




JST-RR Model: Joint Modeling of Ratings and Reviews in Sentiment-Topic Prediction

Qiao Liang, Shyam Ranganathan, Kaibo Wang & Xinwei Deng


To cite this article: Qiao Liang, Shyam Ranganathan, Kaibo Wang & Xinwei Deng (2023) JST-RR Model: Joint Modeling of Ratings and Reviews in Sentiment-Topic Prediction, *Technometrics*, 65:1, 57-69, DOI: [10.1080/00401706.2022.2063187](https://doi.org/10.1080/00401706.2022.2063187)


To link to this article: <https://doi.org/10.1080/00401706.2022.2063187>

 View supplementary material [↗](#)

 Published online: 27 Apr 2022.

 Submit your article to this journal [↗](#)

 Article views: 211

 View related articles [↗](#)

 View Crossmark data [↗](#)



JST-RR Model: Joint Modeling of Ratings and Reviews in Sentiment-Topic Prediction

Qiao Liang^a, Shyam Ranganathan^b, Kaibo Wang^c, and Xinwei Deng^b

^aSchool of Statistics, Southwestern University of Finance and Economics, Chengdu, China; ^bDepartment of Statistics, Virginia Tech, Blacksburg, VA; ^cDepartment of Industrial Engineering, Tsinghua University, Beijing, China

ABSTRACT

Analysis of online reviews has attracted great attention with broad applications. Often times, the textual reviews are coupled with the numerical ratings in the data. In this work, we propose a probabilistic model to accommodate both textual reviews and overall ratings with consideration of their intrinsic connection for a joint sentiment-topic prediction. The key of the proposed method is to develop a unified generative model where the topic modeling is constructed based on review texts and the sentiment prediction is obtained by combining review texts and overall ratings. The inference of model parameters are obtained by an efficient Gibbs sampling procedure. The proposed method can enhance the prediction accuracy of review data and achieve an effective detection of interpretable topics and sentiments. The merits of the proposed method are elaborated by the case study from Amazon datasets and simulation studies.

ARTICLE HISTORY

Received February 2021
Accepted April 2022

KEYWORDS

Generative approach; Joint modeling; Latent Dirichlet allocation; Service analytics; Text mining

1. Introduction

In modern service applications, there are increasing amounts of online reviews generated by customers in recent years. The online reviews often contain both the text reviews and overall ratings. For example, the reviews in Amazon.com contain a review text on customer opinions of products or services, as well as the overall rating score on the general evaluation. Clearly, these user-generated contents can provide valuable information for both customers and online merchants (Liu 2012). Among various research works on analyzing such review data, topic identification (Titov and McDonald 2008b; Blei 2012; Airoidi and Bischof 2016) and sentiment classification (Bai 2011; Taddy 2013; Calheiros, Moro, and Rita 2017) are two major directions. The former aims to extract representing features or aspects of interest from discrete review words, and the latter is to predict the semantic orientation of a review text. With consideration of the inherent dependency between sentiment polarities and topics, a simultaneous detection of correlated topics and sentiments serves as a critical function in the information retrieval of online customer reviews (Titov and McDonald 2008a; Mei et al. 2007; Lin and He 2009).

Note that the existing works mainly focused on topic discovery and sentiment prediction using the review texts only. While the information from the overall ratings has not been integrated to some extent. It is seen that the rating scores provide intuitive orientations of user opinions, which can allow the latent sentiments extracted more appropriately (Li et al. 2015). Moreover, most collected review texts in practice are vague in the sense of low “signal-to-noise ratio” with large amounts of spam content, unhelpful opinions, as well as highly subjective

and misleading information (Lu et al. 2010). In such situations, it is of great importance to consider ratings and review texts in a mutually complement manner for accurate quantification on review sentiments and topics.

The scope of this work is to predict both sentiments and topics from the joint learning of review texts and overall ratings. Typically, the association between textual reviews and overall ratings are prevailing based on the general orientation of review sentiments. For instance, a review stimulated by positive sentiment would present both a higher rating and a positive review text. The sentiment polarities indicated by the overall ratings and textual reviews are closely related, while their relationship varies among different customers. In practice, customers may have different preference and emphasis on different aspects for the same product, and they may give overall ratings based on the partial or whole product aspects discussed in review texts (Li et al. 2015). For example, even a full 5-star rating could be accompanied by negative review content. The dynamic relationship between the overall ratings and the review texts makes it challenging to digest the information in reviews with ratings jointly. As ratings serve as one of the most important metadata of review documents, this problem can be viewed from the perspective of the incorporation of document metadata with the content of the text (Roberts, Stewart, and Airoidi 2016).

To address the aforementioned challenges, we propose a joint sentiment-topic model to accommodate both ratings and review texts. We denote the proposed model as the *JST-RR model*. The proposed method extends the conventional joint sentiment-topic modeling by incorporating the generative process of ratings with textual reviews in a unified framework. Under this framework, the connection between review texts and ratings is

characterized by the latent joint sentiment-topic distribution. We have also developed a weighting mechanism between review words and ratings for a more accurate quantification on review sentiments. The proposed JST-RR model enables an effective identification of topics and sentiments in reviews and a more accurate prediction for review data. Note that the proposed model is weakly supervised with the only supervision from a domain-independent sentiment lexicon. Hence, it can be easily adapted to review mining in various domains or applications.

The remainder of this article is organized as follows. [Section 2](#) reviews the state-of-the-art methods on joint sentiment-topic prediction in review modeling. [Section 3](#) presents the details of the proposed JST-RR model. [Section 4](#) reports the model implementation and performance on the Amazon datasets. [Section 5](#) conducts a simulation study to extensively evaluate the performance of the proposed model. Finally, we conclude this work with some discussion in [Section 6](#). All detailed derivations and additional experimental results are contained in the [supplementary materials](#).

2. Literature Review

This section mainly reviews modeling methods for online review data in sentiment-topic prediction. In the literature, many existing works (Lu, Zhai, and Sundaresan 2009; Brody and Elhadad 2010; Lu et al. 2011) performed topic detection and sentiment classification in a two-stage process. They first detected topics from review texts using traditional topic models such as latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) and probabilistic latent semantic indexing (PLSI) (Hofmann 1999). Then sentiment labels are assigned to specific topics by applying sentiment classification techniques to corresponding review texts. There are several works on detecting topics and sentiments simultaneously from user-generated content (Mei et al. 2007; Titov and McDonald 2008a; Lin and He 2009). For example, Mei et al. (2007) proposed the topic-sentiment mixture (TSM) model for the weblog collection based on the model setting of PLSI. However, the topic-sentiment correlation in the TSM model was not directly constructed but captured through a post-processing of model parameters. With the focus of finding correlated sentiments and topics from texts, the joint sentiment-topic (JST) model (Lin and He 2009) and the Reverse-JST model (Lin et al. 2012) extended the LDA model by constructing an additional sentiment layer conditioning and being conditioned on the topic layer of LDA, respectively. Many follow-up works (Moghaddam and Ester 2011; Li et al. 2013; Dermouche et al. 2015), regarded as variants of the JST and Reverse-JST models, use the same assumption of conditional inter-dependency between topics and sentiments.

However, these works mainly focused on topic discovery and sentiment prediction from review texts only, where the information from the overall ratings has been overlooked to some extent.

For review sentiment prediction, existing methods often employed a supervised learning framework using sentiment labels directly indicated by overall ratings (Pang, Lee, and Vaithyanathan 2002; Blitzer, Dredze, and Pereira 2007; Ye, Zhang, and Law 2009). That is, the ratings were used to supervise the sentiment prediction of corresponding review texts. However, there is still a discrepancy between the sentiment

orientations indicated by review texts and ratings, since customers may give overall ratings based on the partial or whole product aspects discussed in review texts. Considering the complex and dynamic relationship between the overall ratings and the review texts, it is beneficial to construct a joint model of textual reviews and numerical ratings for sentiment-topic prediction. For instance, the models by Wang, Lu, and Zhai (2010, 2011) assumed that the overall ratings were based on ratings of specific aspects or topics extracted from review texts. The aspect identification and rating (AIR) model by Li et al. (2015) followed a reverse assumption that aspect ratings were produced with the prior information of overall ratings. However, these models mainly focused on the detection of aspect ratings and conditioned the joint modeling of textual reviews and overall ratings on the results of aspect ratings.

Motivated by the lack of a general model to characterize the intrinsic connection between review texts and overall ratings, we propose a joint sentiment-topic model to accommodate both overall ratings and review texts in a unified probabilistic framework for accurate prediction of review sentiments and topics.

3. Joint Sentiment-Topic Modeling of Review Texts and Ratings

In this section, we briefly describe the notation and joint sentiment-topic (JST) representation of reviews in [Section 3.1](#). We then detail the proposed JST-RR model for integrating the overall ratings with review words in [Section 3.2](#). The procedure of model inference is constructed in [Section 3.3](#).

3.1. Joint Sentiment-Topic Representation of Reviews

Consider the data consisting of a collection of product review documents $\{d_i, i = 1, \dots, D\}$. For each review document d_i , suppose that it contains N_i words denoted as $w_i = (w_{i1}, \dots, w_{iN_i})$, and it contains M_i rating scores denoted as $r_i = (r_{i1}, \dots, r_{iM_i})$. A review document can be composed of a single review (i.e., $M_i = 1$) or a collection of reviews for extracting review features from various granularity levels. For example, a document is usually defined in recommender systems (McAuley and Leskovec 2013; Ling, Lyu, and King 2014; Yu, Mu, and Jin 2017) as the set of all reviews with ratings of the same product or the same user, and the product-specific or user-specific features are characterized by their corresponding documents. Here, each word in the observed document is assumed to be from the vocabulary indexed by $\{1, \dots, V\}$. Without loss of generality, we assume that the rating $r_{ij} \in \{1, 2, 3, 4, 5\}$ with 5 to be the highest rating and 1 to be the lowest rating.

In a typical joint sentiment-topic modeling framework (Lin and He 2009; Lin et al. 2012; Li et al. 2013), each review document d_i is assumed to be represented by mixtures of sentiment and topic variables that are interdependent. By following the assumption in the general class of mixed membership models (Airoldi et al. 2010, 2015; Manrique-Vallier and Reiter 2012), ratings and words are observational units in the document, and each observational unit belongs to a single cluster that is represented by a latent sentiment label or topic label. Let us

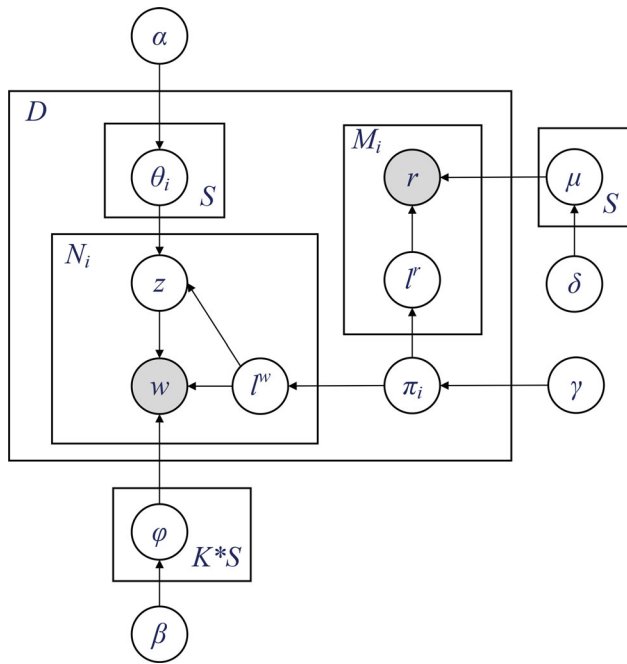


Figure 1. Illustration of the proposed JST-RR model.

denote the sentiment label by $l \in \{1, \dots, S\}$ and the topic label by $z \in \{1, \dots, K\}$. The sentiment of the document d_i follows a multinomial distribution $\text{Multinomial}(\pi_i)$, where the S -dimension prior distribution $\pi_i \sim \text{Dirichlet}(\gamma)$. Conditional on each sentiment label $l \in \{1, \dots, S\}$, the topic follows a multinomial distribution $\text{Multinomial}(\theta_{i,l})$, where the K -dimension prior of topic distribution $\theta_{i,l} \sim \text{Dirichlet}(\alpha_l)$. Typically, the document-level sentiment and topic distributions indicate how likely the current document fits a specific sentiment and topic, providing a quantification on the latent sentiments and topics for unstructured reviews.

3.2. The Proposed JST-RR Model

In this section, we will detail the proposed JST-RR model with the consideration of both ratings and reviews. Based on the JST representation, the proposed JST-RR model integrates the overall ratings with textual words in review documents under a unified probabilistic framework. Figure 1 illustrates a graphical representation of the JST-RR model structure. The notation used in the proposed model is summarized in Table 1.

We consider that each review document d_i is represented by its document-level sentiment distribution $\text{Multinomial}(\pi_i)$ and topic distribution $\text{Multinomial}(\theta_i)$. The key of the JST-RR model is to provide a unified probabilistic generative process for both observed words and ratings in the review documents. That is, each word is assumed to be drawn from the V -dimension multinomial word distribution $\text{Multinomial}(\varphi_{l^w,z})$ conditioned on the word sentiment label l^w and topic label z , where the prior distribution $\varphi_{l^w,z} \sim \text{Dirichlet}(\beta_{l^w,z})$. For the generating process of overall ratings, we consider that the overall ratings provide only a general orientation of sentiments. Each rating is then assumed to be drawn from the five-dimensional multinomial rating distribution $\text{Multinomial}(\mu_{l^r})$ only conditioned on its rating sentiment label l^r , where the prior distribution $\mu_{l^r} \sim \text{Dirichlet}(\delta_{l^r})$.

Table 1. A summary of notation.

Term	Definition
d	Document
w	Word
r	Rating
z	Topic label
l^w	Word sentiment label
l^r	Rating sentiment label
D	Number of documents
V	Vocabulary size
K	Number of topics
S	Number of sentiment classifications
π_i	Coefficient vector of the multinomial sentiment distribution for the i th document
$\theta_{i,l}$	Coefficient vector of the multinomial topic distribution under the sentiment label l for the i th document
$\varphi_{l,z}$	Coefficient vector of the multinomial word distribution under the sentiment label l and topic label z
μ_l	Coefficient vector of the multinomial rating distribution under the sentiment label l
N_i	Number of words in the i th document
$N_{i,l}$	Number of words that are assigned to the sentiment label l in the i th document
$N_{i,l,z}$	Number of words that are assigned to the sentiment label l and topic label z in the i th document
$N_{l,z}$	Number of words that are assigned to the sentiment label l and topic label z in the dataset
$N_{l,z,w}$	Number of times that the word w is assigned to the sentiment label l and topic label z in the dataset
M_i	Number of ratings in the i th document
$M_{i,l}$	Number of ratings that are assigned to the sentiment label l in the i th document
M_l	Number of ratings that are assigned to the sentiment label l in the dataset
$M_{l,r}$	Number of times that the rating r is assigned to the sentiment label l in the dataset

A formal generative process of the review document collection $\{d_i; i = 1, \dots, D\}$ is presented in Algorithm 1. In this framework, words and ratings are jointly generated and used as observations for the estimation of reviews. The hyperparameters β , δ , γ , and α indicate the prior information before the actual words and ratings, that is, the actual data, are observed. The settings of hyperparameters are detailed in Section 4.1 based on a real-world case.

The proposed JST-RR model not only provides a probabilistic and unified framework, but also provides a meaningful manner on how ratings and review texts work in realistic settings. For example, a reviewer on Amazon has an overall sentiment regarding the purchased product, which informs the reviewer's sentiment on the various aspects of the product which are typically represented as "topics" in the model. It is likely that the reviewer has a negative sentiment about one topic while having a positive sentiment on other topics, and this can be reflected by the word sentiment and overall sentiment from the proposed model. In addition, the sequential generative process of observed ratings and review texts is consistent with the real review process on most websites such as *Amazon* and *Tripadvisor*, where customers are first required to give an overall rating score before they write a detailed review text.

3.3. Model Inference

For the inference of the proposed JST-RR model, there are four sets of latent distribution parameters: the document-level sentiment distribution parameter π , the sentiment-specific topic

Algorithm 1: Generative procedure of words and ratings in review documents based on the JST-RR model

- For the entire document collection, first characterize the “topic” and the “sentiment” by the word probability distribution and the rating probability distribution:
 - For each combination of word sentiment label $l^w \in \{1, \dots, S\}$ and topic label $z \in \{1, \dots, K\}$:
 - * Draw sample from word probability distribution $\varphi_{l^w, z} \sim \text{Dirichlet}(\boldsymbol{\beta}_{l^w, z})$.
 - For each rating sentiment label $l^r \in \{1, \dots, S\}$:
 - * Draw sample from rating probability distribution $\boldsymbol{\mu}_{l^r} \sim \text{Dirichlet}(\boldsymbol{\delta}_{l^r})$.
 - For each document $d_i, i = 1, \dots, D$:
 - Draw sample from sentiment probability distribution $\boldsymbol{\pi}_i \sim \text{Dirichlet}(\boldsymbol{\gamma})$.
 - Draw sample from topic probability distribution $\boldsymbol{\theta}_{i, l} \sim \text{Dirichlet}(\boldsymbol{\alpha}_l)$ for each sentiment label $l \in \{1, \dots, S\}$.
 - For each word $w_{ij}, j = 1, \dots, N_i$ in document d_i :
 - * Draw the sentiment assignment $l_{ij}^w \sim \text{Multinomial}(\boldsymbol{\pi}_i)$.
 - * Draw the topic assignment $z_{ij} \sim \text{Multinomial}(\boldsymbol{\theta}_{i, l_{ij}^w})$ conditioned on l_{ij}^w .
 - * Draw a specific word $w_{ij} \sim \text{Multinomial}(\boldsymbol{\varphi}_{l_{ij}^w, z_{ij}})$ conditioned on l_{ij}^w and z_{ij} .
 - For each rating $r_{ij}, j = 1, \dots, M_i$ in document d_i :
 - * Draw the sentiment assignment $l_{ij}^r \sim \text{Multinomial}(\boldsymbol{\pi}_i)$.
 - * Draw a specific rating score $r_{ij} \sim \text{Multinomial}(\boldsymbol{\mu}_{l_{ij}^r})$ conditioned on l_{ij}^r .
-

distribution parameter $\boldsymbol{\theta}$, the joint sentiment/topic-word distribution parameter $\boldsymbol{\varphi}$, and the sentiment-rating distribution parameter $\boldsymbol{\mu}$. Given these latent distributions, we can explicitly express the joint probability of the observed words, ratings, and their sentiment/topic labels in the document collection $\{d_i, i = 1, \dots, D\}$ as

$$\begin{aligned}
 & P(\mathbf{w}, \mathbf{r}, \mathbf{l}^w, \mathbf{l}^r, \mathbf{z} | \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\mu}) \\
 &= \prod_{i=1}^D \prod_{j=1}^{N_i} P(l_{ij}^w, z_{ij}, w_{ij} | \boldsymbol{\pi}_i, \boldsymbol{\theta}_{i, l_{ij}^w}, \boldsymbol{\varphi}_{l_{ij}^w, z_{ij}}) \prod_{j=1}^{M_i} P(l_{ij}^r, r_{ij} | \boldsymbol{\pi}_i, \boldsymbol{\mu}_{l_{ij}^r}) \\
 &= \prod_{i=1}^D \prod_{j=1}^{N_i} P(l_{ij}^w | \boldsymbol{\pi}_i) P(z_{ij} | \boldsymbol{\theta}_{i, l_{ij}^w}) P(w_{ij} | \boldsymbol{\varphi}_{l_{ij}^w, z_{ij}}) \prod_{j=1}^{M_i} P(l_{ij}^r | \boldsymbol{\pi}_i) P(r_{ij} | \boldsymbol{\mu}_{l_{ij}^r}), \quad (1)
 \end{aligned}$$

where the words and ratings are conditionally independent given the document-level sentiments and topics. It is seen that

the observed words are dependent on their latent sentiment and topic assignments, while the ratings are only dependent on their latent sentiment assignments.

Note that there have been several methods in the literature developed for the inference of probabilistic topic models, including Gibbs sampling (Griffiths and Steyvers 2004), variational Bayesian inference (Blei, Ng, and Jordan 2003), and maximum a posteriori (MAP) estimation (Chien and Wu 2008). In this work, we adopt the Gibbs sampling for the model inference because of its promising convergence to the underlying distribution. It is also noted that some advanced algorithms (Hoffman et al. 2013; Srivastava and Sutton 2017) could be adapted to our problem for handling large and complex data. The state transition of the Markov chain formed by the Gibbs sampler is determined by the sampling of the latent variables (i.e., the topic label z and the sentiment label l) given the current values of all other variables and the observed data. The conditional probability of sampling sentiment label l_{ij}^w and topic label z_{ij} for the observed word $w_{ij} = w$ in the document d_i can be written as

$$\begin{aligned}
 & P(l_{ij}^w = l, z_{ij} = z | \mathbf{w}, \mathbf{l}_{-ij}^w, \mathbf{l}^r, \mathbf{z}_{-ij}) \\
 & \propto P(l_{ij}^w = l, z_{ij} = z, w_{ij} = w | \mathbf{w}_{-ij}, \mathbf{l}_{-ij}^w, \mathbf{l}^r, \mathbf{z}_{-ij}) \\
 & = P(l_{ij}^w = l | \mathbf{l}_{-ij}^w, \mathbf{l}^r) \times P(z_{ij} = z | l_{ij}^w = l, \mathbf{l}_{-ij}^w, \mathbf{z}_{-ij}) \\
 & \quad \times P(w_{ij} = w | l_{ij}^w = l, z_{ij} = z, \mathbf{l}_{-ij}^w, \mathbf{z}_{-ij}, \mathbf{w}_{-ij}) \\
 & = \int_{\boldsymbol{\pi}_i} P(l_{ij}^w = l | \boldsymbol{\pi}_i) P(\boldsymbol{\pi}_i | \mathbf{l}_{-ij}^w, \mathbf{l}^r) d\boldsymbol{\pi}_i \\
 & \quad \times \int_{\boldsymbol{\theta}_{i, l}} P(z_{ij} = z | \boldsymbol{\theta}_{i, l}) P(\boldsymbol{\theta}_{i, l} | \mathbf{l}_{-ij}^w, \mathbf{z}_{-ij}) d\boldsymbol{\theta}_{i, l} \times \\
 & \quad \int_{\boldsymbol{\varphi}_{l, z}} P(w_{ij} = w | \boldsymbol{\varphi}_{l, z}) P(\boldsymbol{\varphi}_{l, z} | \mathbf{l}_{-ij}^w, \mathbf{z}_{-ij}, \mathbf{w}_{-ij}) d\boldsymbol{\varphi}_{l, z}. \quad (2)
 \end{aligned}$$

The superscript or subscript $-ij$ hereafter denotes the data quantity excluding the j th position in the document d_i . By integrating out $\boldsymbol{\pi}_i$ (see detailed derivation in Section A of the [supplementary materials](#)), the first term in Equation (2) can be derived as

$$P(l_{ij}^w = l | \mathbf{l}_{-ij}^w, \mathbf{l}^r) = \frac{N_{i, l}^{-ij} + M_{i, l} + \gamma_l}{N_i^{-ij} + M_i + \sum_{l'} \gamma_{l'}}. \quad (3)$$

It represents the probability of sampling $l_{ij}^w = l$ given all other sentiment assignments \mathbf{l}_{-ij}^w of words and \mathbf{l}^r of ratings in the same review document d_i . Here N_i and M_i are the total number of words and ratings in the document d_i , $N_{i, l}$ and $M_{i, l}$ are the number of words and ratings associated with sentiment l in the document d_i . The hyperparameter γ_l can be interpreted as the prior observation counts of the sentiment l assigned with d_i .

From Equation (3) and its derivation in Section A of the [supplementary materials](#), one can see that all the observed ratings and words are treated with the equal weight for the estimation of review sentiments. However, in a typical user review, the number of words is often much larger than the number of ratings, even when multiple ratings are allowed in a particular application. A one-to-many relationship exists between the observed rating and its description words for expressing a particular opinion in a single review. To address these challenges,

we consider to incorporate a weighting mechanism between the observed ratings and words in sentiment estimation. Ratings and words associated with their sentiment assignments in the same review document serve as samples that are independently generated from the same document-level sentiment distribution but with different sample weights. From the perspective of a weighted likelihood for the sentiment assignments of words and ratings (see Remark 1 in the [supplementary materials](#)), Equation (3) can be re-expressed in a more general form:

$$P(l_{ij}^w = l | l_{-ij}^w, \mathbf{r}^w) = \frac{N_{i,l}^{-ij} + \sigma M_{i,l} + \gamma_l}{N_i^{-ij} + \sigma M_i + \sum_{l'} \gamma_{l'}}, \quad (4)$$

where σ is a weighting parameter to indicate the weight of a rating relative to a word in the estimation of review sentiments. When $\sigma = 0$, the document-level sentiment prediction depends only on the review words, which is simplified as the JST model in Lin and He (2009).

Similarly, the second term in Equation (2) can be estimated by integrating out $\theta_{i,l}$, which gives

$$P(z_{ij} = z | l_{ij}^w = l, l_{-ij}^w, \mathbf{z}_{-ij}) = \frac{N_{i,l,z}^{-ij} + \alpha_{l,z}}{N_{i,l}^{-ij} + \sum_{z'} \alpha_{l,z'}}, \quad (5)$$

where $N_{i,l,z}$ is the number of words associated with the sentiment l and topic z in the document d_i , and the hyperparameter $\alpha_{l,z}$ can be interpreted as the prior observation counts of words assigned with the sentiment l and topic z in d_i . For the third term in Equation (2), we can obtain its posterior prediction by integrating out $\phi_{l,z}$ in the same manner as

$$P(w_{ij} = w | l_{ij}^w = l, z_{ij} = z, l_{-ij}^w, \mathbf{z}_{-ij}, \mathbf{w}_{-ij}) = \frac{N_{l,z,w}^{-ij} + \beta_{l,z,w}}{N_{l,z}^{-ij} + \sum_{w'} \beta_{l,z,w'}}, \quad (6)$$

where $N_{l,z}$ is the number of words assigned with the sentiment label l and topic label z in the entire dataset, $N_{l,z,w}$ is the number of times that the word w is associated with the sentiment label l and topic label z in the dataset, and the hyperparameter $\beta_{l,z,w}$ can be interpreted as the prior counts of word w associated with sentiment label l and topic label z in the dataset.

By combining the results in Equations (4)–(6), the expression for the full conditional probability in Equation (2) can be written as

$$\begin{aligned} & P(l_{ij}^w = l, z_{ij} = z | \mathbf{w}, l_{-ij}^w, \mathbf{r}^w, \mathbf{z}_{-ij}) \\ & \propto \frac{N_{i,l}^{-ij} + \sigma M_{i,l} + \gamma_l}{N_i^{-ij} + \sigma M_i + \sum_{l'} \gamma_{l'}} \cdot \frac{N_{i,l,z}^{-ij} + \alpha_{l,z}}{N_{i,l}^{-ij} + \sum_{z'} \alpha_{l,z'}} \\ & \quad \cdot \frac{N_{l,z,w}^{-ij} + \beta_{l,z,w}}{N_{l,z}^{-ij} + \sum_{w'} \beta_{l,z,w'}}. \end{aligned} \quad (7)$$

In a similar manner, we can specify the conditional probability of sampling the sentiment label l_{ij}^r for the observed rating $r_{ij} = r$ in the document d_i as (see detailed derivation in Section A of

the [supplementary materials](#))

$$\begin{aligned} & P(l_{ij}^r = l | \mathbf{r}, l_{-ij}^r, \mathbf{l}^w) \propto P(l_{ij}^r = l, r_{ij} = r | \mathbf{r}_{-ij}, l_{-ij}^r, \mathbf{l}^w) \\ & = P(l_{ij}^r = l | l_{-ij}^r, \mathbf{l}^w) \times P(r_{ij} = r | l_{ij}^r = l, l_{-ij}^r, \mathbf{r}_{-ij}) \\ & = \int_{\boldsymbol{\pi}_i} P(l_{ij}^r = l | \boldsymbol{\pi}_i) P(\boldsymbol{\pi}_i | l_{-ij}^r, \mathbf{l}^w) d\boldsymbol{\pi}_i \\ & \quad \times \int_{\boldsymbol{\mu}_i} P(r_{ij} = r | \boldsymbol{\mu}_i) P(\boldsymbol{\mu}_i | l_{-ij}^r, \mathbf{r}_{-ij}) d\boldsymbol{\mu}_i \\ & = \frac{N_{i,l} + \sigma M_{i,l}^{-ij} + \gamma_l}{N_i + \sigma M_i^{-ij} + \sum_{l'} \gamma_{l'}} \times \frac{M_{l,r}^{-ij} + \delta_{l,r}}{M_l^{-ij} + \sum_{r'} \delta_{l,r'}}, \end{aligned} \quad (8)$$

where M_l is the number of ratings associated with the sentiment l in the dataset, $M_{l,r}$ is the number of times that the rating r is associated with sentiment label l in the dataset, and the hyperparameter $\delta_{l,r}$ can be interpreted as the prior counts of rating r associated with sentiment label l in the dataset.

A sample obtained from the Markov chain in its stable state is used to obtain the posterior estimations of the parameters $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\mu}$ as follows:

$$\begin{aligned} \hat{\pi}_{i,l} &= \frac{N_{i,l} + \sigma M_{i,l} + \gamma_l}{N_i + \sigma M_i + \sum_{l'} \gamma_{l'}}, & \hat{\theta}_{i,l,z} &= \frac{N_{i,l,z} + \alpha_{l,z}}{N_{i,l} + \sum_{z'} \alpha_{l,z'}}, \\ \hat{\phi}_{l,z,w} &= \frac{N_{l,z,w} + \beta_{l,z,w}}{N_{l,z} + \sum_{w'} \beta_{l,z,w'}}, & \hat{\mu}_{l,r} &= \frac{M_{l,r} + \delta_{l,r}}{M_l + \sum_{r'} \delta_{l,r'}}. \end{aligned} \quad (9)$$

For each document d_i , its document-level sentiment distribution parameter $\boldsymbol{\pi}_i$ is approximated based on both N_i words and M_i ratings with a weighting parameter σ , while the topic distribution parameter $\boldsymbol{\theta}_i$ is estimated by only words in the document since ratings are not assigned with topic labels. The Gibbs sampling procedure of making inference of the proposed JST-RR model is summarized in [Algorithm 2](#).

4. Case Study of Amazon Datasets

In this section, we evaluate the performance of the proposed model using three real datasets. The real data are obtained from the publicly available Amazon datasets (McAuley et al. 2015). Specifically, the three datasets are the online reviews of HP laptops, the online reviews of Lenovo laptops, and the online reviews of Dell laptops, which are denoted as *HP*, *Lenovo*, and *Dell*, respectively. For each single review, there is an overall rating that ranges from 1 star to 5 stars.

By defining the review documents at various granularity levels (i.e., from a single review, to a collection of reviews from the same product or the same user), the proposed JST-RR model can be applied for modeling customer opinions on different levels of interest. In this section, we mainly focus on examining the performance of the proposed method for the individual review documents. That is, each document here is based on a single review including a review text and an overall rating.

4.1. Data Preparation and Experiment Settings

For each dataset, we perform data preprocessing in the following steps. First, we convert words into lower cases and remove the punctuation, stop words (e.g., “a,” “and,” “be”), and infrequent

Algorithm 2: Gibbs sampling procedure of JST-RR

Input: Document collection $\{d_i, i = 1, \dots, D\}$, hyperparameters $\beta, \delta, \gamma, \alpha$, and weight parameter σ .

Output: Word distribution parameter ϕ , rating distribution parameter μ , document-level sentiment distribution parameter π and topic distribution parameter θ .

- 1 Assign initial topic/sentiment labels to all words/ratings at random;
- 2 **for each** Gibbs sampling iteration **do**
- 3 **for each** document $d_i, i = 1, \dots, D$ **do**
- 4 **for each** word $w_{ij}, j = 1, \dots, N_i$ in the document d_i **do**
- 5 Exclude w_{ij} associated with its sentiment label l_{ij}^w and topic label z_{ij} from count variables $N_i, N_{i,l}, N_{i,l,z}, N_{l,z}, N_{l,z,w}$;
- 6 Sample a new sentiment-topic combination for w_{ij} based on Equation (7);
- 7 Update count variables $N_i, N_{i,l}, N_{i,l,z}, N_{l,z}, N_{l,z,w}$ by incorporating the new sentiment/topic label of w_{ij} ;
- 8 **end**
- 9 **for each** rating $r_{ij}, j = 1, \dots, M_i$ in the document d_i **do**
- 10 Exclude r_{ij} associated with its sentiment label l'_{ij} from count variables $M_l, M_{l,r}, M_i, M_{i,l}$;
- 11 Sample a new sentiment assignment for r_{ij} based on Equation (8);
- 12 Update count variables $M_l, M_{l,r}, M_i, M_{i,l}$ by incorporating the new sentiment label of r_{ij} ;
- 13 **end**
- 14 **end**
- 15 **end**
- 16 Estimate ϕ, μ, π , and θ based on Equation (9);

words. Second, we stem each word to its root with Porter Stemmer (<http://tartarus.org/martin/PorterStemmer/>). Third, we perform *Negation* by adding a prefix “not_” to the word in negative dependency. For example, in the sentence “I do not like this product,” “not_like” is recognized as a whole to express negative sentiment. Finally, to obtain unbiased training results on sentiment prediction, we balance the number of positive and negative review documents in the dataset. After data preprocessing, the summary statistics of three experimental datasets are listed in Table 2.

In the implementation of the proposed method, we set the number of sentiment polarities $S = 2$ (i.e., positive and negative) and a varying number of topics $K \in \{2, 5, 7, 10, 12, 15, 20\}$. For the setting of hyperparameters, we simply use a symmetric setting for γ and α : $\gamma_l = 3.0/S, l \in \{1, \dots, S\}$; $\alpha_{l,z} = 3.0/(S \times K), l \in \{1, \dots, S\}, z \in \{1, \dots, K\}$. Based on the prior knowledge that a positive polarity is linked to a higher rating score and vice versa, we set $\delta_{l,r} = 10.0 \times r, r \in \{1, 2, 3, 4, 5\}$ for the positive sentiment l , and set $\delta_{l,r} = 10.0 \times (6 - r), r \in \{1, 2, 3, 4, 5\}$ for the negative sentiment l .

Table 2. A description of three Amazon datasets.

Dataset	Number of reviews	Average number of words (review length)
HP	11,655	71.56
Lenovo	4976	71.71
Dell	8438	51.09

For the setting of hyperparameter β , we use an asymmetric prior setting of β for the sentiment-word distribution. Note that many words are commonly treated as positive (e.g., “excellent”) or negative (e.g., “terrible”) regardless of the topics involved. Specifically, we select 1048 positive words and 2149 negative words from the sentiment lexicon MPQA (<http://mpqa.cs.pitt.edu/>) whose polarity orientations are domain independent. For the positive sentiment l , we set elements in β_l to be 0 for the words in negative list, 0.01 for other words. Similarly, for the negative sentiment l , we set elements of β_l to be 0 for the words in positive list, 0.01 for other words. Such a setting of β enables that the words in sentiment lexicons can only be drawn from the word distributions conditioned on their corresponding sentiment labels.

4.2. Quantitative Performance Analysis

The proposed JST-RR model is compared with four alternative methods: JST, RJST, AIR-JST, and AIR-RJST. The JST model in Lin and He (2009) can be treated as a baseline method for modeling topics and sentiments jointly via review texts alone. The RJST (or Reverse-JST) method in Lin et al. (2012) is a variant of JST model where the topic and the sentiment layers are inverted. The last two methods in comparison are denoted as AIR-JST and AIR-RJST based on the related AIR method in Li et al. (2015). The AIR method models observed textual reviews and overall ratings in a generative way by sampling latent sentiments of review texts with the overall ratings as prior parameters. For example, the review sentiment probability π is generated in accordance with its normalized rating r by:

$$\pi \sim \text{Beta}(\lambda r, \lambda(1 - r)).$$

The AIR model is adapted to our experimental settings in this case with two variants: AIR-JST and AIR-RJST, where the sentiment and the topic layers in the two models are inverted.

To quantitatively evaluate the performance of the proposed method, we consider the perplexity based performance measure on the test set. The perplexity is a commonly used metric (Blei, Ng, and Jordan 2003; Li et al. 2015) for evaluating the performance of probabilistic topic models. It measures how well the model fits observed reviews and is derived under the probabilistic framework without requiring manual intervention. Specifically, for a test set of documents $\{d_i, i = 1, \dots, D\}$, the perplexity of observed words $\{w_i, i = 1, \dots, D\}$ in the test set is defined as

$$\text{perplexity}(\{w_i, i = 1, \dots, D\} | \hat{\phi}) = \exp \left\{ - \frac{\sum_{i=1}^D \log P(w_i | \hat{\phi})}{\sum_{i=1}^D N_i} \right\}, \quad (10)$$

where the trained model is described by the word distribution parameter $\hat{\phi}$ that is estimated from the training set. We

employed the importance sampling methods in Wallach et al. (2009) to approximate the probability of the observed words $P(\mathbf{w}_i|\hat{\phi})$ in Equation (10). Since the perplexity values monotonically decrease with the log-likelihood of the test data, a lower perplexity indicates better prediction performance of the proposed model. It is noted that the upper bound of the perplexity in Equation (10) with the worst-case of a random prediction is given by

$$\text{perplexity}(\{\mathbf{w}_i, i = 1, \dots, D\}) = \exp \left\{ - \sum_{w \in V} P(w) \log P(w) \right\},$$

which is determined by the information entropy of words in the test data. Similarly, the perplexity of observed ratings $\{\mathbf{r}_i, i = 1, \dots, D\}$ in the test set can be defined accordingly. As the other four models in comparison only consider the generative process of the observed words, we conduct evaluation mainly based on the word perplexity values. For each experimental dataset, model validation is applied through a 10-fold cross-validation. Moreover, there could be some optimal partition strategies (Joseph and Vakayil 2021) on the experimental dataset for assisting model validation.

For the selection of the tuning parameter σ in the JST-RR model and the prior weight λ in the two AIR models, we adopt the 10-fold cross-validation on the training set in each partition, such that the selected parameters give the average best goodness of fit (indicated by the lowest perplexity values in this study). For example, Figure 2 shows the perplexity values of observed words versus the weight parameter σ by implementing the JST-RR model in the 10-fold cross-validation for a training set of *HP* with topic number $K = 5$. Similar trends of perplexity are also observed in the other cases, and thus omitted here. Generally, a lower perplexity value indicates better model performance in explaining the observed data. When $\sigma = 0$, the proposed JST-RR model becomes the baseline JST model that only focuses on review words. Based on the results in Figure 2, the model performance would benefit from the incorporation of ratings with a proper setting of rating weight σ . Moreover, it is seen from Figure 2 and other cases that $\sigma > 1$ (i.e., assigning larger weights to ratings than review words in sentiment estimation) is desired for better model performance, which conforms to the one-to-many relationship between ratings and review words.

Figure 3 shows the average word perplexity results of five models in the 10-fold cross-validation as well as their percentages against the baseline of RJST model with varying topic numbers in comparison. It is seen that the JST-RR model achieves the best overall performance with the lowest perplexity among all models under a variety of scenarios. In most cases, models that combine both textual reviews and overall ratings (i.e., AIR-JST, AIR-RJST, JST-RR) are superior to the models that only rely on textual reviews (i.e., JST, RJST). It implies that the incorporation of overall ratings can effectively enhance the model prediction accuracy. By taking into account the information on the observed ratings in reviews, the summarization of individual reviews can be more accurate and complete, and the latent topic-sentiment mixtures in the corpus can be more effectively extracted. Compared to the AIR models (e.g., AIR-JST, AIR-RJST), the proposed JST-RR model achieves better performance in capturing the dynamic connection between review words and

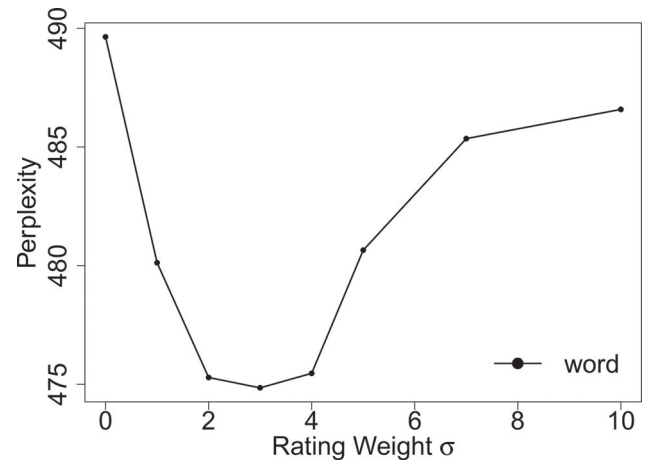


Figure 2. The average word perplexity with topic number $K = 5$ under different values of weight parameter σ in a 10-fold cross-validation for the training set in a partition of *HP* dataset.

ratings, leading to significant improvement in model prediction. For example, the dynamic connection between review words and ratings implies that customers may give overall ratings based on the partial or the whole product aspects discussed in review texts. In this case, even a full 5-star rating could be accompanied by partial negative review content, which can be captured by the sentiment of ratings and the sentiment of words in the proposed JST-RR model.

4.3. Qualitative Performance Analysis

It is also important to examine the effectiveness of the proposed model in the extraction of topics and sentiments from the data. As the estimated word distribution is conditioned on both sentiment and topic assignments, one can refer to the most frequent words (or top words) under each combination of sentiment-topic assignments for understanding the extracted topics with sentiment orientations. Table 3 shows the top positive and negative words under five example topics extracted from the *Dell* dataset. The top words are ranked by their conditional probabilities of occurring under different sentiment polarities $l \in \{1, \dots, S\}$ given the same topic label $z \in \{1, \dots, K\}$:

$$P(l|z, w) = \frac{N_{l,z,w} + \omega}{\sum_{l'=1}^S N_{l',z,w} + S\omega}, \quad (11)$$

where ω is a smoothing parameter (e.g., $\omega = 1$ in Laplace smoothing).

Each extracted topic in Table 3 covers a specific quality aspect of Dell products as well as related services such as battery (topic 1), memory and speed (topic 2), shipping and return (topic 3), network connections (topic 4), and peripherals (topic 5). In terms of sentiment, it can be seen that most of the positive words and negative words under each topic carry the corresponding sentiments well. Some of the words (e.g., “good,” “not_work”) show a general tendency of customer opinions that is independent of topics, and these words tend to appear under multiple topics frequently. Some other words could bear topic-specific sentiments. For example, words such as “crash,” “burn” are frequently used for conveying negative sentiment with respect

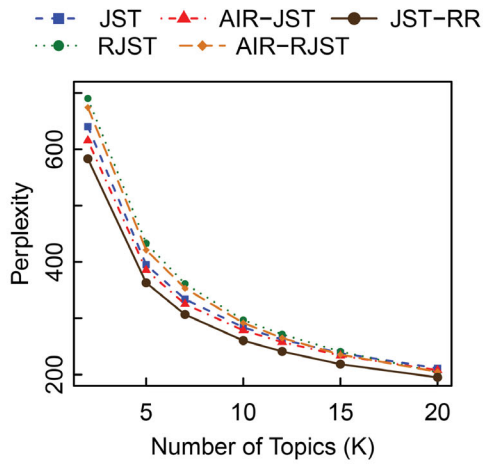
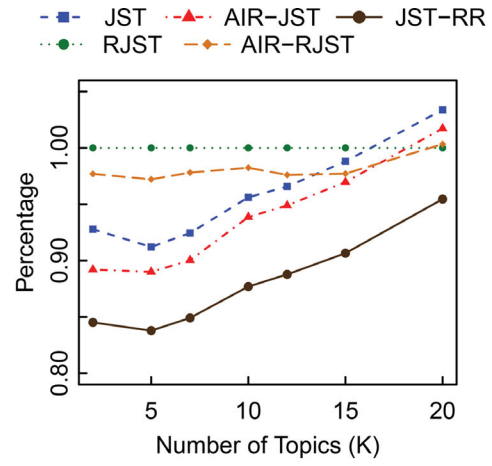
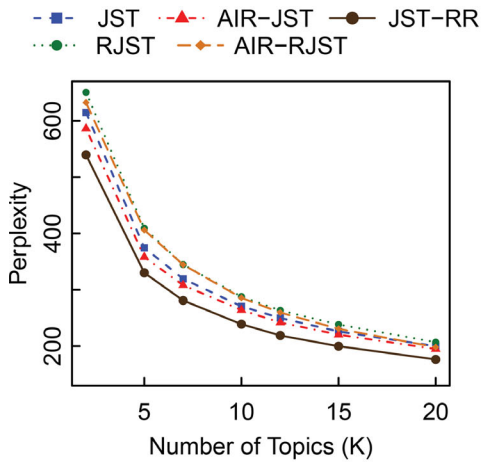
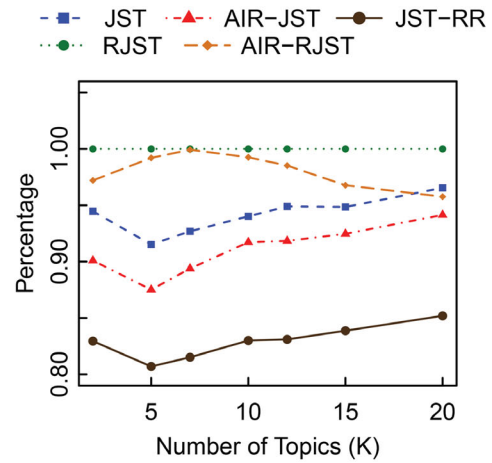
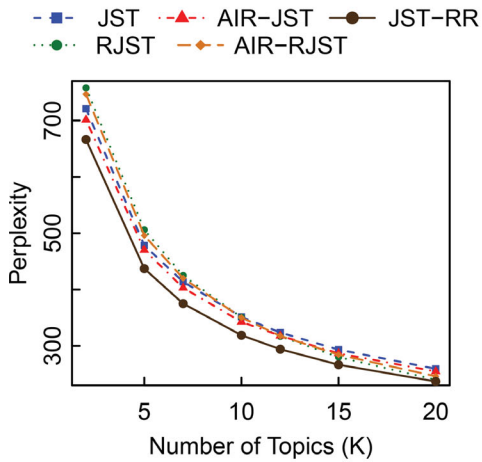
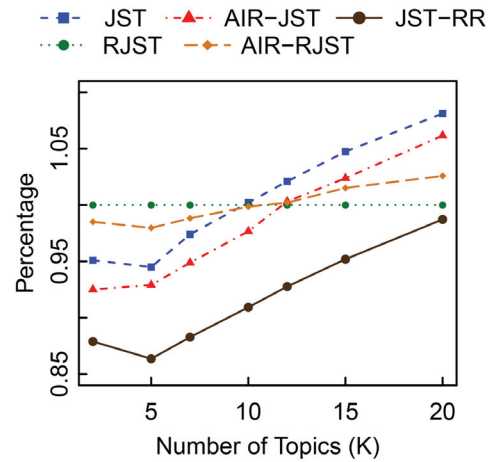
(a) Perplexity on *Lenovo*(b) Percentage of perplexity against RJST on *Lenovo*(c) Perplexity on *Dell*(d) Percentage of perplexity against RJST on *Dell*(e) Perplexity on *HP*(f) Percentage of perplexity against RJST on *HP*

Figure 3. The average results of word perplexity (smaller value indicating better performance) for five methods in the 10-fold cross-validation on three Amazon datasets: *Lenovo*, *Dell*, *HP*. The left column shows the absolute values of word perplexity, and the right column shows the percentages of word perplexity against the baseline of RJST model.

Table 3. Example of topics (e.g., battery, memory and speed, shipping and return, network connection, peripherals) under different sentiment polarities in *Dell* dataset extracted by JST-RR model.

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
like	batteri	gb	problem	work	return	great	connect	good	key
want	use	ram	month	great	refund	love	wireless	like	pad
go	life	processor	time	refurbish	back	need	slow	look	button
know	power	memori	day	good	send	good	tri	nice	mous
come	littl	upgrad	back	new	disappoint	well	wifi	display	click
review	hour	core	fix	well	not_work	school	program	size	time
thought	charg	ghz	repair	thank	bad	recommend	minut	qualiti	annoy
star	replac	intel	start	came	mail	fast	issu	great	back
right	plug	cd	first	recommend	say	happi	back	featur	function
good	cord	hd	get	time	see	product	not_work	bright	thing
better	adapt	cpu	send	servic	que	gift	load	better	tri
sure	charger	dual	burn	packag	sell	perfect	turn	tablet	turn
someth	light	speed	hour	happi	never	purchas	updat	perform	left
back	hot	hdd	bought	fast	buyer	buy	return	color	finger
anoth	quit	bit	warranti	excel	defect	like	driver	model	littl
hope	run	faster	call	unit	miss	thank	boot	resolut	press
make	week	mb	fail	expect	review	easi	fix	weight	cursor
see	minut	machin	issu	day	broken	surf	shut	excel	open
fine	cheap	fast	crash	right	inform	came	frees	love	frees
well	drop	pentium	tech	quickli	sold	basic	network	solid	cheap

NOTE: Topic top words are ranked by their conditional probabilities of occurring under different sentiment polarities given the same topic label.

to the topic of memory and speed (topic 2). It is noted that the extracted results of topics and sentiments in Table 3 are obtained with the only supervision from a domain-independent sentiment lexicon. As a comparison, an extended experiment by considering domain-specific knowledge in the model training can be found in Section B of the [supplementary materials](#).

Moreover, a more general sentiment detection can be examined by the estimated rating distribution. For example, Figures 4(a)–(c) show the estimated rating distribution parameter $\hat{\mu}$ of the three experimental datasets under different sentiment labels with the topic number $K = 5$. It is seen that the positive and negative sentiments are obviously distinguished by their distributions over five rating scores. Such an observation is validated by the results that a positive sentiment tends to produce higher ratings than the negative one, showing consistency with human expectations. Overall, the results above demonstrate that the proposed JST-RR model enables an informative and coherent extraction of both topics and sentiments from the data.

5. Simulation

As model performance varies with the studied review dataset, this section conducts several simulation studies to examine how different characteristics of review corpus, for example, the average review length (or word-rating ratios) and the information value of ratings, will affect the model performance in predicting the document-level sentiment distributions under different model assumptions.

5.1. Simulated Documents

We simulate review documents that are composed of words and ratings with known parameters based on the generative process in Algorithm 1. Specifically, each simulated document is repre-

sented by a random joint sentiment-topic distribution $P(l, z) = \pi_l \theta_{l,z}$ that quantifies how likely the current document is linked to each sentiment and topic label. We let the number of topics $K = 5$ and the number of sentiments $S = 2$. For each review document, we test with the number of ratings $M \in \{1, 2, 3, 4, 5, 7, 10\}$ and the number of words $N \in \{10M, 20M, 30M\}$ for each value of M . Given the sentiment-topic mixtures sampled from $P(l, z)$, a simulated document is generated by sampling words and ratings from the empirical word distribution $\text{Multinomial}(\varphi)$ and rating distribution $\text{Multinomial}(\mu)$, respectively. Without loss of generality, we use the empirical word distribution estimated from the real-world *Dell* dataset in Section 4 for generating the words in simulated documents. In addition, all the ratings are sampled from the empirical rating distribution with parameter μ^{Dell} in Figure 4(a) conditioned on their sentiment assignments.

Accordingly, the rating distribution provides occurrence rules among the observed ratings. For example, based on the rating distribution with parameter μ^{Dell} in Figure 4(a), a positive sentiment is more likely to stimulate a higher rating, while a negative sentiment leads to a lower one. Note that the rating distribution varies with the studied dataset, bringing a variety of information value for model inference, and the simulation data generated with various rating distributions would lead to different results. For conducting a general comparison, our simulation additionally explores two distant cases of rating distributions with the parameters shown in Figure 4(d) and (e). Figure 4(d) represents an extreme case (μ^{diff}) that ratings under two sentiment classifications are totally differentiated. In contrast, Figure 4(e) represents the opposite case (μ^{unif}) that ratings under two sentiment classifications are totally mixed. In practice, the distributions over ratings would range between μ^{diff} and μ^{unif} .

Based on Shannon's concept of information theory, the information gain (IG) on the prediction of sentiments $l \in \{1, \dots, S\}$

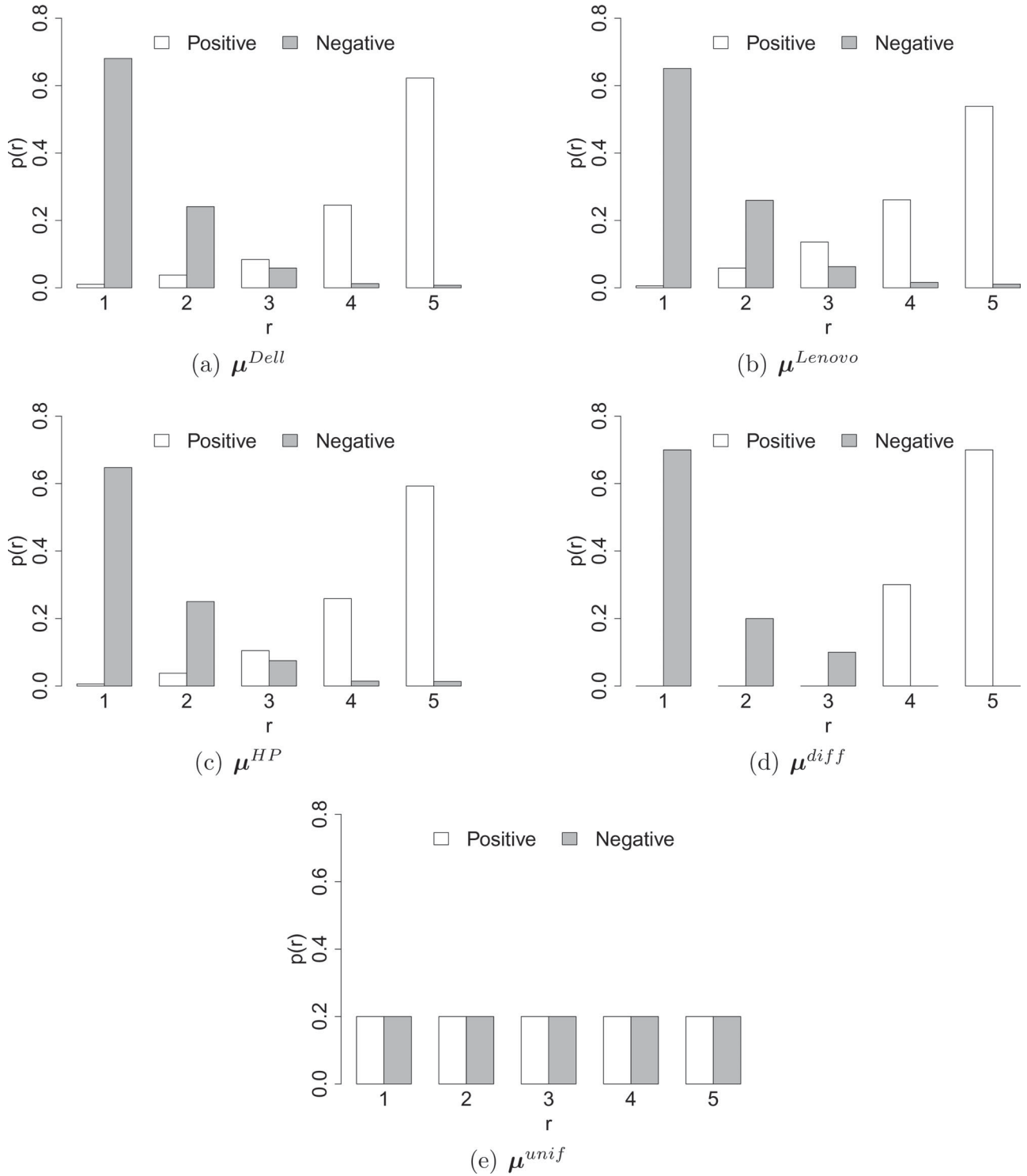


Figure 4. Distributions over ratings under positive and negative sentiments. (a)–(c) are empirical rating distributions of three Amazon datasets: *Dell*, *Lenovo*, *HP*. (d) and (e) present two distinct cases of rating distributions where ratings under two sentiment polarities are totally differentiated and totally mixed, respectively.

given specific ratings $r \in \{1, 2, 3, 4, 5\}$ is defined as

$$\begin{aligned}
 IG(l, r) &= H(l) - H(l|r) = \sum_{r=1}^5 P(r) \sum_{l=1}^S P(l|r) \log P(l|r) \\
 &\quad - \sum_{l=1}^S P(l) \log P(l), \tag{12}
 \end{aligned}$$

which can be regarded as the amount of reduced randomness in predicting a sentiment given a rating. It is easy to show that the information gain in Equation (12) is maximized, namely $H(l|r) = 0$ and $IG(l, r) = H(l)$, in the case of μ^{diff} (Figure 4(d)) where the sentiment prediction is 100% confirmed under each possible rating score. In contrast, it is minimized, namely $IG(l, r) = 0$, at the uniform distribution of

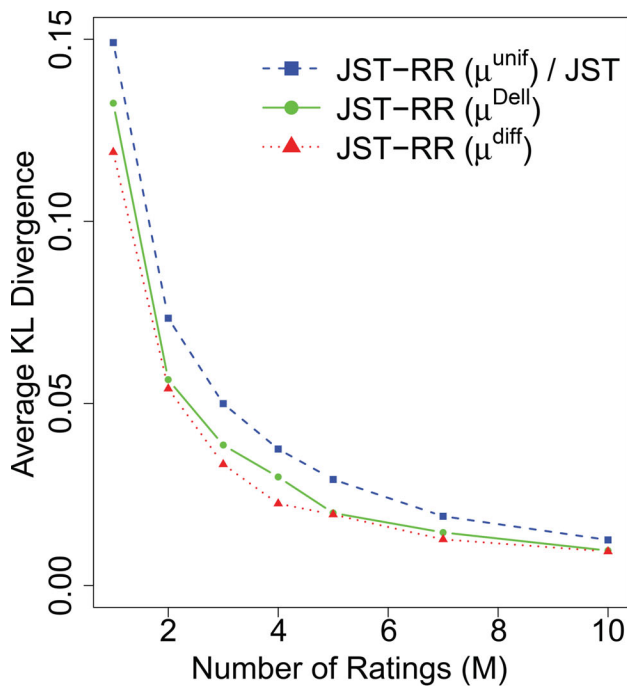
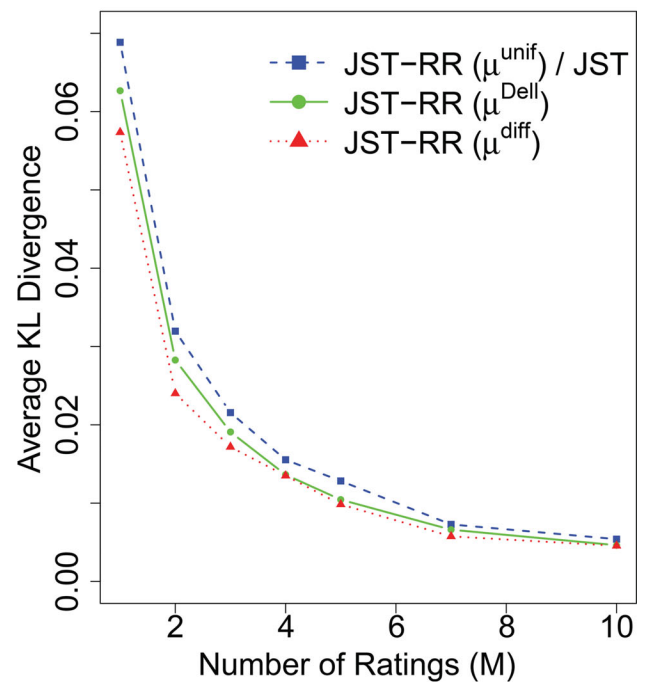
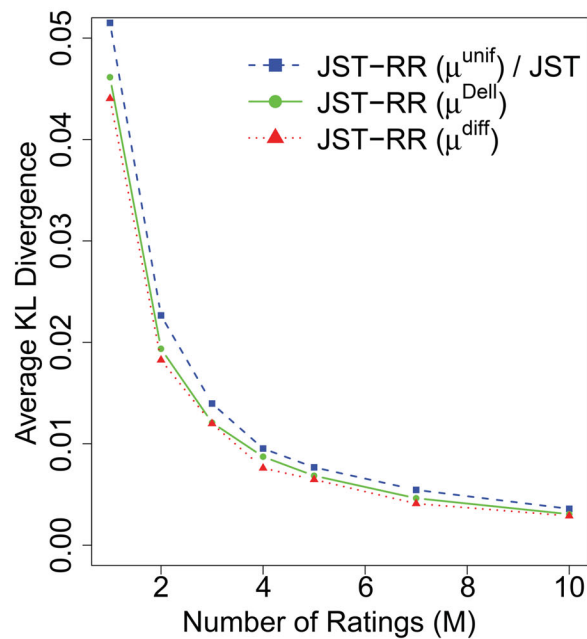
(a) $N/M = 10$ (b) $N/M = 20$ (c) $N/M = 30$

Figure 5. The average KL divergence (smaller value indicating higher accuracy) between the predicted sentiment distribution by various models and the ground truth under a variety of rating number M as well as word-rating ratios ($N/M = 10, 20, 30$).

μ^{diff} (Figure 4(e)). In summary, the information value of ratings in review corpus is measured by the information gain on the prediction of sentiments given various rating distributions,

where μ^{diff} and μ^{unif} represent two distinct cases of the rating's information value reaching its maximum and minimum, respectively.

5.2. Comparison Results

Note that the incorporation of overall ratings mainly makes a difference in the estimation of document-level sentiments. Thus we focus on the accuracy of estimating the sentiment distribution parameter π with the proposed Gibbs sampling algorithm under different model implementations. Specifically, the Kullback Leibler (KL) Divergence (Kullback 1997) is used to evaluate the performance measure of sentiment prediction as

$$D_{\text{KL}}(\hat{\pi}, \pi) = \sum_l \hat{\pi}_l \log \frac{\hat{\pi}_l}{\pi_l}. \quad (13)$$

It measures the distance between the predicted sentiment distribution $\hat{\pi}$ from different models and the target sentiment distribution π (ground-truth). For a general comparison, we consider the following four models:

- JST-RR(μ^{diff}): The JST-RR model applied to simulated documents generated with rating distribution parameter μ^{diff} .
- JST-RR(μ^{unif}): The JST-RR model applied to simulated documents generated with rating distribution parameter μ^{unif} .
- JST-RR(μ^{Dell}): The JST-RR model applied to simulated documents generated with rating distribution parameter μ^{Dell} .
- JST: The JST model only applied to the textual (word) part of simulated documents (baseline).

All the models above are implemented in the same condition. Note that we have not included the implementations of other alternative models (e.g., RJST, AIR-JST, and AIR-RJST) since the simulated data here are based on the generative process of the JST-RR/JST model. The tuning parameter σ is chosen by a 10-fold cross-validation on a separate set of simulated documents.

Figure 5 shows the average results of KL Divergence between the predicted sentiment distribution by various models and the ground truth under different word-rating ratios (e.g., $N/M = 10, 20, 30$). A lower value of KL Divergence indicates higher model accuracy, and each average value of KL Divergence is computed based on $D = 1000$ samples of documents. One can refer to Section C of the [supplementary materials](#) for detailed simulation results plotted in Figure 5.

In general, when N and M (the number of words and the number of ratings) are increased in a document, the document-level sentiment parameters are estimated with higher accuracy. Based on results, the proposed JST-RR model with μ^{diff} appear to achieve the best performance among all scenarios. It indicates that the incorporation of overall ratings in case of a differentiated sentiment-rating distribution with larger information value is helpful for the sentiment prediction. In contrast, the JST-RR model with μ^{unif} is equivalent to the baseline model of JST (i.e., $\sigma = 0$) since the ratings in this case would not contribute to the sentiment prediction. Generally, the improvements of the JST-RR model compared to the baseline model of JST can be explained by the incorporation of informative ratings. When the ratings bring larger information value as in the case of μ^{diff} , the improvements would be more significant. In contrast, when the ratings are noninformative as in the case of μ^{unif} , the improvements are marginal.

By comparing results among different word-rating ratios, it is clearly seen that the improvements in sentiment prediction

become smaller with an increasing word-rating ratio N/M in review documents. For example, the JST-RR model under the word-rating ratio $N/M = 10$ has a significant advantage over the JST model (Figure 5(a)). While the advantage is reduced with the word-rating ratio $N/M = 30$ (Figure 5(c)). It shows that the improvement from complementary ratings in the JST-RR model could be marginal when there are a sufficient amount of words for the document sentiment prediction. In a short summary, the proposed JST-RR model has the advantage for short reviews with insufficient words (or a low word-rating ratio).

6. Discussion

In this work, we propose a joint sentiment-topic model to properly accommodate ratings and review texts. The proposed model characterizes the intrinsic connection between review texts and ratings, leading to accurate prediction on review sentiments and topics. An efficient Gibbs sampling algorithm is developed to make inference for the model parameters. Through the case study on the Amazon datasets, it appears that the proposed JST-RR model can enable an effective identification of latent topics and sentiments in reviews. It is noted that the proposed JST-RR model brings higher improvements in sentiment prediction with a more informative rating distribution and a decreasing word-rating ratio in review documents.

Note that the proposed model is weakly supervised with the only supervision from a domain-independent sentiment lexicon. It can be adapted to other applications easily, such as process monitoring of online products and services (Liang and Wang 2020). For model simplification, only two sentiment polarities (i.e., positive and negative) are considered in the existing experimental settings. In the future work, a neutral sentiment label in addition to the existing sentiment labels is an alternative to separate the background words from sentiment words under each topic. Moreover, one can consider the ratings on some prespecified topics, namely, aspect ratings. In such situations, it is interesting to extend the proposed method to the case where aspect ratings are available, where the topic-sentiment correlation needs to be constructed appropriately by incorporating aspect ratings with review texts. The current proposed method is mainly based on data from one platform, that is, the reviews and ratings from Amazon. Another direction for future research is to incorporate the platform information of reviews into the proposed method such that it can integrate the reviews and ratings of the same or similar products from multiple platforms.

Supplementary Materials

Supplementary document: The pdf file contains: (i) derivations for the model inference in Section 3.3; (ii) extended experiments on Amazon datasets by considering domain-specific knowledge; (iii) and additional simulation results complementing Figure 5.

Code and data: A zip file named “JSTRRexp” contains codes and data to reproduce case study and simulation results in this article.

References

- Airoldi, E. M., and Bischof, J. M. (2016), “Improving and Evaluating Topic Models and Other Models of Text,” *Journal of the American Statistical Association*, 111, 1381–1403. [57]

- Airolidi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (2015), *Handbook of Mixed Membership Models and their Applications*, CRC press. [58]
- Airolidi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C., Love, T., and Shringarpure, S. (2010), “Reconceptualizing the Classification of PNAS Articles,” *Proceedings of the National Academy of Sciences*, 107, 20899–20904. [58]
- Bai, X. (2011), “Predicting Consumer Sentiments from Online Text,” *Decision Support Systems*, 50, 732–742. [57]
- Blei, D. M. (2012), “Probabilistic Topic Models,” *Communications of the ACM*, 55, 77–84. [57]
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, 3, 993–1022. [58,60,62]
- Blitzer, J., Dredze, M., and Pereira, F. (2007), “Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification,” in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 440–447. [58]
- Brody, S., and Elhadad, N. (2010), “An Unsupervised Aspect-Sentiment Model for Online Reviews,” *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 804–812. [58]
- Calheiros, A. C., Moro, S., and Rita, P. (2017), “Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling,” *Journal of Hospitality Marketing & Management*, 26, 675–693. [57]
- Chien, J., and Wu, M. (2008), “Adaptive Bayesian Latent Semantic Analysis,” *IEEE Transactions on Audio, Speech, and Language Processing*, 16, 198–207. [60]
- Dermouche, M., Kouas, L., Velcin, J., and Loudcher, S. (2015), “A Joint Model for Topic-Sentiment Modeling from Text,” in *Association for Computing Machinery Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 819–824. [58]
- Griffiths, T. L., and Steyvers, M. (2004), “Finding Scientific Topics,” *Proceedings of the National Academy of Sciences*, 101, 5228–5235. [60]
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013), “Stochastic Variational Inference,” *The Journal of Machine Learning Research*, 14, 1303–1347. [60]
- Hofmann, T. (1999), “Probabilistic Latent Semantic Indexing,” in *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 50–57. [58]
- Joseph, V. R., and Vakayil, A. (2021), “SPlit: An Optimal Method for Data Splitting,” *Technometrics*, 1–11. [63]
- Kullback, S. (1997), *Information Theory and Statistics*, Mineola, NY: Courier Corporation. [68]
- Li, C., Zhang, J., Sun, J.-T., and Chen, Z. (2013), “Sentiment Topic Model with Decomposed Prior,” in *Proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 767–775. [58]
- Li, H., Lin, R., Hong, R., and Ge, Y. (2015), “Generative Models for Mining Latent Aspects and Their Ratings from Short Reviews,” in *2015 IEEE International Conference on Data Mining*, IEEE, pp. 241–250. [57,58,62]
- Liang, Q., and Wang, K. (2020), “Ratings Meet Reviews in the Monitoring of Online Products and Services,” *Journal of Quality Technology*, 54, 197–214. [68]
- Lin, C., and He, Y. (2009), “Joint Sentiment/Topic Model for Sentiment Analysis,” in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ACM, pp. 375–384. [57,58,61,62]
- Lin, C., He, Y., Everson, R., and Ruger, S. (2012), “Weakly Supervised Joint Sentiment-Topic Detection from Text,” *IEEE Transactions on Knowledge and Data Engineering*, 24, 1134–1145. [58,62]
- Ling, G., Lyu, M. R., and King, I. (2014), “Ratings Meet Reviews, A Combined Approach to Recommend,” in *Proceedings of the 8th ACM Conference on Recommender Systems*, pp. 105–112. [58]
- Liu, B. (2012), “Sentiment Analysis and Opinion Mining,” *Synthesis Lectures on Human Language Technologies*, 5, 1–167. [57]
- Lu, B., Ott, M., Cardie, C., and Tsou, B. K. (2011), “Multi-Aspect Sentiment Analysis with Topic Models,” in *2011 IEEE 11th International Conference on Data Mining Workshops*, IEEE, pp. 81–88. [58]
- Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. (2010), “Exploiting Social Context for Review Quality Prediction,” in *Proceedings of the 19th International Conference on World Wide Web*, ACM, pp. 691–700. [57]
- Lu, Y., Zhai, C., and Sundaresan, N. (2009), “Rated Aspect Summarization of Short Comments,” in *Proceedings of the 18th International Conference on World Wide Web*, pp. 131–140. [58]
- Manrique-Vallier, D., and Reiter, J. P. (2012), “Estimating Identification Disclosure Risk Using Mixed Membership Models,” *Journal of the American Statistical Association*, 107, 1385–1394. [58]
- McAuley, J., and Leskovec, J. (2013), “Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text,” in *Proceedings of the 7th ACM Conference on Recommender Systems*, pp. 165–172. [58]
- McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. (2015), “Image-based Recommendations on Styles and Substitutes,” in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 43–52. [61]
- Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007), “Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs,” in *Proceedings of the 16th International Conference on World Wide Web*, ACM, pp. 171–180. [57,58]
- Moghaddam, S., and Ester, M. (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*, pp. 665–674. [58]
- Pang, B., Lee, L., and Vaithyanathan, S. (2002), “Thumbs up?: Sentiment Classification using Machine Learning Techniques,” in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pp. 79–86. [58]
- Roberts, M. E., Stewart, B. M., and Airolidi, E. M. (2016), “A Model of Text for Experimentation in the Social Sciences,” *Journal of the American Statistical Association*, 111, 988–1003. [57]
- Srivastava, A., and Sutton, C. (2017), “Autoencoding Variational Inference for Topic Models,” arXiv no. 1703.01488. [60]
- Taddy, M. (2013), “Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression,” *Technometrics*, 55, 415–425. [57]
- Titov, I., and McDonald, R. (2008a), “A Joint Model of Text and Aspect Ratings for Sentiment Summarization,” in *Proceedings of ACL-08: HLT*, pp. 308–316. [57,58]
- (2008b), “Modeling Online Reviews with Multi-grain Topic Models,” in *Proceedings of the 17th International Conference on World Wide Web*, pp. 111–120. [57]
- Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009), “Evaluation Methods for Topic Models,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1105–1112. [63]
- Wang, H., Lu, Y., and Zhai, C. (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*, pp. 783–792. [58]
- (2011), “Latent Aspect Rating Analysis Without Aspect Keyword Supervision,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 618–626. [58]
- Ye, Q., Zhang, Z., and Law, R. (2009), “Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches,” *Expert Systems with Applications*, 36, 6527–6535. [58]
- Yu, D., Mu, Y., and Jin, Y. (2017), “Rating Prediction Using Review Texts with Underlying Sentiments,” *Information Processing Letters*, 117, 10–18. [58]