Unifying Topic, Sentiment & Preference in an HDP-Based Rating Regression Model for Online Reviews

Zheng Chen ZC86@DREXEL.EDU
College of Computing & Informatics, Drexel University, Philadelphia, PA, USA

Yong Zhang YCHANG@MAIL.CCNU.EDU.CN
Computer School of Huazhong Normal University, Wuhan, Hubei, China

Yue Shang YS439@DREXEL.EDU
College of Computing & Informatics, Drexel University, Philadelphia, PA, USA

Xiaohua Hu XH29@DREXEL.EDU
College of Computing & Informatics, Drexel University, Philadelphia, PA, USA

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

This paper proposes a new HDP based online review rating regression model named Topic-Sentiment-Preference Regression Analysis (TSPRA). TSPRA combines topics (product aspects), word sentiment and user preference as regression factors, and is able to perform topic clustering, review rating prediction, sentiment analysis and what we invent as "critical aspect" analysis altogether in one framework. TSPRA extends sentiment approaches by integrating the key concept "user preference" in collaborative filtering (CF) models into consideration, while it is distinct from current CF models by decoupling "user preference" and "sentiment" as independent factors. "Critical aspects" is defined as the product aspects seriously concerned by users but negatively commented in reviews. Improvement to such "critical aspects" could be most effective to enhance user experience.

Keywords: Topic Model, Rating Regression, Sentiment, Critical Aspect

1. Introduction

The rapidly growing amount of rated online reviews brings great challenges as well as needs for regression analysis on review ratings. It has been studied in various publications that factors like review texts, product aspects, word sentiments, aspect sentiments, user clusters, and product clusters, have been integrated in the regression analysis. Many recent review models [1], [2], [3], etc. adopt a topic modeling method Latent Dirichlet Allocation (LDA) [4] as the modeling framework. In LDA, product aspects involved in an online review are treated as topics. In future discussion, "topics" and "product aspects" are synonyms and equivalent.

In this paper we build a new review model *Topic-Sentiment-Preference Regression Analysis* (TSPRA) based on the HDP framework. In TSPRA, topics, sentiments and user preferences are considered as regression factors. TSPRA is driven by three major considerations. The first one is to develop an automatic approach to evaluate word sentiments. Until recently, many sentiment analyses still depend on manually annotated corpus for model training. To circumvent laborious annotating efforts and the potential danger that

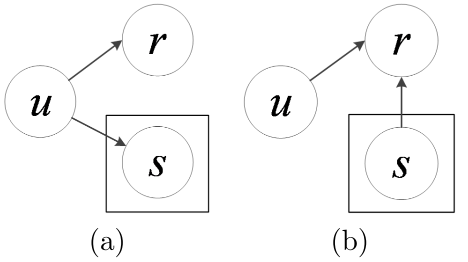


Figure 1: (a) user preference u in two CF models [8] and [9] is designed to cause both the review rating r and word/sentence sentiments s, leading to a strong correlation between u and s; (b) in our model preference u and sentiment s are designed as independent variables that co-determine the review rating.

artificial annotations are done by uninterested persons, exploitation of online reviews and their ratings, which are naturally annotated corpus, has become increasingly popular among researchers to conduct sentiment analysis.

The second consideration is to experiment on a new idea of redefining the concept ”user preferences” that have been addressed by previous collaborative filtering models, such as those discussed in section 2. There is no doubt user preference is a crucial and necessary factor in rating regression. For example, suppose we have a customer in a restaurant that cares more about taste rather than price. If the taste of the restaurant does not meet his expectation, most likely he would not give a good rating, no matter how good the price is. Otherwise he may give a positive rating even if the dinner is a little expensive. However, after careful examination of recent CF papers [1] [8] [9] [10] that include user preference in the modeling, we found the term is not clearly defined and it functions more like topic-level sentiment in the models.

We propose that it might be necessary to distinguish between user preference and sentiment. Intuitively, in the restaurant example, one’s preference in taste does not necessarily mean one must give the restaurant’s taste a high score. However, in CF models mentioned in section 2, user preferences act more like aspect-level sentiments because it is typically modelled as the sole type of factors that directly result in the overall ratings. For example, [8] models both word sentiments and review ratings as consequence of aspect ”interests”, therefore word sentiments exert effect on review ratings through the ”interests” and must have a strong correlation with the ”interests”. Likewise in [9] the aspect scores are designed to cause ratings and sentence sentiments, thus the sentence sentiments must have a strong correlation with the aspect scores. Such correlation contradicts with our intuition and in model design it might also reduce the amount of information that can be unfolded through statistical inference.

The third concern lies in the number of product aspects. If we have many heterogeneous datasets, it would be laborious to find out appropriate topic number for each dataset, while experiment results on topic number based on one or two datasets, as done in [1], might be unwarranted for other datasets. A further development of topic model, the HDP [5] where HDP stands for *Hierarchical Dirichlet Process*, help solve this issue. By taking advantage of the Dirichlet stochastic process we no longer need to specify the topic number for each dataset beforehand.

2. Related Work

Two main approaches are studied in recent publications regarding online review rating regression: the *collaborative filtering* (CF) approaches and the *sentiment* approaches. Collaborative filtering approaches are active research topics in rating regression and prediction. There are models not exploiting review contents, such as [11] [10], and they typically depend on factorization of the user-product rating matrix. More recent models [12] [13] [14] endeavor to incorporate texts as additional resource beyond ratings to add to their predictive power and recognize product aspects. In these models, each aspect is associated with "preference" scores and the overall review rating is viewed as the aggregation of these aspect scores. Further developments unify sentiments into the CF framework. A recent study [9] associates each aspect with a "rating distribution" and views sentence sentiments as the consequence of these rating distributions. Another recent work [8] call the aspect scores as "interests" and combines word sentiments into their model. Both models are similar to our model in comparison to other CF models mentioned above, however, on one hand the "preferences" or "rating distributions" or "interests" feel somewhat like aspect-level sentiment given that they measure if customers like or dislike a product aspect; on the other hand the main objective of CF models is still doing recommendation, therefore sentiment, especially word sentiment, is a subordinate unit of analysis. In all these CF models, [8] is the only CF model that can infer word sentiments.

The other perspective of rating regression is to consider review ratings as a direct consequence of word sentiments. [15] proposes a probabilistic rating regression model that requires aspect keyword supervision, and the follow-up studies [2] removes the keyword supervision by merging LDA into the model to perform keyword assignment. Both papers can be viewed as a further extension to the work dedicated to combine aspects and sentiments – see [16] [17] [3]. Intuitively a key difference between these "sentiment" models and the CF models with sentiment is that in "sentiment" models the review ratings directly result from the aggregation of word sentiments, while the CF approaches treat word sentiments and review ratings as consequences of per-aspect user preferences.

Many of the above-mentioned works are developed from the topic model framework, typically LDA [4], due to the fact that a reviewer usually comments on different aspects of a target item and the strong analogy between product aspects and topics. A direct extension to LDA is [5] which introduces a stochastic process name hierarchical Dirichlet process to infer number of topics from the data. Further improvements targeting rating regression include [18] and [19]. No matter "preference" or "sentiment", they are more meaningful with respect to a certain topic.

3. Model Construction

3.1. Model Specification

Our model consists of four parts: the topic part, the sentiment part, the user preference part and the rating regression part. The topic part is the same as HDP [5], and the rest are our extensions, as shown in Figure 2. These four parts together describe the generative process of observable data: review text and ratings. Due to space constraints the description of HDP (the dashed box of Figure 2) has to be excluded from this paper.

For the sentiment part, $\pi_{K \times V}$ are $K \times V$ sentiment distributions where S is the number of sentiment polarities and V is the vocabulary size. In this paper S is set 3 following the convention of sentiment analysis, indicating each word can be positive, negative or neutral. A word w under topic φ_k is associated with a three-dimensional sentiment distribution π_{kw}, such as 0.2, 0.4, 0.4 where the numbers are the probabilities of the word w being positive, negative or neutral under topic φ_k. We denote the latent sentiment index of word w_{di} as s_{di}, then s_{di} is sampled from the sentiment distribution $\pi_{z_{di} w_{di}}$. We let $s_{di} = -1, 0, +1$ represent negative, neutral and positive sentiment respectively.

The preference part is similar to the sentiment part, where $\psi_{K \times X}$ are $K \times X$ preference distributions over U strengths of preferences and X users. In this paper U is set 2 indicating user preference can be strong or weak. Now an author x under topic φ_k is associated with a two-dimensional preference distribution ψ_{kx}, such as 0.2, 0.8 where the numbers are the probabilities of the author x having strong or weak preference to topic φ_k. We denote the author of review d by x_d, then there is also a preference index u_{di} associated with each word w_{di}, which is sampled from $\psi_{z_{di} x_d}$. We let $u_{di} = 0, 1$ represent strong and weak preference respectively.

Rating regression part is the generative process of the review rating r_d of document d (the rating of an online review is typically a 1-5 scale). For each word, when the latent variables u_{di} and s_{di} are given, the rating variable r_{di} is determined by the association rule in Table 1. In the association rule, μ $(1 \leq \mu \leq 5)$ is a parameter of our model named "neutral rating". For a review rating of 1-5 scale, one might think the middle value $\mu = 3$ is the neutral rating. In fact, this might not be true. We will see later in section 4 that in real world users tend to give positive ratings, and the experiment shows a better μ is larger than 3.

Once all word ratings are determined, the review rating r_d is drawn according to $r_d \sim N\left(r_{di},\ \sigma^2\right)$, where r_{di} is the mean of non-neutral word ratings (i.e. average of word ratings excluding neutral words), and σ^2 is the rating noise, and $N\left(r_{di},\ \sigma^2\right)$ is the normal distribution with r_{di} as mean and σ^2 as variance.

Comparing to HDP, our model adds three more latent variables for each word s_{di}, u_{di} and r_{di}, where r_{di} is generated by association rules given s_{di} and u_{di}. For counting variables, marginal counts are represented with dots. Thus $n_{dt\bullet}$ represents the number of customers in restaurant d at table t, $n_{d\bullet k}$ represents the number of customers in restaurant d eating dish k, $m_{d\bullet}$ represents the number of tables in restaurant d, $m_{\bullet k}$ represents the number of tables serving dish k, and $m_{\bullet\bullet}$ is the total number of tables occupied.

- If $u_{di} = 1$ (strong preference) and $s_{di} = -1$ (negative sentiment), then $r_{di} = 1$.
- If $u_{di} = 0$ (weak preference) and $s_{di} = -1$, then $r_{di} = \frac{1+\mu}{2}$.
- If $u_{di} = 1$ and $s_{di} = +1$ (positive sentiment), then $r_{di} = 5$.
- If $u_{di} = 0$ and $s_{di} = +1$, then $r_{di} = \frac{5+\mu}{2}$
- Otherwise $r_{di} = \mu$. This rating is named neutral word rating, and a word with neutral word rating is named a neutral word.

Table 1: The association rules for rating

- Define the topic prior H for global topic generation as a V-dimensional Dirichlet distribution $\mathrm{Dir}(\beta)$, where V is the vocabulary size. Define G_0 as the global random probability measure distributed according to Dirichlet process $\mathrm{DP}(H, \gamma)$ where γ is the concentration parameter.
- For each topic k
  - Draw a topic distribution over vocabulary $\varphi \sim \mathrm{Dir}(\beta)$
  - For each word w in the vocabulary, draw a distribution over 3-dimensional sentiment $\pi_{kw} \sim \mathrm{Dir}\,(\lambda)$.
  - For each author x, draw a user preference distribution over 2-dimensional strength $\psi_{kx} \sim \mathrm{Beta}(\eta)$.
- For each review d authored by the user x_d
  - Draw a local random probability measure G_d from G_0 distributed according to Dirichlet process $\mathrm{DP}(G_0, \alpha)$ where α is the concentration parameter.
  - For each ith word w_{di}
    - Draw a table t_{di} and a topic z_{di} from G_d and G_0 according to CRF representation of HDP
    - Draw the word $w_{di} \sim \mathrm{Multinomial}(\varphi_{z_{di}})$
    - Draw a sentiment $s_{di} \sim \mathrm{Multinomial}(\pi_{z_{di}w_{di}})$
    - Draw a user preference of author x_d given topic $\varphi_{z_{di}}$, $u_{di} \sim \mathrm{Binomial}(\psi_{z_{di}x_d})$
  - Compute a latent variable r_{di} and sample r_d according to the rules as shown in Table 1.

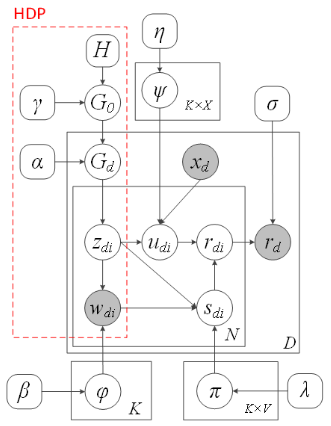
Table 2: The generative process of our regression model, Topic-Sentiment-Preference Regression Analysis (TSPRA).

The above is the complete setup of the generative process assumed by our model, illustrated by the plate diagram in Figure 2, and we summarize the generative process as the following in Table 2,

3.2. Inference

In training, the variables in our model TSPRA are estimated by a collection of reviews and their ratings. We adopt the Gibbs sampling [5] [20] to obtain latent variable estimation under the CRF representation of HDP and Dirichlet priors.

Based on CRF and the generative process illustrated in Figure 2, we have six sets of latent variables: $\mathbf{z}$, $\mathbf{t}$, $\mathbf{k}$, $\mathbf{s}$, $\mathbf{u}$ and $\mathbf{r}$ – the topic indexes, the table indexes of all words, the topic indexes of all tables, the word sentiments, the per-topic user preferences and word ratings. Note that $\mathbf{z}$ is determined by $\mathbf{t}$ and $\mathbf{k}$ as discussed in section 3.1, and $\mathbf{r}$ is determined by $\mathbf{s}$ and $\mathbf{u}$ using rules in Table 1, therefore the whole state space are determined by $(\mathbf{t}, \mathbf{k}, \mathbf{r})$. The conditional distributions of this state space are iteratively re-sampled by the Gibbs sampler described below. Provided a sufficient number of re-sample iterations, the final estimation of latent variables can be viewed as a sample generated from the process in Figure 2, for detailed discussion, please refer to [21].



Constants	
D	The number of reviews.
N	The number of words in review d.
S	The number of sentiment polarities.
X	The number of authors.
U	The number of types of user preference.
K	The number of topics, which is dynamic.
H	The Dirichlet prior for topics
α, γ	The concentration parameters
β, η, λ	The hyper parameters of Dirichlet priors of topics, preferences and sentiments.
σ	The rating noise.

Latent Variables	
G_0	A global random probability measure where topics are sampled, analogous to the “franchise” in CRF representation.
G_d	A random probability measure where the topic of each word in document d is sampled, analogous to a “restaurant” in CRF representation.
π_{kw}	The sentiment distribution associated with the word w under the kth topic.
φ_k	The kth topic.
ψ_{kx}	The user preference distribution associated with the topic k and user x.
z_{di}	The latent topic index for word i in review d.
s_{di}	The latent sentiment index for word i in review d.
u_{di}	The user preference index for word i in review d.
r_{di}	The rating of word i in review d.
t_{di}	The table index of word i in review d.
k_{dt}	The topic index of table t in review d.

Counting Variables	
n_{dtk}	The number of customers (words) in restaurant (document) d at table t eating dish (topic) k
m_{dk}	The number of tables in restaurant d serving dish k
l_{kws}	The number of times that customer (word) w is served by disk (topic) k and labeled by sentiment s
l_{kw}	The number of times that word w labeled by sentiment s in review d
c_{kx0}	The number of times that author x has a weak preference towards topic k
c_{kx1}	The number of times that author x has a strong preference towards topic k

Observations	
w_{di}	The i-th word in review d.
r_d	The rating (1-5) of review d.
x_d	The author of review d.

Figure 2: The graphic presentation of our regression model, Topic-Sentiment-Preference Regression Analysis (TSPRA).

***Re-sample* t.** We first re-sample the table index t_{di} of each word w_{di} according to equation (1) where the tilde ”∼” denotes ”all other variables and parameters”.

$$p\left(t_{di} = t \mid \mathbf{t}^{-t_{di}}, \sim\right) \propto \begin{cases} n_{dt\bullet}^{-di} f_{k_{dt}}^{-w_{di}}(w_{di},\ r_d | z_{di} = k_{dt}, \sim) & t \text{ is previously used} \\ \alpha p\left(w_{di},\ r_d \mid t = t_{new}, \sim\right) & t = t_{new} \end{cases} \tag{1}$$

In the situation when a word chooses an existing table, i.e. ”t is previously used”, the topic of the chosen table is $\varphi_{k_{dt}}$, then the likelihood of observation w_{di} given z_{di} and other variables is estimated using equation (2).

$$\begin{aligned} f_{k_{dt}}^{-w_{di}}\left(w_{di}, r_d | z_{di} = k_{dt}, \sim\right) &= \sum_{u_{di}} \sum_{s_{di}} p(w_{di}, r_d, u_{di}, s_{di} | z_{di}, \sim) \\ = \sum_{u_{di}} \sum_{s_{di}} p\left(w_{di} \mid z_{di}, \sim\right) p(u_{di}&|z_{di}, \sim) p\left(s_{di} \mid z_{di}, w_{di}, \sim\right) p\left(r_d \mid \overline{r_{di}}, \sim\right) \end{aligned} \tag{2}$$

In the situation when a customer chooses a new table, a topic needs to be sampled for the new table using equation (3).

$$p\left(k_{dt^{new}} = k \,|\boldsymbol{\sim}\right) \propto \begin{cases} m_{\bullet k} f_k^{-w_{di}}(w_{di},\ r_d|\varphi_k, \sim) & k \text{ is previously used} \\ \gamma f_{k^{new}}^{-w_{di}}(w_{di},\ r_d|\varphi_{k^{new}}, \sim) & \qquad k{=}k^{new} \end{cases} \tag{3}$$

Hence in equation (1) the likelihood of observing (w_{di}, r_d) given the new table, which is denoted by $p\left(w_{di}, r_d \mid t_{di} = t_{new}, \boldsymbol{\sim}\right)$, can be estimated by equation (4).

$$\begin{aligned} p\left(w_{di}, r_d \mid t_{di} = t_{new}, \boldsymbol{\sim}\right) = \sum_{k=1}^{K} & \frac{m_{\bullet k}}{m_{\bullet\bullet} + \gamma} f_k^{-w_{di}}\left(w_{di},\ r_d|\varphi_k, \sim\right) \\ + & \frac{\gamma}{m_{\bullet\bullet} + \gamma} f_{k^{new}}^{-w_{di}}\left(w_{di},\ r_d|\varphi_{k^{new}}, \sim\right) \end{aligned} \tag{4}$$

The rest of our inference for re-sample of table index t_{di} is how to estimate the likelihood $f_k^{-w_{di}}(w_{di}, r_d|\varphi_k, \sim)$. By equation (2) it is not hard to derive equation (5), in which for notation clarity we simplify w_{di} as w, s_{di} as s, and u_{di} as u.

$$f_k^{-w}\left(w, r_d \mid \varphi_k, \sim\right) \propto \frac{l_{kw} + \beta}{l_{k\bullet} + V\beta} \times \sum_u \sum_s \frac{c_{kxu} + \eta}{c_{kx\bullet} + U\eta} \times \frac{l_{kws} + \lambda}{l_{kw\bullet} + S\lambda} \times e^{-\frac{\left(r_d - \overline{r_{di}}\right)^2}{2\sigma^2}} \tag{5}$$

A special case when $k = k^{new}$ is given by equation (6).

$$f_{k^{new}}^{-w}(w, r_d|\varphi_{k^{new}}, \sim) \propto \frac{1}{V} \times \frac{1}{U} \times \frac{1}{S} \times \sum_u \sum_s e^{-\frac{\left(r_d - \overline{r_{di}}\right)^2}{2\sigma^2}} \tag{6}$$

***Re-sample* k**. We then re-sample the topic index k_{dt} of each table t in review d by equation (7). Note in this step all words associated with table t might switch topic. In the situation when an existing topic is chose for table t in review d, the likelihood of observation $(\mathbf{w}_{dt}, r_d)$ given topic φ_k is estimated by equation (8).

$$p\left(k_{dt} = k \,\middle|\, \mathbf{k}^{-k_{dt}}, \boldsymbol{\sim}\right) \propto \begin{cases} m_{\bullet k} f_k^{-\mathbf{w}_{dt}}(\mathbf{w}_{dt},\ r_d|\varphi_k, \sim) & k \text{ is previously used} \\ \gamma f_{k^{new}}^{-\mathbf{w}_{dt}}(\mathbf{w}_{dt},\ r_d|\varphi_{k^{new}}, \sim) & \qquad k{=}k^{new} \end{cases} \tag{7}$$

The same notation simplification applied to equation (5) also applies here, i.e. we simplify $\mathbf{w}_{dt}$ as $\mathbf{w}$, $\mathbf{s}_{dt}$ as $\mathbf{s}$, u_{dt} as u, and s_{di} as s.

$$f_k^{-\mathbf{w}}\left(\mathbf{w},\ r_d \mid \varphi_k, \sim\right) \propto \left(\prod_{w \in \mathbf{w}} \frac{l_{kw} + \beta}{l_{k\bullet} + V\beta}\right) \times \sum_u \left(\frac{c_{kxu} + \eta}{c_{kx\bullet} + U\eta} \sum_{\mathbf{s}} \prod_{w \in \mathbf{w}} \frac{l_{kws} + \lambda}{l_{kw\bullet} + S\lambda} \times e^{-\frac{\left(r_d - \overline{r_{di}}\right)^2}{2\sigma^2}}\right) \tag{8}$$

Equation (8) is way too complicated to be evaluated in full, especially for the part summing over $\mathbf{s}$, which leads to a sum of $S^{|\mathbf{w}|}$ terms. We might instead use the following approximation.

$$f_k^{-\mathbf{w}}\left(\mathbf{w},\ r_d \mid \varphi_k, \sim\right) \propto \left(\prod_{w \in \mathbf{w}} \frac{l_{kw} + \beta}{l_{k\bullet} + V\beta}\right) \times \frac{c_{kxu} + \eta}{c_{kx\bullet} + U\eta} \times \prod_{w \in \mathbf{w}} \frac{l_{kws} + \lambda}{l_{kw\bullet} + S\lambda} \tag{9}$$

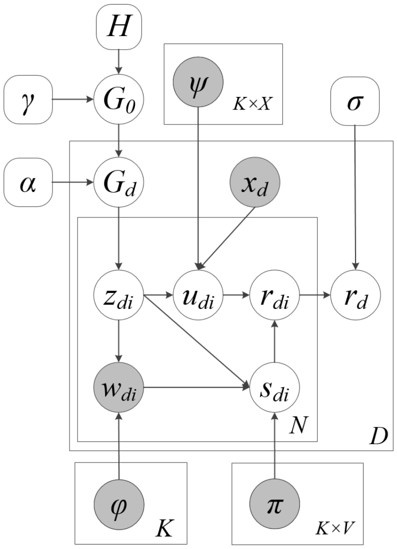


Figure 3: The graphic presentation of our prediction model.

A special case when $k = k^{new}$ is given by equation (10).

$$f_{k^{new}}^{-\mathbf{w}}\left(\mathbf{w},\ r_d \mid \varphi_k, \sim\right) \approx \frac{1}{V^{|\mathbf{w}|}} \times \frac{1}{U} \times \frac{1}{S^{|\mathbf{w}|}} \tag{10}$$

***Re-sample* r (*and* s, u).** In this step we re-sample word ratings r_{di} of each word w_{di}.

$$p\left(r_{di} = r \,\middle|\, \mathbf{r}^{-r_{di}}, \sim\right) \propto \frac{c_{kxu} + \eta}{c_{kx\bullet} + U\eta} \times \frac{l_{kws} + \lambda}{l_{kw\bullet} + S\lambda} \times e^{-\frac{\left(r_d - \overline{r_{di}}\right)^2}{2\sigma^2}} \tag{11}$$

The Gibbs sampling runs for a sufficient number of iterations to repeatedly re-sample each conditional distribution of the state space $(\mathbf{t}, \mathbf{k}, \mathbf{r})$ according to equations (1) to (11). When all iterations are complete, we can estimate the following parameters.

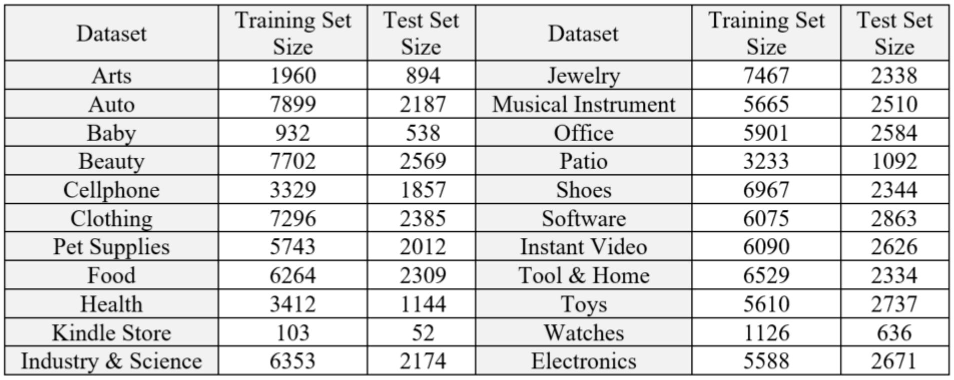
$$\varphi_{kw} = \frac{l_{kw} + \beta}{l_{k\bullet} + V\beta},\ \pi_{kws} = \frac{l_{kws} + \lambda}{l_{kw\bullet} + S\lambda},\ \ \psi_{kxu} = \frac{c_{kxu} + \eta}{c_{kx\bullet} + U\eta} \tag{12}$$

3.3. Prediction

A regression model's performance is commonly evaluated by its ability to predict, so is our regression model. In prediction, our model is to estimate each rating r_d of review d by Gibbs sampling similar to what we described in section 3.2, however, this time r_d is no longer observed, and it needs to be summed out. Meanwhile, the topics (φ), the word sentiments (π) and user preference (ψ) use the inference results of training, and are treated as known.

To sum out r_d, equation (5) is modified as equation (13), where φ_{kw}, ψ_{kxu}, π_{kws} come from the model trained in section 3.2 by equation (12). As shown as Figure 3. We do not modify the architecture of the sampling process but just replace some marginal density estimations to make use of inference results from training.

$$f_k^{-w}\left(w \mid \varphi_k, \sim\right) \propto \sum_{r_d} \sum_{u} \sum_{s} \varphi_{kw} \psi_{kxu} \pi_{kws} \times e^{-\frac{\left(r_d - \overline{r_{di}}\right)^2}{2\sigma^2}} \tag{13}$$



Dataset	Training Set Size	Test Set Size	Dataset	Training Set Size	Test Set Size
Arts	1960	894	Jewelry	7467	2338
Auto	7899	2187	Musical Instrument	5665	2510
Baby	932	538	Office	5901	2584
Beauty	7702	2569	Patio	3233	1092
Cellphone	3329	1857	Shoes	6967	2344
Clothing	7296	2385	Software	6075	2863
Pet Supplies	5743	2012	Instant Video	6090	2626
Food	6264	2309	Tool & Home	6529	2334
Health	3412	1144	Toys	5610	2737
Kindle Store	103	52	Watches	1126	636
Industry & Science	6353	2174	Electronics	5588	2671

Table 3: Training sets and test sets.

In the special case when $k = k^{new}$, the equation is given by (14).

$$f_{k^{new}}^{-w}\left(w, r_d \mid \varphi_{k^{new}}, \sim\right) \propto \frac{1}{V} \times \frac{1}{U} \times \frac{1}{S} \times \int \sum_{u} \sum_{s} e^{-\frac{\left(r_d - \overline{r_{di}}\right)^2}{2\sigma^2}} dr_d \tag{14}$$

Equation (9) is modified as

$$f_k^{-\mathbf{w}}\left(\mathbf{w} \mid \varphi_k, \sim\right) \approx f_k^{-\mathbf{w}}\left(\mathbf{w}, \mathbf{r}_d \mid \varphi_k, \sim\right) \propto \left(\prod_{w \in \mathbf{w}} \varphi_{kw}\right) \times \psi_{kxu} \times \prod_{w \in \mathbf{w}} \pi_{kws} \tag{15}$$

Equation (11) is modified as

$$p\left(r_{di} = r \;\middle|\; \mathbf{r}^{-r_{di}}, \sim\right) \propto \psi_{kxu} \times \pi_{kws} \times \int e^{-\frac{\left(r_d - \overline{r_{di}}\right)^2}{2\sigma^2}} dr_d \tag{16}$$

Other equations of the prediction model's Gibbs sampling is the same as the regression model's Gibbs sampling described in section 3.2. When all iterations complete, we predict r_d as the average of all non-neutral word ratings (17).

$$\widehat{r_d} = \overline{r_{di}} \tag{17}$$

4. Experiments & Evaluations

In this section, we describe the experiments and evaluate the rating prediction performance, effects of parameters, performance of sentiment analysis, and per-aspect sentiment/user preference. Each dataset is a collection of reviews with each review associated with an author ID, a product ID and a rating. The rating is a 1-5 scale to indicate the customer's satisfactory level for the purchase. We put 80% of an author's reviews in the training set and the remaining reviews in the test set. To make user preference effective, we require that each author has at least 3 reviews in the training sets and 1 review in the test set. For datasets that have too many reviews meeting this requirement, we limit the training set to contain no more than 8000 reviews. The size of each training set and test set are shown in Table 3. The Amazon datasets provide realistic reviews for 22 different products, and we expect a comprehensive experiment on them would be convincing. We run the Gibbs sampling for 2000 iterations. We run FLAME for 2000 iterations as well. For the setup of FLAME, we obtained the code from the original author, which is already setup based on the recommended settings described by the section 4.1 and 4.2 of their paper.

Based on experiments of section 4.4 we choose $\mu = 3.5, \sigma^2 = 0.08$. The concentration parameters are set to $\gamma = 1.5$,$\alpha = 1.0$,$\beta = 0.5$,$\eta = 0.5$,$\lambda = 0.5$ following the default values in

[4] [5]. Before the experiment, the review texts are pre-processed to remove punctuations, stop words and lemmatize using Stanford NLP [22].

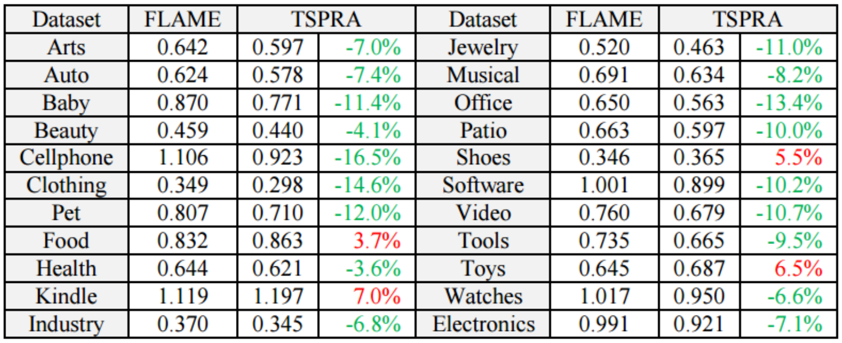
4.1. Prediction Performance Evaluation

For rating prediction we compare our model against a latest state-of-art model FLAME [9]. The performance measures we consider include

1. Absolute error – the absolute value of true rating minus prediction, which is a very straightforward measure. The result is shown in Table 4, indicating our model reduces error in 18 of the 22 datasets, 9 of them achieves more than 10% reduction and with all datasets considered the average reduction is 6.7%.

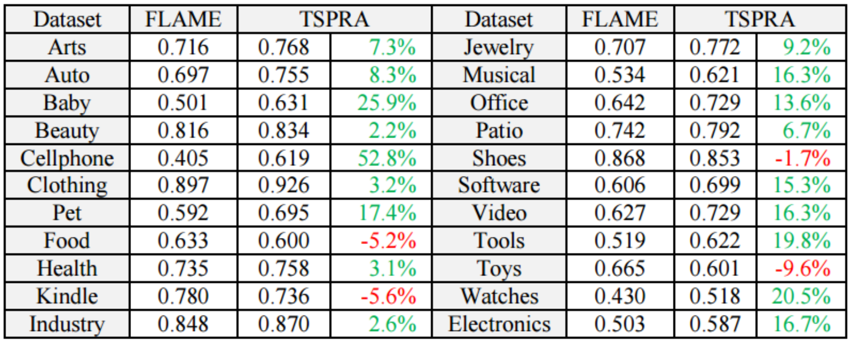
2. Pearson's correlation – measures how well the predicted ratings correlate with their corresponding true ratings. For two reviews with true ratings $r_1 > r_2$, a model with a higher correlation with the true values is in general less likely to give predictions $\hat{r}_1 < \hat{r}_2$. The results in Table 5 show our model performs well in this measure, and outperforms FLAME more than 15% on 10 of the 22 datasets.

3. Number of inverted pairs – counts the actual number of mis-ordered pairs of predicted ratings, i.e. the number of pairs of reviews such that $r_1 > r_2$ but $\hat{r}_1 < \hat{r}_2$. The results in Table 6 shows our model achieves mild better performance in this measure.



Dataset	FLAME	TSPRA		Dataset	FLAME	TSPRA	
Arts	0.642	0.597	-7.0%	Jewelry	0.520	0.463	-11.0%
Auto	0.624	0.578	-7.4%	Musical	0.691	0.634	-8.2%
Baby	0.870	0.771	-11.4%	Office	0.650	0.563	-13.4%
Beauty	0.459	0.440	-4.1%	Patio	0.663	0.597	-10.0%
Cellphone	1.106	0.923	-16.5%	Shoes	0.346	0.365	5.5%
Clothing	0.349	0.298	-14.6%	Software	1.001	0.899	-10.2%
Pet	0.807	0.710	-12.0%	Video	0.760	0.679	-10.7%
Food	0.832	0.863	3.7%	Tools	0.735	0.665	-9.5%
Health	0.644	0.621	-3.6%	Toys	0.645	0.687	6.5%
Kindle	1.119	1.197	7.0%	Watches	1.017	0.950	-6.6%
Industry	0.370	0.345	-6.8%	Electronics	0.991	0.921	-7.1%

Table 4: Rating Prediction Performance by Error.
(lower values indicate better performance)



Dataset	FLAME	TSPRA		Dataset	FLAME	TSPRA	
Arts	0.716	0.768	7.3%	Jewelry	0.707	0.772	9.2%
Auto	0.697	0.755	8.3%	Musical	0.534	0.621	16.3%
Baby	0.501	0.631	25.9%	Office	0.642	0.729	13.6%
Beauty	0.816	0.834	2.2%	Patio	0.742	0.792	6.7%
Cellphone	0.405	0.619	52.8%	Shoes	0.868	0.853	-1.7%
Clothing	0.897	0.926	3.2%	Software	0.606	0.699	15.3%
Pet	0.592	0.695	17.4%	Video	0.627	0.729	16.3%
Food	0.633	0.600	-5.2%	Tools	0.519	0.622	19.8%
Health	0.735	0.758	3.1%	Toys	0.665	0.601	-9.6%
Kindle	0.780	0.736	-5.6%	Watches	0.430	0.518	20.5%
Industry	0.848	0.870	2.6%	Electronics	0.503	0.587	16.7%

Table 5: Rating Prediction Performance by Pearson Correlation
(higher values indicate better performance).

Dataset	FLAME	TSPRA		Dataset	FLAME	TSPRA	
Arts	495	481	-2.8%	Jewelry	2326	2275	-2.2%
Auto	2362	2273	-3.8%	Musical	1468	1451	-1.2%
Baby	338	325	-3.8%	Office	1834	1769	-3.5%
Beauty	1757	1711	-2.6%	Patio	805	829	3.0%
Cellphone	2196	2083	-5.1%	Shoes	1745	1851	6.1%
Clothing	1550	1366	-8.8%	Software	4108	3994	-2.8%
Pet	1799	1749	-2.8%	Video	2764	2716	-1.7%
Food	1727	1756	1.7%	Tools	1753	1735	-1.0%
Health	745	743	-0.3%	Toys	1960	2083	6.3%
Kindle	32	36	12.5%	Watches	537	510	-5.0%
Industry	1072	1010	-5.8%	Electronics	2402	2391	-0.5%

Table 6: Rating Prediction Performance by Inverted Pairs
(lower values indicate better performance)

We notice our model performs best in terms of error reduction on datasets cellphone (-16.5%), clothing (-14.6%), office (-13.4%) and pet (-12.0%). It has been discussed earlier in section 1 that decoupling sentiments and user preferences might bring potential performance improvement and make the model capable of discovering ”critical aspects”, those highly preferred aspects with relatively low sentiments. Later in section 4.3 we demonstrate the discovered obvious critical aspects in these four data sets.

4.2. Sentiment Analysis

In this section we compare the our model's sentiment polarity evaluations to SenticNet, which is first introduced in [23]. It is a public sentiment resource built by non-probabilistic approaches, where each word is also associated with a polarity value ranging from -1 to 1 to quantify from the most negative sentiment to the most positive sentiment. The latest version is SenticNet3 [24]. Due to the unavailability of sentiment analysis results of previous review models, we opt to compare our results with this public sentiment resource.

As explained in section 3.1 and 3.2, our model infers each word's sentiment distribution and stores them in matrix $\pi_{K"V}$. Recall that an element π_{zw} of this matrix is a three dimensional categorical distribution, indicating the probability of word w being positive (let's denote it as π^{+}_{zw}), negative (denoted as π^{-}_{zw}) or neutral (denoted as π^{0}_{zw}) under topic z. As before the noisy neutral probability is ignored and the polarity value of word w is derive by formula (18).

$$\text{polarity}(w) = \frac{\sum_{z=1}^{K} (\pi^{+}_{zw} - \pi^{-}_{zw})}{\sum_{z=1}^{K} (\pi^{+}_{zw} + \pi^{-}_{zw})} \tag{18}$$

The vocabulary of the review texts has a 2524-word intersection with SenticNet3. The polarity distributions of this intersection are shown in Figure 4. From Figure 4 we see most of Sentic3 polarities are squeezed around the neutral polarity 0, while polarity estimations by our model TSPRA are more evenly distributed. Our estimations are right-skewed, which is caused by the unbalanced amount of 5-star review ratings in the datasets, see Table 3. In the future we might want to perform a balanced sampling from the datasets and re-run the experiments.

We demonstrate the top 20 positive words and top 20 negative word identified by our model in Table 7 with comparison of their polarity values to SenticNet3 values, and our estimations are clearly a better match to our intuition in the context of review ratings.

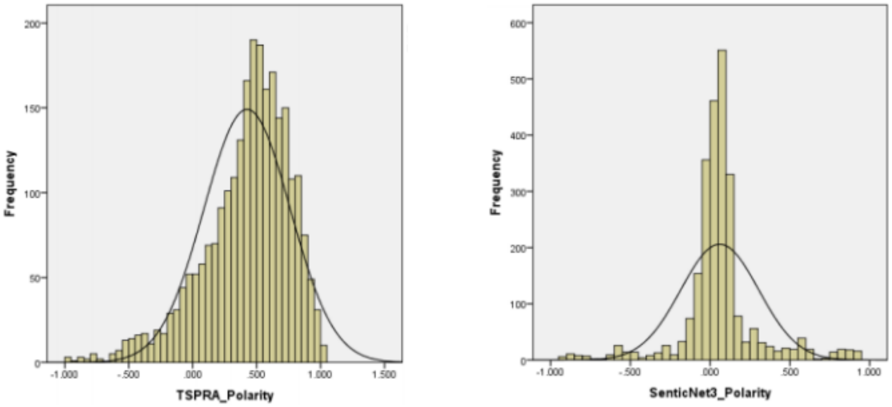


Figure 4: Polarity distributions of our model and SenticNet3.

	TSPRA	SNet3		TSPRA	SNet3
great	0.99	0.86	dreadful	-1.00	-0.29
love	0.99	0.32	rude	-1.00	-0.84
tasty	0.99	0.05	piss	-0.95	-0.04
perfect	0.98	0.58	supervisor	-0.90	0.04
best	0.98	0.22	unreliable	-0.88	0.13
delicious	0.97	0.11	worst	-0.86	-0.12
excellent	0.97	0.50	disappointment	-0.84	-0.06
awesome	0.97	0.78	disappointed	-0.80	-0.17
easy	0.96	0.27	waste	-0.78	-0.06
beautiful	0.95	0.15	awful	-0.78	0.12
well	0.95	0.72	poorly	-0.77	-0.07
nice	0.95	0.16	terrible	-0.76	-0.90
yummy	0.94	0.06	useless	-0.74	-0.26
pleased	0.94	0.45	return	-0.64	0.14
amazing	0.94	0.36	nightmare	-0.62	-0.10
comfort	0.94	0.33	horrible	-0.60	-0.94
wonderful	0.94	0.81	disgust	-0.59	-0.29
classy	0.93	0.10	overprice	-0.56	0.04
comfy	0.93	0.15	error	-0.56	-0.27
happy	0.92	0.30	worse	-0.56	0.13

Table 7: Top 20 positive words and top 20 negative words identified by our model are chosen for comparison.

For example, ”delicious”, ”tasty”, ”comfy” are expressing very positive sentiments in the context of product reviews (mostly food, we think), for which SenticNet3 only assigns quite unreasonable near-zero neutral values 0.05, 0.11, and 0.06. For another example, SenticNet3 assigns the obvious negative word ”unreliable” and ”awful” positive values 0.12, 0.12, while by contrast our model assigns -0.88, -0.78. The experiment of this section confirms that our model is also a good sentiment analysis model in addition to its ability of rating prediction.

4.3. Critical Aspects

In our model TSPRA, the $K \times X$ matrix ψ stores the inferred users' preferences distribution over product aspects. Recall that an element ψ_{zx} of this matrix is a 0-1 distribution, indicating the probability of user x (let's denote it as ψ_{zx}^{+}) having strong preference to product aspect z, and the probability of weak preference to product aspect z. Thus the average aspect preference of a product aspect z can be calculated by formula (19), and its value ranges from 0 to 1.

$$\text{preference}(z) = \frac{\sum_{x=1}^{X} \psi_{zx}^{+}}{X} \tag{19}$$

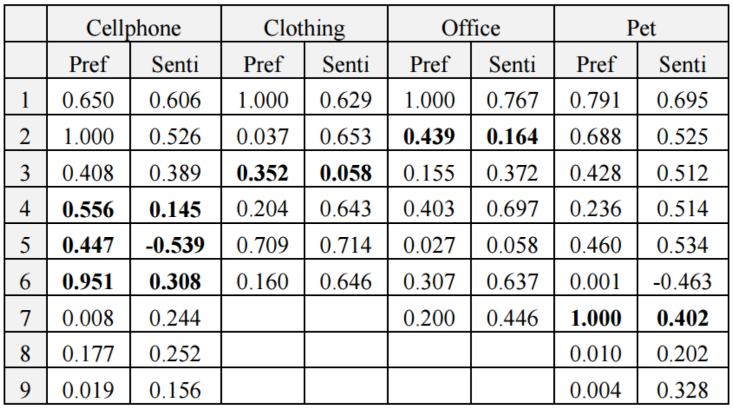
Similar to formula (18) in section 4.2, formula (20) is used to calculate aspect sentiment and its value ranges from -1 to 1.

$$\text{sentiment}(z) = \frac{\sum_{w=1}^{V} (\pi_{zw}^{+} - \pi_{zw}^{-})}{\sum_{w=1}^{V} (\pi_{zw}^{+} + \pi_{zw}^{-})} \tag{20}$$

We then identify those aspects with minimum preference 0.3 and preference-sentiment ration being negative or larger than 2. These aspects are concerned by users yet quite negatively or less positively commented. We define these aspects as ”*critical aspects*”, meaning that business decision makers should really pay attention to these aspects, and improvement to these aspects could be most effective to enhance user experience.

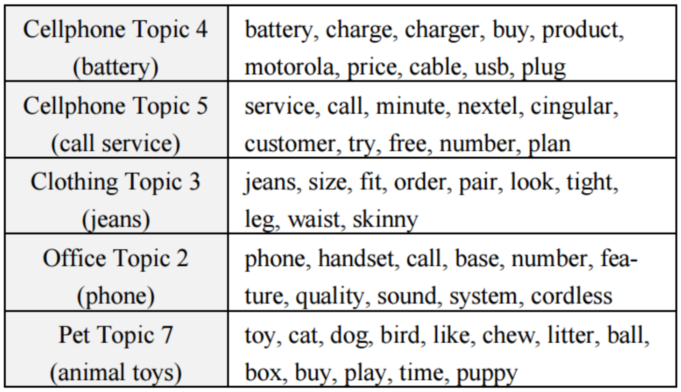
The results for three datasets ”cellphone”, ”clothing” and ”office” are shown in Table 8, where 4 critical aspects can be observed. Most frequent words of the 5 critical aspects are shown in Table 9. Based on both tables we can clearly see ”battery” and ”phone call service” are aspects of ”cellphone” that are of relatively high user concern but plagued by negative sentiment. Aspect ”jeans” of product ”clothing”, and aspect ”phone” of product ”office” are also such critical aspects.

At the end we examine the Pearson's correlation between aspect preferences and aspect sentiments, which is a weak value 0.349 and it confirms our decouple of these two types of variables in our model is valid. As discussed in introduction, a distinction between TSPRA and collaborative filtering models like [8, 9] is that we define ”preference” as how much customers care of an aspect and by definition it is independent from word sentiments.



	Cellphone		Clothing		Office		Pet	
	Pref	Senti	Pref	Senti	Pref	Senti	Pref	Senti
1	0.650	0.606	1.000	0.629	1.000	0.767	0.791	0.695
2	1.000	0.526	0.037	0.653	**0.439**	**0.164**	0.688	0.525
3	0.408	0.389	**0.352**	**0.058**	0.155	0.372	0.428	0.512
4	**0.556**	**0.145**	0.204	0.643	0.403	0.697	0.236	0.514
5	**0.447**	**-0.539**	0.709	0.714	0.027	0.058	0.460	0.534
6	**0.951**	**0.308**	0.160	0.646	0.307	0.637	0.001	-0.463
7	0.008	0.244			0.200	0.446	**1.000**	**0.402**
8	0.177	0.252					0.010	0.202
9	0.019	0.156					0.004	0.328

Table 8: Demonstration of critical aspect.



Cellphone Topic 4 (battery)	battery, charge, charger, buy, product, motorola, price, cable, usb, plug
Cellphone Topic 5 (call service)	service, call, minute, nextel, cingular, customer, try, free, number, plan
Clothing Topic 3 (jeans)	jeans, size, fit, order, pair, look, tight, leg, waist, skinny
Office Topic 2 (phone)	phone, handset, call, base, number, feature, quality, sound, system, cordless
Pet Topic 7 (animal toys)	toy, cat, dog, bird, like, chew, litter, ball, box, buy, play, time, puppy

Table 9: Most frequent words of the critical aspects in Table 8.

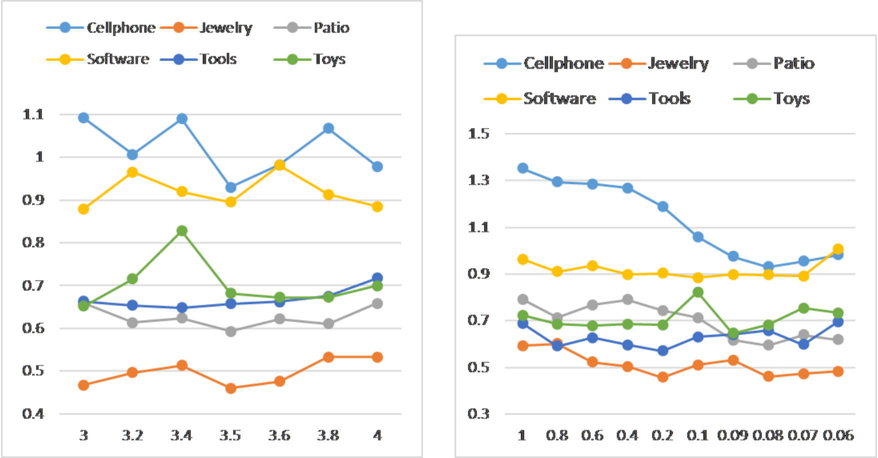
4.4. Experiments on Parameters

The neutral rating μ and rating noise σ are two main parameters introduced by our model TSPRA. In this section we experiment on the effects of both parameters in terms of error. We choose 6 datasets with different average review ratings to test the parameter μ by fixing σ^2= 0.08, and then we test σ^2 by fixing $\mu = 3.5$ on the same 6 datasets. The results are shown in Figure 5. As we can see there is no strong agreement on the best parameter among the datasets. That is not surprising since it is reasonable that different products have somewhat different neutral rating and rating noise. However, we do observe from the above charts that the model archives relatively good results on all six data sets when μ is around 3.5 and σ^2 is around 0.08. Hence we recommend $\mu = 3.5, \sigma^2 = 0.08$ as the default values, as what we have done in section 4.1.

Also we find $\sigma^2 = 0.08$ is actually making sense. The parameter value of σ^2 reflects our belief how the majority of users give the final rating based on the average of non-neutral word ratings (the neutral words are ignored when review rating is estimated, see section 3.1 and 3.2). Under normal distribution, 68% of final ratings vary from $(r_{di}-\sigma, r_{di}+\sigma)$ and 95% of final ratings vary from $(r_{di}-2\sigma, r_{di}+2\sigma)$. When we set $\sigma^2 = 0.08$ and thus $\sigma \approx 0.283$, it actually indicates 68% of review ratings are in the range of $(r_{di}-0.283, r_{di}+0.283)$ and 95% of review ratings are in the range of $(r_{di}-0.566, r_{di}+0.566)$. In plain words this implies users are rounding their review ratings. For example, if a user wants to rate a product as 3.7, then due to limit of the 1-5 scale, the user has to round the rating to 4.0.

5. Conclusion

We have proposed a model that considers topics, sentiments and user preferences as integrated factors to explain review texts and ratings. Two distinctions between our model and previous models are the de-correlation between preference and sentiment, and the adoption of HDP framework that automatically identifies the topic number of each dataset. Our model outperforms the FLAME model proposed by [9] in terms of error, Pearson's correlation and number of inverted pairs. Further experiment on sentiment analysis show our model can yield very reasonable word sentiments, most of them make more sense in comparison to a public sentiment resources SenticNet3 [24] constructed via non-probabilistic approaches. A third experiment shows our model is able to capture the "critical aspects" –



Varies μ with fixed $\sigma^2 = 0.08$ Varies σ^2 with fixed $\mu = 3.5$

Figure 5: Error varies with neutral rating parameter μ and noise parameter σ^2.

the negatively commented product aspects concerned a lot by users. Improving the critical aspects would be most effective to enhance overall user experience. Last but not the least, the correlation between ”preference” and ”sentiment” in our model is weak, verifying our claim that these two concepts are largely independent factors.

References

1. Xu, Y., W. Lam, and T. Lin. *Collaborative Filtering Incorporating Review Text and Co-clusters of Hidden User Communities and Item Groups.* in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management.* 2014. ACM.

2. Wang, H., 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.* 2011. ACM.

3. Jo, Y. and A.H. Oh. *Aspect and sentiment unification model for online review analysis.* in *Proceedings of the fourth ACM international conference on Web search and data mining.* 2011. ACM.

4. David M. Blei, A.Y.N., Michael I. Jordan, *Latent dirichlet allocation.* Journal of Machine Learning Research, 2003.

5. Teh, Y.W., et al., *Hierarchical dirichlet processes.* Journal of the american statistical association, 2006. **101**(476).

6. Buntine, W.L. and S. Mishra. *Experiments with non-parametric topic models.* in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2014. ACM.

7. Wang, C., J.W. Paisley, and D.M. Blei. *Online variational inference for the hierarchical Dirichlet process.* in *International conference on artificial intelligence and statistics.* 2011.

8. Diao, Q., et al. *Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS).* in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2014. ACM.

9. Wu, Y. and M. Ester, *FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering.* 2015.

10. Yu, K., et al. *Fast nonparametric matrix factorization for large-scale collaborative filtering.* in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* 2009. ACM.

11. Paterek, A. *Improving regularized singular value decomposition for collaborative filtering.* in *Proceedings of KDD cup and workshop.* 2007.

12. Ganu, G., N. Elhadad, and A. Marian. *Beyond the Stars: Improving Rating Predictions using Review Text Content.* in *WebDB.* 2009.

13. Agarwal, D. and B.-C. Chen. *fLDA: matrix factorization through latent dirichlet allocation.* in *Proceedings of the third ACM international conference on Web search and data mining.* 2010. ACM.

14. McAuley, J. and J. Leskovec. *Hidden factors and hidden topics: understanding rating dimensions with review text.* in *Proceedings of the 7th ACM conference on Recommender systems.* 2013. ACM.

15. Wang, H., 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.* 2010. ACM.

16. Mei, Q., et al. *Topic sentiment mixture: modeling facets and opinions in weblogs.* in *Proceedings of the 16th international conference on World Wide Web.* 2007. ACM.

17. Zhao, W.X., et al. *Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid.* in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* 2010. Association for Computational Linguistics.

18. Mcauliffe, J.D. and D.M. Blei. *Supervised topic models.* in *Advances in neural information processing systems.* 2008.

19. Rabinovich, M. and D. Blei. *The inverse regression topic model.* in *Proceedings of The 31st International Conference on Machine Learning.* 2014.

20. Griffiths, T.L. and M. Steyvers, *Finding scientific topics.* Proceedings of the National Academy of Sciences, 2004. **101**(suppl 1): p. 5228-5235.

21. Suess, E.A. and B.E. Trumbo, *Introduction to probability simulation and Gibbs sampling with R.* 2010: Springer Science & Business Media.

22. Manning, C.D., et al. *The Stanford CoreNLP natural language processing toolkit.* in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* 2014.

23. Cambria, E., et al. *SenticNet: A Publicly Available Semantic Resource for Opinion Mining.* in *AAAI fall symposium: commonsense knowledge.* 2010.

24. Cambria, E., D. Olsher, and D. Rajagopal. *SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis.* in *Twenty-eighth AAAI conference on artificial intelligence.* 2014.