Using Joint Sentiment Topic Detection with Bigrams to improve the classification in Weakly Supervised Sentiment Analysis

PCD.Kalaivaani
Faculty, Department of Computer Science and Engineering
Kongu Engineering College,Perundurai, Erode – 638052, Tamil Nadu, India

Dr.R.Thangarajan
Professor, Department of Computer Science and Engineering
Kongu Engineering College,Perundurai, Erode – 638052, Tamil Nadu, India

Abstract — Online reviews evolve rapidly over time, which demands much more efficient and flexible algorithms for sentiment analysis than the current approaches. Current approaches detect the overall sentiment of a document, without performing an in–depth analysis to discover. We propose a Document level sentiment classification in conjunction with topic detection and topic sentiment analysis of bigrams simultaneously. This model is based on the weakly supervised Joint Sentiment-Topic model, and this extends the Latent Dirichlet Allocation by adding the sentiment layer. Bigrams are considered in order to increase the accuracy of sentiment analysis. A sentiment thesaurus is created with positive and negative lexicons and this is used to find the sentiment polarity of the bigrams. This model can be shifted to other domains. This is verified experimentally through four different domains which even outperforms the existing semi-supervised approaches.

Keywords-Sentiment Analysis, Opinion Mining, Joint Sentiment topic (JST) model, Sentiment Classification, Maximum Entropy.

I. INTRODUCTION

Sentiment analysis or opinion mining aims to determine the attitude of a speaker or a writer with respect to some topic. The attitude may be his or her judgment, evaluation, emotions and opinions. Sentiments are the subjective information and not facts. This is useful in determining whether a text contains positive or negative sentiments. Most of the current opinion mining research is focused on business and e-commerce applications, such as review of products and movie reviews. Few researches have tried to understand opinions in the social and geopolitical context [1].

Machine learning techniques is been widely used for sentiment classification at various levels. It does not require prior training. These techniques provide the overall sentiment of a document, and do not concentrate the topics in the document. These observations have thus motivated the problem of using weakly supervised approach for domain-independent sentiment mining

The main aim of our work is to classify the text using sentiment analysis on the review (opinion) of the user to evaluate the product. In this paper we have proposed a hybrid model which uses the LDA topic modeling to extract the topic and sentiment of the text simultaneously. In this model we have used bigrams which are formed by combining the adjacent words. And the formed bigrams are checked with sentiment lexicons to extract the essential bigrams, this is done to eliminate the overhead of input. The use of bigrams instead of unigrams can improve the text categorization as some words give different polarity meaning when considered as bigrams. For e.g. phrases such as “not good” or “not durable” will give different polarity meanings when it is taken as unigrams. Then the sentiment sensitive thesaurus has been used to identify the word’s sentiment.

The rest of the paper is organized as follows. Section II describes the related work. Section III presents the hybrid model. Section IV presents experimental setup and the results are discussed in Section V and VI. Finally section VII concludes the paper and outlines future work.

II. RELATED WORK

Machine learning techniques have been widely deployed for sentiment classification. B.Pang and Lee experimented with Naïve Bayes, maximum entropy classification and support vector machines to classify movie reviews. They compared Naïve Bayes, Maximum Entropy and SVM achieved highest accuracy (83 percent). But it can be tested with only movie domain [2]. Christopher and Dorbin investigated the density-based algorithm and proposed the scalable distance-based algorithm (SDC) for analyzing Web opinions. Even though SDC achieves good performance in clustering Web opinions, it has few drawbacks. In this predefined number of clusters is not required [3]. Corpus-based techniques and Dictionary-based techniques which are used in previous semantic orientation approach finds the co-occurrence pattern of words to determine their opinions. Corpus-based techniques depend on a huge corpus to calculate the mathematical information needed to decide sentiment orientation. So it might not be efficient as it is dictionary based [4].

Swati and Manali worked with movie reviews and proposed a new hybrid approach with rule based classification, supervised learning, and machine learning method for sentiment analysis [5]. This produced maximum accuracy. Another novel approach for solving domain dependency problem annotated in-domain data and a lexicon based system for classifying reviews. But it produced only 60 percent of accuracy [6]. Blitzer experimented with structural correspondence (SCL) algorithm to solve domain transfer problem for opinion mining, Candidate pivot features were selected based on the frequent words. They achieved only 46 percent of accuracy [7]. Shoushan Li and Chengqing proposed a task called multi domain sentiment classification which aims to improve the performance by training data from multiple domains. It produced 83 percent of accuracy [8].

The Multi-Grain Latent Drichlet Allocation model (MG-LDA) used in another research. It can be used to build topics. The limitation of MG-LDA is it is topic based, it does not consider the associations between topics and sentiments [9]. Titov and McDonald proposed the Multi-Aspect Sentiment (MAS) model for sentiment extraction and this is based on supervised learning technique [8]. Schutze applied dimensionality reduction techniques to overcome computational intensity and over fitting in solving document routing, which is a problem of statistical text categorization. They used single words and two-word phrases called bigrams that were chosen by term frequency as an evaluation measure. This showed that a reduced feature space was beneficial for document routing [10].

Jensen and Martinez used conceptual and contextual features, such as synonyms, hypernyms, and bigrams, to improve text classification. They reported that introducing these features decreased the error by 33% [11]. All of the aforementioned work shares some similar limitations: 1) Most of the previous researches are based on supervised learning. It requires labeled data for training and it produces poor performance. 2) They focused only on classifying opinions but do not concentrate on topics which reduce the effectiveness of classification. 3) Most of the works used lexicon or Word net for classification. 4) Only unigrams were considered for sentiment analysis which does not give accurate sentiment analysis for negated words.

III. HYBRID MODEL

The hybrid model aims to overcome the two problems in sentiment analysis, first is the portability of the classification model to various domains, and second is the loss of accuracy in classification by using unigrams. In this paper we propose a hybrid model which is based on LDA. In this model we used bigrams instead of unigrams because bigrams give different polarity meaning. Topics and sentiments are extracted simultaneously with the help of Gibbs sampling method. Sentiment layer is added between the document layer and the topic layer, this model contains three layers first is the sentiment labels associated with the documents, second is the topics associated with the sentiment labels and third is the layer of words associated with both sentiment labels and topics. Prior information is given to the model using sentiment sensitive thesaurus, this thesaurus contains the positive and negative lexicons. Thus polarity of a bigram can be found by comparing it with the lexicons in the thesaurus. Finally the sentiment labels are found for each bigram.

The architectural framework of the hybrid model consist of pre-processing steps such as stop word removal and stemming, bigram generation, topic word extraction sentiment label generation, incorporation of prior information and finally the prediction. The architecture diagram of the hybrid model is shown in Figure. 1

In the stop word removal process numbers, punctuations and words which are not of much importance in NLP, (a, and, the, is,…) are removed, in the stemming inflected (or derived) words are reduced to their stem, base or root form ("argue", "argued", "argues", "arguing" the base or root word is "argu").
In the generation of bigrams the following steps are carried out

- Find the list of lexicon (S)
- Preprocessed documents (stemmed and stop words removed)
- Combine the adjacent words
- For each pair of (W1 + W2) check if any one of W1 or W2 is present in S then add to the bigram list (B)
- Output B

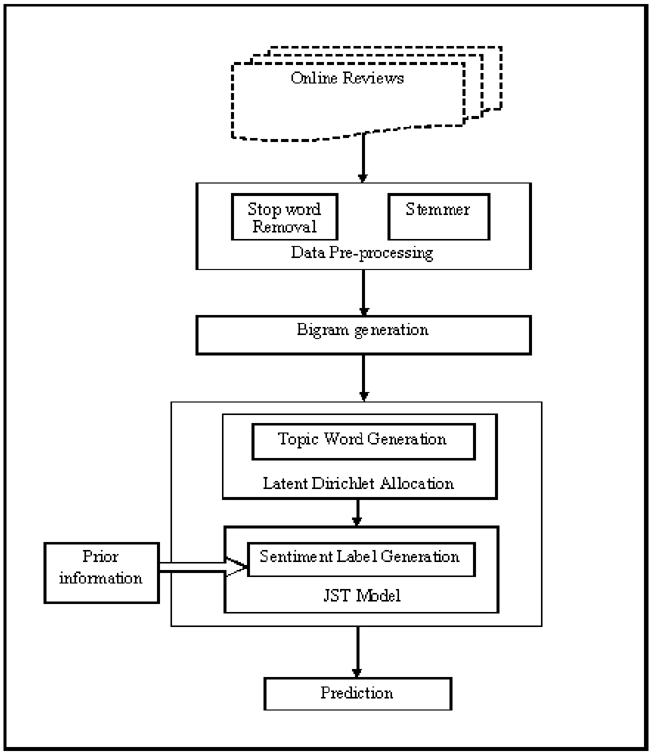

Figure 1. Architecture diagram of Hybrid Model

In ordered to learn the topic representation of each document and the words associated to each topic the following procedure is carried out.

1. choose a fixed number of topics and go through each document, and randomly assign each word in the document to one of the topics to improve the classification
2. Go through each word in the document
3. For each topic t compute => P(topic t | document d)
   compute => P(word w | topic t)
4. Reassign w with probability of
   P((topic t | document d)* P(word w | topic t))

These steps are repeated many times in ordered to get a good result. This is an outline of the Latent Dirichlet Allocation which classifies the words in the document under various topics. The LDA model is used as a base for JST model.

The JST model is used to find the sentiment label and the topic associated with each bigram. Gibbs sampling procedure is used in JST. The requirements for Gibbs sampling are α, β, γ of a corpus, where **α** is the prior observation count for number of times topic **j** is associated with sentiment label **l, β** is taken as a constant value of

0.01 and γ is taken as **(0.05*L) / S**, here L is the length of the document and S is the number of sentiment labels. It is ensured that the sentiment and topic labels are assigned for all bigrams in the corpus. Then the following steps are carried out [1].

```
Initialize 3 matrices $S \times T \times V$ matrix$\varphi$; $D \times S \times T$ matrix$\emptyset$, and $D \times S$ matrix $\pi$
For i = 1 to max Gibbs sampling iterations do
    For all documents do
        For all words do
            Find the variables $N_{k,j,i}, N_{k,j}, N_{d,k,j}$ , $N_{d,k}$ and $N_d$ where $N_{k,j,i}$ is the number of times word I appeared in topic j and sentiment label k, $N_{k,j}$ is the number of times words are assigned to topic j and sentiment label k, $N_{d,k,j}$ is the number of times a word from document d being associated with topic j and sentiment label k $N_{d,k}$ is the number of times sentiment label k is assigned to words in document d, $N_d$ is the number of words in document d
            Sample a new sentiment–topic pair ~l and ~z
            Update the variables $N_{k,j,i}, N_{k,j}, N_{d,k,j}$ , $N_{d,k}$ and $N_d$ using the new sentiment label ~l and topic label ~z
        end for
    end for
    for every 25 iterations do
        Update the hyper parameter $\alpha$ with the maximum likelihood estimation.
    end for
    for every 100 iterations do
        Update the matrix ,$\emptyset$ and $\pi$ with new sampling results;
    end for
end for
```

New sentiment topic pairs are sampled using equation (1)

$$\text{New sentiment topic pair} = \varphi_{i,j} * \Theta_{d,k,j} * \Pi_{d,k} \qquad (1)$$

Where $\varphi_{ij} = \frac{N_{k,j,i}+\beta}{N_{k,j}+V\beta}$

$$\Theta_{d,k,j} = \frac{N_{d,k,j,}+\alpha_{k,j}}{N_{d,k}+\Sigma_j \alpha_{k,j}}$$

$$\Pi_{d,k} = \frac{N_{d,k}+\gamma}{N_{d,}+S_\gamma}$$

IV. INCORPORATION OF PRIOR INFORMATION

The efficient way for sentiment analysis can be done by incorporating prior information and extracting the word from topic word distributions generated from a Dirichlet distribution. First, the λ matrix of size S x V, which is used to encode word prior sentiment information into JST model, is initialized with all the elements taking a value of 1. For each term w ∈ {1, . . . ,V} in the corpus vocabulary and for each sentiment label l ∈ {1, . . . ,S}, if w is found in the sentiment lexicon, the element λ_{lw} is updated as in equation (2)

$$\lambda_{lw} = \begin{cases} 1, & \text{if } S(w)=l, \\ 0, & \text{otherwise,} \end{cases} \qquad (2)$$

Where the function S(w) returns the prior sentiment label of w in sentiment lexicon, i.e., neutral, positive, or negative. For example, the word "excellent" with the index in the vocabulary has a positive sentiment polarity. Equation (4.2) shows the corresponding row vector in λ is [0, 1, 0] with its element representing neutral, positive, and negative prior polarity. Then for each topic j ∈ {1, . . . ,T}, multiplying λ_{li} with $\beta_{lji,}$ only the value of β_{lposji} is retained, and β_{lneuji} andβ_{lnegji} are set to 0. Thus, "excellent" can only be drawn from the positive topic word distributions generated from a Dirichlet distribution with parameter β_{lpos} Finally after completing all the steps we obtain the polarity of the words i.e. Positive or negative. Performance of the classification is evaluated and it is observed that accuracy is increased while bigrams are used instead of unigrams. The example scenario of our work is shown in the Figure 2.

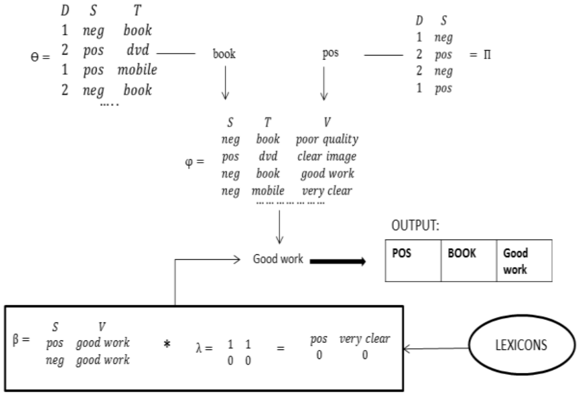

Figure 2. Example scenario

V. EXPERIMENTAL SETUP

5.1 Datasets Description

For experiment, Multi domain sentiment dataset is used. The Book and DVD reviews corpus has 2000 reviews that were classified in terms of the overall orientation as being either positive or negative (1000 positive and 1000 negative reviews). The benchmark review dataset is collected from Cornell University. The dataset statistics before and after removing the punctuations and stop words, and after stemming the derived words is shown in the Table-1

Table-1 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

5.2 Sentiment Thesaurus

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

VI. EXPERIMENTAL RESULTS

A set of experiments were conducted on the hybrid model with multiple number of documents. A novel method is proposed to use the 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. Sentiment labels are extracted from thesaurus. Sentiment topics are chosen randomly by reading each sentence in the documents. These topics are reassigned many times using Gibbs sampling method. The sentiment labels are found for each word using the thesaurus. The thesaurus contains both labeled and unlabeled data from different domains. Few extracted topics and sentiment labels for bigrams are shown in Table-2.

Table-2 Extracted Positive and Negative Words Bigrams

	CAMERA	**BOOK**
Negative Senti. Labels	Poor Quality	No interest
	Small Picture	Read deeply
	less clear	Read Bore
Positive Senti. Labels	Screen Clarity	Low Cost
	Good resolution	Short story
	Easy handle	Very interest

The count of words in the input and the count of bigrams obtained after implementing the Joint Sentiment Topic model are shown in the Table-3

Table-3 Count of Words

Number of Words In Input	3,64,244
Number of Bigrams Formed	1,81,984
Number of Positive Bigrams Classified	70,993
Number of Negative Bigrams Classified	1,10,991
Number of Topics	2

VII. CONCLUSION AND FUTURE WORK

In this work the approaches to sentiment classification favour supervised learning. JST model targets sentiment and topic detection simultaneously in a weakly supervised fashion. The usage of bigrams for classification instead of unigrams improves the efficiency of sentiment classification. The experiments conducted on data sets across different domains reveals that the model behaves differently when sentiment prior knowledge is incorporated. For general domain sentiment classification, incorporation of a small amount of domain independent prior knowledge in the JST model achieved better and comparable performance compared to existing semi–supervised approaches despite using no labelled documents. The topics and topic sentiments detected by JST are indeed coherent and informative. In future, incremental learning technique is implemented for the JST parameters on facing the new data other than the corpus bigrams.

REFERENCES

[1] 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

[2] 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

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

[4] 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

[5] 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

[6] J. Blitzer, M. Dredze and F. Pereira, "Biographies, Bollywood, Boomboxes 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] Titov and R. McDonald, "Modeling Online reviews with Multi-Grain Topic Models", Proc. 17th Int'l Conf. World Wide Web, pp. 111-120, 2008

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

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

[11] 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

[12] 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.

[13] 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

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

[15] 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

[16] 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

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

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