Sentiment-Aspect Analysis through Semi-Supervised Topic Modeling

Yong Heng Chen[1], Wanli Zuo[2], Hao Yue [1] andYaojin Lin[1]

[1]*College of Computer Science, Minnan Normal University, Zhang zhou 363000, China*
[2]*College of Computer Science and Technology, Jilin University, Changchun, China*
Corresponding author: Yongheng Chen email: cyh771@163.com

Abstract

Sentiment analysis based on the aspects of products or services is designed to explore subjective information such as attitudes and opinions in user-generated reviews. Although a great many of approaches have been proposed in detecting aspects and the relevant aspect-specific sentiments, most of them detect the latent aspects with no proper classifying them or classify them employing unsupervised topic modeling without predicting the sentiment towards these aspects. This paper proposes a novel sentiment-aspect analysis probabilistic modeling framework consisting of Seeding words extraction and semi-supervised topic (SST) model based on Sentence-LDA. More specifically, the proposed methodology starts by capturing seeding words from the websites inherent semi-structured information about products or services description. Then, it employs the captured seeding words to instruct the discovery of aspects and relevant sentiment of products or services simultaneously. Experimental results show that significant improvements have been achieved by the proposed method with respect to other state-of-the-art methods.

Keywords: Bayesian Models, Gibbs Sampling, Variational Expectation-Maximization, Topic Modeling

1 Introduction

As e-commerce gets ever-increasing fashionable, customer-generated reviews are becoming an increasingly important portion of e-commerce. Since reviews are representatively written by customers who have previously bought some product or service and expressed emotions of praise and criticism to specific aspects, these user-generated reviews offer a plenty of information that can be greatly beneficial to evaluate the customer's attitudes and opinions. Identifying these attitudes and opinions implied in customer-generated reviews is a core problem in natural language understanding or sentiment analysis. We can understand the significance of such information from the following two viewpoints, consumer and company. From the viewpoint of consumers, potential consumers often seek attitudes and opinions from other users by reading their reviews and use these extracted information for their own purchasing decisions before making purchases. From the viewpoint of company, company can research the consumer's sentiment in user-generated reviews to acquire consumer's feedback towards its products or services; while the regulation can be adjusted with response to the consumer's sentiment.

Driven by the demand of gleaning insights into such great amounts of customer-generated reviews, work on new methodologies for automated sentiment analysis and extracting latent information has bloomed splendidly [1]. This task is quite challenging because it is difficult for a human being to extract statistical sentiment information

automatically from a massive set of online reviews. Thus, in recent years there is surging interest on the research topic of sentiment analysis and opinions mining [2]. For purpose of assess customer favor, sentiments concerning particular aspects (for instance, the Screen aspect of a computer and the transportation of a restaurant) of a commodity or server are more useful than a general sentiment about commodity or server. As a consequence, aspect-specific sentiment mining, which directs at extracting customer sentiments embedded in reviews for specific aspect of product or server, has received considerable prevalent in more recent years [3-4].

Among various considerable works of aspect-specific sentiment mining, one of them is to obtain latent aspects of commodity or server that have been discussed in reviews. Existing relevant research has presented numerous approaches for aspect extraction. A few ways have been proposed based on frequent-string noun and phrases as commodity or server aspects [5-6]. Other ways are to employ supervised sequence labeling [7-9]. Nevertheless, a well-known disadvantage of these methods is that they do not categorize the extracted latent aspects on the basis of semantic information.

Recently, because topic models can simultaneously extract aspects of commodity or server and cluster semantically related aspects, they have received considerable attention. A topic model has been used to detect topic-based content presentations, each documents are modeled as a mixture of probabilistic topics. Probabilistic Latent Semantic Indexing (PLSI) is proposed by Hofmann, which is one of topic model and employs topics represented by latent variables to connect documents and words [10]. Blei *et al.* proposed Latent Dirichlet Allocation (LDA), which develops the generative model to accomplish the ability of concluding generalizing the topic distributions so that research can also employ LDA model to create unseen document [11].

The sentiment extraction based on aspect-specific using topic modeling consists mainly two parts: detecting aspect and relevant sentiment terms (*i.e.*, positive and negative sentiment polarity words) simultaneously [12-13, 20]; detecting respectively aspect and relevant sentiment terms by two step [14]. Although these topic models have sufficiently capacity to detect the aspect distribution for a review, as an unsupervised model, the extracted aspects may be incomprehensible to the end costumer. Chan *et al.* explain that the main cause is that the objective functions of unsupervised methods may be at odds with human judgments about products [15]. However words of a document are limited to be produced merely by a sub collection of the topics relying the document's labels, focused and labeled topics, some alterations of Latent Dirichlet Allocation incorporating supervised label are been put forward in existing literature [16-18]. The supervised Latent Dirichlet Allocation (sLDA) model emphasizes the prediction issue through deducing the most predictive potential topics of labeled document [16]. The Dirichlet-Multinomial regression (DMR) topic model is put forward by Mimno *et al.*, which includes a log-linear prior on the document-topic distributions, where the prior is a function of the observed document features [17]. In order to restrict the labeled-document to adopt merely topics related with the subset of the document's labels, two novel partially supervised productive models of labeled document, which associates each topic with a label, are proposed by Ramage *et al.* [18]. Nevertheless, majority costumers frequently have a tendency to make reference to many aspects in reviews, a number of reviews have alike, even of the same, aspect labels. These similar aspect labels affect the capability of extracting aspects for these methods. Domain-specific information is considered and been integrated into model to instruct the aspect distributions of words [19]. A. Mukherjee *et al.* [3] proposed relevant aspects from commodity or server reviews through the seeding words constructed by handmade definition.

Motivated by these observations, in this paper, we are centered on sentence-level sentiment classification for domain-specific documents in association with aspect discovery and aspect-based sentiment analysis. Our method is distinct from other state-of-the-art aspect-based sentiment models in that:

(1) Our proposed method automatic extracts the seeding words information provided in commerce websites, and adopts the constructed seeding words to find aspects.

(2) Our semi-supervised topic (SST) modeling can achieve the discovery of aspect and aspect-based sentiment simultaneously, and incorporate sentiment polarity labels to realize sentiment classification.

(3) For improving the performance of aspect detection, our method adopts firstly the constructed seeding words to extract and recombine the sentences. The goal is to concentrate the distribution of aspect and enrich co-occurrence context. Secondly, our method realizes SST model at sentence level. It is considered because there are not more aspects embed in a single sentence.

The proposed method is appraised employing a Net book reviews. Empirical results demonstrate significant improvements accuracy of aspect extraction of our proposed model and dominance over other state-of-the-art baselines by large margins.

2. Proposed Approach

2.1. Seeding Words and Regrouping

Common problem with the current unsupervised models is the shortage of monitoring at the aspect-word level, which will result in a many of non-relevant noise words with specific aspect. In addition, some aspects extracted by these unsupervised models cannot be easily interpretability to costumers. In practical application, in specific commodity or server reviews some words is belonging distinctly to some aspects, for instance, "Core" belonging to "CPU" of computer. So these domain-specific words may be utilized to supervise the detection or generation of some aspects.

Moreover, every commodity or server review may imply a lot of latent aspects. As a consequence, traditional unsupervised topic models based on co-occurring have not the capabilities to classify the semantic irrelevant aspects. The classification of the extracted aspects tends to be un-conspicuous, *i.e.*, one extracted aspect may be made up of many aspects. For boosting the performance of aspect extraction, the detecting seeding words are used to extract the relevant sentences to make up new document. The goal is to concentrate the distribution of aspect and enrich co-occurrence context. However, the process through artificial intervention will be an extremely dull and time-consuming task. Most commerce websites offers the introduction information about commodity or server. The introduction information involves aspect and related expressions. In place of artificial intervention, we can employ this semi-structured information to detect seeding words collection, and further recombine semantic relevant sentences.

Given the aspect category, we extract helpful seeding words from related expressions, and recombine sentences in reviews by Algorithm 1.

Algorithm 1. Seeding words detection and sentences recombining

Input: description information I:{$i_1,i_2...,i_k$}, vocabulary V:{$v_1,v_2,...,v_N$}, reviews R:{$r_1, r_2, ..., r_m$}
Output: set of seeding words {{c_1},{c_2},...,{c_k}}, Recombined reviews R:{$r_1,r_2,...,r_{m+k}$}
//Seeding words extraction
1 calculate $tf_{j,i}$ and $tfi_{j,i}$ of each token w_j in introduction information of aspect i, and save the
Max(tf_i) and Max(tfi_i)
2 for each aspect t
    for each word $w_j \in i_t$
      if Max(tfi_i==1) and $tfi_{j,i}$= Max(tfi_i) or
        Max(tfi_i==1) and $tfi_{j,i}$= Max(tf_i) or
       $tfi_{j,i}$==1 then
         insert w_j in c_t
//Combining sentences
3 for each review r_d
    for each sentence s_d
      set q=0
      for each word $w \in s_d$
        for each collection c_t
          if $w \in c_t$ then
            set q=1 and copy sentence s_d to r_{m+t}
      if q==1 then
        delete s_d from r_d

The set c_i of seeding word is that the words in c_i constantly occurring in aspect i but scarcely ever occurring in other aspects. So we use frequency (tf) and tfidf, expressed as tf and tfi, to achieve seeding words extraction. We create new documents according aspect through detect each sentence in review if it contains seeding words.

2.2. SST Model

Table 1. Notation Used in the Paper

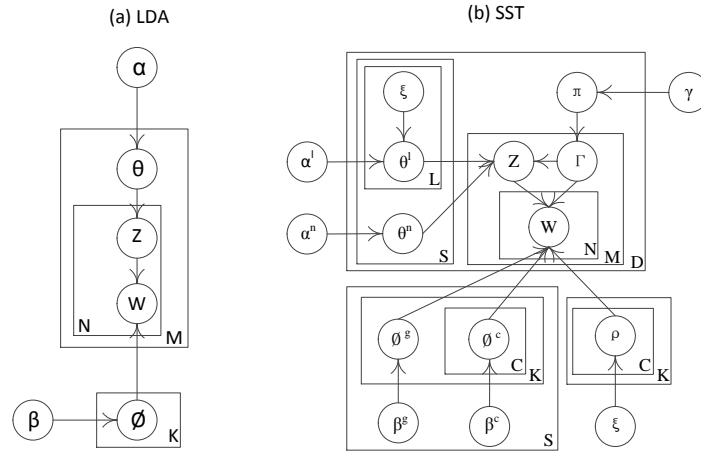| SYMBOL	DESCRIPTION
K	the number of aspects
D	total number of reviews (including aspect-based reviews)
V	total number of unique words
α	K-vector of dirichlet priors for review
β	V-vector of dirichlet priors for aspect
r	s-vector of dirichlet priors for sentiment
m_i	the number of total words in sentence i
m_{iw}	the number of word w in sentence i
z_i	the
s_i	the sentiment of sentence i
z_{-i}	the
s_{-i}	the sentiment of sentence i
s	the number of different sentiment labels
C	seeding words collection
θ	the multinomial distribution of over aspects
Ø	the multinomial distribution of over words
π	the multinomial distribution over sentiments
C_{dj}^{DS}	the number of sentences that are assigned sentiment j in review d
C_{djk}^{DST}	the number of sentences that are assigned sentiment j and aspect k in review d
C_{jkw}^{STW}	the number of word w that are assigned sentiment j and aspect k

Figure 1. SST is shown with One Inspiration Model: (A) LDA Model (B) SST Model

Before presenting the semi-supervised topic (SST) model based on Sentence-LDA, let us review the basic Latent Dirichlet Allocation model. A glossary of notations used in the paper is summarized in Table 1, and the graphical model representations of LDA and SST models are shown in Figure 1. Latent Dirichlet Allocation model, Figure 1(a), a completely unsupervised algorithm, is a hierarchical Bayesian network that product a document making use of a mixture of topics (or aspects). The topic and document distributions are conditionally separation because of the words that can be observed. Moreover, the direct connection between documents and words is interrupted [21]. Rather, by introducing additional potential variable z, which represents the responsibility of a especial topic in employing that word in the document, *i.e.* the topic(s) that the document is centralized on, the connection is managed [22]. The generative process of the topic model specifies a probabilistic sampling procedure that describes how words in documents can be generated based on the hidden topics. To deal with unseen documents the generative model of LDA is generalized. For LDA model the words in a document is exchangeable, this is the same for the documents in a corpus. The generative process of the topic model specifies a probabilistic sampling procedure that describes how words in documents can be generated based on the hidden topics [23]. Since, as an unsupervised model, LDA offers no obvious way of incorporating domain-specific information, we propose Latent Dirichlet Allocation based model: the semi-supervised topic (SST) model based on Sentence-LDA. SST model merges domain-specific information, *i.e.*, the extracted seeding words introduced above, and adds a sentiment layer further for incorporating aspect and aspect –based sentiment simultaneously from reviews. A graphical model of SST is represented in in Figure 1(b).

In our formalization of SST model, assume we have a set of D reviews and each review is denoted as a vector of word token, $w_d=\{w_{d1},w_{d2},\ldots,w_{dNd}\}$, where N_d is the number of words included in review d. The value of D equals m plus k, *i.e.*, the sum of original reviews and new constructed aspect-based reviews. Each word in the review is an item from a vocabulary index with V distinct terms denoted by $\{1, 2,\ldots,V\}$. Everyone of new constructed reviews $d \in \{r_{D-1},r_{D-2},\ldots,r_D\}$ introduced above is attached a aspect label ξ. A review d is related with a multinomial distribution θ over aspect and each aspect is related with a multinomial distribution $\emptyset$ over words. θ and $\emptyset$ have a Dirichlet prior with hyper parameters α and β, respectively. The review-aspect Dirichlet prior of new constructed review d is denoted as: $\alpha^l \times \xi^{ld}$, where $\alpha^l=(\alpha_1,\alpha_2,\ldots,\alpha_k)^T$, and the K-dimension vector $\xi^{ld}=(\xi^{ld}_1,\xi^{ld}_2,\ldots,\xi^{ld}_K)$.

Let M_d and S denote the number of sentences included in review d and different sentiment labels, respectively. Every sentence included in review d is attached a sentiment label l, which can be obtained from the review-sentiment distribution π_d. r is the hyper parameters of π. Let K is the number of distinct aspect. For a seeding word w contained into the collection of seeding words $\{\{c_1\},\{c_2\},...,\{c_k\}\}$, we use $r^w_j=1$ to denote that this word is part of c_j, Let ρ denotes the relevance degree between word and aspect. The relevance degree between word and aspect is calculated through Bernoulli with relaxation factor λ.

Given seeding words sets $C=\{\{c_1\},\{c_2\},...,\{c_k\}\}$ of k aspects, the formal definition of the generative process in SST model corresponding to the graphical model is as follows.

1. For each sentiment label $s\in\{1,...,S\}$
    For each topic $j\in\{1,...,K\}$
      Choose a distribution for general words:
        $\phi^g_{s,j} \sim Dir(\beta^g_s)$
      Choose a distribution for seeding words:
        $\phi^c_{s,j} \sim Dir(\beta^c_s)$
2. For each topic $j\in\{1,...,K\}$
    For each seeding word $w\in\{\{c_1\},\{c_2\},...,\{c_k\}\}$
      If $w\in c_j$
        Choose a relevant weight: $\rho^w_j = 1 + tf_{w,j}\lambda$
      Else
        Choose a relevant weight: $\rho^w_j = 1$
3. For each document $d\in\{1,...,D\}$
    Choose the document's sentiment distribution: $\pi_d \sim Dir(\gamma)$
    For each sentiment label $s\in\{1,...,S\}$
      If $d\in\{r_{D-1},r_{D-2},...,r_D\}$
        Choose an aspect distribution of labeled document d:
          $\theta^l_{ds} \sim Dir(\alpha^l \times \xi^{l_d}_d)$
      Else
        Choose an aspect distribution of unlabeled document d:
          $\theta^n_{ds} \sim Dir(\alpha^n)$
    For each sentence $m\in\{1,...,M_d\}$
      Choose a sentiment: $\Gamma_m \sim Multinomial(\pi_d)$
      If $d\in\{r_{D-1},r_{D-2},...,r_D\}$
        Choose an aspect: $z_m \sim Multinomial(\theta^l_{d\Gamma_m})$
      Else
        Choose an aspect: $z_m \sim Multinomial(\theta^u_{d\Gamma_m})$
      For each word w_i
        If $w_i\notin C$
          Draw word : $w_i \sim Multinomial(\phi^g_{sz_m})$
        If $w_i\in C$ and $r^{w_i}_{z_i} \neq 1$
          Draw word : $w_i \sim Multinomial(\phi^c_{sz_m})$
        Else
          Draw word : $w_i \sim Multinomial(\phi^c_{sz_m}, \rho^{w_i}_{z_m})$

Similar to the original LDA model, exact inference for our model is intractable. A variety of algorithms have been used estimate the parameters of topic models, from basic expectation-maximization[24], to approximate inference expectation propagation[25], and Gibbs sampling [26]. We use Gibbs Sampling here for posterior inference over parameters, since it is easy to extend and it has been proved to be quite effective in avoiding local optima.

At each transition step of the Markov chain, the predictive sentiment and aspect of the ith sentence are chosen according to the conditional probability as follows.

$$p(z_i = k, s_i = j \mid z_{-i}, s_{-i}, w) =$$

$$
\begin{cases}
\bullet \, if \quad d \in \{r_{D-1}, r_{D-2}, \dots, r_D\} \quad w \in C \\[2mm]
\dfrac{C_{dj}^{DS} + \gamma_j}{\sum_{j'=1}^{S} C_{dj'}^{DS} + S\gamma_j} \; \dfrac{C_{jkw}^{STW} + \beta^c}{\sum_{w'=1}^{V} C_{jkw'}^{STW} + V\beta^c} \; \dfrac{C_{djk}^{DST} + \alpha^l \times \xi^{l_d}}{\sum_{k'=1}^{K} C_{djk'}^{DST} + K(\alpha^l \times \xi^{l_d})} \; \rho_k^w \\[4mm]
\bullet \, if \quad d \in \{r_{D-1}, r_{D-2}, \dots, r_D\} \quad w \notin C \\[2mm]
\dfrac{C_{dj}^{DS} + \gamma_j}{\sum_{j'=1}^{S} C_{dj'}^{DS} + S\gamma_j} \; \dfrac{\Gamma(\sum_{w=1}^{V} C_{jkw}^{STW} + \beta^g)}{\Gamma(\sum_{w'=1}^{V} C_{jkw'}^{STW} + V\beta^g + m_i)} \prod_{w=1}^{v} \dfrac{\Gamma(C_{jkw}^{STW} + \beta^g + m_w)}{\Gamma(C_{jkw}^{STW} + \beta^g)} \\[4mm]
\dfrac{C_{djk}^{DST} + \alpha^l \times \xi^{l_d}}{\sum_{k'=1}^{K} C_{djk'}^{DST} + K(\alpha^l \times \xi^{l_d})} \\[4mm]
\bullet \, else \quad d \in \{r_1, r_2, \dots, r_D\} \\[2mm]
\dfrac{C_{dj}^{DS} + \gamma_j}{\sum_{j'=1}^{S} C_{dj'}^{DS} + S\gamma_j} \; \dfrac{\Gamma(\sum_{w=1}^{V} C_{jkw}^{STW} + \beta^g)}{\Gamma(\sum_{w'=1}^{V} C_{jkw'}^{STW} + V\beta^g + m_i)} \prod_{w=1}^{v} \dfrac{\Gamma(C_{jkw}^{STW} + \beta^g + m_w)}{\Gamma(C_{jkw}^{STW} + \beta^g)} \\[4mm]
\dfrac{C_{djk}^{DST} + \alpha^n}{\sum_{k'=1}^{K} C_{djk'}^{DST} + K\alpha^n}
\end{cases}
$$

The notations are given in Table 1, with a minor exceptional use of notation that C_{dj}^{DS}, C_{djk}^{DST} and C_{jkw}^{STW} in this expression exclude the sentence i. A sample obtained from the Markov chain can be used to approximate probability of sentiment j in review d is

$$\pi_{dj} = \frac{C_{dj}^{DS} + \gamma_j}{\sum_{j'=1}^{S} C_{dj'}^{DS} + S\gamma_j}.$$

The approximate probability of aspect k for sentiment j in review d is given by:

$$\theta_k^l = \frac{C_{djk}^{DST} + \alpha^l \times \xi^{l_d}}{\sum_{k'=1}^{K} C_{djk'}^{DST} + K(\alpha^l \times \xi^{l_d})}$$

$$\theta_k^n = \frac{C_{djk}^{DST} + \alpha^n}{\sum_{k'=1}^{K} C_{djk'}^{DST} + K\alpha^n}$$

The approximate probability of word w in sentiment j and aspect k can be derived by:

$$
\phi_k^w =
\begin{cases}
\rho_k^w \; \dfrac{C_{jkw}^{STW} + \beta^c}{\sum_{w'=1}^{V} C_{jkw'}^{STW} + V\beta^c} & w \in C \\[4mm]
\dfrac{C_{jkw}^{STW} + \beta^g}{\sum_{w'=1}^{V} C_{jkw'}^{STW} + V\beta^g} & w \notin C
\end{cases}
$$

3. Experiences

In this section, we present the experimental setup of aspect extraction and aspect-specific sentiment analysis based on the supervision of constructed seeding words to compare the performances of our proposed method with two existing state-of-the-art semi-supervised topic models, SAS [3], JST [20] and ASUM [13].

3.1. Dataset and Settings

In this paper we use a data set of Net book (N.B) reviews scrawled from Amazon.com for evaluation. This data set was preprocessed for deleting extremely common words, non-English alphabets, removing the words that frequency less than six times, reviews containing less than 4 words and sentences that are not shorter than 50 words in the data set. In addition, the data set also was segmented into sentences by ".", "!", "?" and "newline". Each word was stemmed with Poster Stemmer in order to reduce the vocabulary size and address the issue of data sparseness. After preprocess this data set contains 35,287 reviews, 241,794 sentences and total of 4,325,769 words.

Table 2. Data Distribution of N.B Reviews

Corpus	Reviews	Words	Sentences		
N.B	35,287	4,325,769	241,794		
			Sentence length		
			≤ 4	≤ 12	≤ 20
			35.00%	88.05%	98.12%
Positive words	good, nice, excellent, positive, fortunate, correct, superior, amazing, attractive, awesome, best, comfortable, enjoy, fantastic, favorite, fun, glad, great, happy, impressive, love, perfect, recommend, satisfied, thank, worth				
Negative words	bad, nasty, poor, negative, unfortunate, wrong, inferior, annoying, complain, disappointed, hate, junk, mess, not good, not like, not recommend, not worth, problem, regret, sorry, terrible, trouble, unacceptable, upset, waste, worst, worthless				

The sentiment estimation needs more effort than aspect detection, since costumers express their sentiments by more sensitive method while aspect can be merely detected by the co-occurrence of words. One method to improve the accurateness of sentiment estimation can be carried by integrating sentiment lexicon. In this paper, two sentiment lexicons, Paradigm and Paradigm+, are been used. Paradigm is the sentiment words for defining the positive and negative semantic orientation, containing seven positive word and seven negative words. Paradigm+ is the extension of Paradigm lexicon through adding other affective words and general evaluative words. Table 2 shows data distribution and sentiment words adopted in our experiments for corpus.

In SST model, α^l was set as $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, and each α_k in α^l was initialized as the vector $50/(k+\eta)$, η is random number between 0 and 5. The symmetric prior α^n was set as $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, and each α_k in α^n was initialized as the $50/k$. β^c and β^g were set as 0.1 and 0.15, respectively. The relaxation factor λ was set as 0.8.

3.2. Sentiment-Aspect Extraction

Before using SST model to carry aspects and relevant sentiment extraction, we firstly construct seeding words from related expressions of N.B corpus by using Algorithm 1. Examples of the aspects coupled with a sentiment label, positive or negative sentiment

label used in this experiment, are extracted under supervision of constructed seeding words in Table 3. Due to space limitations, we select four aspects coupled with extracted seeding words to show. Since SAS model do not consider sentiment classification, we merely compare our proposed SST with JST and ASUM topic models in sentiment-aspect extraction experiment.

In comparing the results produced by SST, JST and ASUM, a distinct difference is that the aspects generated under supervision of constructed seeding words by SST are seldom mixed together. However, since JST and ASUM are unsupervised sentiment-aspect analysis models and completely dependent on co-occurrence of words, the purity of extracted aspects is not enough. For example, the aspect Battery with positive sentiment label generated by JST and ASUM contain "fast" and "CPU" which belong to the aspect CPUs and with negative sentiment label contain "slow" and "CPU" which also belong to the aspect CPUs. The above analysis demonstrates the effectiveness of seeding words of SST in extracting mixture of aspects from a corpus.

Table 3. Aspects Extracted by SST Under different Sentiment Labels

Aspect {seed words}	Sentiment label	Topic models		
		SST	JST	ASUM
CPUs{ Core, Frequency, L2 cache, cpu, LGA }	Positive	**Fast**, Excellent, Worth, Cache, Hyperthreading, Good, Cpu, LGA, Affordable,Pleasantly	Core, Cpu, Memory, Power, Pleasantly, Good, Big, Cool, Affordable	Cpu, Core, Good, Memory, Battery, Hyperthreading, Big, Power,Pleasantly
	Negative	Hiccups, Frequency, Cpu, Unimpressive, Not good, Low, Slow, Single,Regret	Cpu, Memory, Small, Not good, Latency, Slow, Regret, Low	Core, Small, Slow, Memory, Small, Cpu, Regret, Latency, Slow
Memory{ ram, DDR, extra, Channel, memory }	Positive	DDR, Ram, Good, Nice, Reliable, Memory, Cool ,Smoothly, Memory	DDR, Memory, Core, Fast, Reliable, Capacity, Power, Smoothly, Good	Ram, Memory, DDR, CPU, Smoothly, Good, Excellent, Nice, Reliable
	Negative	**Small**, Latency, Channel, Problems, Trouble, Memory, annoying, worst, bad	Memory, Small, Slow, annoying, Cpu, DDR, bad, Trouble, Core	Ram, Small, Memory, annoying, Cpu, Slow, worst, Core, Latency
Battery{ battery, Capacity, last, Voltage }	Positive	Power, Battery, **Big**, High, Capacity, Voltage, Last, Long, Exceed	Battery, Long, Capacity, ***Fast***, Exceed, CPU, Good, High, Power	Battery, Last, Worth, CPU, Exceed, ***Fast***, Good, Power, High
	Negative	Battery, Short, Capacity, Unfortunately, **Fast**, Disappoint, Model, Waste	Fast, Battery, ***Slow***, Disappoint, Short, Latency, CPU, Last, Low	Battery, Fast, CPU, ***Slow***, Unfortunately, Waste, Not good, Low, Short
System{ Bit Version, Packaging, System, Version, Model, Os}	Positive	System, OEM, Comfortable, worth, Desktop, Stable, **Fast**, **Small**, XP	System, Operating, Stable, Button, XP, Small, Screen, Version, Fast	System, Packaging, Stable, Fast, OEM, setting, Delete, worth, menu
	Negative	System, OEM, Worry, unacceptable, Package, Trouble, OS, Operating, **Big**	System, worry, sensitivity, OS, Unacceptable, Big, OEM, Trouble, stumble	OS, System, Big, OEM, Trouble, view, Version, unacceptable, Worry

Another important task in review analysis is distinguishing the various sentiments label for one word in the context. It is because a word may invoke various sentiments depending on the aspect as much as context. For example, the words "Fast", "Small" and "Big" are probable in both positive sentiment and negative sentiment. In aspects CPUs and System, the word "Fast" is used positively, whereas the word "Fast" is used negative in aspect Battery. In aspect System, the word "Small" is used positively, whereas this

word is used negative in aspect Memory. In aspect Battery, the word "big" is used positively, whereas this word is used negative in aspect System. The above analysis demonstrates the power of our topic model that a word is not limited to one sentiment.

3.3. Predictive Power

The density measurement, expressing the potential configuration of data, is the intention of document modeling. Measuring the model's universal performance on formerly unobserved document is general method to estimate that. Perplexity is a canonical measure of goodness that is used in language modeling to measure the likelihood of a held-out test data to be generated from the potential distributions of the model. Better universal performance is manifested by a lower perplexity and the higher the likelihood on a held-out document.

We divided the D=35,287 reviews into a training set of 35,000 reviews and a test set of 287 reveiws. Formally, given the test set M, the perplexity value can be calculated as follows.

$$Perplexity(M^{est}) = \exp\{-\frac{\sum_{d=1}^{M}(p(w_d \mid D_{train}))}{\sum_{d=1}^{M} N_d}\}$$

In this experiment, we compared SAS, JST, and our presented SST model. The averaged perplexity as a function of the number of Aspects trained on N.B dataset is presented in Figure 2.
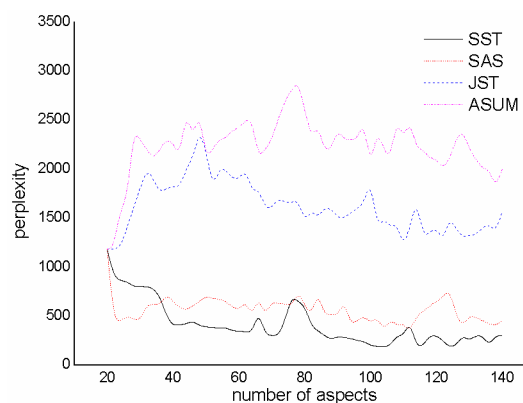

Figure 2. Perplexity of Values versus Number of Aspects

From above figure, we find that SAS presents obvious lower perplexity value than JST and ASUM. This is because that JST and ASUM models based on co-occurring do not have the capability to classify the semantic irrelevant aspects. However, SAS model, viewed as partially supervised clustering model, has the capacity to incorporate seeding words information within corpus into it model during the process of review clustering, this capacity promotes the generation of review's aspect. So the effect of review clustering is obviously improved with respect to unsupervised JST and ASUM models.

However, SST model further develops the generalization performance of review clustering of SAS model through making use of the relevant sentences to make up new documents and employing Sentence-LDA. These improvements concentrate the distribution of aspect and enrich co-occurrence context, and boost model's predictive log-likelihood of SST with respect to SAS model. One can see shown in Figure 2 that SST has a relatively stable and lower perplexity value than SAS model.

3.4. Sentiment Classification

This section offers a different method to quantitatively compare SST model and other models mentioned above. we use the probabilistic sentiment labels, only considering positive and negative, distribution π_{dj} of review to evaluate the quality of sentiment-aspects detected by SST model. If the negative sentiment has equal or a higher probability in π_{dj} distribution, the review d is set to be negative. Otherwise, the review d will be set to be positive. N.B corpus uses 5-stars rating system. In this paper, a review with 1 or 2-stars is considered as negative. A review is considered as positive with 4 or 5-stars. Reviews with 3-stars will not been considered. Similar as the sentiment-aspect extraction experiment, we merely compare our proposed SST with JST and ASUM topic models in sentiment classification experiment.

We predict the sentiment classification for test reviews and estimate the error in terms of the accuracy, L1 error (the difference between predicted and true classification) and average L1 distance over the testing data set to the right, and right classification respectively. Figure 3 shows the predictive results of sentiment classification.
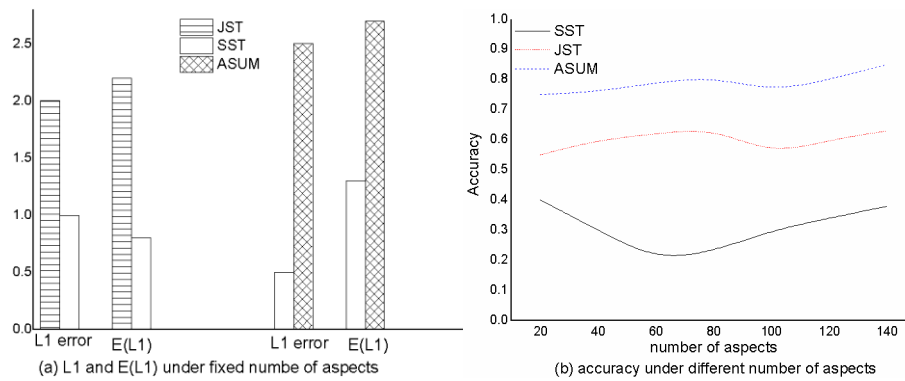

Figure 3. Sentiment Classification Accuracy

Figure 3(a) illustrates the results of sentiment classification predictive under fixed number of aspects for testing set. As shown in it, the accuracy of SST model always performs better than or as well as JST and ASUM models. L1 error of SST decrease 25% error with respect to JST for sentiment classification predictive, and E(L1) of SST 55% with respect to JST. The prediction of sentiment classification shows similar experiments, when we compare L1 and E(L1) of SST with ASUM. Figure 3(b) illustrates the results of sentiment classification in terms of accuracy under different number of aspects for testing set. As shown in it, in all settings, the accuracy of sentiment classification of SST model outperforms the other two unsupervised models.

4. Conclusions and Future Work

In this paper, we propose a sentiment-aspect analysis probabilistic modeling framework which consists of seeding word extraction and semi-supervised topic (SST) model based on Sentence-LDA. This model deals with the problem of discovering aspects and aspect-specific sentiment under the supervision of abstracted seeding word. In place of artificial intervention, this framework firstly adopts semi-structured description information to detect seeding words collection. In order to concentrate the distribution of aspect and enrich co-occurrence context, this framework secondly recombine semantic relevant sentences. Thirdly, this framework employs proposed SST model to realize the detection of aspect and aspect–based sentiment simultaneously from reviews. SST model provides expressively optimized performance in the fields of perplexity, and sentiment classification accuracy compared to other state-of-the-art models.

From an ordinary perspective, our framework can also be applied to other dataset except N.B corpus. However, during practical testing, we do not test SST model on other dataset. In addition, another of the limitations of SST model is that it does not consider dealing with reviews by an on-line fashion. Thus, possible future directions for this work include test our method on other dataset and achieve online fashion to perform detection and analysis of aspects.

Acknowledgment

References

[1] C. Lin, Y. He, Everson R. and Ruger S., "Weakly Supervised Joint Sentiment-Topic Detection from Text", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 6, **(2012)**, pp. 1134–114.

[2] Moghaddam S. and Ester M., "Aspect-based opinion mining from online reviews", Tutorial at SIGIR Conference, **(2012)**.

[3] A. Mukherjee and B. Liu, "Aspect extraction through semi-supervised modeling", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, Association for Computational Linguistics, vol. 1, **(2012)**, pp. 339–348.

[4] S. Moghaddam and M. Ester, "The flda model for aspect-based opinion mining: addressing the cold start problem", Proceedings of the 22nd International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, **(2013)**, pp. 909–918.

[5] Y. Wu, Q. Zhang, X. Huang and L. Wu, "Phrase dependency parsing for opinion mining", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. vol. 3, **(2009)**, pp. 1533–1541.

[6] G. Qiu, B. Liu, J. Bu and C. Chen, "Opinion word expansion and target extraction through double propagation", Computer Linguist, vol. 37, no. 1, **(2011)**, pp. 9–27.

[7] Jakob N. and Gurevych I., "Extracting Opinion Targets in a Single-and Cross-Domain Setting with Conditional Random Fields", Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), **(2010)**.

[8] Yu J., Zha Z. J., Wang M. and Chua T. S., "Aspect ranking: identifying important product aspects from online consumer reviews", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), **(2011)**.

[9] Choi Y. and Cardie C., "Hierarchical sequential learning for extracting opinions and their attributes", Proceedings of Annual Meeting of the Association for Computational (ACL), **(2010)**.

[10] T. H., "Probabilistic latent semantic indexing", Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SI-GIR'99), **(1999)**, pp. 50–57.

[11] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, **(2003)**, pp. 993–1022.

[12] Brody S. and Elhadad S., "An Unsupervised Aspect-Sentiment Model for Online Reviews", Proceedings of the Annual Conference of the North American Chapter of the ACL (NAACL), **(2010)**.

[13] Jo Y. and Oh A., "Aspect and sentiment unification model for online review analysis", ACM Conference in Web Search and Data Mining (WSDM), **(2011)**.

[14] Zhao W., Jiang J., Yan Y. and Li X., "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid", Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), **(2010)**.

[15] Chan J., Boyd-Graber J., Wang C. Gerrish S. and Blei D., "Reading tea leaves: How humans interpret topic models", Neural Information Processing Systems (NIPS), **(2009)**.

[16] D. M. Blei and J. M., "Supervised Topic Models", In NIPS, **(2007)**.

[17] D. M and A. M., "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression", In Uncertainty in Artificial Intelligence, **(2008)**, pp. 411–418.

[18] D. Ramage, C. D. Manning and S. Dumais, "Partially labeled topic models for interpretable text mining", Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **(2011)**, pp. 457–465.

[19] H. Wang, Y. Lu and C. Zhai, "Latent aspect rating analysis on review text data: a rating regression approach", Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **(2010)**, pp. 783–792.

[20] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis", Proceeding of the 18th ACM conference on Information and knowledge management, **(2009)**, pp. 375–384.

[21] X. R. W., "Structured Topic Models: Jointly Modeling Words and Their Accompanying Modalities", University of Massachusetts Amherst. Computer Science, **(2009)**.

[22] X. W, A. M., "Topics over Time: A Non-Markov continuous time model of topical trends", Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **(2006)**, pp. 424–433.

[23] L. A, D. B., "On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking", ICDM Data Mining, **(2008)**.

[24] T. Hofmann, "Probabilistic latent semantic indexing", Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99), **(1999)**, pp. 50-57.

[25] T. Minka and J. Lafferty., "Expectation-propagation for the generative aspect model", Proceedings the 18th Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, **(2002)**.

[26] G. Heinrich, "Parameter estimation for text analysis", Technical report, Fraunhofer IGD and University of Leipzig, Germany, **(2009)**.