains [15]. McDonald and Hannan used fine-to-coarse sentiment analysis to identify the sentiment of the document and all of its subcomponents, whether at the paragraph, sentence, phrase or word level [14]. Another researcher used novel approach to the problem of system portability across different domains by developing a sentiment annotation system that integrates a corpus-based classifier with a lexicon-based system trained on Word Net [5]. Online reviews evolve rapidly over time which demands much more efficient and flexible algorithms for sentiment classification than the current approaches can offer. These observations have thus motivated the problem of using weakly supervised approaches for domain-independent sentiment classification.

In this paper we focus on classifying positive and negative sentiments on document for different domains in association with topic detection and sentiment analysis based on the proposed weakly supervised naïve bayes algorithm. The weakly supervised nature of this model performs well in different domains. Experiments have been conducted with the algorithm to detect positive and negative sentiment of the document by using sentiment sensitive thesaurus on multi- domain sentiment (MDS) datasets.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 presents the naïve bayes algorithm. Section 4 presents experimental setup and the results are discussed in Section 5. Finally section 6 concludes the paper and outlines future work.

## II. RELATED WORK

Sentiment analysis aims to determine the speaker or a writer with respect to some topic or the overall contextual polarity of a document. Machine learning techniques have been widely deployed for sentiment classification at various levels. B.Pang and Lee experimented with three machine learning methods (Naïve Bayes, maximum entropy classification and support vector machines) to classify movie reviews. They compared Naïve Bayes, Maximum Entropy and SVM and achieved highest accuracy (83 percent) using SVM. But it can be tested with only movie domain [12]. Another research used unsupervised learning algorithm for classifying reviews as recommended or not recommended by calculating the semantic orientation of a phrase. PMI-IR algorithm was applied to estimate the semantic orientation of a phrase. It can be calculated by comparing its similarity to a positive reference word with its similarity to a negative reference word [18]. Christopher and Dorbin investigated the density-based algorithm and proposed the scalable distance-based algorithm (SDC) for analyzing Web opinions. Although SDC achieves good performance in clustering Web opinions, it has own limitations. SDC does not require a predefined number of clusters and two parameters used for identifying clusters have impacts on micro and macro accuracy [3]. Corpus-based techniques and Dictionary-based techniques used in previous semantic orientation approach. Corpus-based technique was used to find co-occurrence pattern of words to determine their opinions. Corpus-based techniques depend on a large corpus to calculate the mathematical information needed to decide sentiment orientation. So it might not be efficient as dictionary-based techniques [20]. Swati and Manali [16] tested with movie reviews and proposed a new hybrid approach involving rule based classification, supervised learning, and machine learning method for analyzing opinions. It produced maximum accuracy. Another research proposed a novel approach for solving domain dependency problem. They used annotated in-domain data and a lexicon-based system for classifying reviews. But it will produce only 60 percent of accuracy [5]. Blitzer experimented with structural correspondence (SCL) algorithm to solve domain transfer problem for sentiment classification. Candidate pivot features were selected based on frequent words in both source and target domains and pivots were then chosen based on mutual information between candidate features and the source labels. They achieved only 46 percent of accuracy [6]. Shoushan Li and Chengqing proposed a new task called multi domain sentiment classification that aims to improve performance through training data from multiple domains. Feature-level and classifier level approaches were used for classification. It produced 83 percent of accuracy [15]. The Multi-Grain Latent Drichlet Allocation model (MG-LDA) used in another research. It can be used to build topics that are representative of ratable aspects of customer reviews by allowing terms being generated from either global topic or local topic. The limitation of MG-LDA is it still purely topic based without considering the associations between topics and sentiments [8]. Titov and McDonald proposed the Multi-Aspect Sentiment (MAS) model for sentiment text extraction based on supervised learning techniques [15].

All of the aforementioned work shares some similar limitations: 1) Most of the previous research based on supervised learning. It requires labeled data for training and it will produce poor performance. 2) They focused only on classifying opinions without considering topics which reduce the effectiveness. 3) Mostly all work used lexicon or Word net for classification.

## III. HYBRID MODEL

Hybrid model for sentiment classification focuses on the challenge of training a classifier from one or more domains and applying the trained classifier on different domains. This hybrid model must overcome two problems. First we identify source domain features to the target domain features. Second we require a learning framework for these features. In this paper, we propose a hybrid model to overcome both problems. A Hybrid model for sentiment classification is based on LDA [19]. LDA model was designed by the assumption that documents are mixture of topics, where a topic is a probability distribution over words. To generate a word in a document, we first choose a distribution over a mixture of T topics for the document. Following that we will choose a topic randomly from the topic distribution and generates a word from that topic corresponding to topic-word distribution. The existing work has three layers, where topics are joined with documents, and words are joined with topics. To classify the document sentiment, we have to add an additional sentiment layer between the document and topic layers. The architecture diagram of a hybrid model is shown in Figure 1.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

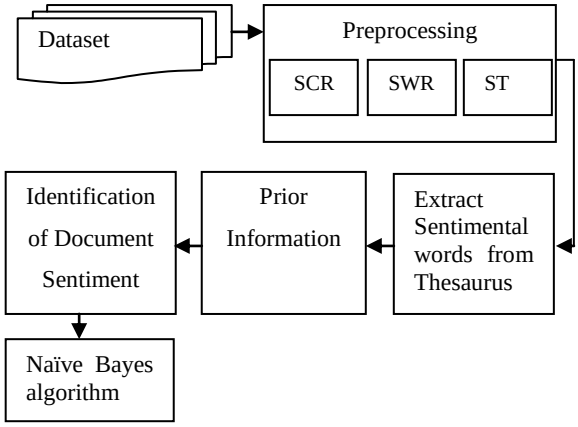


**Fig 1 Architecture diagram**

Note: SCR-Special Characters Removal, SWR-Stop Words removal, ST-Stemming

Consider a database which contains collection of documents doc1, doc2, doc3..... docn. Each document contains Nd words denoted by wrd1, wrd2……wrdNd and each word in the document is an item from a vocabulary index with V distinct terms denoted by 1, 2,…V. Let l be the number of distinct labels and t be the number of topics. To generate a word w, first one select label l from document distribution δd then select a topic from topic distribution μdoc,l. Finally generate a word from corpus distribution φdoc.

First we calculate posterior distribution to obtain the distribution of δ, μ, φ. Gibbs sampling procedure was used to estimate posterior distribution by sampling the variables [1].

Let the subscript –t denote a quantity that excludes data from tth position, the posterior distribution is calculated as follows,

$$p((t=i,l=m/wrd,t^{-t},l^{-t})\alpha,\beta,\gamma)$$
$$\alpha \frac{G_{l,t,wrd}^{-t}+\beta}{G_{l,t}^{-t}+V\beta}\cdot\frac{G_{doc,l,t}^{-t}+\alpha_{l,t}}{\sum_t \alpha_{l,t}}\cdot\frac{G_{doc,l}^{-t}+\gamma}{G_d^{-t}+S\gamma} \quad (1)$$

where V is the size of the vocabulary, l indicates sentiment labels, Gt,l,wrd is the number of times a word wrd occurred with label l and topic t. Gl,t is the number of times words are assigned to topic t and sentiment label l. Gdoc,l,t is the number of times a word from document doc and being associated with t and sentiment label l. We set the symmetric prior β = 0.01, the symmetric prior γ = (0.05 x L)/S, where L is the average document length, S is the total number of sentiment labels, and the average of 0.05 allocates 5 percent of probability mass for mixing [1]. The procedure of document sentiment classification is as follows,

1. Multiple Datasets (MDS) is used for classification. forms of a word to its lemma. Then use simple word filter First datasets are preprocessed.

2. We create sentiment sensitive thesaurus that aligns different words that express the same sentiment in different domains.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

3. Extract sentimental words from created thesaurus.

4. Classify positive and negative sentiments by comparing the words with thesaurus.

5. Calculate the probability distribution of a hybrid model.

6. Finally we identify the document sentiment by comparing the count of sentiments.

## IV. EXPERIMENTAL SETUP

### *A. Datasets Description*

In this paper, we used multiple data sets (MDS) to classify the document sentiment. Movie review datasets are used in most of the existing research. MDS data set contains four different types of product reviews crawled from Amazon.com including Book, DVD, Electronics and Kitchen. Normally data set contains punctuation, numbers, non alphabet characters and stop words.

**TABLE I. DATA SET STATISTICS**

| Datasets | Book | DVD | Electronic | Kitchen |
|---|---|---|---|---|
| Doc Length(+) | 156 | 150 | 100 | 90 |
| Doc Length(*) | 116 | 110 | 85 | 70 |
| Vocabulary size(+) | 20,020 | 19,800 | 10,550 | 9875 |
| Vocabulary size(*) | 18,970 | 20,564 | 9490 | 8560 |

Note: (+) denotes before preprocessing and (*) denotes after preprocessing

So we first perform preprocessing to remove special characters and stop words. Then Stemming is performed to find the root words. The data sets before and after preprocessing is shown in table I.

### *B. Sentiment Thesaurus*

We create sentiment sensitive thesaurus to classify the sentiments from multiple documents. We use labeled data from multiple source domains and unlabeled data from source and target domains to represent the distribution of features. We use lexical elements and sentiment elements to represent a user review. Next for each lexical element, we measure its relatedness to other lexical elements and group related lexical elements to create a sentiment sensitive thesaurus. We split the review into individual sentences and conduct part-of-speech (POS) tagging and lemmatization using RASP system [2]. Lemmatization is the process of normalizing the inflected based on POS tags to filter function words, retaining only nouns, verbs, adjectives and adverbs. Finally we extract lexical elements from both source and target domains. From the lexical words we classify the sentiments. The main problem in sentiment classification is that features that appear in source domain do not always appear in target domain. For that reason, even if we train a classifier using labeled data from the source domains, the trained system cannot be used to classify test domain. So we use feature expansion method to solve this problem.

### *C. Classification of Sentiments*

To classify document sentiment, we perform preprocessing on data sets to remove special characters and symbols. Then extract lexical elements from the created thesaurus. Next we extract positive sentiment words by comparing each word in our input document with thesaurus. Thesaurus contains sentimental words and also defined the polarity for each word. Similarly negative sentiments are also classified. The main use of Hybrid model is to detect sentiments and topics from text. Sentiments may be varied according to the topics. So by

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

detecting sentiments and topics will improve the efficiency of mining results. After the classification of positive and negative sentiments, we must identify the sentiment labels and sentiment topics from our documents. Probability distribution of Hybrid model is calculated by using the count of sentiment labels, topics and document length. Finally document sentiment is classified based on the probability of sentiment labels. We define that document d is classified as a positive sentiment document if its probability of positive sentiment label is greater than its probability of negative sentiment label, and vice versa.

## V. EXPERIMENTAL RESULTS

Results of document sentiment classification and topic extraction are discussed in this section. We conducted a set of experiments on Hybrid model with multiple numbers of documents. We propose a method to use created thesaurus to expand feature vectors at train and test times in a binary classifier. The performance of the method depends on the sentiment sensitive thesaurus we use for feature expansion. Examples of Sentiment labels and sentiment topics extracted from documents are shown in table II. Sentiment labels extracted from thesaurus. Sentiment topics are chosen randomly by reading each sentence in the documents. A few examples are shown for books and DVD data sets. Similarly we can get sentiments and topics for all data sets. The thesaurus contains both labeled and unlabeled data from different domains.

**TABLE II. EXAMPLES OF SENTIMENT LABELS AND TOPICS**

| | Book | DVD | Electronic | Kitchen |
|---|---|---|---|---|
| Positive Sentiment | Best<br>Redemption<br>Enchanment<br>Deserve<br>Better | -<br>Charming<br>Animated<br>Honest<br>Adequate<br>Decent | Excellent<br>Redundant<br>Perfect<br>Sturdy<br>Large | Pleasant<br>Ready<br>Fancy<br>Support<br>Strong |
| Negative Sentiment | Confusion<br>Allusions<br>Boring<br>Corrupt<br>Disorder | Angry<br>Horror<br>Worst<br>Hard<br>Broke | Miserable<br>Pretend<br>Suspicious<br>Dead<br>Empty | Irritating<br>Problem<br>Cheap<br>Mad<br>Difficult |

Extract the lexical words from the thesaurus. These words are related to sentimental analysis task. From that, we classify positive and negative sentiments by comparing each positive or negative word with thesaurus. Then count the sentiments and find the document sentiment. Hybrid model performs well in all domains when compared to existing supervised and semi-supervised learning techniques. Example of positive and negative sentiments classified by a Hybrid model is shown in table III.

**TABLE III. CLASSIFICATION OF POSITIVE AND NEGATIVE SENTIMENTS**

| Datasets | Sentiment Labels | Sentiment Topics |
|---|---|---|
| Books | Potential<br>Simple<br>Clearly<br>Disstateful<br>Bad | Best Chrichton novels<br>Horrible book<br>Medicine of the future<br>Shallow self-indulgence<br>Disappointing mess |
| DVD | Predictable<br>Adequate<br>Honest<br>Perfect | Best Animated film<br>Reminiscing<br>Not too impressed<br>The Bone Collector |

More number of sentiments is detected from our input documents. Some of the examples are shown in the above table. These words are used further for classifying document sentiment. It can be tested with both unigrams and bigrams of words. These words are identified with the help of constructed thesaurus.

The naïve bayes algorithm gives the accurate result than the normal LDA method for both the classes book and dvd. Figure 2 shows the accuracy and result for book with proposed naïve bayes method and the normal JST, reverse JST. Figure 3 shows the error rate of the same classes.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

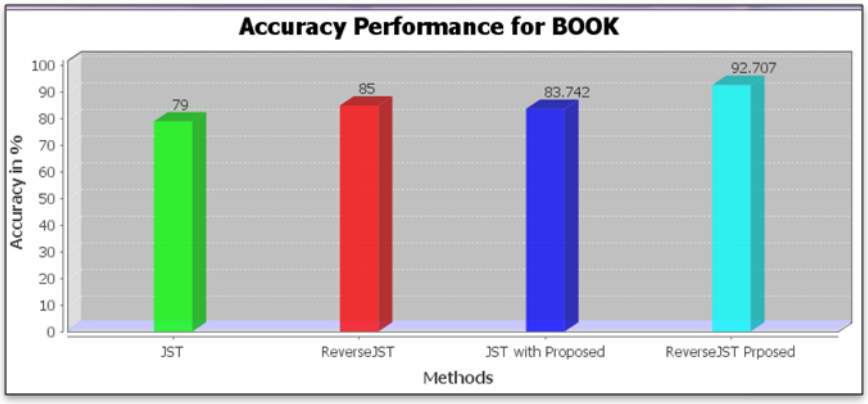


**Fig 2 Accuracy Performance of book**

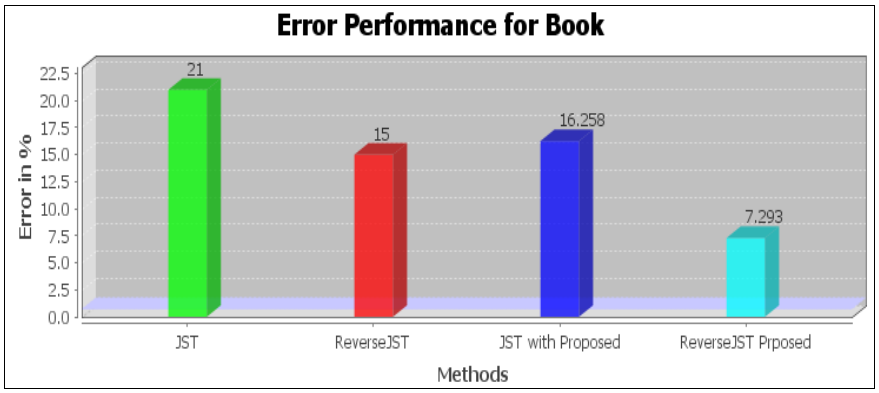


**Fig 3 Error Performance for book**

## VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a naïve bayes classifier to overcome the problem of domain dependency by using an automatically extracted sentiment sensitive thesaurus. We use labeled data from multiple source domains and unlabeled data from source and target domains to compute the relatedness of features and construct a sentiment sensitive thesaurus to solve feature mismatch problem. This naïve bayes algorithm is created based on weakly supervised learning techniques. So it is portable to all other domains. It produces good performance results which demonstrate the flexibility of hybrid model for sentiment analysis task.

There are several directions we plan to investigate in the future. One is, after identifying document sentiment we will also identifies the semantic orientation of specific components of the review which will help to improve the mining results.

## REFERENCES

[1] Chenghua Lin, Yulan He, and Richard Everson, "Weakly Supervised Joint Sentiment Topic Detection from Text", Proc. IEEE Transactions on Knowledge and Data Engineering, pp. 1134-1145, 2012.

[2] Danushka Bollegala, David Weir, and John Carroll, "Cross Domain Sentiment Classification using a Sentiment Sensitive Thesaurus", Proc. IEEE Transactions on Knowledge and Data engineering, 2012.

[3] Christopher, Yang and Tobun, "Analyzing and Visualizing Web Opinion Development and Social Interactions with Density Based Clustering", IEEE Transactions 2011.

[4] A. Aue and M. Gamon, "Customizing Sentiment Classifiers to New Domains: A Case Study", Proc. Recent Advances in Natural Language Processing (RANLP), 2005.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

[5] Alina Andreevskaia and Sabine Bergler, "When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging", Proceedings of ACL-08, June 2008.

[6] J. Blitzer, M. Dredze and F. Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment classification," Proc. Assoc. for Computational Linguistics (ACL), pp. 440-447, 2007.

[7] Hsinchun Chen and David Zimbra, "AI and Opinion Mining", Proc. IEEE Intelligent Systems, pp. 74-80, 2010.

[8] T. Li, Y. Zhang and V. Sindhwani, "A Non-Negative Matrix Tri-Factorization Approach to Sentiment Classification with Lexical Prior Knowledge", Proc Joint Conf. Conf. 47th Ann. Meeting of the ACL and the Processing of the Fourth Int'l Joint. Conf Natural Language AFNLP, pp. 244-252, 2009.

[9] C. Lin and Y. He, "Joint Sentiment/Topic Model for Sentiment Analysis", Proc . 18th Conf. Information and Knowledge Management (CIKM), pp. 375-384, 2009.

[10] S. Lacoste - Julien, F. Sha and M. Jordan, "Disc LDA: Discriminative Learning for Dimensionality Reduction and Classification", Proc. Neural Information Processing Systems (NIPS), 2008.

[11] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs Up: Sentiment Classification using Machine Learning Techniques", Proc. ACL Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 79-86, 2002.

[12] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts", Proc. 42th Ann. Meeting on Assoc. for Computational Linguistics (ACL), pp. 271278, 2004.

[13] Ryan McDonald and Kerry Hannan Tyler Neylon Mike Wells Jeff Reynar, "Structured Models for Fine-to- Coarse Sentiment Analysis", Proc. Assoc. for Computational Linguistics (ACL), pp. 432-439, 2007.

[14] Shoushan Li and Chengqing Zong, "Multi-Domain Sentiment Classification", Proc. Assoc. Computational Linguistics – Human Language Technology (ACL- HLT), pp. 257-260, 2008.

[15] Swati A. Kawathekar and M. Kshirsagar, "Sentiment Analysis using Hybrid Approach involving Rule-Based and Machines Method", Proc. IOSR, Vol. 2 Issue 1, Jan.2012, pp. 055-058.

[16] Titov and R. McDonald, "Modeling Online reviews with Multi-Grain Topic Models", Proc. 17th Int'l Conf. World Wide Web, pp. 111-120, 2008.

[17] P. D. Turney, "Thumbs Up or Thumbs Down?: A Semantic Orientation Applied to Unsupervised Classification of Reviews", Proc. Assoc. for Computational Linguistics (ACL '01), pp. 417-424, 2001.

[18] Titov and McDonald, "A Joint Model of Text and Aspect Ratings for Sentiment Summarization", Proc.Asso. Computational Linguistics – Human Language Technology (ACL-HLT), pp. 308-316, 2008.

[19] Yan Dang, Yulei Zhang, and Hsinchun Chen, "A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews", Proc. Intelligent Systems IEEE, pp. 46-53, 2010.