Review Mining for Estimating Users' Ratings and Weights for Product Aspects

Feng Wang * and Li Chen
Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
e-mail: {fwang,lichen}@comp.hkbu.edu.hk

Abstract.
Fine-grained opinions are often buried in user reviews. The opinionated aspects may also be associated with different weights by reviewers to represent the aspects' relative importance. As the opinions and weights provide valuable information about users' preferences for products, they can facilitate the generation of personalised recommendations. However, few studies to date have investigated the three inter-connected tasks in a unified framework: aspect identification, aspect-based rating inference and weight estimation. In this paper, we propose a unified framework for performing the three tasks, which involves 1) identifying the product aspects mentioned in a review, 2) inferring the reviewer's ratings for these aspects from the opinions s/he expressed in a review, and 3) estimating the reviewer's weights for these aspects. The relationship among these three tasks is inherently dependent in that the output of one task adjusts the accuracy of another task. We particularly develop an unsupervised model to ***C**ollectively estimate **A**spect **R**atings and **W**eights* (shorted as CARW), which performs all of the three tasks by enhancing each other mutually. We conduct experiments on three real-life datasets to evaluate the CARW model. Experimental results show that the proposed model can achieve better performance than the related methods regarding each task.

Keywords: Review Mining, Aspect Identification, Aspect-based Rating Inference, Weight Estimation

1. Introduction

With the explosive growth of e-commerce and social media over the past two decades, review writing has become popular. Reviews enable users to express their opinions about products and services, such as hotels, restaurants and digital cameras. The opinions embedded in these reviews provide valuable information for other consumers. Many consumers rely on online reviews to make informed purchase decisions, especially when they know little about the products [7,17]. Indeed, the body of a review often contains the reviewer's detailed opinions about the multi-faceted aspects of a product. For example, a hotel review may convey the reviewer's opinions about *food quality*, *service*, and *ambience*. Therefore, it is meaningful to automatically extract these fine-grained aspect opinions from reviews, which has been referred as *aspect-based opinion mining* [32]. Specifically, the goal of aspect-based opinion mining is to discover the set of aspects mentioned in the reviews of a product and their associated user sentiments.

However, existing approaches to aspect-based opinion mining have some limitations that restrict their use in practice. Some methods require a set of labeled entities to be prepared in advance for identifying the aspects from reviews [14,18,29]. This requirement makes it hard to be applied in different product domains. Moreover, for the task of aspect-based rating inference (i.e., the opinion quantification), many of the related works are based on a sentiment lexicon [5,13,39], which contains a static sentiment score for each word without considering the aspect it is related to. For example, although the word "friendly" can be a strong positive opinion word for the "service" aspect, but not for the "value" aspect in hotel reviews.

Another meaningful task related to the aspect-based opinion mining is to estimate the weights that review-

*Corresponding author. e-mail: fwang@comp.hkbu.edu.hk.

ers place on different aspects of a product from their written reviews. These weights reveal the preferences of the reviewers for aspects [35]. For example, consider the following hotel review: *"The food is delicious. However the ambience and service is not so good."* It can be seen that this review expresses negative opinions about *ambience* and *service* aspects but a positive opinion about the *food* aspect. Given that the reviewer's overall rating for this hotel is 4 (in the range of [1, 5]), it can imply that the *food* aspect is more important than the other aspects to the reviewer.

In this paper, we are interested in investigating the relationship among the three tasks: *aspect identification*, *aspect-based rating inference*, and *aspect-based weight estimation*. To the best of our knowledge, most of related works have just focused on one or two of these tasks [13,23,41]. In our view, these tasks are essentially inter-connected between each other. The accuracy of aspect identification can influence the performance of aspect-based rating inference. Therefore, errors may be accumulated if the tasks are performed separately.

In this paper, we develop a unified framework to improve the three tasks simultaneously. We aim not only to identify the aspects mentioned in product reviews and reviewers' opinions about these aspects at a fine granularity, but to derive reviewers' weights for these aspects (see Figure 1). An example of the expected output is shown in Figure 2. The main challenge is how to minimise error propagation when performing the three tasks. Error propagation occurs when errors caused by an upstream sub-task propagate to and adversely affect the performance of downstream sub-tasks. We address this problem by using shared representations to create dependencies between the tasks and thereby recast them as three components of a joint learning task. This enables knowledge transfer between tasks. Specifically, we propose a unified unsupervised CARW (shorted from *Collectively estimate Aspect Ratings and Weights*) model. The data-sparsity problem is also solved by discovering cluster-level preferences for accommodating reviewers' preference similarity.

The reminder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our problem statement and notations used in this paper. Section 4 presents the details of our proposed model. In Section 5, we show the results of our experimental evaluations, and Section 6 concludes the paper and discusses directions for future work.

2. Related Work

Researchers are paying increasing attention to methods of extracting information from reviews that indicates users' opinions of aspects about products [32]. In this section, we describe the existing literatures on aspect identification, aspect-based rating inference and aspect-based weight estimation.

2.1. Aspect Identification

Most of the earliest attempts to identify aspects are frequency-based [2,13,19,31], for which some constraints are applied to identify the high-frequent nouns or noun phrases as aspect candidates. For example, in [13] and [19], the aspects are extracted by using an association rule miner. In [31], the noun phrases that occur more frequently in general English than in product reviews are discarded. As the main limitation of the frequency-based approaches, the low-frequent aspects are often ignored. To overcome this problem, some methods construct a set of rules to identify aspects, which can be called rule-based approaches [19,30,38]. In [19], a set of predefined Part-of-Speech (POS) patterns are used to extract aspects from reviews. For example, a POS pattern such as 'ADJ NN' is applied to identify the noun word "manual" in the phrase 'good_ADJ manual_NN' as the aspect. The limitation of these methods is that they will produce non-aspects that match with the relation patterns. Furthermore, the frequency- and rule-based approaches require the manual effort of tuning various parameters, which limits their generalization in practice.

To address the problems mentioned above, some model-based approaches that automatically learn the model parameters from the data have been proposed. Some of these models are based on supervised learning techniques, such as the Hidden Markov Model (HMM) and Conditional Random Field (CRF). For example, in [14], a system named *OpinionMiner* is developed to extract aspects and associated opinions based on lexicalized HMM (L-HMMs), which can integrate POS information in the HMM framework, but the model does not consider the interaction between sequence labels. In [29] and [18], they extend the CRF model to extract aspects and corresponding opinions from review texts. Although the supervised model-based methods overcome the limitation of frequency- and rule-based methods, these models require a set of manually labeled entities for training the model.

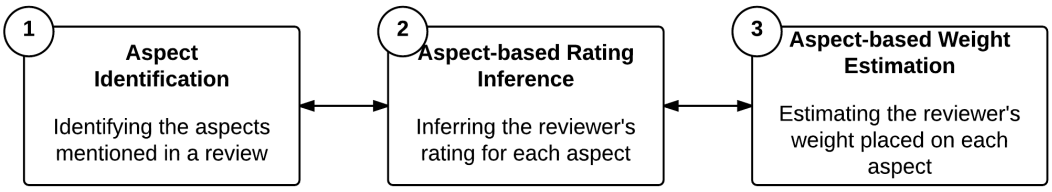

Fig. 1. The inter-connected three tasks related to aspect-based opinion mining.

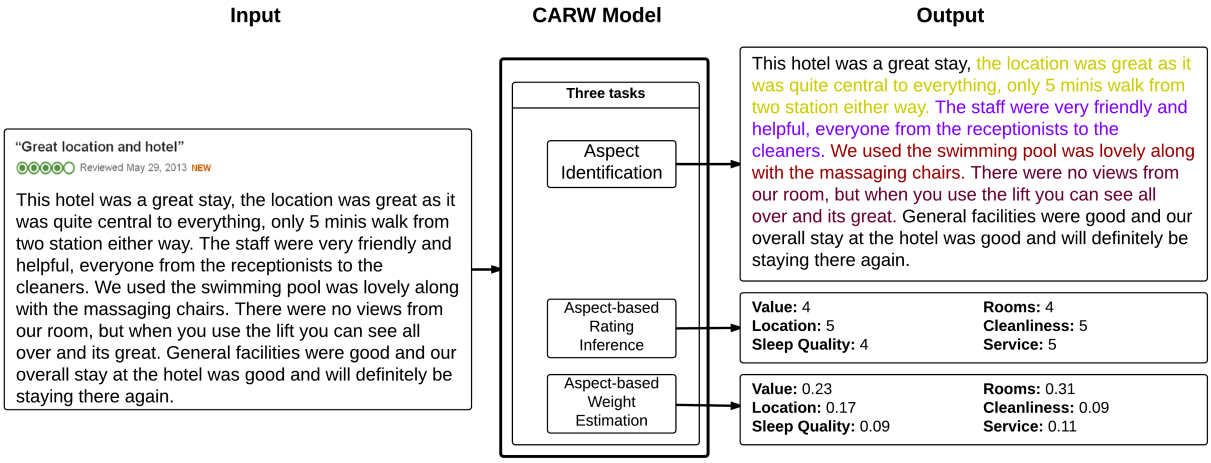

Fig. 2. An example of the input and output of our model in the hotel domain.

The unsupervised model-based methods, on the other hand, are primarily based on statistical topic models, such as the Probabilistic Latent Semantic Analysis (PLSA) [12] and Latent Dirichlet Allocation (LDA) [3] model. For example, in [22,23,33,5, 21,42,27], they adopt topic models to learn *latent* topics that correlate directly with aspects. The basic idea behind this model is that documents (i.e., reviews or review sentences) are represented as a small number of latent topics (here, the topics can be referred as aspects), where topics are associated with a distribution over the words. More concretely, [23] proposes a topic modeling method, called structured PLSA that models the dependency structure of phrases in short comments. In this method, each phrase is represented as a pair of head terms and modifier terms, and the head term is about an aspect (e.g., the head term *picture* and the modifier term *great* in the phrase 'great picture'). The basic idea of this method is that the head terms associated with similar set of modifier terms are more likely to share similar semantic meaning. [33] proposes a topic model, named as MG-LDA, based on LDA for discovering aspects from reviews. In the MG-LDA model, two types of topics including global topics and local topics are separately modeled. The global topics are related to the background description of a product in reviews, and the local topics are related to the aspects of the product. [5] applies the LDA model at sentence-level to identify the local topic of each sentence as the aspect. In [42], a topic model called MaxEnt-LDA is devised that can leverage the POS tags of words to distinguish aspects, opinions, and background words by integrating a discriminative maximum entropy (Max-Ent) with the LDA model. To alleviate the cold-start problem, [27] assumes that each reviewer (and item) has a set of distributions over aspects and aspect-based ratings. In contrast with supervised learning models, the unsupervised methods do not require the labeled training data.

2.2. *Aspect-based Rating Inference*

The existing rating inference methods can be categorised into two groups: lexicon-based and supervised learning approaches. Lexicon-based approaches

use a sentiment lexicon, in which each word is associated with an orientation (*positive* or *negative*) or rating [13,5,39]. The critical issue is how to construct such a sentiment lexicon. Typically, a small-scale set of seed words is first constructed manually. Then some techniques are applied to enlarge this seed set to include more words. For example, [13] enlarges the sentiment lexicon by identifying the synonyms or antonyms of a seed word. [5] propagates the polarity score across a conjunction graph, which is built over adjective words with a set of seed words and their polarities. For the supervised learning approaches [4], because a classifier, which is trained from labeled data in one domain, will perform poorly in another domain, some recent researches leverage the overall rating associated with each review to learn individual classifier or called rating predictor for each aspect [23,40]. For example, [40] proposed a semi-supervised method to train a classifier by treating the overall ratings as sentiment labels.

In addition, some works [25,15,26] have used topic modeling techniques to simultaneously identify aspects and infer the rating for each aspect. In the work of [25], a review is assumed to be generated by sampling words from a set of topic distributions and two sentiment distributions which correspond to positive and negative, respectively. In [15], each review is assumed to have a distribution over sentiments and each sentiment have a distribution over aspects. Then, the words from the review are generated based on the aspect's and the corresponding sentiment's distributions. However, [25,15] purely estimate the polarity of sentiment (i.e., positive and negative) expressed on aspects, which is different from the numerical rating that we aim to infer.

2.3. Aspect-based Weight Estimation

So far only few studies have been conducted to uncover the weights the reviewer places on aspects [1, 41,28,36,37]. In [1], the authors study how the opinions expressed in reviews affect the product demand. In particular, the hedonic regression model, which has been commonly used in econometrics, is adopted to identify the weight of each aspect by using product demand as an objective function. But the derived weights are common to all of the reviewers without considering their individual preferences. [41] uses the Probabilistic Regression Model (PRM) to estimate aspect-based weights. Concretely, the overall rating is assumed to be drawn from a Gaussian distribution with the mean as the product of the aspect-based ratings and the aspect-based weights. For each review, given the inferred aspect ratings, the aspect weights with the most likely posterior probability are inferred with the occurrence frequency as the priori knowledge. In [36], the PRM is also used to estimate the aspect weights. The novelty of this method is that a probabilistic graphic model is introduced to concurrently estimate both the rating and the weight of each aspect. As an extension of [36], [37] introduces a statistical topic model to identify aspects and estimate aspect ratings and weights, which is similar to the objective of our proposed model. However, their model is limited when there are only a few number of reviews posted. Hence, it suffers from the review sparsity phenomenon.

2.4. Limitations of Related Work

We summarise the limitations of the three branches of related works and indicate the novelty of our proposed CARW model in comparison with them in Table 1. Moreover, relative to previous work [35,6], we propose a unified framework to perform the three tasks, aspect identification, aspect-based rating inference, and aspect-based weight estimation, simultaneously so as to reduce the error propagation.

3. Problem Statement

Formally, in this paper, we assume that we have a set of U users, which can be denoted as $\mathcal{U} = \{u_1, \ldots, u_U\}$, and a set of M products (such as hotels or digital cameras), which can be denoted as $\mathcal{M} = \{m_1, \ldots, m_M\}$. Then, we let $\mathcal{R} = \{r_{ij} | u_i \in \mathcal{U} \text{ and } m_j \in \mathcal{M}\}$ be a set of reviews that have been posted for certain products. Typically, when writing a review r_{ij}, the user u_i also assigns an overall rating $y_{ij} \in \mathbb{R}^+$ (say from 1 to 5) to express the overall quality of the reviewed product m_j. We also assume that there are W unique words $\mathcal{W} = \{w_1, \ldots, w_W\}$ occurring in all of the reviews. The major notations used throughout the paper can be found in Table 2.

The research problems that we have been engaged in solving are as follows:

1. **Aspect identification**: The goal of this task is to extract aspects mentioned in a review. An aspect is an attribute or a component of the product, such as a hotel's "service", "location" and "food". We assume that there are A aspects men-

Table 1
The novelty of our proposed CARW model in comparison to the related work

Task	Related work	Core ideas	Limitations	Novelty of CARW
Aspect identification	Frequency based [13,19,31,2]	Identifying the frequently occurring nouns and noun phrases as aspect candidates.	1. Some low-frequency nouns are ignored. 2. Various parameters (like thresholds) need to be manually tuned.	1. Performing the task in an unsupervised manner. 2. Fewer parameters need to be tuned. 3. The synonyms are grouped automatically in the model.
	Rule based [19]	Constructing a set of POS patterns to identify aspects.	Non-aspects matched with the POS patterns are produced.	
	Supervised model based [14,29,18]	Learning a model (e.g., classifier) based on labeled data.	It requires manually labeled data for training models.	
	Topic model based [22,23,33,5,21,42]	Mapping co-occurring words in texts to aspects.	Some auxiliary information is discarded (e.g., the sentiment score of the aspect).	
Aspect-based rating inference	Lexicon based [13,5]	Using a sentiment lexicon to infer the ratings.	1. The sentiment score of a word is the same no matter what the related aspect is. 3. Not all of the sentiment words are included in the lexicon.	1. The sentiment scores of words are learned from the data automatically. 2. Each word can have different sentiment scores related to different aspects. 3. The learned sentiment scores can be numerical ratings (e.g., in the range of $[1, 5]$).
	Supervised model based [23]	Using the overall rating to learn individual classifier/rating predictor for each aspect.	The aspect ratings share the same value as the corresponding overall rating in the training process.	
	Topic model based [25,15]	Considering the positive and negative words as two distinct topics in the topic model.	1. Some useful sentiment-indicating information (e.g., overall rating) is not considerred. 2. Only the binary polarity ratings (e.g., positive and negative) are considered.	
Aspect-based weight estimation	Probabilistic regression model (PRM) based [41,36,37]	Using a linear regression model.	1. It is difficult to learn each reviewer's aspect-level weights when there is review sparsity. 2. It requires that the aspect ratings are available.	1. Does not require that the aspect ratings are available. 2. Reviewers are clustered for alleviating the problem of review sparsity.
	Latent class regression model (LCRM) based [6]	1. Using a linear regression model. 2. Considering the cluster-wise behaviors behind all of the reviewers.	It requires that the aspect ratings are available.	

tioned in reviews, $\mathcal{A} = \{a_1, \ldots, a_A\}$. An aspect can be denoted by $a_i = \{w | w \in \mathcal{W}, A(w) = i\}$, where $A(\cdot)$ is a mapping function from a word to an aspect. For example, for hotel reviews, words such as "price", "value" and "worth" can be mapped to aspect "price".

2. **Aspect-based rating inference**: We use an A-dimensional vector $\mathbf{v}_{ij} \in \mathbb{R}^{A \times 1}$ to represent the aspect-based ratings (e.g., the range of rating can be from 1 to 5). Each element v_{ijk} of $\mathbf{v}_{ij}$ is a score value indicating the reviewer's sentiment toward aspect a_k. The task of *aspect-based rat-*

ing inference is then to estimate the vector $\mathbf{v}_{ij}$ given a review r_{ij} and the associated overall rating y_{ij}.

3. **Aspect-based weight estimation**: This task aims to estimate the non-negative weights $\boldsymbol{\alpha}_i$ (the degree of importance) that the user u_i places on aspects $\mathcal{A}$. The aspect-based weights enable system to generate recommendations tailored to individual user's preferences.

We emphasize the identification of aspects, and the estimation of aspect-based ratings $\mathbf{v}_{ij}$ of review r_{ij}, and the reviewer u_i's weights $\boldsymbol{\alpha}_i$ on aspects, with a unified model. We expect that this model will reduce the error propagation among the three tasks. Moreover, when deriving the aspect weights of each reviewer, we propose to integrate the Latent Class Regression Model (LCRM) into a probabilistic graphic model, so as to address the review sparsity problem. In the next section, we present the details of our proposed model.

4. Our Methodology

In this section, we propose an unsupervised model that can collectively perform the three tasks *aspect identification*, *aspect-based rating inference* and *aspect-based weight estimation* simultaneously, called CARW. Before presenting details of this model, we first list our some assumptions:

- The text describing a particular aspect is generated by sampling words from a topic model (i.e., a multinomial word distribution) corresponding to the aspect. For example, the words "service", "staff" and "waiter" are frequently used to describe the aspect "service" in the hotel reviews.
- The rating for an aspect is determined based on the words describing the corresponding aspect. For example, if the review text says "the staff are very friendly and helpful", we can infer the rating for the aspect "service" as 5 (within the range $[1, 5]$) because the opinion expression "very friendly and helpful" indicates a strong positive sentiment.
- The overall rating is regarded as the weighted combination of aspect ratings where the weight reflects the relative emphasis of each aspect. Following this assumption, the overall rating has a linear relationship with the aspect ratings, and the ratings for different aspects are independent with each other. Although the assumption of independence may not be true in reality, this assumption can help to maintain the model's simplicity [41].
- Each product has a distribution over the aspects representing how often different aspects are discussed in reviews of that product.
- Each product has a rating distribution over aspects that represents how well the product is evaluated on different aspects by reviewers.
- Each reviewer belongs to a cluster so reviewers in the same cluster share similar aspect-based weights.

Based on the above assumptions, to generate a review text, we first sample the aspects expressed in that review conditioned on the aspect distribution of the corresponding product m_j. Following the basic Latent Dirichlet Allocation (LDA) model, this distribution follows a multinomial distribution θ_j with prior Dirichlet distribution $Dir(\gamma)$, denoted as $\theta_j \sim Dir(\gamma)$. The aspect-based ratings expressed in a review are then sampled conditioned on the rating distribution of the corresponding product. For the sake of simplicity, we define the aspect rating distribution of product m_j as a multivariate Gaussian distribution $\mathbf{v}_j \sim \mathcal{N}(\vartheta_j, \eta_j^2 I)$. The aspect-based weights $\boldsymbol{\alpha}_i$ of reviewer u_i are sampled conditioned on the cluster s/he belongs to and the weight distribution associated with that cluster. The aspect weight distribution is also defined by following a multivariate distribution $\boldsymbol{\alpha}_i \sim N(\mu_k, \Sigma_k)$, given that the user u_i belongs to the k-th cluster (denoted as $c_i = k$). The overall rating y_{ij} is sampled based on the aspect-based weights $\boldsymbol{\alpha}_i$ of the reviewer and the aspect-level ratings $\mathbf{v}_{ij}$ that follow a Gaussian distribution, denoted as $y_{ij} \in \mathcal{N}(\boldsymbol{\alpha}_i^T \mathbf{v}_{ij}, \sigma^2)$. We use $z_{ijl} = k$ to indicate that the l-th word in review r_{ij} belongs to the k-th aspect. Finally, the words appearing in a review are sampled based on the mapped aspects and their ratings. Figure 3 shows the graphical model.

4.1. Model Inference and Parameters Learning

Formally, for each review r_{ij} of product m_j given by reviewer u_i, the log-posterior probability of the latent variables (note that the latent variables include 1) aspect ratings vector $\mathbf{v}_{ij}$, 2) the word's topic/aspect identification $\mathbf{z}_{ij}$, and 3) reviewer's cluster membership c_i) is conditioned on the model parameters $\Phi = \{\boldsymbol{\pi}_{1:U}, \boldsymbol{\alpha}_{1:U}, \boldsymbol{\theta}_{1:M}, \boldsymbol{\vartheta}_{1:M}, \boldsymbol{\eta}_{1:M},$

Table 2
Notations used in this paper

Notation	Description
$\mathcal{U} = \{u_1, \ldots, u_U\}$	the set of users (reviewers), and U is the number of users.
$\mathcal{M} = \{m_1, \ldots, m_M\}$	the set of products, and M is the number of products.
$\mathcal{R} = \{r_{ij} \vert u_i \in \mathcal{U} \text{ and } m_j \in \mathcal{M}\}$	the set of user-item pairs, where $r_{ij} \in \mathcal{R}$ indicates that user u_i wrote a review to product m_j, and R denotes the total number of reviews.
$\mathcal{A} = \{a_1, \ldots, a_A\}$	the set of aspects, and A is the number of aspects.
r_{ij}	the review written by user u_i for item m_j.
$y_{ij} \in \mathbb{R}^+$	the overall rating associated with review r_{ij}.
$\mathbf{v}_{ij} \in \mathbb{R}^A$	the aspect ratings inferred from review r_{ij} over A aspects $\{v_{ij1}, \ldots, v_{ijA}\}$.
$\mathbf{w}_{ij}$	the words occurring in review r_{ij}, and w_{ijl} denotes the l-th word in review r_{ij}.
$\mathbf{z}_{ij}$	the aspect assignment of each word in review r_{ij}, and $z_{ijl} = k$ denotes the l-th word that is assigned to k-th aspect.
$\mathcal{W} = \{w_1, \ldots, w_W\}$	the corpus of words, and W is the number of words.
$c_i \in \{1, \cdots, C\}$	the cluster membership of reviewer u_i ($c_i = k$ denotes that reviewer u_i belongs to k-th cluster), and C is the number of clusters.
$\boldsymbol{\alpha}_i \in \mathbb{R}^{A \times 1}$	the aspect weights reviewer u_i places on A aspects.
$\boldsymbol{\pi}_i \in \mathbb{R}^{C \times 1}$	the prior cluster distribution of reviewer u_i.

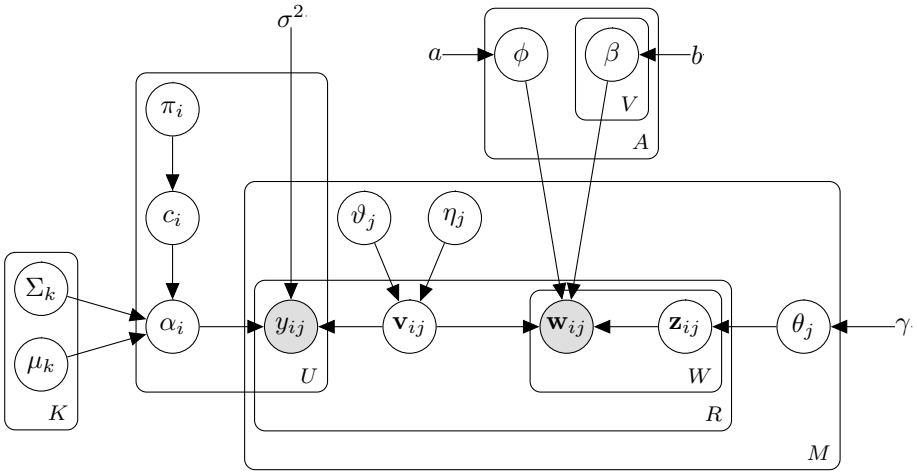

Fig. 3. The graphical plate notation for our CARW model

$\boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}, \phi, \beta\}$ and the hyperparameters $\{\tau, \sigma, \gamma\}$:

$$\begin{aligned}\mathcal{L}(\Phi; r_{ij}) &= \log P(\mathbf{z}_{ij}, \mathbf{v}_{ij}, c_i | \mathbf{w}_{ij}, y_{ij}, \Phi, \tau, \gamma) \\ &\propto \log P(\mathbf{w}_{ij}, y_{ij} | \mathbf{z}_{ij}, \boldsymbol{\alpha}_i, \mathbf{v}_{ij}, \Phi) \\ &\quad + \log P(\mathbf{z}_{ij}, \mathbf{v}_{ij}, c_i | \Phi, \tau, \gamma) \\ &= \log P(\mathbf{w}_{ij} | \mathbf{z}_{ij}, \mathbf{v}_{ij}, \phi, \beta) \\ &\quad + \log P(y_{ij} | \mathbf{v}_{ij}, \boldsymbol{\alpha}_i, \sigma^2) \\ &\quad + \log P(\mathbf{v}_{ij}, \mathbf{z}_{ij} | \theta_j, \vartheta_j, \eta_j) \\ &\quad + \log P(c_i | \pi_i, \alpha_i)\,. \end{aligned} \tag{1}$$

In the above equation, the log-likelihood probability of the observed words $\mathbf{w}_{ij}$ given the aspect assignments $\mathbf{z}_{ij}$ and ratings $\mathbf{v}_{ij}$ is defined as

$$\log P(\mathbf{w}_{ij} | \mathbf{v}_{ij}, \mathbf{z}_{ij}, \phi, \beta) = \sum_{l=1:z_l=z_{ijl}}^{N} (\phi_{z_l w_l} + \beta_{z_l v_{z_l} w_l}), \tag{2}$$

where N is the number of words contained in a review, w_l and z_l indicate the l-th word and the corresponding word's aspect assignment, respectively, and v_{z_l} denotes the rating for aspect z_l. Note that ϕ_{z_l} is indexed by aspect z_l, indicating which words are associated with the aspect. Alternatively, $\beta_{z_l v_{z_l}}$ is indexed by aspect z_l and the rating for that aspect is v_{z_l}, so that we can learn the opinion score associated with each word for every aspect.

As mentioned above, given the rating for each aspect in a review and the associated reviewer's weight on the aspect, the observed overall rating is assumed to be drawn from a Gaussian distribution around $\boldsymbol{\alpha}_i^T \mathbf{v}_{ij}$. Formally, the log-likelihood of the observed overall rating y_{ij} given the aspect weights $\boldsymbol{\alpha}_i$ and aspect ratings $\mathbf{v}_{ij}$ is defined as

$$\log P(y_{ij}|\boldsymbol{\alpha}_i, \mathbf{v}_{ij}, \sigma^2) = \mathcal{N}(y_{ij}|\boldsymbol{\alpha}_i^T \mathbf{v}_{ij}, \sigma^2) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(y_{ij} - \sum_{k=1}^{A} \alpha_{ik} \cdot v_{ijk})^2, \tag{3}$$

where α_{ik} and v_k denote the weight and the rating of the k-th aspect, respectively.

The log-likelihood of the probability of aspect ratings $\mathbf{v}_{ij}$ and the words' aspect assignments $\mathbf{z}_{ij}$ with regard to a review of product m_j is defined as

$$\log P(\mathbf{v}_{ij}, \mathbf{z}_{ij}|\theta_j, \vartheta_j) = \log P(\mathbf{z}_{ij}|\theta_j) + \log P(\mathbf{v}_{ij}|\vartheta_j, \eta_j), \tag{4}$$

where the probability of aspect assignment of each word $P(\mathbf{z}|\theta_j)$ follows a multinomial distribution with parameter θ_j, denoted as $\mathbf{z}_{ij} \sim Multinomial(\theta_j)$, and the aspect-based ratings $\mathbf{v}_{ij}$ follow a multivariate Gaussian distribution with mean as ϑ_j and covariance matrix as $\eta_j I$, denoted as $\mathbf{z}_{ij} \sim \mathcal{N}(\vartheta_j, \eta_j I)$. The mean rating ϑ_j reflects how much most of reviewers enjoy the product, and the variance parameter η_j shows whether the reviewers agree with each other in terms of their opinions about that product as well as the aspects.

According to the assumptions that we mentioned at the beginning of this section, within the framework of latent class regression model (LCRM), the reviewer's aspect weight can be drawn from a multivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$ given that the reviewer belongs to a cluster k. We expect that this clustering procedure could enhance a reviewer's weight estimation by considering the inner-similarity among reviewers within the same cluster. Formally, the aspect-weight probability of the reviewer u_i belonging to cluster k (denoted as $c_i = k$) is defined as

$$\log P(c_i|\pi_i, \boldsymbol{\alpha}_i) = \log \frac{\pi_{ik} P(\boldsymbol{\alpha}_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{C} \pi_{ik} P(\boldsymbol{\alpha}_i|\mu_k, \Sigma_k)}, \tag{5}$$

where π_{ik} is the prior probability of the reviewer u_i belonging to the k-th cluster.

We now show how to learn the model's parameters Φ and the hidden variables $\mathbf{v}, \mathbf{z}, \mathbf{c}$, with regard to each review and each reviewer so as to maximize the log-posterior probability as defined in Eqn 1. In this work, the optimization proceeds by coordinating ascent on hidden variables including $\{\mathbf{v}, \mathbf{z}, \mathbf{c}\}$[1] and model parameters Φ, i.e., by alternately performing the following operations:

1. Update hidden variables with fixed parameters

$$(\hat{\mathbf{v}}, \hat{\mathbf{z}}, \hat{c_i}) = \arg\max_{(\mathbf{v}, \mathbf{z}, c_i)} \mathcal{L}(\Phi; r_{ij}). \tag{6}$$

For each review, the aspect ratings $\mathbf{v}$ and the words' aspect assignments are updated as

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}} \left[\sum_{k=1}^{A} \sum_{l=1}^{N} \delta(z_l = k) \log P(w_l|z_l, v_k, \phi, \beta) + \log P(y_{ij}|\alpha_i, \mathbf{v}, \sigma^2) + \log P(\mathbf{v}|\vartheta_j) \right], \tag{7a}$$

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} \left[\sum_{k=1}^{A} \sum_{l=1}^{N} \delta(z_l = k) \log P(w_l|z_l, \hat{v}_k, \phi, \beta) + \log P(\mathbf{z}|\theta_j) \right], \tag{7b}$$

where $\delta(z_l = k)$ is an indicator function denoting that the l-th word is relevant to the k-th aspect.

Specifically, for updating each word's aspect assignment z_l using above equation 7b, the pa-

[1] In the following, for the sake of simplicity, we use notation without index to represent parameters.

rameter $\phi_{z_l w_l}$ that indicates how likely the word w_l is assigned to aspect k is calculated as:

$$\phi_{z_l w_l | z_l = k} = \frac{n_{-l,k}^{(w_l)} + a}{n_{-l,k}^{(\cdot)} + Wa}, \tag{8}$$

where $n_{-l,k}^{(\cdot)}$ is the total number of words assigned to the k-th aspect, which does not include the current one; $n_{-l,k}^{(w_l)}$ is the total times of word w_l assigned to the k-th asepct; and a is a hyperparameter that determines how this multinomial distribution is smoothed. The parameter $\beta_{z_l v_{z_l} w_l}$ is calculated via:

$$\beta_{z_l v_{z_l} w_l | v_{z_l} = t, z_l = k} = \frac{n_{-l,t,k}^{(w_l)} + b}{n_{-l,t,k}^{(\cdot)} + Wb}, \tag{9}$$

where $n_{-l,t,k}^{(\cdot)}$ is the total number of words assigned to aspect k and aspect rating t; $n_{-lt,k}^{(w_l)}$ is the total times of word w_l assigned to aspect k and aspect rating t; and b is a hyperparameter for smoothing the multinomial distribution.

For each reviewer, his/her cluster membership is updated according to

$$\hat{c}_i = \arg\max_{c_i} \left[\log P(\alpha_i | c_i) + \log P(c_i | \pi_i) \right], \tag{10}$$

and the cluster-level aspect weight prior (μ_c, Σ_c) can be updated according to

$$\hat{\mu}_k = \frac{1}{U_c} \sum_{i=1}^{U} \delta(c_i = c) \alpha_i \tag{11}$$

$$\hat{\Sigma}_k = \frac{1}{U_c} \sum_{i=1}^{U} \left[(\alpha_i - \hat{\mu}_c)(\alpha_u - \hat{\mu}_c)^T \right], \tag{12}$$

where U_c denotes the set of reviewers who belong to cluster c.

2. Update parameters with fixed hidden variables

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\alpha}}) = \arg\max_{(\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\pi}, \boldsymbol{\alpha})} \sum_{r_{ij} \in \mathcal{R}} \mathcal{L}(\Phi; r_{ij}),$$

so as to update the aspect distribution for product m_j:

$$\hat{\theta}_j = \arg\max_{\theta_j} \sum_{r_{ij} \in \mathcal{R}} \log P(\hat{\mathbf{z}} | \theta_j), \tag{13}$$

update the aspect-based ratings distribution for product m_j as

$$\hat{\vartheta}_j = \arg\max_{\vartheta_j} \sum_{r_{ij} \in \mathcal{R}} \log P(\hat{\mathbf{v}} | \vartheta_j), \tag{14}$$

and update the aspect-based weights for reviewer u_i as

$$\hat{\alpha}_i = \arg\max_{\alpha_i} \sum_{r_{ij} \in \mathcal{R}} \log P(y_{ij} | \hat{\mathbf{v}}, \alpha_i) + \log P(\alpha_i | \hat{c}_i). \tag{15}$$

Algorithm 1 gives the pseudo-code of the model's inference process.

Algorithm 1 The optimization procedure of our proposed CARW model

```
initialize the hidden latent variables {z, v} and c_i randomly
initialize the model parameters Φ randomly
repeat
    1. update hidden variables with fixed parameters
    for each review r_ij do
        update the aspect ratings v via Eqn 7a
        update the words' aspect assignments z via Eqn 7b
    end for
    for each reviewer u_i do
        update the cluster membership c_i via Eqn 10
    end for
    2. update parameters with fixed hidden variables
    for each product m_j do
        update the aspect distribution via Eqn 13
        update the aspect rating distribution via Eqn 14
        update the aspect weights via Eqn 15
    end for
until convergence
```

5. Experiment and Results

5.1. Aspect Identification Task

In this section, we conduct an experiment to validate how the CARW model performs in terms of the

aspect identification task. We first describe the review data set we used for evaluation, the compared methods and evaluation metrics.

5.1.1. Description of the Dataset

With the goal of evaluating the quality of the identified aspects from reviews, we use a publicly available restaurant review dataset collected from CitySearch[2], originally used in [11]. After excluding short reviews (say with less than 50 words), we have 28,323 reviews posted by 19,408 reviewers for 3,164 restaurants (on average 1.46 reviews *per* reviewer). As the ground-truth, we use 1,490 labeled sentences which were classified into three main aspects (*food*, *service*, *ambiance*). To check for the classification agreement, each of the sentence was annotated by three different annotators. We also use a set of seed words related to each aspect as prior knowledge to guide the model learning. Table 3 shows the seed words, which are the same as ones used in [20]. As for the main parameters, they are set as $\sigma = 0.1, \gamma = 0.5, a = 0.01, b = 0.01, A = 4, C = 50$, thorough experimental trials.

Table 3
Seed words for four main aspects in restaurant reviews

Aspect	Seed words
food	food, chicken, beef, steak
service	service, staff, waiter, reservation
ambiance	ambiance, atmosphere, room, experience
price	price, value, quality, worth

5.1.2. Compared Methods and Evaluation Metrics

The frequency-based method used in [13] is treated as the baseline method. In this method, two phases are performed for the task of aspect identification. The first is a POS tagger implemented in the package CoreNLP[3] to identify frequent nouns (and noun phrases) as the aspect candidates. The second is to compute the candidate's lexical similarity to the seed words. The lexical similarity is determined via WordNet [9].

In addition, we implemented three different topic models to be compared with our CARW model: LDA based [3], Local LDA based [5] and MG-LDA [33]. The standard LDA model only considers the word co-occurrence patterns in review contents. In contrast, Local LDA model assumes that aspects are more likely discovered from sentence-level word co-occurrence patterns. The property of MG-LDA model is that it distinguishes between broad topics and fine-grained ratable topics [33]. To maintain comparability with the three models, we use the seed words as contained in Table 3 to guide the process of model learning. We also compared to the supervised SVM classifier [34], which was trained on unigram word features.

In order to test whether the outcome of our aspect-based weight estimation (see Section 5.3) can be beneficial from the accuracy improvement on aspect identification, we also compared our CARW model to a variation $\text{CARW}_{\textit{fixed\_weights}}$. In the $\text{CARW}_{\textit{fixed\_weights}}$ model, the weight for each aspect is fixed with as a constant value (e.g., $1/7$ when there are 7 aspects).

The evaluation metrics include precision (P), recall (R), and F-1 score, as they have been widely used for evaluating labeling accuracy [10]. In our case, for each aspect, the metric *precision* represents the proportion of correctly classified sentences among all of the classified ones. Formally, considering a specific aspect, *precision* is defined as

$$Precision = \frac{|IdentifiedAspects \cap TrueAspects|}{|IdentifiedAspects|} \tag{16}$$

For each aspect, metric *recall* refers to the proportion of correctly classified sentences among all of the sentences annotated with that aspect. Formally, *recall* is defined as

$$Recall = \frac{|IdentifiedAspects \cap TrueAspects|}{|TrueAspects|} \tag{17}$$

Another metric is the harmonic mean of precision and recall, termed as the $F1$ score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

5.1.3. Analysis of Results

Table 4 reports the experiment results. We can observe that our proposed unsupervised CARW model produces results comparable to those by the supervised model SVM. Additionally, CARW outperforms the other unsupervised models (i.e., LDA, Local LDA and MG-LDA) in terms of $F1$ metric for "Food", "Service" and "Ambiance" aspects. With regard to *preci-*

[2] http://www.citysearch.com
[3] http://nlp.stanford.edu/software/corenlp.shtml

sion, CARW beats Local LDA for "Service" and "Ambiance" aspects. In terms of the *recall* metric, the performance of CARW is better than the others for "Food" aspect.

In addition, the better performance of CARW relative to CARW$_{fixed_weights}$ indicates that the aspect weights learned in our joint model can empirically benefit from the task of aspect identification.

We also report the quantitative analysis of the experiment results. Specifically, Appendix Table 10 gives the aspect-related words and the associated sentiment words resulted from our CARW mode. From this table, we can see the frequency-based method shows the worst performance. As introduced before, the frequency-based method uses a set of seed words as the discriminator to identify which aspect a sentence refers to. Hence, its performance should be sensitive to the quality of constructed seed words.

5.2. Aspect-based Rating Inference Task

To evaluate the performance of CARW model in performing the task of inferring aspect-based rating, we use two datasets since they contain ground-truth aspect ratings in each review.

5.2.1. Description of the Dataset

The first dataset contains a set of hotel reviews from Tripadvisor.com[4] [36]. In this dataset, in addition to the overall rating, each hotel review is associated with ratings for seven aspects: *value*, *room*, *location*, *cleanness*, *check-in/front desk*, *service* and *business service*. To ensure that each review includes all aspects, we remove those reviews in which any of the seven aspect ratings is missing or which has less than 50 words. Thus, there are 53,696 reviews (given by 45,744 reviewers for 1,455 hotels) for the evaluation. In this dataset, we set $\sigma = 0.1, \gamma = 0.5, a = 0.01, b = 0.01, A = 7, C = 120$ via experimental trials. Another dataset is a subset of the beer review dataset used in [24] that includes four aspects: *feel*, *look*, *smell* and *taste*, which were collected from BeerAdvocate[5]. We use a subset of 7,015 beer reviews in our experiment. For this dataset, we set $\sigma = 0.1, \gamma = 0.5, a = 0.01, b = 0.01, A = 4, C = 50$. The statistical descriptions of the two datasets are shown in Table 7. The seed words for hotel reviews are shown in Table 5, and the seed words for beer reviews are shown in Table 6.

[4] http://www.tripadvisor.com
[5] http://beeradvocate.com

5.2.2. Compared Methods and Evaluation Metrics

We implemented a lexicon-based method as the baseline [35]. In this method, each aspect rating is estimated based on the words that describe that aspect in the review. Concretely, the rating of aspect A_k in review r_{ij} is computed as

$$\hat{v}_{ijk} = \frac{\sum_{w \in W_k(r_{ij})} opinion(w)}{|W_k(r_{ij}))|}, \tag{19}$$

where $W_k(r_{ij})$ denotes the set of words in the review r_{ij} that are relevant to aspect A_k, and $opinion(w)$ denotes the word's sentiment score according to the sentiment lexicon SentiWordNet [8].

We also implemented two related methods, *local prediction* and *global prediction*, as introduced in [23]. Note that they both assume that the results of aspect identification are known before they conduct the rating inference task. Thus, in the experiment, the results of aspect identification created by our CARW model are used as inputs to these compared methods. Specifically, in the *local prediction* method [23], all of the aspects are assumed to share the same ratings with the overall rating. It means that only a single rating classifier is learned by using the overall rating as the target label. For each aspect, the trained classifier is applied to estimate its rating. In contrast, the *global prediction* [23] method first learns a rating classifier for each rating level (from 1 to 5 in our case) of the aspect based on the Native Bayes classifier. For example, for the 2-star rating classifier, the phrases occurring in reviews with the overall rating 2 are used as the training corpus. Then, the Native Bayes classifier was trained based on a unigram language model.

In this experiment, we test whether the aspect weights learned from CARW model can in turn enhance the accuracy of inferring aspect-based ratings. Similar to the aspect identification task, we take CARW$_{\text{fixed_weights}}$ model as the baseline.

One evaluation metric is L_1 error, which measures the absolute difference between the estimated ratings and real ratings as defined in Eqn 20:

$$L_1 = \frac{\sum_{(i,j) \in \mathcal{R}} |\mathbf{v}_{ij} - \mathbf{v}^*_{ij}|}{R \times A}, \tag{20}$$

in which $\mathbf{v}_{ij}$ and $\mathbf{v}^*_{ij}$ denote the estimated aspect ratings vector and real aspect ratings vector regarding review r_{ij}, respectively.

In addition to the L_1 measure, we use three other metrics according to [36]. The first metric is ρ_{aspect},

Table 4
Comparison results regarding aspect identification task (P: Precision, R: Recall, F: F1 score)

	Food			Service			Ambiance		
	P	R	F	P	R	F	P	R	F
Frequency-based	0.575	0.329	0.466	0.514	0.515	0.514	0.239	0.285	0.260
LDA	0.646	0.554	0.597	0.469	0.494	0.481	0.126	0.179	0.148
MG-LDA	0.888	0.772	0.826	0.637	0.648	0.642	0.609	0.876	0.719
Local LDA	**0.969**	0.775	0.861	0.731	**0.810**	**0.768**	0.573	**0.892**	0.698
CARW	0.802	**0.970**	**0.878**	**0.864**	0.682	0.762	**0.853**	0.720	**0.780**
$\text{CARW}_{\textit{fixed\_weights}}$	0.653	0.642	0.647	0.501	0.523	0.512	0.412	0.514	0.457
SVM	0.814	**0.975**	**0.887**	**0.874**	0.670	0.759	**0.860**	0.538	0.662

Table 5
Seed words for seven aspects in hotel reviews

Aspect	Seed words
value	value, price, quality, worth
room	room, suite, view, bed
location	location, traffc, minute, restaurant
cleanliness	clean, dirty, maintain, smell
check-in/front desk	stuff, check, help, reservation
service	service, food, breakfast, buffet
business service	business, center, computer, internet

Table 6
Seed words for four aspects in beer reviews

Aspect	Seed words
feel	silky, velvety, mouthfeel, body, watery
look	beauty, dark, gorgeous, appearance, light
smell	sweet, malt , smell, nose , smell
taste	taste, hops, bitter, bland, chocolate

Table 7
Statistical descriptions of hotel and beer review datasets

	Hotel dataset	Beer dataset
#Products	1,455	1,000
#Reviews	53,696	7,015
#Reviewers	45,744	964
#Avg. reviews *per* reviewer	1.17	7.28

which is the average Pearson correlation between the estimated ratings and real ratings across all aspects within each review, formally defined as

$$\rho_{aspect} = \frac{\sum_{(i,j)\in\mathcal{R}} \rho_{\mathbf{v}_{ij},\mathbf{v}^*_{ij}}}{|\mathcal{R}|}, \tag{21}$$

where $\rho_{\mathbf{v}_{ij},\mathbf{v}^*_{ij}}$ is the Pearson correlation between the estimated aspect ratings vector $\mathbf{v}_{ij}$ and real ratings vector $\mathbf{v}^*_{ij}$ regarding review r_{ij}. This metric can measure how well the estimated aspect-based ratings can preserve the ranking of aspects based on their real ratings.

The second metric is ρ_{review}, which is the average Pearson correlation between the estimated ratings and real ratings for each aspect across all products, defined as

$$\rho_{review} = \frac{\sum_{k=1}^{A} \rho(\overrightarrow{\mathbf{v}_k}, \overrightarrow{\mathbf{v}^*_k})}{A}, \tag{22}$$

where $\overrightarrow{\mathbf{v}_k}$ and $\overrightarrow{\mathbf{v}^*_k}$ are respectively the average of estimated aspect-based ratings and the average real aspect-based ratings across all products regarding aspect A_k. This metric measures how well the estimated ratings can be used for ranking in terms of each aspect.

The third one is *MAP@10* which measures how well the estimated aspect-based ratings preserve the top products on the top positions in the ranking list, defined as

$$MAP@10 = \frac{\sum_{A_i\in\mathcal{A}} \sum_{m_j\in Rel(A_i)} \frac{\sigma(rank(m_j)<10)}{rank(m_j)}}{A}, \tag{23}$$

where $Rel(A_i)$ denotes the set of relevant products (here, we treat the top-100 products according to their on their real aspect ratings as the relevant products), and $rank(m_j)$ indicates the ranking position according to the estimated aspect ratings, and $\sigma(\cdot)$ is an indicator function that ensures only the top-10 products are considered.

5.2.3. Analysis of Results

We report the results of running different methods on hotel and beer reviews in Table 8. From this table, we can see that CARW model outperforms the other methods in both datasets. For the hotel reviews, CARW outperforms the second-best method (i.e., the global prediction) by 17% in terms of the L_1 metric and by 44% in respect to the ρ_{aspect} metric. For the beer reviews, similar trends appear. What's more, the better performance of CARW against CARW$_{fixed_weights}$ indicates that the aspect weight estimation can be helpful for improving the accuracy of aspect rating estimation.

In Tables 11 and 12 (see Appendix), we show the aspect-related words and associated sentiment words in descending order of their sentiment scores as returned by CARW model.

5.3. Aspect-based Weight Estimation Task

As for the third task, aspect-based weight estimation, because we do not have ground-truth data, we implemented a recommender system that incorporates the estimated aspect-based weights into the process of generating recommendations, so as to indirectly measure the accuracy of our method. In this experiment, we use the same datasets of the second task.

5.3.1. Recommendation Method and Evaluation Procedure

For a user whose aspect weights are $\boldsymbol{\alpha}_u$, the score of a product m_j can be computed as

$$score(u, m_j) = \sum_{k=1}^{A} \alpha_{uk} \times opinion(m_j, k), \quad (24)$$

where α_{uk} denotes the user's weight on the k-th aspect, and $opinion(m_j, k)$ indicates the average opinion value on the k-th aspect of the product m_j based on its reviews, calculated via $avg_{(i,j)\in\mathcal{R}}[\mathbf{v}_{ijk}]$. Then, the products with highest scores are recommended to the user.

The following procedure is conducted to perform the evaluation:

1. Choose reviewers who have posted at least 5 reviews. In this step, 1000 reviewers who satisfy this criterion are chosen for each dataset.
2. Treat each reviewer as a simulated user whose aspect-based weights are estimated by the tested method (e.g., CARW).
   - For each tested user, the reviewed products (with overall rating above 4) are used for testing, and taken as relevant products when we evaluate the recommendations.
   - The products are ranked according to their scores (via Eqn 24) that consider both the aspect-based weights and the aspect-based ratings.

5.3.2. Compared Methods and Evaluation Metrics

One compared method is based on *probabilistic regression model* (PRM) [41], which is a linear regression model, that learns the weights for individual reviewers. For the PRM-based model, we apply CARW to identify aspects and estimate aspect ratings as inputs to estimate aspect weights. The only difference between CARW and PRM is thus that the reviewers are clustered in CARW according to their aspect weights so that their inter-similarity can be accommodated.

To evaluate the recommendation accuracy, we measure how well the ranking returned by the recommender agrees with the user's own ranking. The first metric is the widely used *MAP* metric, which takes the top 10 candidates into account, as defined in Eqn 23. Another metric, the *Kendall rank correlation coefficient* [16], computes the fraction of pairs with the same order in both system's ranking and user's ranking. Formally, it is defined as

$$Kendall = \frac{\#concordant\ pairs - \#disordant\ pairs}{\frac{1}{2}M(M-1)}, \quad (25)$$

where $\#concordant\ pairs$ ($\#disordant\ pairs$) denotes the number of pairs of products with the same (different) order between the product ranking resulted from the Eqn 24 and the product ranking resulted from the overall ratings given by the user, and M is the total number of products contained in the dataset.

5.3.3. Analysis of Results

As shown in Table 9, the recommendations based on the aspect weights estimated by CARW model are more accurate than PRM-based method on both datasets. Specifically, for hotel recommendations, CARW achieves higher *Kendall* value 0.610 (vs. 0.526 by PRM) and *MAP@10* value 0.0033 (vs. 0.0016 by PRM). For beer recommendations, CARW also achieves better performance in terms of both metrics. Thus, we can conclude that the clustering-based approache

Table 8
Evaluation of the estimated aspect ratings on hotel and beer reviews

	Hotel reviews				Beer reviews			
	L_1	ρ_{aspect}	ρ_{review}	$MAP@10$	L_1	ρ_{aspect}	ρ_{review}	$MAP@10$
Lexicon-based	1.401	0.112	0.201	0.208	1.712	0.028	0.103	0.198
Local prediction	1.343	0.230	0.534	0.297	1.302	0.211	0.245	0.263
Global prediction	1.243	0.231	0.561	0.298	1.503	0.232	0.246	0.263
CARW	**1.061**	**0.413**	**0.647**	**0.308**	**1.081**	**0.235**	**0.310**	**0.278**
$CARW_{fixed\_weights}$	1.316	0.234	0.551	0.283	1.301	0.210	0.257	0.257

Table 9
Evaluation of the recommendation accuracy for the third task of aspect weight estimation

	Hotel reviews		Beer reviews	
	Kendall	*MAP@10*	*Kendall*	*MAP@10*
PRM	0.526	0.0016	0.510	0.0012
CARW	**0.610**	**0.0033**	**0.582**	**0.0023**

CARW is able to facilitate the generation of better recommendations than PRM.

6. Conclusion

In this paper, we propose a unified CARW model that can simultaneously 1) identify the aspects mentioned in reviews, 2) infer the aspect-based ratings based on the sentiments expressed on identified aspects, and 3) estimate the aspect-based weights placed on aspects by a reviewer. The three tasks are addressed in an unsupervised manner, so that the CARW model can be feasibly applied across different domains by minimizing the training effort. From the experimental results, we can conclude that CARW outperforms the related methods regarding all of the three tasks. In addition, we demonstrate that the three tasks can be complementary to each other and be improved simultaneously through the unified model.

In the future, we will try to improve our model by parallelizing its learning process to reduce the time consumption. In addition, we will apply the proposed model to other domains (such as digital camera, cars) to validate its generalized effectiveness.

7. Acknowledgements

This research work was supported by Hong Kong Research Grants Council under Project ECS/HKBU211912 and China National Natural Science Foundation under Project NSFC/61272365.

Appendix

Table 10, Table 11 and Table 12.

References

[1] N. Archak, A. Ghose, and P. G. Ipeirotis. Show me the money!: Deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Multi-facet rating of product reviews. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 461–472, Berlin, Heidelberg, 2009. Springer-Verlag.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

[4] E. Boiy and M. F. Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558, Oct. 2009.

[5] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 804–812, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[6] L. Chen and F. Wang. Preference-based clustering reviews for augmenting e-commerce recommendation. *Knowledge Based Systems*, 50(0):44 – 59, 2013.

[7] J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2006.

[8] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC'06, pages 417–422. 2006.

[9] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[10] J. M. Francis, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pages 249–252, 1999.

[11] G. Ganu, N. Elhadad, and A. Marian. Beyond the stars: Improving rating predictions using review text content. In *12th*

Table 10
Top 20 aspect related words and their associated sentiment words for restaurant reviews returned by our CARW model

Aspect	Aspect words ϕ	Sentiment words with high value β
food	food, place, steak, restaurant, chicken, order, table, best, friend, menu, eat, wine, dinner, dish, meal, beef, nice, delicious, dessert, appetizer	perfect, love, amazing, favorite, highly, derful, friendly, best, delicious, fantastic, outstanding, excellent, great, die, superb, attentive, cozy, incredible, romantic, family
service	service, staff, waiter, reservation, place, restaurant, time, order, table, wait, come, friend, night, people, delicious, excellent, ask, friendly, seat, think	asd, inn, ethiopian, heaven, royalty, genius, oasis, sicilian, genuine, hooked, greenwich, unassuming, virgil, derfully, innovative, vegan, authenticity, recomend, art, marvelous
ambiance	experience, atmosphere, room, ambiance, place, restaurant, wait, come, friend, eat, wine, night, drink, say, people, delicious, nice, bar, look, friendly	amazing, incredible, die, family, perfect, delicious, favorite, fantastic, derful, efficient, superb, friendly, comfortable, great, helpful, awesome, love, relax, excellent, reasonable
price	price, worth, quality, value, place, restaurant, just, order, wine, dinner, eat, night, dish, drink, nice, people, seat, love, dessert, bar	worth, beef, price, waiter, food, value, service, chicken, breakfast, reservation, quality, staff, recommend, steak, sprinkle, international, franchise, flay, younger, pro

Table 11
Top 20 aspect related words and their associated sentiment words for hotel reviews returned by our CARW model

Aspect	Aspect words ϕ	Sentiment words with high value β
value	hotel, price, stay, great, value, worth, night, day, quality, make, like, beach, pool, area, resort, free, bathroom, people, excellent, recommend	helpful, florence, excellent, perfect, highly, paris, fantastic, friendly, comfortable, value, love, modern, distance, spacious, great, derful, central, nyc, recommend, amsterdam
room	room, bed, view, hotel, suite, stay, night, day, nice, make, pool, resort, say, book, bar, little, need, comfortable, come, desk	florence, helpful, excellent, perfect, friendly, derful, paris, great, fantastic, highly, distance, recommend, love, modern, staff, city, comfortable, stay, quiet, london
location	location, restaurant, minute, hotel, stay, traffic, nice, day, time, walk, like, place, pool, area, small, friendly, want, look, trip, street	location, helpful, excellent, florence, fantastic, comfortable, perfect, love, spacious, modern, friendly, recommend, highly, paris, great, derful, london, quiet, stay, lovely
cleanliness	clean, hotel, smell, stay, dirty, maintain, place, make, like, beach, people, bathroom, floor, excellent, look, use, helpful, little, best, need	helpful, perfect, florence, excellent, paris, comfortable, fantastic, derful, friendly, great, highly, recommend, love, stay, definitely, spacious, quiet, clean, distance, definately
check-in/front-desk	check, hotel, help, reservation, stay, staff, nice, time, make, area, book, friendly, say, people, excellent, use, helpful, best, desk , ask	florence, paris, helpful, perfect, excellent, highly, modern, spacious, comfortable, distance, fantastic, friendly, london, great, superb, fabulous, lovely, derful, attraction, recommend
service	service, breakfast, food, hotel, buffet, stay, place, like, area, resort, friendly, book, bar, little, people, use, helpful, need, free,	helpful, perfect, florence, friendly, comfortable, spacious, distance, great, excellent, derful, love, attraction, fantastic, recommend, paris, definitely, modern, quiet, highly, city
business service	hotel, internet, business, center, stay, great, staff, good, place, beach, want, say, book, friendly, bar, people, excellent, use, trip, free	florence, helpful, excellent, perfect, fantastic, friendly, highly, paris, comfortable, great, derful, distance, modern, recommend, love, spacious, square, quiet, stay, definitely

Table 12
Top 20 aspect related words and their associated sentiment words for beer reviews returned by our CARW model

Aspect	Aspect words ϕ	Sentiment words with high value β
feel	mouthfeel, bottle, carbonation, alcohol, light, smooth, poured, drink, body, beer, medium, dry, sweet, feel, thin, finish, like, full, tongue, creamy	perfect, silky, amazing, velety, incredible, exceptional, perfect, flat, thin, absolute, velvet, water, weak, watery, thin, bland, disappoint, macro, bad, bearing
look	head, dark, body, beer, nice, color, like, carbonation, glass, white, thin, brew, black, tastes, appearance, clear, pour, golden, hops, pale	beautiful, perfect, massive, amazing, pitch, huge, gorgeous, forever, pitch, incredible, yellow, cheap, macro, water, soda, miller, bud, lime, poor, horrible
smell	sweet, hops, smell, malt, caramel, beer, nose, light, coffee, alcolhol, like, sweetness, slight, hints, fruity, spicy, yeast, fruit, aroma, finish	amazing, awesome, fantastic, incredible, wonderful, absolutely, exceptional, perfect, beutiful, good, weak, nothing, cheap, skunky, bland, macro, water, adjunct, stale, corn
taste	malt, taste, hop, flavor, chocolate, sweet, caramel, bitter, coffee, bitterness, light, finish, fruit, smell, alcohol, hint, strong, citrus, dark, sweetness	amazing, delicious, perfect, wonderful, incredible, absolutely, awesome, outstanding, fantastic, bourbon, truly, bland, weak, watery, metallic, corn, boring, macro, disappointing, skunk

International Workshop on the Web and Databases, WebDB '09, 2009.

[12] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[13] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[14] W. Jin, H. H. Ho, and R. K. Srihari. Opinionminer: A novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1195–1204, New York, NY, USA, 2009. ACM.

[15] Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824, New York, NY, USA, 2011. ACM.

[16] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[17] Y. Kim and J. Srivastava. Impact of social influence in e-commerce decision making. In *Proceedings of the 9th International Conference on Electronic Commerce*, ICEC '07, pages 293–302. ACM, 2007.

[18] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang, and H. Yu. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 653–661, Stroudsburg, PA, USA, 2010.

[19] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.

[20] B. Lu, M. Ott, C. Cardie, and B. K. Tsou. Multi-aspect sentiment analysis with topic models. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 81–88, Washington, DC, USA, 2011. IEEE Computer Society.

[21] Y. Lu, H. Duan, H. Wang, and C. Zhai. Exploiting structured ontology to organize scattered online opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 734–742. Association for Computational Linguistics, 2010.

[22] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 121–130, New York, NY, USA, 2008. ACM.

[23] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 131–140, New York, NY, USA, 2009. ACM.

[24] J. McAuley, J. Leskovec, and D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 1020–1025, Washington, DC, USA, 2012. IEEE Computer Society.

[25] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 171–180, New York, NY, USA, 2007. ACM.

[26] S. Moghaddam and M. Ester. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 665–674, New York, NY, USA, 2011. ACM.

[27] S. Moghaddam and M. Ester. The FLDA model for aspect-based opinion mining: Addressing the cold start problem. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 909–918, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[28] J. Parker, A. Yates, N. Goharian, and W. G. Yee. Efficient estimation of aspect weights. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1057–1058, New York, NY, USA, 2012. ACM.

[29] L. Qi and L. Chen. Comparison of model-based learning methods for feature-level opinion mining. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, pages 265–273, Washington, DC, USA, 2011. IEEE Computer Society.

[30] G. Qiu, B. Liu, J. Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27, Mar. 2011.

[31] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin. Red opal: Product-feature scoring from reviews. In *Proceedings of the 8th ACM conference on Electronic Commerce*, EC '07, pages 182–191, New York, NY, USA, 2007. ACM.

[32] B. Snyder and R. Barzilay. Multiple aspect ranking using the good grief algorithm. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 300–307, Rochester, New York, April 2007. Association for Computational Linguistics.

[33] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 111–120, New York, NY, USA, 2008. ACM.

[34] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, 2000.

[35] F. Wang and L. Chen. Recommending inexperienced products via learning from consumer reviews. In *Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence*, WI'12, pages 596–603, Washington, DC, USA, 2012. IEEE Computer Society.

[36] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 783–792, New York, NY, USA, 2010. ACM.

[37] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 618–626, New York, NY, USA, 2011. ACM.

[38] Y. Wu, Q. Zhang, X. Huang, and L. Wu. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1533–1541, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[39] F. Xianghua, L. Guo, G. Yanyan, and W. Zhiqiang. Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge Based System*, 37:186–195, Jan. 2013.

[40] A. Yates, N. Goharian, and W. G. Yee. Semi-supervised probabilistic sentiment analysis: Merging labeled sentences with unlabeled reviews to identify sentiment. In *Proceedings of the 76th ASIST Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, ASIST '13, pages 81:1–81:10, Silver Springs, MD, USA, 2013. American Society for Information Science.

[41] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua. Aspect ranking: Identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1496–1505, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[42] W. X. Zhao, J. Jiang, H. Yan, and X. Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 56–65, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.