IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 4, Ver. I (July – Aug. 2015), PP 103-108
www.iosrjournals.org

Domain Independent Joint Sentiment And Topic Detection Using AFINN And Other Lexicons

Supriya Paul[1,] Sachin deshmukh[2]
[1]*Department of Computer Science and IT Dr Babasaheb Ambedkar Marathwada UniversityAurangabad, India*
[2]*Department of Computer Science and IT Dr Babasaheb Ambedkar Marathwada UniversityAurangabad, India*

Abstract: *Sentiment analysis and opinion mining field concentrates on automatically classifying sentiments of documents. But detecting mere sentiments does not provide one the sufficient knowledge concealed in the text. In this paper, we have evaluated joint sentiment and topic detection model (JST) to detect sentiment and topic simultaneously from text. JST is evaluated for movie reviews and product reviews with the help of domain independent prior information. AS no labeled data is required for training JST, it becomes highly portable to any domain. JST provides user more information regarding the text than mere sentiments.*

Keywords: *sentiment analysis; topic detection; Latent Dirichlet Allocation (LDA); opinion mining; Joint sentiment and Topic (JST)*

I. Introduction

Text analysis is the study of techniques that extracts information from text data. Explosion of data around the world demands newer and efficient techniques to procure useful information from it. Mostly user generated data is in the form of reviews on various websites or blogs or posts or comments. Contents of such posts vary in terms of information. In studies it is found that online reviews have bigger influence on customer and companies compared to traditional media [17]. These posts can be processed to gain hidden knowledge.

One of the text analysis research area is sentiment analysis i.e. to detect whether the given text orientation is positive, negative or neutral. Many researches done for sentiment analysis mainly consist of supervised learning models trained on labeled corpora [2], [3], [4], [5], [6], [7]. The chief shortcomings of such methods are such labeled corpora is not that easy to obtain and model trained on one domain does not produce agreeable results with other domains. Moreover apart from variety of domains and big size of data to be processed, user-generated data like online reviews is rapidly changing over time. Thus it is necessary to search more and more efficient and flexible methods. This observation inspires domain independent sentiment classification.

However mere sentiment classification is not sufficient for gleaning insights into such user generated data. Some of the researches have suggested that a review can be represented as a mixture of topics [1]. Users are interested in sentiment orientation of topic along with overall sentiment of document. Sentiment along with topic information provides more knowledge to user.

In this paper we are implementing an innovative joint sentiment and topic detection model (JST) which is based on latent dirichlet allocation (LDA) algorithm which is used for topic detection. In JST, LDA is extended by adding additional sentiment layer to it. JST assumes that any word drawn from text belongs to a sentiment and topic pair. Experiment has been conducted with movie review dataset along with multi-domain sentiment data set using AFINN and other lexicons set as prior information.

Paper consists of section 2 of related work. In section 3, methodology is explained and section 4 consists of experimental setup. Section 5 has result analysis followed by conclusion.

II. Related Work

There are few attempts in sentiment and topic detection field. Topic Sentiment Mixture (TSM) model [10] considers sentiment and topic as two different language models. Any selected word is assumed to be coming from either sentiment or topic model. TSM is essentially based on the probabilistic latent semantic indexing (pLSI) [7] model with an extra background component and two additional sentiment subtopics. Thus word sampled from background component model is either conditioned on sentiment or topic unlike JST where word drawn is from joint distribution which is conditioned on both sentiment and topic. Also for detecting sentiment of document TSM need to perform post-processing while JST discover sentiment of document along with topic.

Multi-Grain Latent Dirichlet Allocation model (MG-LDA) [11] allows terms being generated from either a global topic or a local topic. AS MG-LDA is purely topic based, the authors Titov and McDonald proposed Multi-Aspect Sentiment (MAS) model which can aggregate sentiment text for the sentiment summary of each rating aspect extracted from MG-LDA [12]. Basic difference between MAS and JST is that MAS work

DOI: 10.9790/0661-1741103108

with supervised approach as it requires predefined aspects while JST does not need it. MAS model was designed for sentiment text extraction or aggregation while JST is best for sentiment classification. In Leveraging Sentiment Analysis for Topic

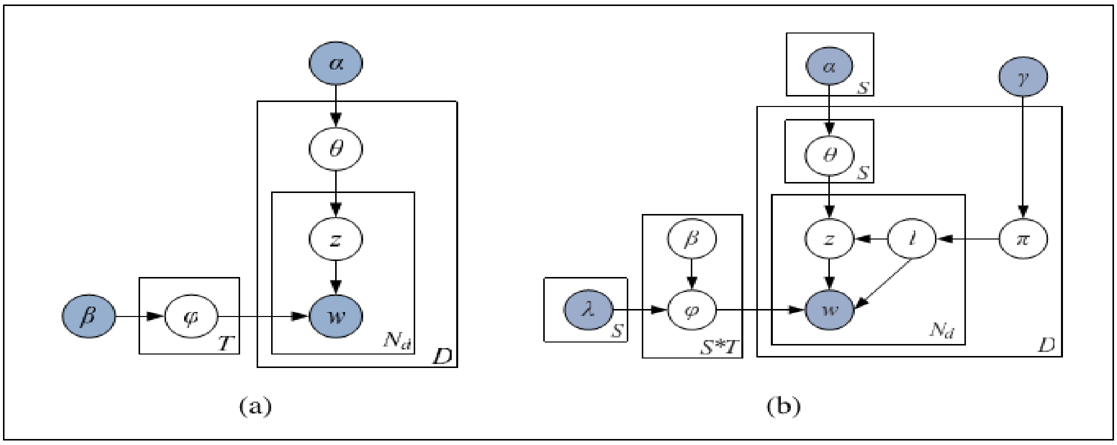

Fig 1: LDA and JST

Detection (STD), the sentiment classification component computes the sentiment polarity of each snippet and creates sentiment taxonomy. Based on the result of this component, the topic detection component further identifies the most significant information related to each sentiment category. Again in this model sentiments are detected first then sentiment topics are searched [19].

Aspect and Sentiment Unification Model (ASUM) unifies aspects and sentiment and discovers pairs of {aspect, sentiment}, which they call senti-aspects. ASUM capture important aspects that are closely coupled with a sentiment. ASUM models sentence-LDA to detect aspects, they assume all words in a single sentence are drawn from single aspect, while JST models LDA for whole document [20].

III. Methodology

Fig 1a shows Latent Dirichlet Allocation model. LDA considers each document as bag of words. For generating a word in document one first choose distribution over topic then one can choose a topic and draw a word from the topic distribution. LDA is three layer hierarchical architecture, where topics are associated with document and words are associated with topic [9]. In JST, LDA is modified by adding sentiment layer between document and topic as shown in graphical model in fig 1b. Thus JST is four layer model, in which sentiment labels are associated with documents, topics are associated with sentiments and words are associated with both sentiment labels and topics [1].

JST considers a corpus with D documents denoted as C ={d_1,d_2,d_3,….,d_d}. Each document consists of N_d words given as d={w_1,w_2,….,w_{Nd})} and each word in document belongs to a vocabulary index with V distinct terms denoted by {1, 2… V}. Also, consider S is number of sentiment labels, T is total number of topics. The graphical model of JST approach as shown in figure 1b can be defined as follows:

- For every l (sentiment label) ϵ {1…., S}
  - For every topic j ϵ {1…., T}, draw φ_{lj}~ Dir (λ_l X βTlj).
- For every document d, choose a distribution π_d ~ Dir (γ).
- For every l ϵ {1….S} under document d, choose a distribution $\theta_{d,l}$~ Dir (α).
- For every word w_i in document d
  - choose l_i ~ Mult (π_d),
  - choose z_i~ Mult ($\theta_{d,l}$),
  - choose a word w_i from φ_{lizi} which is a multinomial distribution over words conditioned on both sentiment label li and topic z_i.

The hyper parameters α and β in JST is the number of times topic j associated with sentiment label l is sampled from a document and the number of times words sampled from topic j are associated with sentiment label l, respectively. The hyper parameter γ is number of times sentiment label l sampled from a document before any word from the corpus is observed. β and γ are symmetric priors whereas α is asymmetric prior. π is per-document sentiment distribution, θ is per-document sentiment label specific topic distribution, and φ is per corpus joint sentiment-topic word distribution.

A. Assimilating model priors

Some modifications to Phan's Gibbs LDA++ package have been done for implementing JST. Here we are using a matrix λ of size S X V, which is considered as transformation matrix which modifies the Dirichlet prior β of size S X T X V, so that word prior sentiment polarity can be captured. To incorporate prior knowledge into JST model, first λ is initialized by assigning 1 to each element. Then for each term w ∈{1,…,V} in corpus vocabulary and for each sentiment label $l \in \{1, \dots, S\}$, if w is found in list of sentiment lexicon then element λ_{lw} is updated as follows:

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

Where the function S (w) returns prior sentiment label of w in a sentiment lexicon, i.e., neutral, positive, or negative. After this for each topic $j \in \{1, \dots, T\}$, multiplying λ_{li} with β_{lji} only the value of β with corresponding sentiment label will be retained and for other sentiments β will be 0.

B. Model Inference

We need to gain distributions of π, θ, and φ, for that we need to assign word tokens to topics and sentiment labels. The sampling distribution for a word given the remaining topics and sentiment labels is $P(z_t = j, l_t = k | \boldsymbol{w}, \boldsymbol{z}^{-t}, \boldsymbol{l}^{-t}, \boldsymbol{\alpha}.\beta, \gamma)$where $\boldsymbol{z}^{-t}$and $\boldsymbol{l}^{-t}$ are vectors of assignments of topics and sentiment labels for all the words in the collection except for the word at position in document d. The joint probability of the words, topics, and sentiment label assignments can be factored into the following three terms:

P(w,z,l)=P(w|z,l) P(z,l)= P(w|z,l) P(z|l) P(l) (2)

For the first term, by integrating out φ, we obtain

$$P(w|z,l) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^{S\times T} \prod_k \prod_j \frac{\prod_i \Gamma(N_{k,j,i} + \beta)}{\Gamma(N_{kj} + V\beta)}, \qquad (3)$$

Where $N_{k,j,i}$ is the number of times word i appeared in topic j and with sentiment label k, $N_{k,j}$ is the number of times words are assigned to topic j and sentiment label k, and Γ is the gamma function.
For the second term probability of a topic for given sentiment label, by integrating out θ, we obtain

$$P(z|l) = \left(\frac{\Gamma(\sum_{j=1}^{T} \alpha_{k,j})}{\prod_{j=1}^{T} \Gamma(\alpha_{k,j})}\right)^{D\times S} \prod_d \prod_k \frac{\prod_j \Gamma(N_{d,k,j} + \alpha_{k,j})}{\Gamma(N_{d,k} + \sum_j \alpha_{k,j})}, \qquad (4)$$

Where D is the total number of documents in the collection, $N_{d,k,j}$is the number of times a word from document d being associated with topic j and sentiment label k, and $N_{d,k}$ is the number of times sentiment label k being assigned to some word tokens in document d.
For the third term probability, by integrating out π, we obtain

$$P(l) = \left(\frac{\Gamma(S\gamma)}{\Gamma(\gamma)^S}\right)^D \prod_d \frac{\prod_k \Gamma(N_{d,k} + \gamma)}{\Gamma(N_d + S\gamma)} \qquad (5)$$

Where N_d is the total number of words in document d.

We are using Gibbs sampling to estimate the posterior distribution by sampling the variables of interest, $\boldsymbol{z}^{-t}$and $\boldsymbol{l}^{-t}$ here, from the distribution over the variables given the current values of all other variables and data. Letting the superscript -t denote a quantity that excludes data from t^th^ position, the conditional posterior for $\boldsymbol{z}^{-t}$and $\boldsymbol{l}^{-t}$ by marginalizing out the random variables φ, θ and π is

$$P(z_t = j, l_t = k | w, z^{-t}, l^{-t}, \alpha.\beta, \gamma) \alpha \frac{N_{k,j,w_t}^{-t} + \beta}{N_{k,j}^{-t} + V\beta} \cdot \frac{N_{d,k,j}^{-t} + \alpha_{k,j}}{N_{d,k}^{-t} + \sum_j \alpha_{k,j}} \cdot \frac{N_{d,k}^{-t} + \gamma}{N_d^{-t} + S\gamma}$$

(6)

Samples obtained from the Markov chain will then use to approximate the per-corpus sentiment-topic word distribution for given sentiment k, topic j and word i as follows,

$$\varphi_{k,j,i} = \frac{N_{k,j,i} + \beta}{N_{k,j} + V\beta}. \qquad (7)$$

The approximate per-document sentiment label specific topic distribution for document d, sentiment label k and topic j is

$$\theta_{d,k,j} = \frac{N_{d,k,j} + \alpha_{k,j}}{N_{d,k} + \sum_j \alpha_{k,j}}. \qquad (8)$$

Finally, the approximate per-document sentiment distribution is

$$\pi_{d,k} = \frac{N_{d,k} + \gamma}{N_d + S\gamma}. \quad (9)$$

The pseudo code for the Gibbs sampling procedure of JST is shown in Algorithm below.

Algorithm: Procedure of Gibbs sampling for JST model.

Input: corpus, α, β, γ

Output: sentiment and topic label assignment for all word tokens in the corpus.

```
Initialize S X T X V matrix Φ, D X S X T matrix Θ, D X S matrix π.
for i = 1 to maximum Gibbs sampling iterations do
    for all documents d = [1, D] do
        for all terms t = [1, $N_D$] do
            Exclude term t associated with topic label z and sentiment label l from variables Nd , Nd,k , Nd,k,j  Nk,j and Nk,j,I; Sample a new sentiment-topic pair l  and z using above (6);
            Update variables Nd , Nd,k , Nd,k,j, Nk,j and Nk,j,I using the new sentiment label $\tilde{l}$  and topic label  $\tilde{z}$;
        end for
    end for
    for every 25 iterations do
        Using Maximum Likelihood Estimation update hyper parameter α;
    end for
    for every 100 iterations do
        Update matrices θ, Φ, and π with new Sampling results;
    end for
end for
```

IV. Experimental Setup

A. Dataset

1) Movie review (MR) data set[1]

The MR data set contains 1,000 positive and 1,000 negative movie reviews with average of 30 sentences each document

2) Multi-domain sentiment (MDS) data set[2]

MDS data set is crawled from Amazon.com which includes reviews of four different products book, DVD, electronics and kitchen appliances. Each file contains a pseudo XML scheme for encoding the reviews. Most of the fields are self-explanatory. The reviews have a unique ID field that isn't very unique.

Both data sets are first preprocessed in which punctuation, non-alphabet characters, numbers and stop words are removed ange the default, adjust the template as follows.

B. Assimilating prior knowledge

Two subjectivity lexicons are used to label words sentiment prior providing the documents to JST. Both datasets are freely available and domain independent. First is sentiment datasets curated by Bing Liu and Minquing Hu of the University of Illinois at Chicago[3] and second is AFINN[4] word list of affective lexicon. We also added few words of our own to

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data

[2] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index 2.html

[3] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets

[4] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

Table I Sentiment Classification Results For Different Domains

	Movie Review	**Multi domain sentiment dataset (MDS)**			
		Book	**DVD**	**Electronics**	**Kitchen**
Accuracy	61.07	64	65	65.5	58.5

Table II Topic Detection From Movie Review And MDS Dataset

	MR	Book	DVD	Electronics	Kitchen
Positive sentiment	know	will	much	product	first
	jackie	book	comedy	used	set
	screen	can	can	recommend	oven
	movie	gets	movie	one	like
	chan	really	series	sound	price
	plays	well	best	wireless	cup
	election	certainly	film	bag	product
	island	food	best	great	coffee
	like	reading	disk	quality	well
	film	author	great	highly	knives
Negative sentiment	stalked	story	james	discs	well
	teen	people	time	back	first
	like	hard	movie	machine	water
	though	crime	actually	gets	blade
	film	book	da	sandisk	good
	story	john	heads	us	get
	stalker	diet	movie	power	money
	don	really	dvd	just	ice
	make	t	one	mb	time
	s	well	vinci	panasonic	used

cover more sentiment labels.

C. Hyper parameter setting

There are three hyper parameter required in implementation of JST. We have set β =0.01, and γ =(0.05 X L)/S. α is initialized as(0.05 X L)/(T X S) and later learned from data by using maximum likelihood estimation [18] and updated every 50 iterations during Gibbs sampling procedure. For both γ andα, L denotes average document length, S indicate number of sentiment labels and T denotes number of topics.

D. Document sentiment classification

The document sentiment can be classified as the probability of a sentiment label given a document P(l|d). This experiment will only considers the probability of positive and negative labels for a given document, while the neutral label probability is ignored. A document d will be classified as a positive if the probability of a positive sentiment label P(lpos|d) is greater than its probability of negative sentiment label P(lneg|d), and vice versa.

V. Result Analysis

We need to analyze results in two sections, document level sentiment classification and topic detection. We have conducted experiments for 1, 3, 5, 8, 10 topics for both movie review and MDS dataset.

The sentiment classification results of JST at document level with prior information extracted from AFINN[3] word list and sentiment dataset[1] curated by Bing Liu and Minquing Hu. As we can see from table accuracy of movie review is slightly more than any other dataset and DVD domain have less accuracy. Basically this difference is due to size of dataset. Movie reviews are larger than any MDS dataset, while DVD contains small size of reviews. It indicates bigger the size of data, more improved results in case of JST. Results are quite different from the results of JST using MPQA and appraisal lexicons [1].Our accuracy is quite below of their accuracy. This underlines the importance of better subjectivity lexicons. We can say that results of JST may not be dependent on domain of dataset but it certainly depends upon size of review and subjectivity lexicons used for sentiment analysis.

Topic detection: We have applied JST on MR and MDS dataset to extract topics under sentiments and evaluated the results, which is our second goal. As shown below, table contains 10 topics from each type of dataset under positive and negative sentiment labels.

We can observe in above table all the topics are relevant to their respective domains as well as sentiments. For example under positive sentiment one can guess that reviewer has remarked Jackie Chan's movies in positive way in movie reviews. While in case of books, reviewer certainly has praised the author. While under negative sentiments the documents indicates negative sentiments towards stalker or teen based movies. Then in case of topics in other domains like DVD one can guess that da vinci got critical perception.

VI. Conclusions

We have modeled JST using AFINN and Bing Liu and Minquing Hu's sentiment lexicons to learn sentiments and topics from various domains simultaneously. We observe that JST delivers almost same accuracy for various domains. But it is certainly affected by size of the documents or corpus and lexicons used for assimilating prior information. In future we can find more appropriate sentiment lexicons. In addition, JST can be modeled for n gram to improve the accuracy.

Acknowledgment

Authors are grateful to the department of computer science and IT, Dr Babasaheb Ambedkar Marathwada University, Aurangabad for providing infrastructure and guidance. We also thank to the staff of the department for their constant guidance and support.

References

[1] Lin, Chenghua; He, Yulan; Everson, Richard and R¨uger, Stefan (2012). "Weakly-supervised joint sentiment-topic detection from text". IEEE Transactions on Knowledge and Data Engineering, 24(6), pp. 1134–1145.

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

[3] 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.

[4] 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. 271-278, 2004.

[5] C. Whitelaw, N. Garg, and S. Argamon, "Using Appraisal Groups for Sentiment Analysis,"Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM), pp. 625-631, 2005.

[6] A. Kennedy and D. Inkpen, "Sentiment Classification of Movie Reviews Using Contextual Valence Shifters," Computational Intelligence, vol. 22, no. 2, pp. 110-125, 2006.

[7] 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.

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

[9] D.M. Blei, A.Y. Ng and M.I. Jordan," Latent Dirichlet Allocation, J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[10] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs," Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 171-180, 2007.

[11] I. Titov and R. McDonald, "Modeling Online Reviews with MultiGrain Topic Models," Proc. 17th Int'l Conf. World Wide Web, pp. 111-120, 2008.

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

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

[14] T. Minka, "Estimating a Dirichlet Distribution", technical report, MIT, 2003.

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

[16] T. Hofmann, "Probabilistic Latent Semantic Indexing", Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 50-57, 1999.

[17] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", J. Foundations and Trends in Information Retrieval, vol. 2, nos. 1/2, pp. 1-135, 2008.

[18] H. Wallach, D. Mimno, and A. McCallum, "Rethinking LDA: Why Priors Matter, "Proc. Topic Models: Text and Beyond Workshop Neural Information Processing Systems Conf., 2009.

[19] Keke Cai, Scott Spangler, Ying Chen, Li Zhang," Leveraging Sentiment Analysis for Topic Detection", International Conference on Web Intelligence and Intelligent Agent Technology(IEEE/WIC/ACM), pp.265-271, 2008.

[20] Yohan Jo, Alice ho,"Aspect and Sentiment Unification Model for Online Review Analysis", WSDM'11, February 9–12, 2011.

[21] Supriya Paul and Sachin N Deshmukh."Analysis of Techniques of Sentiment and Topic Detection." International Journal of Computer Applications 116(14):1-4, April 2015.