Sentiment-Aspect Extraction based on Restricted Boltzmann Machines

Linlin Wang[1], **Kang Liu**[2*], **Zhu Cao**[1], **Jun Zhao**[2] **and Gerard de Melo**[1]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{ll-wang13, cao-z13}@mails.tsinghua.edu.cn,
{kliu, jzhao}@nlpr.ia.ac.cn, gdm@demelo.org

Abstract

Aspect extraction and sentiment analysis of reviews are both important tasks in opinion mining. We propose a novel sentiment and aspect extraction model based on Restricted Boltzmann Machines to jointly address these two tasks in an unsupervised setting. This model reflects the generation process of reviews by introducing a heterogeneous structure into the hidden layer and incorporating informative priors. Experiments show that our model outperforms previous state-of-the-art methods.

1 Introduction

Nowadays, it is commonplace for people to express their opinion about various sorts of entities, e.g., products or services, on the Internet, especially in the course of e-commerce activities. Analyzing online reviews not only helps customers obtain useful product information, but also provide companies with feedback to enhance their products or service quality. Aspect-based opinion mining enables people to consider much more fine-grained analyses of vast quantities of online reviews, perhaps from numerous different merchant sites. Thus, automatic identification of aspects of entities and relevant sentiment polarities in Big Data is a significant and urgent task (Liu, 2012; Pang and Lee, 2008; Popescu and Etzioni, 2005).

Identifying aspect and analyzing sentiment words from reviews has the ultimate goal of discerning people's opinions, attitudes, emotions, etc. towards entities such as products, services, organizations, individuals, events, etc. In this context, *aspect-based opinion mining*, also known as *feature-based opinion mining*, aims at extracting and summarizing particular salient aspects of entities and determining relevant sentiment polarities from reviews (Hu and Liu, 2004). Consider reviews of computers, for example. A given computer's components (e.g., *hard disk*, *screen*) and attributes (e.g., *volume*, *size*) are viewed as aspects to be extracted from the reviews, while sentiment polarity classification consists in judging whether an opinionated review expresses an overall positive or negative opinion.

Regarding aspect identification, previous methods can be divided into three main categories: rule-based, supervised, and topic model-based methods. For instance, association rule-based methods (Hu and Liu, 2004; Liu et al., 1998) tend to focus on extracting product feature words and opinion words but neglect connecting product features at the aspect level. Existing rule-based methods typically are not able to group the extracted aspect terms into categories. Supervised (Jin et al., 2009; Choi and Cardie, 2010) and semi-supervised learning methods (Zagibalov and Carroll, 2008; Mukherjee and Liu, 2012) were introduced to resolve certain aspect identification problems. However, supervised training requires hand-labeled training data and has trouble coping with domain adaptation scenarios.

Hence, unsupervised methods are often adopted to avoid this sort of dependency on labeled data. Latent Dirichlet Allocation, or LDA for short, (Blei et al., 2003) performs well in automatically extracting aspects and grouping corresponding representative words into categories. Thus, a number of LDA-based aspect identification approaches have been proposed in recent years (Brody and Elhadad, 2010; Titov and McDonald, 2008; Zhao et al., 2010). Still, these methods have several important drawbacks. First, inaccurate approximations of the distribution over topics may reduce the computational accuracy. Second, mixture models are unable to exploit the co-occurrence of topics to yield high probability predictions for words that are sharper than the distributions predicted by in-

[*] Corresponding Author: Kang Liu (kliu@nlpr.ia.ac.cn)

dividual topics (Hinton and Salakhutdinov, 2009).

To overcome the weaknesses of existing methods and pursue the promising direction of jointly learning aspect and sentiment, we present the novel Sentiment-Aspect Extraction RBM (SERBM) model to simultaneously extract aspects of entities and relevant sentiment-bearing words. This two-layer structure model is inspired by conventional Restricted Boltzmann machines (RBMs). In previous work, RBMs with shared parameters (RSMs) have achieved great success in capturing distributed semantic representations from text (Hinton and Salakhutdinov, 2009).

Aiming to make the most of their ability to model latent topics while also accounting for the structured nature of aspect opinion mining, we propose replacing the standard hidden layers of RBMs with a novel heterogeneous structure. Three different types of hidden units are used to represent aspects, sentiments, and background words, respectively. This modification better reflects the generative process for reviews, in which review words are generated not only from the aspect distribution but also affected by sentiment information. Furthermore, we blend background knowledge into this model using priors and regularization to help it acquire more accurate feature representations. After m-step Contrastive Divergence for parameter estimation, we can capture the required data distribution and easily compute the posterior distribution over latent aspects and sentiments from reviews. In this way, aspects and sentiments are jointly extracted from reviews, with limited computational effort. This model is hence a promising alternative to more complex LDA-based models presented previously. Overall, our main contributions are as follows:

1. Compared with previous LDA-based methods, our model avoids inaccurate approximations and captures latent aspects and sentiment both adequately and efficiently.

2. Our model exploits RBMs' advantage in properly modeling distributed semantic representations from text, but also introduces heterogeneous structure into the hidden layer to reflect the generative process for online reviews. It also uses a form of regularization to incorporate prior knowledge into the model. Due these modifications, our model is very well-suited for solving aspect-based opinion mining tasks.

3. The optimal weight matrix of this RBM model can exactly reflect individual word features toward aspects and sentiment, which is hard to achieve with LDA-based models due to the mixture model sharing mechanism.

4. Last but not the least, this RBM model is capable of jointly modeling aspect and sentiment information together.

2 Related Work

We summarize prior state-of-the-art models for aspect extraction. In their seminal work, Hu and Liu (2004) propose the idea of applying classical information extraction to distinguish different aspects in online reviews. Methods following their approach exploit frequent noun words and dependency relations to extract product features without supervision (Zhuang et al., 2006; Liu et al., 2005; Somasundaran and Wiebe, 2009). These methods work well when the aspect is strongly associated with a single noun, but obtain less satisfactory results when the aspect emerges from a combination of low frequency items. Additionally, rule-based methods have a common shortcoming in failing to group extracted aspect terms into categories.

Supervised learning methods (Jin et al., 2009; Choi and Cardie, 2010; Jakob and Gurevych, 2010; Kobayashi et al., 2007) such as Hidden Markov Models, one-class SVMs, and Conditional Random Fields have been widely used in aspect information extraction. These supervised approaches for aspect identification are generally based on standard sequence labeling techniques. The downside of supervised learning is its requirement of large amounts of hand-labeled training data to provide enough information for aspect and opinion identification.

Subsequent studies have proposed unsupervised learning methods, especially LDA-based topic modeling, to classify aspects of comments. Specific variants include the Multi-Grain LDA model (Titov and McDonald, 2008) to capture local rateable aspects, the two-step approach to detect aspect-specific opinion words (Brody and Elhadad, 2010), the joint sentiment/topic model (JST) by Lin and He (2009), the topic-sentiment mixture model with domain adaption (Mei et al., 2007), which treats sentiment as different topics, and MaxEnt-LDA (Zhao et al., 2010), which integrates a maximum entropy approach into LDA.

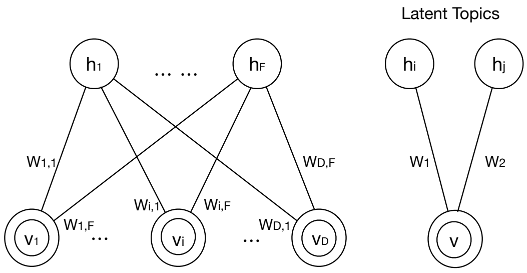

Figure 1: RBM Schema

However, these LDA-based methods can only adopt inaccurate approximations for the posterior distribution over topics rather than exact inference. Additionally, as a mixture model, LDA suffers from the drawbacks mentioned in Section 1 that are common to all mixture models.

3 Model

In order to improve over previous work, we first introduce a basic RBM-based model and then describe our modified full model.

3.1 Basic RBM-based Model

Restricted Boltzmann Machines can be used for topic modeling by relying on the structure shown in Figure 1. As shown on the left side of the figure, this model is a two-layer neural network composed of one visible layer and one hidden layer. The visible layer consists of a softmax over discrete visible units for words in the text, while the hidden layer captures its topics. More precisely, the visible layer is represented as a $K \times D$ matrix **v**, where K is the dictionary size, and D is the document length. Here, if visible unit i in **v** takes the k-th value, we set $v_i^k = 1$. The hidden layer can be expressed as $h \in \{0, 1\}^F$, where F is the number of hidden layer nodes, corresponding to topics. The right side of Figure 1 is another way of viewing the network, with a single multinomial visible unit (Hinton and Salakhutdinov, 2009).

The energy function of the model can be defined as

$$\begin{aligned} E(\mathbf{v}, h) = &-\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ij}^k h_j v_i^k \\ &-\sum_{i=1}^{D}\sum_{k=1}^{K} v_i^k b_i^k - \sum_{j=1}^{F} h_j a_j, \end{aligned} \tag{1}$$

where W_{ij}^k specifies the connection weight from the i-th visible node of value k to the j-th hidden node, b_i^k corresponds to a bias of v_i^k, and a_j corresponds to a bias of h_j.

The probability of the input layer **v** is defined as

$$P(\mathbf{v}) = \frac{1}{Z}\sum_h \exp(-E(\mathbf{v}, h)), \tag{2}$$

where Z is the partition function to normalize the probability.

The conditional probabilities from the hidden to the visible layer and from the visible to the hidden one are given in terms of a softmax and logistic function, respectively, i.e.

$$\begin{aligned} P(\ v_i^k = 1 \mid\ h) &= \frac{\exp\left(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k\right)}{\sum_{q=1}^{K} \exp\left(b_i^q + \sum_{j=1}^{F} h_j W_{ij}^q\right)}, \\ P(\ h_j = 1 \mid\ \mathbf{v}) &= \sigma\left(a_j + \sum_{i=1}^{D}\sum_{k=1}^{K} v_i^k W_{ij}^k\right), \end{aligned} \tag{3}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic function.

3.2 Our Sentiment-Aspect Extraction model

While the basic RBM-based method provides a simple model of latent topics, real online reviews require a more fine-grained model, as they consist of opinion aspects and sentiment information. Therefore, aspect identification is a different task from regular topic modeling and the basic RBM-based model may not perform well in aspect extraction for reviews.

To make the most of the ability of the basic RBM-based model in extracting latent topics, and obtain an effective method that is well-suited to solve aspect identification tasks, we present our novel Sentiment-Aspect Extraction RBM model.

3.2.1 Generative Perspective

From a generative perspective, product reviews can be regarded as follows. Every word in a review text may describe a specific aspect (e.g. “expensive” for the *price* aspect), or an opinion (e.g. “amazing” for a positive sentiment and “terrible” for a negative one), or some irrelevant background information (e.g. “Sunday”). In a generative model, a word may be generated from a latent aspect variable, a sentiment variable, or a background variable. Also, there may exist certain relations between such latent variables.

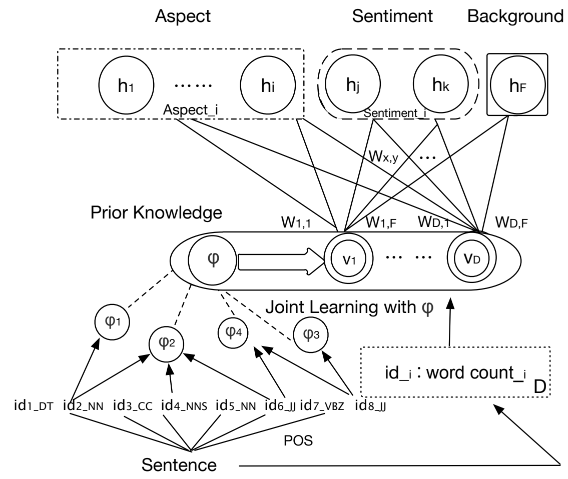

Figure 2: Sentiment-Aspect Extraction Model

3.2.2 Structure

To simulate this generative process for reviews, we adapt the standard RBM structure to reflect the aspect-sentiment identification task.

Undirected Model. Our Sentiment-Aspect Extraction model structure is illustrated in Figure 2.

Compared to standard RBMs, a crucial difference is that hidden units now have a heterogeneous structure instead of being homogeneous as in the standard basic RBM model. In particular, we rely on three types of hidden units, representing aspect, sentiment, and background, respectively. The first two types are self-explanatory, while the background units are intended to reflect the kind of words that do not contribute much to the aspect or sentiment information of review documents. Since the output of the hidden units is a re-encoding of the information in the visible layer, we obtain a deeper representation and a more precise expression of information in the input reviews. Thus, this approach enables the model to learn multi-faceted information with a simple yet expressive structure.

To formalize this, we denote $\widehat{v}^k = \sum_{i=1}^{D} v_i^k$ as the count for the k-th word, where D is the document length. The energy function can then be defined as follows:

$$E(\mathbf{v}, h) = -\sum_{j=1}^{F}\sum_{k=1}^{K} W_j^k h_j \widehat{v}^k - \sum_{k=1}^{K} \widehat{v}^k b^k - \sum_{j=1}^{F} h_j a_j, \tag{4}$$

where W_j^k denotes the weight between the k-th visible unit and the j-th hidden unit.

The conditional probability from visible to hidden unit can be expressed as:

$$P(h_j = 1|\mathbf{v}) = \sigma(a_j + \sum_{k=1}^{K} \widehat{v}^k W_j^k). \tag{5}$$

In an RBM, every hidden unit can be activated or restrained by visible units. Thus, every visible unit has a potential contribution towards the activation of a given hidden unit. The probability of whether a given visible unit affects a specific hidden unit is described as follows (cf. appendix for details):

$$\begin{aligned} P(h_j = 1 \mid \widehat{v}^k) =& P(h_j = 1 \mid h_{-j}, \widehat{v}^k) \\ =& \sigma(a_j + W_j^k \widehat{v}^k). \end{aligned} \tag{6}$$

Under this architecture, this equation can be explained as the conditional probability from visible unit k to hidden unit j (softmax of words to aspect or sentiment). According to Eq. 6, the conditional probability for the k-th word feature towards the j-th aspect or sentiment $p(h_j = 1 \mid v_k)$ is a monotone function of W_j^k , the (k, j)-th entry of the optimal weight matrix. Thus, the optimal weight matrix of this RBM model can directly reflect individual word features toward aspects and sentiment.

Informative Priors. To improve the ability of the model to extract aspects and identify sentiments, we capture priors for words in reviews and incorporate this information into the learning process of our Sentiment-Aspect Extraction model. We regularize our model based on these priors to constrain the aspect modeling and improve its accuracy. Figure 3 provides an example of how such priors can be applied to a sentence, with ϕ_i representing the prior knowledge.

Research has found that most aspect words are nouns (or noun phrases), and sentiment is often expressed with adjectives. This additional information has been utilized in previous work on aspect extraction (Hu and Liu, 2004; Benamara et al., 2007; Pang et al., 2002). Inspired by this, we first rely on Part of Speech (POS) Tagging to identify nouns and adjectives. For all noun words, we first calculate their term frequency (TF) in the review corpus, and then compute their inverse document frequency (IDF) from an external Google n-gram corpus[1]. Finally, we rank their TF$*$IDF

[1] http://books.google.com/ngrams/datasets

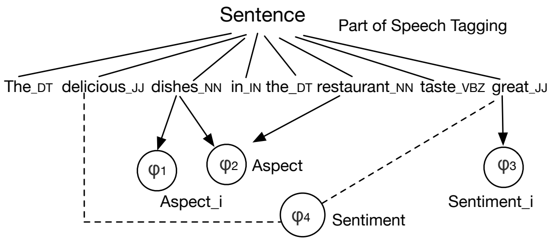
Sentence
Part of Speech Tagging
The_DT delicious_JJ dishes_NN in_IN the_DT restaurant_NN taste_VBZ great_JJ
φ1
φ2 Aspect
φ3
Aspect_i
Sentiment_i
φ4 Sentiment

Figure 3: Prior Feature Extraction

values and assign them an aspect prior probability p_{A,v_k}, indicating their general probability of being an aspect word. This TF-IDF approach is motivated by the following intuitions: the most frequently mentioned candidates in reviews have the highest probability of being an opinion target and false target words are non-domain specific and frequently appear in a general text corpus (Liu et al., 2012; Liu et al., 2013). For all adjective words, if the words are also included in the online sentiment resource SentiWordNet[2], we assign prior probability p_{s,v_k} to suggest that these words are generally recognized as sentiment words.

Apart from these general priors, we obtain a small amount of fine-grained information as another type of prior knowledge. This fine-grained prior knowledge serves to indicate the probability of a known aspect word belonging to a specific aspect, denoted as p_{A_j,v_k} and an identified sentiment word bearing positive or negative sentiment, denoted as p_{S_j,v_k}. For instance, "salad" is always considered as a general word that belongs to the specific aspect *food*, and "great" is generally considered a *positive* sentiment word.

To extract p_{A_j,v_k}, we apply regular LDA on the review dataset. Since the resulting topic clusters are unlabeled, we manually assign top k words from the topics to the target aspects. We thus obtain fine-grained prior probabilities to suggest these words as belonging to specific aspects. To obtain p_{S_j,v_k}, we rely on SentiWordNet and sum up the probabilities of an identified sentiment word being positive or negative sentiment-bearing, respectively. Then we adopt the corresponding percentage value as a fine-grained specific sentiment prior.

It is worthwhile to mention that the priors are not a compulsory component. However, the procedure for obtaining priors is generic and can easily be applied to any given dataset. Furthermore, we only obtain such fine-grained prior knowledge for a small amount of words in review sentences and rely on the capability of model itself to deal with the remaining words.

3.2.3 Objective Function

We now construct an objective function for our SERBM model that includes regularization based on the priors defined above in Section 3.2.2. Suppose that the training set is $S = \mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^{n_s}$, where n_s is the number of training objects. Each element has the form $\mathbf{v}^i = (v_1^i, v_2^i, \dots, v_K^i)^D$, where $i = 1, 2, \dots, n_s$, and these data points are assumed to be independent and identically distributed.

We define the following novel log-likelihood function $\ln L_S$, with four forms of regularization corresponding to the four kinds of priors:

$$\begin{aligned}\ln L_S = \ln \prod_{i=1}^{n_s} P(\mathbf{v}^i) - \sum_{i=1}^{n_s} \Bigg[\\ \lambda_1 \ln \prod_{j=1}^{F_1-1} \prod_{k \in R_1} \left[P(h_j = 1 \mid \widehat{v}^k) - p_{A_j,v_k} \right]^2 \\ + \lambda_2 \ln \prod_{k \in R_2} \left[\sum_{j=1}^{F_1} P(h_j = 1 \mid \widehat{v}^k) - p_{A,v_k} \right]^2 \\ + \lambda_3 \ln \prod_{j=F_2}^{F_2+1} \prod_{k \in R_3} \left[P(h_j = 1 \mid \widehat{v}^k) - p_{S_j,v_k} \right]^2 \\ + \lambda_4 \ln \prod_{k \in R_4} \left[\sum_{j=F_2}^{F_2+1} P(h_j = 1 \mid \widehat{v}^k) - p_{S,v_k} \right]^2 \Bigg]\end{aligned} \tag{7}$$

Here, $P(h_j = 1 \mid \widehat{v}^k)$ stands for the probability of a given input word belonging to a specific hidden unit. We assume all $\lambda_i > 0$ for $i = 1 \dots 4$, while F_1 and F_2 are integers for the offsets within the hidden layer. Units up to index F_1 capture aspects, with the last one reserved for miscellaneous *Other Aspects*, while units from F_2 capture the sentiment (with $F_1 = F_2 + 1 < F$ for convenience).

Our goal will be to maximize the log-likelihood $\ln L_S$ in order to adequately model the data, in accordance with the regularization.

3.2.4 Training

We use Stochastic Gradient Descent (SGD) to find suitable parameters that maximize the objective function. Given a single training instance $\mathbf{v}$ from

[2] http://sentiwordnet.isti.cnr.it

the training set S, we obtain

$$\begin{aligned}\frac{\partial \ln L}{\partial \theta} &= \frac{\partial \ln P(\mathbf{v})}{\partial \theta} \\ &- \lambda_1 \sum_{j=1}^{F_1-1} \sum_{k \in R_1} \frac{\partial \ln \left[P(h_j = 1 \mid \widehat{v}^k) - p_{A_j,v_k}\right]^2}{\partial \theta} \\ &- \lambda_2 \sum_{k \in R_2} \frac{\partial \ln \left[\sum_{j=1}^{F_1} P(h_j = 1 \mid \widehat{v}^k) - p_{A,v_k}\right]^2}{\partial \theta} \\ &- \lambda_3 \sum_{j=F_2}^{F_2+1} \sum_{k \in R_3} \frac{\partial \ln \left[P(h_j = 1 \mid \widehat{v}^k) - p_{S_j,v_k}\right]^2}{\partial \theta} \\ &- \lambda_4 \sum_{k \in R_4} \frac{\partial \ln \left[\sum_{j=F_2}^{F_2+1} P(h_j = 1 \mid \widehat{v}^k) - p_{S,v_k}\right]^2}{\partial \theta}\end{aligned} \tag{8}$$

where $\theta = \{W, a_j, b_i\}$ stands for the parameters. Given N documents $\{\mathbf{v}^n\}_{n=1}^N$, the first term in the log-likelihood function with respect to W is:

$$\frac{1}{N} \sum_{n=1}^{N} \frac{\partial \ln P(\mathbf{v}^n)}{\partial W_j^k} = E_{D_1}[\hat{v}^k h_j] - E_{D_2}[\hat{v}^k h_j]. \tag{9}$$

Here, $D_1[\cdot]$ and $D_2[\cdot]$ represent the expectation with respect to the data distribution and the distribution obtained by this model, respectively. We use Contrastive Divergence (CD) to approximate $E_{D_2}[\hat{v}^k h_j]$ (Hinton and Salakhutdinov, 2009). Due to the m steps of transfer between input and hidden layers in a CD-m run of the algorithm, the two types of hidden units, aspect and sentiment, will jointly affect input reviews together with the connection matrix between the two layers.

Finally, we consider the partial derivative of the entire log-likelihood function with respect to the parameter W. Denoting $\ln \frac{\partial L}{\partial W}$ as ∇W, in each step we update ∇W_j^k by adding

$$\begin{aligned}&\lambda \left[P(h_j = 1 | \mathbf{v}^{(0)}) v_k^{(0)} - P(h_j = 1 | \mathbf{v}^{(\mathrm{cdm})}) v_k^{(\mathrm{cdm})}\right] \\ &- \lambda_1 \sum_{j=1}^{F_1-1} \sum_{k \in R_1} \frac{2 G_j \widehat{v}^k}{(1 + G_j)^2 (\frac{1}{1+G_j} - p_{A_j,v_k})} \\ &- \lambda_2 \sum_{k \in R_2} \frac{2 \widehat{v}^k}{\sum_{j=1}^{F_1} \frac{1}{(1+G_j)} - p_{A,v_k}} \sum_{j=1}^{F_1} \frac{G_j}{(1 + G_j)^2} \\ &- \lambda_3 \sum_{j=F_2}^{F_2+1} \sum_{k \in R_3} \frac{2 G_j \widehat{v}^k}{(1 + G_j)^2 (\frac{1}{1+G_j} - p_{S_j,v_k})} \\ &- \lambda_4 \sum_{k \in R_4} \frac{2 \widehat{v}^k}{\sum_{j=F_2}^{F_2+1} \frac{1}{(1+G_j)} - p_{S,v_k}} \sum_{j=F_2}^{F_2+1} \frac{G_j}{(1 + G_j)^2},\end{aligned}$$

where $G_j = e^{-(a_j + W_j^k \widehat{v}^k)}$ for convenience, and $v^{(\mathrm{cdm})}$ is the result from the CD-m steps.

4 Experiments

We present a series of experiments to evaluate our model's performance on the aspect identification and sentiment classification tasks.

4.1 Data

For this evaluation, we rely on a restaurant review dataset widely adopted by previous work (Ganu et al., 2009; Brody and Elhadad, 2010; Zhao et al., 2010), which contains 1,644,923 tokens and 52,574 documents in total. Documents in this dataset are annotated with one or more labels from a gold standard label set $S = \{$*Food*, *Staff*, *Ambience*, *Price*, *Anecdote*, *Miscellaneous*$\}$. Following the previous studies, we select reviews with less than 50 sentences and remove stop words. The Stanford POS Tagger[3] is used to distinguish noun and adjective words from each other.

We later also rely on the Polarity dataset v2.0[4] to conduct an additional experiment on sentiment classification in order to better assess the model's overall performance. This dataset focuses on movie reviews and consists of 1000 positive review documents and 1000 negative ones. It has also been used in the experiments by Lin & He (2009), among others.

4.2 Aspect Identification

We first apply our novel model to identify aspects from documents in the restaurant review dataset.

4.2.1 Experimental Setup

For the experimental setup, we use ten hidden units in our Sentiment-Aspect Extraction RBM (SERBM), where units 0–6 capture aspects, units 7–8 capture sentiment information, and unit 9 stores background information. In particular, we fix hidden units 0–6 to represent the target aspects *Food*, *Staff*, *Ambience*, *Price*, *Ambience*, *Miscellaneous*, and *Other Aspects*, respectively. Units 7–8 represent *positive* and *negative* sentiment, respectively. The remaining hidden unit is intended to capture irrelevant background information.

Note that the structure of our model needs no modifications for new reviews. There are two cases for datasets from a new domain. If the new

[3] http://nlp.stanford.edu/software/tagger.shtml

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

Method	RBM	RSM	SERBM
PPL	49.73	39.19	21.18

Table 1: Results in terms of perplexity

dataset has a gold standard label set, then we assign one hidden unit to represent each label in the gold standard set. If not, our model only obtains the priors p_{A,v_k} and p_{S,v_k}, and the aspect set can be inferred as in the work of Zhao et al. (2010).

For evaluation, following previous work, the annotated data is fed into our unsupervised model, without any of the corresponding labels. The model is then evaluated in terms of how well its prediction matches the true labels. As for hyperparameter optimization, we use the perplexity scores as defined in Eq. 10 to find the optimal hyperparameters.

As a baseline, we also re-implement standard RBMs and the RSM model (Hinton and Salakhutdinov, 2009) to process this same restaurant review dataset and identify aspects for every document in this dataset under the same experimental conditions. We recall that RSM is a similar undirected graphical model that models topics from raw text.

Last but not the least, we conduct additional comparative experiments, including with LocLDA (Brody and Elhadad, 2010), MaxEnt-LDA (Zhao et al., 2010) and the SAS model (Mukherjee and Liu, 2012) to extract aspects for this restaurant review dataset under the same experimental conditions. In the following, we use the abbreviated name MELDA to stand for the MaxEnt LDA method.

4.2.2 Evaluation

Brody and Elhadad (2010) and Zhao et al. (2010) utilize three aspects to perform a quantitative evaluation and only use sentences with a single label for evaluation to avoid ambiguity. The three major aspects chosen from the gold standard labels are $\mathcal{S} = \{\textit{Food, Staff, Ambience}\}$. The evaluation criterion essentially is to judge how well the prediction matches the true label, resulting in Precision, Recall, and F_1 scores. Besides these, we consider perplexity (PPL) as another evaluation metric to analyze the aspect identification quality. The average test perplexity PPL over words is defined as:

$$\exp\left(-\frac{1}{N}\sum_{n=1}^{N}\frac{1}{D_n}\log P(v_n)\right), \qquad (10)$$

Aspect	Method	Precision	Recall	F_1
food	RBM	0.753	0.680	0.715
	RSM	0.718	0.736	0.727
	LocLDA	**0.898**	0.648	0.753
	MELDA	0.874	0.787	0.828
	SAS	0.867	0.772	0.817
	SERBM	0.891	**0.854**	**0.872**
staff	RBM	0.436	0.567	0.493
	RSM	0.430	0.310	0.360
	LocLDA	0.804	**0.585**	0.677
	MELDA	0.779	0.540	0.638
	SAS	0.774	0.556	0.647
	SERBM	**0.819**	0.582	**0.680**
ambi -ence	RBM	0.489	0.439	0.463
	RSM	0.498	0.441	0.468
	LocLDA	0.603	**0.677**	0.638
	MELDA	0.773	0.588	0.668
	SAS	0.780	0.542	0.640
	SERBM	**0.805**	0.592	**0.682**

Table 2: Aspect identification results in terms of precision, recall, and F_1 scores on the restaurant reviews dataset

where N is the number of documents, D_n represents the word number, and v_n stands for the word-count of document n.

Average perplexity results are reported in Table 1, while Precision, Recall, and F_1 evaluation results for aspect identification are given in Table 2. Some LDA-based methods require manual mappings for evaluation, which causes difficulties in obtaining a fair PPL result, so a few methods are only considered in Table 2.

To illustrate the differences, in Table 3, we list representative words for aspects identified by various models and highlight words without an obvious association or words that are rather unspecific in bold.

4.2.3 Discussion

Considering the results from Table 1 and the RBM, RSM, and SERBM-related results from Table 2, we find that the RSM performs better than the regular RBM model on this aspect identification task. However, the average test perplexity is greatly reduced even further by the SERBM method, resulting in a relative improvement by 45.96% over the RSM model. Thus, despite the elaborate modification, our SERBM inherits RBMs' ability in modeling latent topics, but significantly outperforms other RBM family models

Aspect	RSM	RBM	Loc-LDA	ME-LDA	SAS	SERBM
Food	**great**	menu,drink	chicken	chocolate	food,menu	salad,cheese
	dessert	food,pizza	menu,salad	dessert	dessert	dessert
	beef	chicken	**good**	cream	drinks	chicken
	drink,BBQ	seafood	fish	ice,cake	chicken	sauce
	menu	**good**	drinks	desserts	cheeses	rice,pizza
	delicious	sandwich	wine,sauce	**good**	beers,salad	food
	good	soup	rice	bread	delicious	dish
	fish	flavor	cheese	cheese	rice	sushi,menu
Staff	service	staff	service	service	staff,**slow**	service
	room	**helpful**	staff,waiter	staff,**food**	waitress	staff,friendly
	slow	waiter	attentive	wait,waiters	attentive	waitress
	table	friendly	**busy**	waiter	**helpful**	waitstaff
	quick	**good**,attentive	**slow**,friendly	**place**	service	attentive
	waitress	**slow**,service	**table**	restaurant	**minutes**	waitresses
	friendly	restaurant	wait	waitress	wait,friendly	servers
	waiter	**minutes**	**minutes**	waitstaff	waiter	**minutes**
Ambience	atmosphere	place	**great**	room	place	atmosphere
	music	atmosphere	atmosphere	**dining**	decor	atmosphere
	place	cozy	**wonderful**	tables	**great**	scene
	dinner	**door**	**music**	**bar**	**good**	place
	romantic	**cute**	**seating**	place	romantic	**tables**
	room	**bar**	experience	decor	**tables**	outside
	comfortable	**great**	relaxed	scene	bar	area
	tables	**seating**	**bar**	space	decor	ambiance
	good	experience	room	area	**great**	outdoor
	ambiance	romantic	outside	**table**	**music**	romantic,cozy

Table 3: Aspects and representative words

on the aspect identification task.

In Table 2, we also observe that SERBM achieves a higher accuracy compared with other state-of-the-art aspect identification methods. More specifically, it is evident that our SERBM model outperforms previous methods' F_1 scores. Compared with MELDA, the F_1 scores for the SERBM lead to relative improvements of 5.31%, 6.58%, and 2.10%, respectively, for the *Food*, *Staff*, and *Ambience* aspects. Compared with SAS, the F_1 scores yield relative improvements by 6.73%, 5.1%, and 6.56%, respectively, on those same aspects. As for Precision and Recall, the SERBM also achieves a competitive performance compared with other methods in aspect identification.

Finally, we conclude from Table 3 that the SERBM method has the capability of extracting word with obvious aspect-specific features and makes less errors compared with other models.

4.3 Sentiment Classification

We additionally conduct two experiments to evaluate the model's performance on sentiment classification.

4.3.1 Comparison with SentiWordNet

We assign a sentiment score to every document in the restaurant review dataset based on the output of SERBM's sentiment-type hidden units. To analyze SERBM's performance in sentiment classification, we compare these results with SentiWordNet[5], a well-known sentiment lexicon. For this SentiWordNet baseline, we consult the resource to obtain a sentiment label for every word and aggregate these to judge the sentiment information of an entire review document in terms of the sum of word-specific scores. Table 4 provides a comparison between SERBM and SentiWordNet, with Accuracy as the evaluation metric.

We observe in Table 4 that the sentiment

[5] http://sentiwordnet.isti.cnr.it

Method	SentiWordNet	SERBM
Accuracy	0.703	**0.788**

Table 4: Accuracy for SERBM and SentiWordNet

classification accuracy on the restaurant review dataset sees a relative improvement by 12.1% with SERBM over the SentiWordNet baseline.

4.3.2 Comparison with JST

We additionally utilize the Polarity dataset v2.0 to conduct an additional sentiment classification experiment in order to assess SERBM's performance more thoroughly. We compare SERBM with the advanced joint sentiment/topic model (JST) by Lin & He (2009). For the JST and the Trying-JST methods only, we use the filtered subjectivity lexicon (subjective MR) as prior information, containing 374 positive and 675 negative entries, which is the same experimental setting as in Lin & He (2009). For SERBM, we use the same general setup as before except for the fact that aspect-specific priors are not used here.

Table 5 provides the sentiment classification accuracies on both the overall dataset and on the subsets for each polarity, where pos. and neg. refer to the positive and negative reviews in the dataset, respectively.

Method	overall	pos.	neg.
JST(%)	84.6	96.2	73
Trying-JST(%)	82	89.2	74.8
SERBM(%)	**89.1**	**92.0**	**86.2**

Table 5: Accuracy for SERBM and JST

In Table 5, we observe that SERBM outperforms JST both in terms of the overall accuracy and for the positive/negative-specific subsets. SERBM yields a relative improvement in the overall accuracy by 5.31% over JST and by 8.66% over Trying-JST.

5 Conclusion

In this paper, we have proposed the novel Sentiment-Aspect Extraction RBM (SERBM) model to jointly extract review aspects and sentiment polarities in an unsupervised setting. Our approach modifies the standard RBM model by introducing a heterogeneous structure into the hidden layer and incorporating informative priors into the model. Our experimental results show that this model can outperform LDA-based methods.

Hence, our work opens up the avenue of utilizing RBM-based undirected graphical models to solve aspect extraction and sentiment classification tasks as well as other unsupervised tasks with similar structure.

Appendix

The joint probability distribution is defined as

$$p_\theta(\mathbf{v}, h) = \frac{1}{Z_\theta} e^{E_\theta(\mathbf{v}, h)}, \tag{11}$$

where Z_θ is the partition function. In conjunction with Eq. 1, we obtain

$$E_\theta(\widehat{v}_k, h) = -b_i \widehat{v}^k - \sum_{j=1}^{F} a_j h_j - \sum_{j=1}^{F} h_j W_j^k \widehat{v}^k \tag{12}$$

Then, we can obtain the derivation in Eq. 6.

$$\begin{aligned} & P(h_j = 1 \mid \widehat{v}^k) \\ =& P(h_j = 1 \mid h_{-j}, \widehat{v}^k) \\ =& \frac{P(h_j = 1, h_{-j}, \widehat{v}^k)}{P(h_{-j}, \widehat{v}^k)} \\ =& \frac{P(h_j = 1, h_{-j}, \widehat{v}^k)}{P(h_j = 1, h_{-j}, \widehat{v}^k) + P(h_j = 0, h_{-j}, \widehat{v}^k)} \\ =& \frac{\frac{1}{Z} e^{-E(h_j=1, h_{-j}, \widehat{v}^k)}}{\frac{1}{Z} e^{-E(h_j=1, h_{-j}, \widehat{v}^k)} + \frac{1}{Z} e^{-E(h_j=0, h_{-j}, \widehat{v}^k)}} \\ =& \frac{e^{-E(h_j=1, h_{-j}, \widehat{v}^k)}}{e^{-E(h_j=1, h_{-j}, \widehat{v}^k)} + e^{-E(h_j=0, h_{-j}, \widehat{v}^k)}} \\ =& \frac{1}{1 + e^{-E(h_j=0, h_{-j}, \widehat{v}^k) + E(h_j=1, h_{-j}, \widehat{v}^k)}} \\ =& \sigma(a_j + W_j^k \widehat{v}^k) \end{aligned} \tag{13}$$

Acknowledgments

The research at IIIS was supported by China 973 Program Grants 2011CBA00300, 2011CBA00301 and NSFC Grants 61033001, 61361136003, 61450110088. The research at CASIA was supported by the National Basic Research Program of China Grant No. 2012CB316300 and NSFC Grants 61272332 and 61202329.

References

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of ICWSM 2007*.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of NAACL-HLT 2010*, pages 804–812. Association for Computational Linguistics.

Yejin Choi and Claire Cardie. 2010. Hierarchical sequential learning for extracting opinions and their attributes. In *Proceedings of ACL 2010*, pages 269–274. Association for Computational Linguistics.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *Proceedings of WebDB 2009*, pages 1–6.

Geoffrey Hinton and Ruslan Salakhutdinov. 2009. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems (NIPS 2009)*, pages 1607–1614.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD 2004*, pages 168–177, New York, NY, USA. ACM.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with Conditional Random Fields. In *Proceedings of EMNLP 2010*, pages 1035–1045. Association for Computational Linguistics.

Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. A novel lexicalized HMM-based learning framework for Web opinion mining. In *Proceedings of ICML 2009*, pages 465–472.

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of EMNLP-CoNLL*, pages 1065–1074.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 375–384. ACM.

Bing Liu, Wynne Hsu, and Yiming Ma. 1998. Integrating classification and association rule mining. In *Proceedings of KDD 1998*, pages 80–86. AAAI Press.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the Web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Kang Liu, Liheng Xu, and Jun Zhao. 2012. Opinion target extraction using word-based translation model. In *Proceedings of EMNLP-CoNLL 2012*, pages 1346–1356.

Kang Liu, Liheng Xu, Yang Liu, and Jun Zhao. 2013. Opinion target extraction using partially-supervised word alignment model. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, pages 2134–2140. AAAI Press.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on the World Wide Web (WWW 2007)*, pages 171–180. ACM.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of ACL 2012*, pages 339–348.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*, pages 79–86. Association for Computational Linguistics.

Ana-Maria Popescu and Orena Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP 2005*. Springer.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of ACL-IJCNLP 2009*, pages 226–234. Association for Computational Linguistics.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on the World Wide Web (WWW 2008)*, pages 111–120. ACM.

Taras Zagibalov and John Carroll. 2008. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In *Proceedings of COLING 2008*, pages 1073–1080.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of EMNLP 2010*, pages 56–65. Association for Computational Linguistics.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international Conference on Information and Knowledge Management (CIKM 2006)*, pages 43–50. ACM.