A Sentiment-aligned Topic Model for Product Aspect Rating Prediction

Hao Wang
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
hwa63@sfu.ca

Martin Ester
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
ester@sfu.ca

Abstract

Aspect-based opinion mining has attracted lots of attention today. In this paper, we address the problem of product aspect rating prediction, where we would like to extract the product aspects, and predict aspect ratings simultaneously. Topic models have been widely adapted to jointly model aspects and sentiments, but existing models may not do the prediction task well due to their weakness in sentiment extraction. The sentiment topics usually do not have clear correspondence to commonly used ratings, and the model may fail to extract certain kinds of sentiments due to skewed data. To tackle this problem, we propose a sentiment-aligned topic model(SATM), where we incorporate two types of external knowledge: product-level overall rating distribution and word-level sentiment lexicon. Experiments on real dataset demonstrate that SATM is effective on product aspect rating prediction, and it achieves better performance compared to the existing approaches.

1 Introduction

Online reviews have become an important source of information for consumers. People tend to read reviews to help them compare products, and make informed decisions. As the volume of product reviews continues to grow, it is often impossible to read all of them, which calls for efficient methods for opinion mining. Nowadays, for each product, many websites aggregate the overall rating of reviews, and display its distribution. However, this cannot provide detailed information. For example, two products may have similar overall rating distributions, while people talk about different unsatisfactory aspects. This problem has inspired a new line of research on aspect-level opinion mining(Hu and Liu, 2004).

An aspect refers to a rateable feature, such as staff and location in hotel reviews, or size and battery for digital camera reviews. In this paper, we deal with the problem of *product aspect rating prediction*. The input is a collection of products, and each product is associated with a set of reviews. The goal is to extract the corpus-level aspects, and predict the aspect ratings for each product. This kind of fine-grained sentiment analysis will help users efficiently digest the reviews, and gain more insight into the product quality.

The product aspect rating prediction problem usually involves two subtasks: aspect extraction and sentiment identification(Titov and McDonald, 2008b). Given some text, we would like to know what aspects it talks about, and what kind of sentiments are expressed. For example, given a sentence "the room is filthy", we would like to know that it talks about the aspect "room". Also, "filthy" is a sentiment word, and it expresses strongly negative sentiment towards the aspect "room".

Topic models(Blei et al., 2003; Hofmann, 1999) have been popular in aspect-based opinion mining(Liu, 2012). Existing works have used topic models to extract only aspects(Titov and McDonald, 2008a; Brody and Elhadad, 2010; Chen et al., 2013), or jointly model aspects and sentiments (Mei et al., 2007; Lin and He, 2009; Li et al., 2010; Jo and Oh, 2011; Moghaddam and Ester, 2011; Lakkaraju et al., 2011; Sauper et al., 2011; Mukherjee and Liu, 2012; Lazaridou et al., 2013; Moghaddam and Ester, 2013; Kim et al., 2013). In the joint modelling approaches, a sentiment topic is usually modelled as a sentiment label-word distribution, analogous to the topic-word distribution in standard topic models. However, the difference is that the sentiment topics need to be ordered. If the model is to be applied for aspect rating prediction, the sentiment topics should have clear cor-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1192–1202,
October 25-29, 2014, Doha, Qatar.

respondence to the ratings. Suppose there are 5 sentiment topics with sentiment labels from 1 to 5. The sentiment topic with label i is expected to correspond to the rating i on the 1-5 rating scale. For example, the sentiment topic with label 5 should have high probability over positive sentiment words, so it expresses highly positive sentiment, which matches our natural interpretation for the rating 5. In this case, sentiment labels and ratings are *aligned*. However, in a standard topic model, the learned sentiment topics may not have clear correspondence with different ratings. Also, if the positive reviews are dominant in the data, the topic model may fail to capture the negative sentiments with any sentiment topic, so no sentiment labels are matched with low ratings. If the sentiment labels are not correctly aligned to the ratings, we cannot use these sentiment labels to predict aspect ratings. Consequently, the aspect rating prediction accuracy is compromised, and the method is less practical. We call this the *sentiment label alignment* problem. To tackle this problem, models in the literature usually use some seed words for each sentiment topic to define Dirichlet priors with asymmetric concentration parameter vectors (Sauper et al., 2011; Kim et al., 2013), or use seed words to initialize word assignment to sentiment topic(Lin and He, 2009), or both(Li et al., 2010; Jo and Oh, 2011). However, these seed words are usually arbitrarily selected, and how to define asymmetric priors is not clear, especially when we would like to capture more than two (positive and negative) kinds of sentiments.

In this paper, we propose a sentiment-aligned topic model(SATM) for product aspect rating prediction, which focuses the sentiment label alignment problem. We use two kinds of external knowledge: the product overall rating distribution, and a sentiment lexicon. For each product, the overall rating distribution is available on most online review websites. It provides the big picture of the product-level sentiments. In SATM, for each product and each aspect, we define a multinomial distribution over sentiment labels, with prior parameterized by the overall rating distribution. Sentiment lexicon is constructed by linguistic experts, and every word in the lexicon is associated with a sentiment polarity score(Taboada et al., 2011). We treat the polarity score as an extra word feature in a semi-supervised framework. By incorporating both product-level and word-level knowledge into the model, the sentiment labels can be aligned with ratings, and the extracted sentiment topics can capture different kinds of sentiments, ranging from highly positive to highly negative. Experiments on a TripAdvisor dataset demonstrate that our method can effectively deal with the sentiment label alignment problem, and outperforms state-of-the-art methods in terms of product aspect rating prediction accuracy.

2 Related work

Several methods have been proposed for product aspect rating prediction, and many of them are based on topic models.

In (Lu et al., 2009), the authors studied the problem of generating an aspect rating summary for short comments. The text was first preprocessed into phrases of the format <headterm, sentiment word>, and the headterms are clustered by Structured PLSA to find K major aspects. Then, phrase ratings are predicted by either Local Prediction or Global Prediction, and they are aggregated to get aspect ratings. The method in (Brody and Elhadad, 2010) also first uses topic models to find aspects. Then, for each aspect, it extracts all the relevant adjectives, and builds a conjunction graph. A label propagation algorithm(Zhu and Ghahramani, 2002) is used on the graph to learn the sentiment polarity score of adjective words. Although this approach is not proposed for aspect rating prediction, it can be used for this task if the polarity scores of adjective words are aggregated for each aspect. All the methods above perform aspect extraction and sentiment identification separately, while our approach takes a joint modelling approach so that different subtasks can potentially reinforce with each other. To demonstrate this, we use these methods as baselines in our experiments.

Wang et al. worked on the *Latent Aspect Rating Analysis* problem(Wang et al., 2010; Wang et al., 2011), the task of inferring aspect ratings for each review and the relative weights reviewers have placed on each aspect. In (Wang et al., 2010), aspect keywords are provided as user input, and a two-stage method, called Latent Rating Regression(LRR), is proposed. The first stage uses a bootstrapping algorithm to obtain more related words for each aspect, and segments the document content. In the second stage, the overall rating is "generated" as weighted combination of the latent aspect ratings, and LRR is used to infer both the

weights and aspect ratings. Their follow-up work (Wang et al., 2011) does not need keyword specification from users, and replaces the bootstrapping method with a topic model. However, both methods implicitly require that each review talks about all aspects, which is not always true due to the data sparsity in online reviews.

In (Moghaddam and Ester, 2011), ILDA was proposed for product aspect rating prediction. Later, it was extended to FLDA (Moghaddam and Ester, 2013) to address the cold start problem, when there are few reviews associated with a product. Similar to (Lu et al., 2009), in ILDA and FLDA, a preprocessing step parses the text into phrases of the format <headterm, sentiment word>, and a review is modelled as a bag of phrases. We also adopt this assumption in our model. The method in (Sauper et al., 2011; Sauper and Barzilay, 2013) does not use phrases, but instead uses "snippets", and an snippet is a short sentence or phrase. However, the sentiment label alignment problem is not well addressed in these models, which limits their practicality. ILDA and FLDA did not deal with this problem. The model in (Sauper et al., 2011; Sauper and Barzilay, 2013) follows the most common approach of using seed words to define asymmetric priors. It supports only two kinds of sentiment topics: positive and negative, while how to define asymmetric priors for more sentiment topics becomes unclear. More importantly, the prior approach may not work well in practice(see Experiment Section). Lakkaraju et al. try to tackle the sentiment label alignment problem by assuming that the overall rating is generated as response variable(Lakkaraju et al., 2011), with the sentiment topic proportions as features. However, how the sentiment labels are related to ratings is still unknown until learned, and we may not get the desired alignment. Lazaridou et al. attempt to connect sentiment labels with ratings by Kronecker symbol, but this method only applies to three sentiment polarities: -1(negative), 0(neutral), $+1$(positive), and it does not explore the word-level lexicon, which is also an important source of knowledge.

Another line of research on product aspect rating prediction or summarization does not use topic models, but relies mainly on word frequency and grammatical relations(Hu and Liu, 2004; Popescu and Etzioni, 2005; Blair-goldensohn et al., 2008), or specialized review selection(Long et al., 2014). In this case, the extracted aspect words need to be clustered manually. For example, *picture* and *photo* may refer to the same aspect in digital camera reviews. By comparison, topic modelling approaches extract aspect words and cluster them simultaneously.

Our method incorporates the product overall rating distributions and sentiment lexicons into the model, so it is also related to topic models which use observed features or domain knowledge(Mimno and McCallum, 2008; Andrzejewski et al., 2009; Andrzejewski et al., 2011). Mimno et al. introduces two general frameworks to integrate observed features into the generative process: downstream and upstream topic models(Mimno and McCallum, 2008). In the context of aspect-based opinion mining, MaxEnt-LDA(Zhao et al., 2010) integrates a discriminative maximum entropy component to help separate aspect words and sentiment words. The SAS model (Mukherjee and Liu, 2012) uses seed words to provide guidance for aspect discovery, and MC-LDA (Chen et al., 2013) uses must-links and cannot-links to extract coherent aspects. However, MaxEnt-LDA, SAS and MC-LDA cannot be used for aspect rating prediction, since they fail to identify the sentiment polarity of sentiment words.

3 Method

3.1 Preliminaries

We first introduce several key concepts used in our model.

Products: Let $P = \{P_1, P_2, \dots\}$ be a set of products. Each product P_i is associated with a set of reviews $D_i = \{d_1, d_2, \dots d_{N_i}\}$, and also an overall rating distribution Y_i. Y_i is a multinomial distribution on R ratings. It is available on most online review websites, and usually $R = 5$.

Aspects: An aspect is a rateable feature of a product, and each aspect is modelled as a distribution over aspect words. The number of aspects is predefined as K.

Sentiment topics: A sentiment topic is modelled as a distribution over sentiment words, and each sentiment topic is associated with a sentiment label. To make it consistent with commonly used rating scale, we assume there are R sentiment labels, corresponding to the R ratings. The challenge is that sentiment labels with higher values are expected to be associated with sentiment topics which express more positive sentiments, so that

we can match sentiment labels with ratings.

Phrases: An opinion phrase $f =< h, m >$ is a pair of aspect word h and sentiment word m, such as $< room, filthy >$(Lu et al., 2009; Moghaddam and Ester, 2011). For each product P_i, we first parse the related reviews D_i into phrases F_i, and each product can be modelled as a bag of phrases.

Sentiment lexicons : A sentiment lexicon L is a list of sentiment words, and each word $m \in L$ is associated with a sentiment polarity score s_m. s_m can take T values. Note that the lexicon L usually only covers a small subset of sentiment words in the whole vocabulary.

Sentiment association: The sentiment label takes R values, and there are T different values for the polarity score in the sentiment lexicon. However, the relation between sentiment labels and polarity scores are unknown. If we have training instances where a sentiment word m is associated with both a sentiment label r_m and polarity score s_m, we can build a classifier, where the explanatory variable for the classifier is a sentiment label, and outcome is the polarity score. In this case, $H(s_m|r_m)$ can be interpreted as the probability of observing a polarity score s_m, given its sentiment label r_m. We refer to this probability H as sentiment association. This is a key component in our model. It naturally bridges the gap between sentiment labels and polarity scores, and captures the uncertainty in their relations. Note that H can be trained independent of the topic model part. For each training instance, suppose the sentiment word is $m \in L$, we need to know its sentiment label r_m and polarity score s_m. s_m can be retrieved directly from the sentiment lexicon, and r_m can be either manually or automatically annotated. For example, suppose the word m appears in review d, we can assign the overall rating of d as its sentiment label. In this case, each word $m \in L$ can be associated with multiple training instances that have the same value for s_m but different sentiment labels r_m. We adopt this approach to automatically annotate sentiment labels, and details are described in the Experiments section.

3.2 Problem definition

The product aspect rating prediction problem can be defined as follows. The input is a set of products P. Each product P_i has a bag of phrases F_i, and an overall rating distribution Y_i over R ratings. The output is the K corpus-level aspects, and for each product, we predict its ratings on the K aspects, also in the $[1, R]$ rating scale. We assume products in P are in the same category so they share the same aspects.

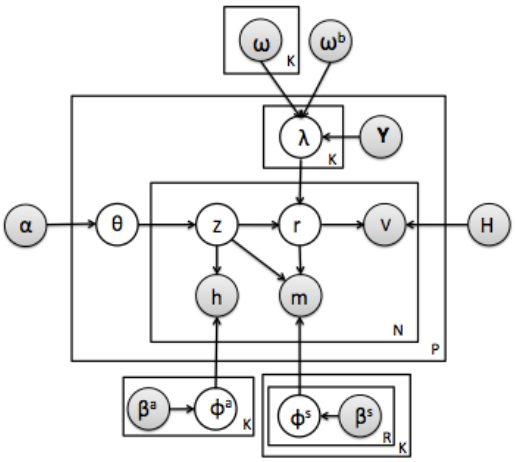

Figure 1: Graphical model of SATM

3.3 The SATM model

We introduce the Sentiment-aligned Topic Model(SATM) in this section, and its graphical representation is shown in Figure 1. Note that the sentiment association H is observed, because it is trained independently of the topic model part.

At the word level, each observed phrase $< h, m >$ is associated with two latent variables: aspect z and sentiment label r. Aspect z models what aspect this phrase talks about, and r determines the sentiment of m. If m is in the sentiment lexicon, we assume r is also responsible for generating a word feature v_m, based on the sentiment association H, which is equal to its polarity score s_m in the lexicon. In this case, the observed data becomes $(< h, m >, v_m)$, and the latent sentiment label r is responsible for generating both word m, and word feature v_m. For example, for the phrase <room, filthy>, we observe a word feature $v = -5$, since the sentiment polarity score for the word "filthy" is -5. Given H, sentiment labels 1 or 2 are more likely to generate a word feature -5. Also, people tend to use "filthy" to express low ratings, like 1 or 2, so the sentiment labels and ratings can be aligned.

At the product level, for each product p and each aspect k, we define a multinomial distribution $\lambda_{p,k}$ over R sentiment labels. Since Y_p already gives us the big picture about the overall senti-

ment expressed on this product, we assume $\lambda_{p,k}$ is drawn from a dirichlet distribution $Dir(\boldsymbol{\pi_{p,k}})$ with asymmetric concentration parameters, where $\boldsymbol{\pi_{p,k}} = f(\boldsymbol{Y_p}, \boldsymbol{\omega_k}, \omega^b)$. We can use a linear parametrization, and set

$$f(\boldsymbol{Y_p}, \boldsymbol{\omega_k}, \omega^b) = \omega_k^1 \boldsymbol{Y_p} + \omega_k^0 + \omega^b \quad (1)$$

ω_k^1 captures the influence of the product overall rating distribution, and can favour certain sentiment labels in the prior. ω_k^0 and ω^b are the aspect-specific and corpus-level bias, respectively. Through this linear parametrization, we build a direct matching between sentiment label i and rating i. For example, for a product p, if its overall rating distribution Y_p has high probability over rating 4, for aspect k, we assume its product-aspect-sentiment label distribution also has high probability on sentiment label 4 in the prior. The actual aspect rating is affected by both the text which talks about aspect k, and also the prior.

To sum up, we assume the generative process as follows:

- For each aspect $k = 1, 2, \ldots K$,
  - draw an aspect-word distribution $\phi_k^a \sim Dir(\boldsymbol{\beta}^a)$
  - For each sentiment label $r = 1, 2, \ldots R$, draw an aspect-sentiment label-word distribution $\phi_{k,r}^s \sim Dir(\boldsymbol{\beta}^s)$
- For each product $p \in P$,
  - draw a product-aspect distribution $\theta_p \sim Dir(\boldsymbol{\alpha})$
  - for each aspect k, draw a product-aspect-sentiment label distribution $\lambda_{p,k} \sim Dir(\boldsymbol{\pi_{p,k}})$ where $\boldsymbol{\pi_{p,k}} = f(\boldsymbol{Y_p}, \boldsymbol{\omega_k}, \omega^b)$
- For each phrase $f =< h, m >$ of product p,
  1. Draw an aspect z from θ_p
  2. Draw a sentiment label r from $\lambda_{p,z}$
  3. Draw an aspect word h from ϕ_z^a
  4. Draw a sentiment word m from $\phi_{z,r}^s$. If $m \in L$, generate a word feature v_m based on H.

By integrating out $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$, the joint probability can be defined as:

$$P(\boldsymbol{z}, \boldsymbol{r}, \boldsymbol{h}, \boldsymbol{m}, \boldsymbol{v} | \boldsymbol{\alpha}, \boldsymbol{\beta^a}, \boldsymbol{\beta^s}, \boldsymbol{\pi}, H) = P(\boldsymbol{z}|\boldsymbol{\alpha}) P(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{\pi}) P(\boldsymbol{h}|\boldsymbol{z}, \boldsymbol{\beta^a}) P(\boldsymbol{m}|\boldsymbol{z}, \boldsymbol{r}, \boldsymbol{\beta^s}) P(\boldsymbol{v}|\boldsymbol{r}, H) \quad (2)$$

3.4 Inference

We use Gibbs Sampling(Griffiths and Steyvers, 2004) to estimate the posterior distribution given the observed data.

We jointly sample the aspect z and sentiment label r for the ith phrase $< h, m >$ of product p, given the assignments of other phrases:

$$P(z_i = k, r_i = l | \boldsymbol{z}_{-i}, \boldsymbol{r}_{-i}, \boldsymbol{h}, \boldsymbol{m}, \boldsymbol{v}) \propto (n_{p,k} + \alpha) \frac{n_{k,h}^a + \beta^a}{\sum_{h'} (n_{k,h'}^a + \beta^a)} \frac{n_{p,k,l} + \pi_{p,k,l}}{\sum_{l'} (n_{p,k,l'} + \pi_{p,k,l'})} \frac{n_{k,l,m}^s + \beta^s}{\sum_{m'} (n_{k,l,m'}^s + \beta^s)} g(m, l) \quad (3)$$

where $g(m, l) = H(v_m | l)$ if $m \in L$. In this case, when we sample the sentiment label r for this phrase, the probability of generating word feature v_m from r is also considered. For example, the word "excellent" has a word feature value $v_m = 5$. Based on H, the probability of generating a word feature 5 is higher for sentiment labels with larger values. If $m \notin L$, there is no $g(m, l)$ term, since no word feature is associated with this phrase. In Equation 3, $n_{p,k}$ is the number of times aspect k is assigned to phrases of product p, and $n_{k,h}^a$ is the number of times aspect word h is assigned to aspect k. $n_{p,k,l}$ is the number of times sentiment label l is assigned to aspect k for product p, and $n_{k,l,m}^s$ is the number of times sentiment word m is assigned to aspect k and sentiment label l. All these counts exclude assignments for the current phrase $< h, m >$.

Based on the samples, we can estimate $\lambda_{p,k,r}$ as:

$$\lambda_{p,k,r} = \frac{n_{p,k,r} + \pi_{p,k,r}}{\sum_{r'} (n_{p,k,r'} + \pi_{p,k,r'})} \quad (4)$$

Since sentiment labels and atings are aligned, the aspect rating t_{pk} of product p on aspect k can be simply calculated as the expectation of $\boldsymbol{\lambda_{p,k}}$:

$$t_{pk} = \sum_r \lambda_{p,k,r} \cdot r \quad (5)$$

4 Experiments

In this section, we describe the experiments and analyze the results.

4.1 Dataset

We use the TripAdvisor dataset[1](Wang et al., 2010) for evaluation, since in this dataset, reviews are not only associated with overall ratings, but also with ground truth aspect ratings on 7 aspects: *value, room, location, cleanliness, check in/front desk, service, business service*. All the ratings in the dataset are in the range from 1 star to 5 stars. We first remove reviews with any missing aspect ratings or very short reviews(less than three sentences). Then we adopt the dependency parser technique to identify opinion phrases, and collect phrases with adjective sentiment words. The dependency parser can deal with conjunctions, negations and bigram aspect words, and it results in the best performance according to (Moghaddam and Ester, 2012). Some sample phrases are shown in Table 1. All words are converted into lower case, and we remove phrases containing words that appear no more than 10 times or stop words. Since we are only interested in product-level aspect rating prediction, for each product, we aggregate all the review overall ratings to get the overall rating distribution. The statistics of the dataset is shown in Table 2. The average rating is the rating averaged over all reviews and all products. As we can see, positive reviews are dominant in the data, which raises the challenge of discovering negative sentiment topics.

Sentences	Phrases
The room, facing the courtyard, was large and comfortable.	<room, large>, <room, comfortable>
The room was not really clean.	<room, no_clean>
Internet access was available.	<Internet access, available>

Table 1: Sample extracted phrases

#Products	#Reviews	Avg rating	#Phrases
1850	61306	4.03	740982

Table 2: Statistics of the dataset

4.2 Metrics

We use three evaluation metrics for comparison.

RMSE: Root-mean-square error is used to measure the difference between the predicted aspect ratings and ground truth aspect ratings. It is defined as:

$$RMSE = \sqrt{\frac{\sum_p \sum_k (t_{pk} - \hat{t}_{pk})^2}{|P| \times K}} \quad (6)$$

where t_{pk} is the predicted aspect rating for product p on aspect k, and $\hat{t}_{pk}$ is the ground truth.

Precision@N: For each aspect k, we rank the hotels based on their predicted aspect ratings, and get the top N results. A hotel is considered *relevant* if its ground truth aspect rating is in the top 10% of the ground truth aspect ratings of all hotels. Precision@N is defined as the percentage of the top N results that are relevant:

$$Precision@N = \frac{|\{\text{relevant hotels}\} \cap \{\text{top N ranked hotels}\}|}{N} \quad (7)$$

We use $N = 10$, and the result is averaged over K aspects.

$\boldsymbol{\rho_{hotel}}$: Pearson correlation across hotels(Wang et al., 2010) is defined as:

$$\boldsymbol{\rho_{hotel}} = \frac{\sum_k \rho(\boldsymbol{t_k}, \hat{\boldsymbol{t}}_{\boldsymbol{k}})}{K} \quad (8)$$

where $\boldsymbol{t_k}$ is the predicted aspect rating vector for all hotels on aspect k, and $\hat{\boldsymbol{t}}_{\boldsymbol{k}}$ is the corresponding ground truth vector. $\rho(\boldsymbol{t_k}, \hat{\boldsymbol{t}}_{\boldsymbol{k}})$ is the Pearson correlation between these two vectors. It measures how the predicted ratings of aspect k can preserve the order in the ground truth(Wang et al., 2010). If we can predict an aspect-specific ranking similar to the ground truth, we can use the predicted aspect ratings to answer questions like "Is hotel a better than hotel b on aspect k?"

4.3 Baselines

The first three baselines are **Local Prediction**, **Global Prediction** and **Graph Propagation**. They all separate aspect extraction and sentiment identification. For each phrase $f =< h, m >$ from review d of product p, we first find the aspect assignment of this phrase. Then, we use three methods to get the phrase rating. Local Prediction(Lu et al., 2009) simply uses the overall rating of d as its phrase rating. Global Prediction(Lu et al., 2009) trains a multi-class classifier to classify the sentiment word m into a rating category $r \in 1, 2 \ldots R$,

[1] http://sifaka.cs.uiuc.edu/~wang296/Data/index.html

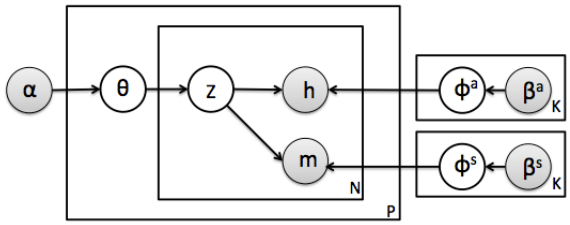

Figure 2: Method for aspect extraction in Local Prediction, Global Prediction and Graph Propagation

then assigns r as the phrase rating. Graph Propagation(Brody and Elhadad, 2010) builds a conjunction graph for sentiment words, and uses a Label Propagation algorithm on the graph to learn the sentiment polarity score for each sentiment word. The score of m is set as phrase rating. Finally, we aggregate all the phrases of each aspect to predict the aspect ratings. To apply these methods in our experiments, in the aspect extraction step, we adapt our model to extract only aspects, as shown in Figure 2. In this simplified model, no sentiment labels is involved, and the latent aspect explains both the aspect word and sentiment word.

ILDA(Moghaddam and Ester, 2011) was proposed for aspect rating prediction, but it fails to deal with the sentiment label alignment problem, so it cannot be directly used for this task. We adopt the common approach of providing seed words to set priors for each sentiment topic.

LRR(Wang et al., 2010) was proposed to predict aspect ratings for each review, but it can also be used to predict product aspect ratings by aggregating all the reviews of a product into a single "h-review"(Wang et al., 2011). First, we can run a topic model to learn aspects, and annotate each sentence with an aspect. Then LRR is applied on the annotated sentences to predict aspect ratings. This approach provided the best result, according to (Wang et al., 2011). In the first step we use the sentence-LDA(Jo and Oh, 2011) to annotate sentences, which is slightly different from the original method, but still provides a good analogy.

We also test two simplified version of the SATM model. First, we remove the part which involves sentiment lexicons, so we only use the product overall rating distribution. We call this method **SATM-O**. Second, we use only sentiment lexicons, ignoring the influence of overall rating distribution. We call it **SATM-L**. These two baselines can help us identify how the sentiment lexicon and overall rating distribution can improve the results, if used separately.

Our last baseline simply uses the overall rating of a hotel as its aspect ratings. For each hotel, its overall rating is defined as the average overall rating of its reviews. This method is referred to as **Overall**.

4.4 Experimental Setup

For all topic modelling based approaches, the number of aspects is set to 7. Since we can evaluate aspect rating prediction only on the predefined aspects, we need to ensure the discovered aspects match the predefined aspects. To do this, we adopt the common approach of providing a few seed words for each aspect as priors, as in (Wang et al., 2010). The seed words are listed in Table 3. There may be better methods to use seed words for aspect discovery (Jagarlamudi et al., 2012; Mukherjee and Liu, 2012), and it would be interesting to combine their methods with ours. However, this is beyond the scope of this paper, and we list it as future work.

Aspects	Seed words
Value	value, price, worth
Room	room, rooms
Location	location
Cleanliness	room, dirty, smelled, clean
Check in/front desk	staff
Service	service, breakfast, food
Business service	internet, wifi

Table 3: Seed words for aspect discovery

We use 5 sentiment labels in SATM, SATM-L and SATM-O, as this is the number of distinct ratings. The lexicon L used in our experiment is part of (Taboada et al., 2011) where words are associated with polarity scores in the range $[-5,-1] \cup [1,5]$. We observe that words with polarity score 1 and -1 express too weak sentiments, so we discard them in our experiment. To get training instances for sentiment association H, we treat each appearance of word $m \in L$ in the data as one training instance. The polarity score s_m is directly retrieved from L, and the sentiment label r_m is the overall rating of review d where m appears. This approach avoids the need for manual annotation of sentiment labels, and the annotation result captures the characteristics of the dataset. How-

ever, all training instances in a review will have the same sentiment label, which means that we assume all sentiment words in a review express the same sentiment, no matter what aspects they talk about. This is not true, thus will introduce noise to the training. To reduce noise, for words with positive polarity score, we ignore their appearance in reviews with rating 1 and 2, since we assume positive sentiment words rarely express negative sentiments, even if they appear in negative reviews. Therefore, $H(s_m|r_m) = 0$ for $r_m = 1, 2$ and s_m in the range $[2, 5]$. A similar method is used to deal with words with negative polarity score.

For Global Prediction, in (Lu et al., 2009), the prior for the multi-class classifier is uniform, while in our experiment, for product p, we used product overall rating distribution on r as the prior for rating category r, which achieves better results than uniform prior.

The Graph Propagation method requires a small set of sentiment words as seeds, from which the algorithm can learn sentiment score for other words. The method in (Brody and Elhadad, 2010) constructs these seed words based on morphology in an unsupervised way, and can only support two kinds of sentiment: positive and negative. In our experiment, since the sentiment lexicon is available, the sentiment seed words are from the lexicon, and we update the polarity score for those not in the lexicon.

For ILDA, since we need to provide seed words as priors for sentiment topics, we have two options, and we use both for experiment. First, we can employ the common approach of using two sentiment labels(R=2, positive and negative). Then, words with positive polarity scores in lexicon L are used as priors for the positive sentiment topic, and similarly words with negative polarity scores for negative sentiment topic. An alternative approach is to use 5 sentiment labels(R=5). It provides finer grained sentiment extraction, but raises the question of how to choose seed words for each sentiment topic. To do this, we use the full sentiment lexicon in (Taboada et al., 2011), where sentiment words have polarity score in the range of $[-5, -1] \cup [1, 5]$. We divide the lexicon, and use words with polarity score 4 and 5 as prior for the sentiment topic with label 5. Then, words with polarity score 2 and 3 are used for the sentiment topic with label 4, and so on.

For all topic modelling based approaches, we set the number of iterations for Gibbs Sampling to 3000, and take samples from the markov chain every 50 iterations after a burn-in period of 1000 iterations. In SATM and SATM-O, for all aspects k, we need to choose the parameters $\boldsymbol{\omega_k}$ and also w^b. We use a small portion of dataset with ground truth to choose the best value, and we set $\omega_k^1 = 20$, $w_k^0 = 0.01, w^b = 0$. Automatically learning these parameters are feasible. One possible option is to use stochastic EM sampling scheme, as in (Mimno and McCallum, 2008). For the LRR implementation[2], we use the default parameters included in the package, and train the model with seed words provided by the author(Wang et al., 2010).

4.5 Results

The experimental results are listed in Table 4. For RMSE, the smaller the better, while for the other two measures, the larger the better. Graph Propagation, ILDA and SATM-L do not use the overall ratings(except for training sentiment association H), so we group them together. Similarly we group Local Prediction, Global Prediction, SATM-O and SATM. The Overall method is a special baseline that does not do any aspect based prediction. For the LRR method, after the first step of sentence annotation, we notice that sentence-LDA fails to annotate the "h-review" of some hotel with all 7 aspects, mainly because these hotels are associated with less reviews. In this case, the LRR model will fail in the second step, so we do not include LRR in Table 4. Instead, we compared our method with LRR on a subset of products that comment on all aspects based on the sentence annotation. There are 1533 hotels in this subset, and the result is shown in Table 5. Note that our experimental results for LRR are far worse than those reported in the original paper(Wang et al., 2011). We believe this maybe due to different parameter settings, or due to the choice of different reviews.

We observe that SATM achieves the best RMSE value, i.e., it produces the most accurate aspect rating prediction. The Overall method does better in ranking all the hotels(ρ_{hotel}), but SATM is better at ranking top hotels($P@10$). When we compare the results of SATM with SATM-L and SATM-O, we find that the good performance of SATM is mainly due to the use of the overall rating distribution. On one hand, this is reasonable, since in-

[2] http://sifaka.cs.uiuc.edu/~wang296/Codes/LARA.zip

Sentiment label	Top sentiment words
1	old, dirty, worn, older, dark, stained, broken, dated, outdated, bad
2	small, tiny, little, noisy, single, double, uncomfortable, smaller, larger, narrow
3	large, double, big, mini, hard, main, huge, twin, single, jacuzzi
4	nice, comfortable, modern, clean, new, good, great, flat, big, comfy
5	large, huge, great, beautiful, big, lovely, separate, spacious, wonderful, excellent

Table 6: Top sentiment words for aspect "room" with different sentiment labels

Methods	RMSE	P@10	ρ_{hotel}
ILDA,R=2	1.202	0.30	0.193
ILDA,R=5	1.096	0.257	0.222
Graph Propagation	0.718	0.271	0.442
SATM-L	0.774	0.443	0.483
Local Prediction	0.572	0.486	0.761
Global Prediction	0.625	0.30	0.778
SATM-O	0.429	0.80	0.841
SATM	**0.384**	**0.814**	0.854
Overall	0.415	0.80	**0.863**

Table 4: Experimental results except LRR

Methods	RMSE	P@10	ρ_{hotel}
LRR	1.018	0.3	0.404
SATM	**0.373**	**0.829**	**0.849**

Table 5: Experimental comparison with LRR

tuitively aspect ratings usually do not diverge too far from the overall rating, especially for hotels with higher overall ratings. As we can see from the result of Overall, the overall rating has good correlation with aspect ratings, and using overall rating only is already a strong predictor for aspect ratings. Also, in most cases, methods using overall ratings(Overall and the four methods in the middle of Table 4) are better than others(first four methods). On the other hand, we should not rely only on the overall rating distribution. By incorporating the sentiment lexicon, for RMSE, SATM achieves 10% improvement over SATM-O and 7% improvement than Overall. Also, the overall rating may not always be a good aspect rating predictor, depending on the dataset.

To take a closer look at cases where the overall rating is not a good aspect rating predictor, we evaluate the RMSE on different subsets of hotels. We divide the hotels into different overall rating ranges: [1.2), [2,3), [3,4) and [4,5]. The results are shown in Table 7. Going from the [4,5] group to [1,2) group, the overall rating becomes less and less reliable to predict aspect ratings, and the gain of SATM increases compared to SATM-O and Overall. For a hotel with higher overall rating(good hotel), its aspect ratings are closer to the overall rating. This matches our intuition that good hotels are expected to be good on most aspects, if not on all aspects. For a hotel with average and lower overall rating, the average difference between aspect ratings and overall rating is larger. In this case, the overall rating can not tell us the whole story, which calls for aspect based prediction. Our method achieves the best RMSE gain on this group of hotels.

Methods	[1,2)	[2,3)	[3,4)	[4-5]
Local Prediction	0.789	0.772	0.621	0.456
Global Prediction	1.013	0.884	0.584	0.567
SATM-O	0.703	0.564	0.446	0.359
SATM	**0.606**	**0.494**	**0.394**	0.332
Overall	0.735	0.612	0.431	**0.320**

Table 7: RMSE on hotels with different overall rating ranges

4.6 Qualitative analysis

To provide a qualitative analysis, we can list the top words for the aspect-sentiment label-word distributions. In Table 6, we list them for the aspect "room", with 5 different sentiment labels. We observe that, as the sentiment label value increases, the sentiment topics express more and more positive sentiments. This means the sentiment labels and ratings are indeed aligned, so that we can use these sentiment labels to predict ratings.

5 Conclusion and future work

In this paper, we proposed a sentiment aligned topic model(SATM) for product aspect rating prediction. By incorporating the overall rating distribution and a sentiment lexicon, our SATM model can align sentiment labels with ratings. Experiments on a TripAdvisor dataset demonstrate the effectiveness of SATM on aspect rating prediction.

In SATM, for each product and each aspect, the multinomial distribution over sentiment labels has

prior parameterized by product overall rating distribution. We assume linear dependency, but it will be interesting to explore other dependencies. Another direction is to learn the parameters $\boldsymbol{\omega_k}$ automatically, so that $\boldsymbol{\omega_k}$ can be different for different k, capturing the influence of the overall rating on different aspects.

Finally, we assume each phrase is associated with one latent aspect. However, aspects may be correlated. For example, the phrase <room, filthy> gives us information about the aspect *room* and also the aspect $cleanliness$. To deal with this problem, we can relax the assumption that one phrase talks about one aspect, or we can model correlation among aspects.

Acknowledgments

This research is supported by NSERC Discovery Grant. The authors thank Dr. Maite Taboada for providing the sentiment lexicon.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 25–32, New York, NY, USA. ACM.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1171–1177. AAAI Press.

Sasha Blair-goldensohn, Tyler Neylon, Kerry Hannan, George A. Reis, Ryan Mcdonald, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Samuel Brody and Noemie Elhadad. 2010. 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. Association for Computational Linguistics.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malú Castellanos, and Riddhiman Ghosh. 2013. Exploiting domain knowledge in aspect extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1655–1667.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April.

Thomas Hofmann. 1999. 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. ACM.

Minqing Hu and Bing Liu. 2004. 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. ACM.

Jagadeesh Jagarlamudi, Hal Daumé, III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 204–213, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824, New York, NY, USA. ACM.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice H. Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA*.

Himabindu Lakkaraju, Chiranjib Bhattacharyya, Indrajit Bhattacharya, and Srujana Merugu. 2011. Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA*, pages 498–509.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1630–1639.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment analysis with global topics and local dependency. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*.

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 '09, pages 375–384, New York, NY, USA. ACM.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Chong Long, Jie Zhang, Minlie Huang, Xiaoyan Zhu, Ming Li, and Bin Ma. 2014. Estimating feature ratings through an effective review selection approach. *Knowl. Inf. Syst.*, 38(2):419–446.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. 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. ACM.

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 World Wide Web*, WWW '07, pages 171–180, New York, NY, USA. ACM.

David M. Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *the Conference on Uncertainty in Artificial Intelligence*, pages 411–418.

Samaneh Moghaddam and Martin Ester. 2011. 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. ACM.

Samaneh Moghaddam and Martin Ester. 2012. On the design of lda models for aspect-based opinion mining. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 803–812, New York, NY, USA. ACM.

Samaneh Moghaddam and Martin Ester. 2013. 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.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 339–348, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christina Sauper and Regina Barzilay. 2013. Automatic aggregation by joint modeling of aspects and values. *J. Artif. Int. Res.*, 46(1):89–127, January.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2011. Content models with attitude. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 350–358, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June.

Ivan Titov and Ryan McDonald. 2008a. 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. ACM.

Ivan Titov and Ryan T. McDonald. 2008b. A joint model of text and aspect ratings for sentiment summarization. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 308–316.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. 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. ACM.

Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. 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. ACM.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. 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. Association for Computational Linguistics.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University.