



Ratings meet reviews in the monitoring of online products and services

Qiao Liang & Kaibo Wang

To cite this article: Qiao Liang & Kaibo Wang (2022) Ratings meet reviews in the monitoring of online products and services, Journal of Quality Technology, 54:2, 197-214, DOI: [10.1080/00224065.2020.1829216](https://doi.org/10.1080/00224065.2020.1829216)

To link to this article: <https://doi.org/10.1080/00224065.2020.1829216>



Published online: 22 Oct 2020.



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



Article views: 258



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 3 View citing articles [↗](#)



Ratings meet reviews in the monitoring of online products and services

Qiao Liang and Kaibo Wang 

Tsinghua University, Beijing, China

ABSTRACT

User-generated content including both review texts and user ratings provides important information regarding the customer-perceived quality of online products and services. This article proposes a modeling and monitoring method for online user-generated content. A unified generative model is constructed to combine words and ratings in customer reviews based on their latent sentiment and topic assignments, and a two-chart scheme is proposed for detecting shifts of customer responses in dimensions of sentiments and topics, respectively. The proposed method shows superior performance in shift detection, especially for the sentiment shifts in customer responses, based on the results of simulation and a case study.

KEYWORDS

control charts; joint sentiment-topic model; statistical process control; user-generated content

1. Introduction

Because of the rise of e-commerce worldwide, large amounts of online user-generated content including both product review texts and ratings have become available in recent years. On some large online shopping websites like Amazon and Taobao, users are encouraged to give a general rating score (e.g., 1–5 star) for the products they bought and write a review text to complement the rating. The user-generated content provides important information regarding the customer-perceived quality, with user ratings indicating the latent sentiment polarities and review texts containing topics (or quality characteristics) behind the rating. They have been extensively used for various purposes in previous research, for example, to evaluate various dimensions of service quality (Duan et al. 2013; Sperkova, Vencovsky, and Bruckner 2015), to make predictions on a user's preferences and build recommender systems (Ling, Lyu, and King 2014; Xu, Lam, and Lin 2014).

The monitoring of customer responses based on these user-generated content data is highly desired, as it helps to figure out the current state of online products as well as their service process and plays an important part in the quality control of the after-sales stage. Statistical process control (SPC) is an efficient tool for monitoring various quality characteristics and has been widely used in many cases (Montgomery 2012; Woodall and Montgomery 2014). A

conventional SPC framework is designed for the mechanical process, in which the control charts serve as tools for monitoring responses of the equipment and the mechanized environments. This study explores the applications of SPC methods on the user-generated content, with its focus on the monitoring of customer responses to the online service process. Applications of this research make it possible to quickly detect the hidden evolution in customer opinions and improve the customer perceived quality of online products and services.

Although the use of text data has attracted some interest in process-monitoring applications (e.g., Ashton, Evangelopoulos, and Prybutok 2014, 2015; Lo 2008), the research in this area is still shallow. Our previous work in Liang and Wang (2019) applied review texts to the monitoring of online products and services by using the quality characteristics extracted via text processing algorithms. Despite that review texts can serve individually as an efficient tool for evaluating the product quality and interpreting the topics behind, they could be unstable due to the low density of valid information. For example, customers might not be patient enough to give a detailed review text to accompany the rating, and most collected review texts in real applications are short or suffer from 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). Moreover, each review text needs to be preprocessed by filtering out the nonwords and stop words in it, leaving only the informative words. By contrast, ratings tend to show higher availability and lower noise in spite of lacking topic-related information, and the latent topics/sentiments can be extracted more appropriately with the help of the overall ratings (Li et al. 2015). In consideration of the circumstances above, both types of data are combined and mutually complemented in this study for jointly improving the process monitoring of online products and services.

The focus of this article is to develop a modeling and phase-II monitoring approach for online customer reviews composed of both text words and user ratings. Compared with the continuous or categorical quality characteristics monitored in the conventional SPC framework, it takes more effort to characterize and integrate the features embedded in review texts and ratings for the task of monitoring. The contribution of this article is three-fold. First, we propose a novel method to combine review texts seamlessly with user ratings with a joint generative sentiment-topic model, and an efficient Gibbs sampling method is derived for model inference. Second, a two-chart scheme is proposed for simultaneously detecting shifts in dimensions of user sentiments and topics. Third, a comparison between words and ratings in reviews is demonstrated, and an insight into their properties in the classification and estimation of latent sentiments is achieved.

The remainder of this paper is organized as follows. Section 2 presents the existing methods for modeling review texts and ratings. Section 3 proposes a method for monitoring product review words and ratings jointly in both offline and online stages. Section 4 performs a case study to show the implementation of the proposed method in practice. Section 5 conducts a simulation study to evaluate the performance of the proposed method and compare it with another alternative. Section 6 discusses the informative comparison between words and ratings in recognizing sentiments, and Section 7 concludes the article.

2. Related work

2.1. Text modeling

Different from other structured data that can be directly used for monitoring, the documents of review texts need to be processed into quantitative results through text modeling algorithms. A group of methods that enable the classification of sentiments and the extraction of topics simultaneously in text

documents have been extensively studied recently. Mei et al. (2007) proposed the topic sentiment mixture (TSM) model, the first unified probabilistic approach to model topics and sentiments simultaneously, based on the probabilistic latent semantic indexing (pLSI) model (Hofmann 1999). To solve the problem in the TSM model that induced distributions of the sentiment words are universal and independent of topics, Titov and McDonald (2008b) built topics to represent the rating aspects of customer reviews based on the proposed multi-grain latent Dirichlet allocation (MG-LDA) model, and they extended the model in their subsequent research work, namely the multi-aspect sentiment (MAS) model (Titov and McDonald 2008a), by aggregating sentiment text for the sentiment summary of each rating aspect extracted from MG-LDA. Unlike MAS, which works on a supervised setting with each aspect required to be rated, the joint sentiment-topic (JST) model proposed by Lin and He (2009) is fully unsupervised. The JST model is an extended version of the state-of-the-art topic model, latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003), by adding a sentiment layer to the three-layer structure of LDA. It defines a four-layer probabilistic structure in which each document d is quantified by a multinomial distribution π_d over sentiments and a set of multinomial distributions $\theta_{d,l}$ over topics conditioned on each sentiment label l , and each word in the document is sampled from a multinomial distribution $\phi_{l,z}$ over vocabulary $\{1, \dots, V\}$ conditioned on the word sentiment label l and topic label z . The generative process of documents $d = 1, \dots, D$ composed of individual words in the JST model is presented as follows (see graphical model in Figure 1(a)):

- For each combination of sentiment label $l \in \{1, \dots, S\}$ and topic label $z \in \{1, \dots, K\}$:
 - Let the word counts under sentiment l and topic z follow a multinomial distribution over vocabulary $\{1, \dots, V\}$ with coefficient vector $\phi_{l,z} \sim \text{Dirichlet}(\beta_{l,z})$.
- For each document $d \in \{1, \dots, D\}$:
 - Let the sentiment assignment counts of words in d follow a multinomial distribution over sentiments $\{1, \dots, S\}$ with coefficient vector $\pi_d \sim \text{Dirichlet}(\gamma)$.
 - Conditioned on each sentiment label $l \in \{1, \dots, S\}$, let the topic assignment counts of words in d follow a multinomial distribution over topics $\{1, \dots, K\}$ with coefficient vector $\theta_{d,l} \sim \text{Dirichlet}(\alpha_l)$.
 - For the i th word $w_i, i = 1, \dots, n_d$:

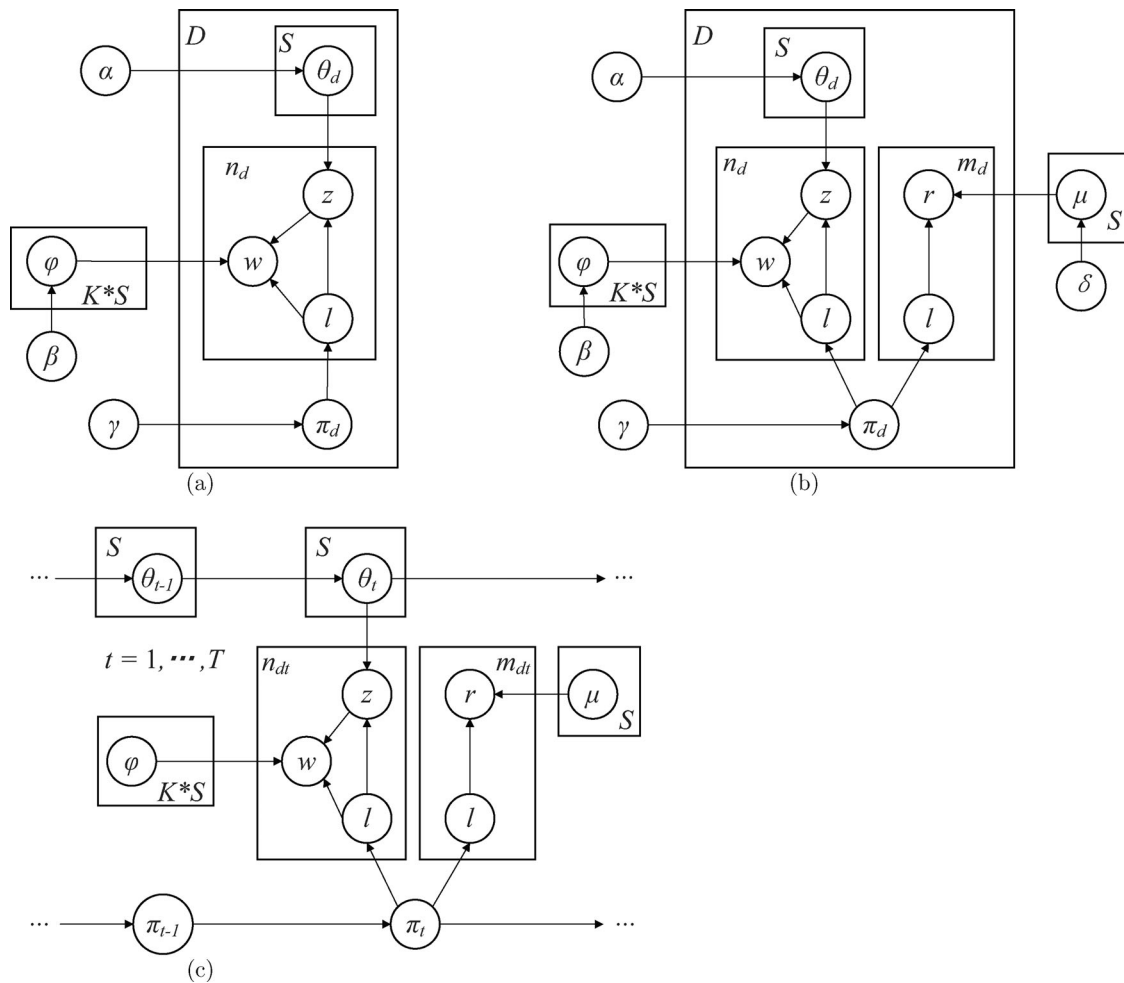


Figure 1. Graphical representations of (a) JST, (b) JST-RMR, and (c) sequential JST-RMR.

- Sample the word sentiment assignment $l_i \sim \text{Multinomial}(\pi_d)$.
- Sample the word topic assignment $z_i \sim \text{Multinomial}(\theta_{d,l_i})$ conditioned on the sentiment l_i .
- Given the word sentiment label l_i and topic label z_i , sample a specific word from vocabulary: $w_i \sim \text{Multinomial}(\phi_{l_i,z_i})$.

The reverse joint sentiment-topic (reverse-JST) model is like a twin version of the JST model that exchanges the sampling sequence of topics and sentiments in the generative process (Lin et al. 2012). Both models map each document to its document-level distributions over sentiments and topics, explaining the co-occurrence rule among words based on the latent sentiment and topic labels they share.

2.2. Ratings meet reviews

The joint modeling of ratings and review texts is more challenging due to the entirely different data

types they present. Most efforts of current research are focused on the recommender systems (e.g., Diao et al. 2014; Ling, Lyu, and King 2014; Xu, Lam, and Lin 2014;) that achieve better predictive accuracy for user ratings with the help of incorporating the review texts, or on the aspect-based opinion mining (e.g., Li et al. 2015; Wang, Lu, and Zhai 2011) that identifies aspects from review texts and reveals ratings on these aspects. A common method (e.g., McAuley and Leskovec 2013; Wang and Blei 2011) is to model user ratings with the latent features of users and product items through matrix factorization, and item features are aligned with the item topic distribution extracted from review texts through the topic models like LDA. A shortage of such an approach is that there is an obvious discrepancy between the item topic distribution obtained from the topic model and the item feature vector from the matrix factorization. Another type of method (e.g., Ling, Lyu, and King 2014; Wang, Lu, and Zhai 2011) is to combine the generating process of both review texts and ratings in a joint probabilistic model, which avoids bridging the gap

between the topic distribution and the feature vector of items. For example, the ratings meet reviews (RMR) model proposed by Ling, Lyu, and King (2014) extended the probabilistic topic model of LDA by incorporating the generative process of ratings that was also decided by their latent topic labels. By connecting ratings and review texts based on the same item topic distribution, the RMR model provided a novel way to combine a review topic model seamlessly with a rating model.

3. Methodology

This study focuses on the joint modeling and monitoring of text words and rating scores in online customer reviews. In this section, we will introduce the proposed method in the offline training stage that prepares for the modeling of review words and ratings and in the online monitoring stage that detects shifts in customer responses throughout the process.

3.1. Offline training

The stage of offline training aims to model words and ratings based on their latent sentiment/topic labels. As we have introduced in Section 2.2, the RMR model (Ling, Lyu, and King 2014) proposed a joint generative structure for combining a topic model seamlessly with a rating model. It assigned a topic label to each observed rating, and this label explained the latent dimension valued by users in the recommender system. Similarly, we plan to incorporate user ratings with review words through a joint generative model in this study.

3.1.1. Model formulation

In order to model review words and ratings jointly, we propose a joint sentiment-topic model in scenarios that ratings meet reviews (JST-RMR). Assume that we have a corpus made of a collection of documents $d = 1, \dots, D$. Each document d is defined as a collection of n_d words and m_d ratings, where each word is an observation from the vocabulary $\{1, \dots, V\}$ and each rating is an item from the given rating scales $\{1, \dots, R\}$. Let S be the number of sentiment labels and K be the number of topics. We follow the assumption of the JST model introduced in Section 2.1 that each document d is represented by an S -dimension multinomial sentiment distribution π_d and a set of K -dimension multinomial topic distributions $\theta_{d,l}$ conditioned on each sentiment label $l = 1, \dots, S$. The document-level sentiment and topic distributions

indicate how likely the current document fits a specific sentiment and topic, respectively. Each word is assumed to be generated from the V -dimension multinomial word distribution $\phi_{l,z}$ conditioned on the word sentiment label l and topic label z , while each rating is generated from the R -dimension multinomial rating distribution μ_l only conditioned on its sentiment label l , considering that ratings provide only the general orientation of sentiments. Similar to the earlier probabilistic generative models such as JST and LDA, the observations in a document (including both words and ratings in this study) are conditionally independent given the document-level distributions over sentiments and topics.

In the offline training stage, each customer review that is composed of a review text and a rating with given scales (e.g., 1 to 5) is treated as a document. A formal generative process of the collection of documents (or customer reviews) $d = 1, \dots, D$ is presented as follows (see graphical model in Figure 1(b)):

- For each combination of sentiment label $l \in \{1, \dots, S\}$ and topic label $z \in \{1, \dots, K\}$:
 - Let the word counts under sentiment l and topic z follow a multinomial distribution over vocabulary $\{1, \dots, V\}$ with coefficient vector $\phi_{l,z} \sim \text{Dirichlet}(\beta_{l,z})$.
- For each sentiment label $l \in \{1, \dots, S\}$:
 - Let the rating counts under sentiment l follow a multinomial distribution over rating scales $\{1, \dots, R\}$ with coefficient vector $\mu_l \sim \text{Dirichlet}(\delta_l)$.
- For each document $d \in \{1, \dots, D\}$:
 - Let the sentiment assignment counts of words and ratings in d follow a multinomial distribution over sentiments $\{1, \dots, S\}$ with coefficient vector $\pi_d \sim \text{Dirichlet}(\gamma)$.
 - Conditioned on each sentiment label $l \in \{1, \dots, S\}$, let the topic assignment counts of words in d follow a multinomial distribution over topics $\{1, \dots, K\}$ with coefficient vector $\theta_{d,l} \sim \text{Dirichlet}(\alpha_l)$.
 - For the i th word w_i in document d :
 - Sample the word sentiment assignment $l_i \sim \text{Multinomial}(\pi_d)$.
 - Sample the word topic assignment $z_i \sim \text{Multinomial}(\theta_{d,l_i})$ conditioned on the sentiment l_i .
 - Given the word sentiment label l_i and topic label z_i , sample a specific word from vocabulary: $w_i \sim \text{Multinomial}(\phi_{l_i, z_i})$.

- For the rating in document d :
 - Sample the rating sentiment assignment $l \sim \text{Multinomial}(\boldsymbol{\pi}_d)$.
 - Given the rating sentiment label l , sample a specific rating scale ...

The hyperparameters $\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}$, and $\boldsymbol{\alpha}$ provide the prior information before any actual words and ratings are observed. For example, γ_l and $\alpha_{l,z}$ can be interpreted as the prior observation counts of the sentiment l sampled from a document and the topic z associated with sentiment l sampled from a document, respectively. Similarly, $\beta_{l,z,w}$ and $\delta_{l,r}$ can be treated as the prior number of times that the word w is associated with sentiment label l and topic label z in the corpus and that the rating r is associated with sentiment label l in the corpus, respectively.

Compared with the generative process in the JST model (see Section 2.1), JST-RMR incorporates a mixture of multinomial distributions over rating scales. Individual words and ratings in a document are connected based on the same document-level sentiment distribution.

3.1.2. Model inference

A target of the application of JST-RMR in the offline training stage is to estimate the joint sentiment/topic-word distribution $\boldsymbol{\phi}$ and the sentiment-rating distribution $\boldsymbol{\mu}$ from the corpus of customer reviews. For model inference, we use Gibbs sampling (Griffiths and Steyvers 2004) for sequentially sampling each variable of interest (i.e., latent topic label z and sentiment label l) from the distribution over that variable given the current values of all other variables and the observed data. For example, we need to specify the following conditional probability of sampling the sentiment label l_i and topic label z_i for the observed word w_i in the document d :

$$P(l_i, z_i | \mathbf{w}, \mathbf{z}^{-i}, \mathbf{I}^{-i}) \propto P(l_i | \mathbf{I}^{-i}) P(z_i | l_i, \mathbf{I}^{-i}, \mathbf{z}^{-i}) P(w_i | l_i, z_i, \mathbf{I}^{-i}, \mathbf{z}^{-i}, \mathbf{w}^{-i}). \quad [1]$$

The superscript i hereafter denotes the data quantity excluding its i th position. Specifically, the first term in Eq. [1] can be presented as

$$P(l_i | \mathbf{I}^{-i}) \propto \int_{\boldsymbol{\pi}_d} P(l_i | \boldsymbol{\pi}_d) P(\boldsymbol{\pi}_d | \mathbf{I}^{-i}) d\boldsymbol{\pi}_d, \quad [2]$$

which can be estimated by the posterior distribution of $\boldsymbol{\pi}_d$:

$$P(\boldsymbol{\pi}_d | \mathbf{I}^{-i}) \propto P(\boldsymbol{\pi}_d) P(\mathbf{I}^{-i} | \boldsymbol{\pi}_d). \quad [3]$$

Because $P(\boldsymbol{\pi}_d)$ is Dirichlet($\boldsymbol{\gamma}$) and conjugate to $P(\mathbf{I}^{-i} | \boldsymbol{\pi}_d)$, the posterior is also a Dirichlet distribution with its mean value as

$$P(l_i | \mathbf{I}^{-i}) = \frac{n_{d,l_i}^{-i} + m_{d,l_i} + \gamma_{l_i}}{n_d^{-i} + m_d + \sum_l \gamma_l}, \quad [4]$$

where n_d and m_d are the total number of words and ratings in the document d , $n_{d,l}$ and $m_{d,l}$ are the number of times that the sentiment label l has been assigned to the words and ratings, respectively, in the document d .

Similarly, the second term in Eq. [1] can be represented as the posterior estimation of θ_{d,l_i} :

$$P(z_i | l_i, \mathbf{I}^{-i}, \mathbf{z}^{-i}) = \frac{n_{d,l_i,z_i}^{-i} + \alpha_{l_i,z_i}}{n_{d,l_i}^{-i} + \sum_z \alpha_{l_i,z}}, \quad [5]$$

where $n_{d,l,z}$ is the number of times that a word in the document d is associated with the sentiment label l and topic label z . For the third term in Eq. [1], its posterior estimation is obtained by integrating out $\boldsymbol{\phi}$ as

$$P(w_i | l_i, z_i, \mathbf{I}^{-i}, \mathbf{z}^{-i}, \mathbf{w}^{-i}) = \frac{n_{l_i,z_i,w_i}^{-i} + \beta_{l_i,z_i,w_i}}{n_{l_i,z_i}^{-i} + \sum_w \beta_{l_i,z_i,w}}, \quad [6]$$

where $n_{l,z}$ is the number of words assigned with the sentiment label l and topic label z in the corpus and $n_{l,z,w}$ is the number of times that the word w is associated with the sentiment label l and topic label z in the corpus.

By combining the results in Eq. [4], Eq. [5], and Eq. [6], the expression for the full conditional probability in Eq. [1] is

$$P(l_i, z_i | \mathbf{w}, \mathbf{z}^{-i}, \mathbf{I}^{-i}) \propto \frac{n_{d,l_i}^{-i} + m_{d,l_i} + \gamma_{l_i}}{n_d^{-i} + m_d + \sum_l \gamma_l} \cdot \frac{n_{d,l_i,z_i}^{-i} + \alpha_{l_i,z_i}}{n_{d,l_i}^{-i} + \sum_z \alpha_{l_i,z}} \cdot \frac{n_{l_i,z_i,w_i}^{-i} + \beta_{l_i,z_i,w_i}}{n_{l_i,z_i}^{-i} + \sum_w \beta_{l_i,z_i,w}}. \quad [7]$$

With a similar procedure, we sample the sentiment label l_i for the observed rating r_i in the document d according to the following conditional probability:

$$P(l_i | \mathbf{r}, \mathbf{I}^{-i}) \propto P(l_i | \mathbf{I}^{-i}) P(r_i | l_i, \mathbf{r}^{-i}, \mathbf{I}^{-i}) = \frac{n_{d,l_i} + m_{d,l_i}^{-i} + \gamma_{l_i}}{n_d + m_d^{-i} + \sum_l \gamma_l} \cdot \frac{m_{l_i,r_i}^{-i} + \delta_{l_i,r_i}}{m_{l_i}^{-i} + \sum_r \delta_{l_i,r}}, \quad [8]$$

where m_l is the number of times that a rating is assigned to the sentiment label l in the corpus and $m_{l,r}$ is the number of times that the rating r is associated with the sentiment label l in the corpus.

When the Markov chain performed by the Gibbs sampler becomes stable, the distribution of words under the sentiment label $l \in \{1, \dots, S\}$ and the topic label $z \in \{1, \dots, K\}$ is approximated as

$$\hat{\phi}_{l,z,w} = \frac{n_{l,z,w} + \beta_{l,z,w}}{n_{l,z} + \sum_w \beta_{l,z,w}}, \quad w = 1, \dots, V, \quad [9]$$

and the distribution of ratings under the sentiment label $l \in \{1, \dots, S\}$ is

$$\hat{\mu}_{l,r} = \frac{m_{l,r} + \delta_{l,r}}{m_l + \sum_r \delta_{l,r}}, \quad r = 1, \dots, R. \quad [10]$$

The Gibbs sampling procedure of the JST-RMR model is shown in Algorithm 1.

Algorithm 1 Gibbs sampling procedure of JST-RMR in the offline stage

Input: Prior parameters $\beta, \delta, \gamma, \alpha$.

Output: Word distribution ϕ and rating distribution μ .

1: Assign initial topic/sentiment labels to all words/ratings at random.

2: **for** each Gibbs sampling iteration **do**

3: **for** each document $d \in \{1, \dots, D\}$ **do**

4: **for** each word w in d **do**

5: Exclude w associated with its sentiment label l and topic label z from variables $n_d, n_{d,l}, n_{d,l,z}, n_{l,z}, n_{l,z,w}$.

6: Sample a new sentiment-topic combination for w based on Eq. [7].

7: Update variables $n_d, n_{d,l}, n_{d,l,z}, n_{l,z}, n_{l,z,w}$ by incorporating the new sentiment/topic label of w .

8: **end for**

9: **for** rating r in d **do**

10: Exclude r associated with its sentiment label l from variables $m_b, m_{l,r}, m_d, m_{d,l}$.

11: Sample a new sentiment assignment for r based on Eq. [8].

12: Update variables $m_b, m_{l,r}, m_d, m_{d,l}$ by incorporating the new sentiment label of r .

13: **end for**

14: **end for**

15: **end for**

16: Estimate ϕ and μ based on Eq. [9] and Eq. [10].

3.2. Online monitoring

In this stage, daily collected words and ratings in customer reviews are represented by their latent distributions over sentiments and topics through a sequential version of JST-RMR model, and these distributions are subsequently monitored for controlling the

ongoing process of online products and services. Specifically, we focus on the quantitative evolution (Dermouche et al. 2014) that indicates the change in proportion of data discussing a specific sentiment-topic combination, and this evolution represents the shift in customer opinions on quality concerns.

3.2.1. Sequential JST-RMR model

For the purpose of daily monitoring, we define a corpus composed of documents $d_t, t = 1, \dots, T$, in which each document d_t is a collection of n_{d_t} words and m_{d_t} ratings gathered from customer reviews in the t th day. In this stage, we treat a document in one day as the unit to be monitored. With a sequential JST-RMR model, the documents are represented by their latent document-level sentiment distribution π and topic distribution θ . A sequential generative process of documents d_1, \dots, d_T with sequential JST-RMR model is described as follows (see graphical model in Figure 1(c)):

- For the t th document $d_t, t = 1, \dots, T$:
 - Let the sentiment assignment counts of words and ratings in d_t follow a multinomial distribution over sentiments $\{1, \dots, S\}$ with coefficient vector $\pi_t \sim \text{Dirichlet}(\rho n_{d_t} + m_{d_t}) \pi_{t-1}$.
 - Conditioned on each sentiment label $l \in \{1, \dots, S\}$, let the topic assignment counts of words in d_t follow a multinomial distribution over topics $\{1, \dots, K\}$ with coefficient vector $\theta_{t,l} \sim \text{Dirichlet}(\rho n_{d_t} \pi_{t-1,l} \theta_{t-1,l})$.
 - For the i th word $w_i, i = 1, \dots, n_{d_t}$:
 - Sample the word sentiment assignment $l_i \sim \text{Multinomial}(\pi_t)$.
 - Sample the word topic assignment $z_i \sim \text{Multinomial}(\theta_{t,l_i})$ conditioned on the sentiment l_i .
 - Given the word sentiment label l_i and topic label z_i , sample a specific word from vocabulary: $w_i \sim \text{Multinomial}(\phi_{l_i, z_i})$.
 - For the i th rating $r_i, i = 1, \dots, m_{d_t}$:
 - Sample the rating sentiment assignment $l_i \sim \text{Multinomial}(\pi_t)$.
 - Given the rating sentiment label l_i , sample a specific rating scale $r_i \sim \text{Multinomial}(\mu_{l_i})$.

Documents are temporally correlated under the Markov assumption. π_t and θ_t are assumed to be generated from their prior Dirichlet distributions conditioned on π_{t-1} and θ_{t-1} , and a weighting parameter ρ is used for adjusting the influence of the prior. Specifically, the first document d_1 as well as π_1 and θ_1

are generated from the target sentiment distribution $\pi^{(0)}$ and topic distribution $\theta^{(0)}$, which supply the prior information at the beginning of online monitoring. The word distribution ϕ and rating distribution μ estimated in the offline stage (see Section 3.1) are directly used in this process.

According to the generative model above, documents are represented by two sets of latent variables, namely, the document-level sentiment distribution π and the document-level topic distribution θ conditioned on sentiments. Gibbs sampling is used in the same way for approximating these variables at each time stamp. For example, the conditional posterior for sampling the sentiment label l_i and the topic label z_i for the i th observed word w_i in the document d_t is obtained by

$$\begin{aligned} & P_t(l_i, z_i | \mathbf{z}^{-i}, \mathbf{I}^{-i}, \phi) \\ & \propto P_t(l_i | \mathbf{I}^{-i}) P_t(z_i | l_i, \mathbf{z}^{-i}, \mathbf{I}^{-i}) P_t(w_i | l_i, z_i, \phi) \\ & = \frac{n_{d_t, l_i}^{-i} + m_{d_t, l_i} + \rho(n_{d_t}^{-i} + m_{d_t}) \pi_{t-1, l_i}}{(n_{d_t}^{-i} + m_{d_t})(1 + \rho)} \\ & \quad \frac{n_{d_t, l_i, z_i}^{-i} + \rho n_{d_t}^{-i} \pi_{t-1, l_i} \theta_{t-1, l_i, z_i}}{n_{d_t, l_i}^{-i} + \rho n_{d_t}^{-i} \pi_{t-1, l_i}} \cdot \varphi_{l_i, z_i, w_i}. \end{aligned} \quad [11]$$

Similarly, the i th observed rating r_i in the document d_t is assigned with its sentiment label l_i according to the following probability:

$$\begin{aligned} P_t(l_i | \mathbf{I}^{-i}, \mu) & \propto P_t(l_i | \mathbf{I}^{-i}) P_t(r_i | l_i, \mu) \\ & = \frac{n_{d_t, l_i} + m_{d_t, l_i}^{-i} + \rho(n_{d_t} + m_{d_t}^{-i}) \pi_{t-1, l_i}}{(n_{d_t} + m_{d_t}^{-i})(1 + \rho)} \cdot \mu_{l_i, r_i}. \end{aligned} \quad [12]$$

A sample obtained from the Markov chain performed by the Gibbs sampler can be used to approximate the posterior distribution of sentiments for the t th document as follows:

$$\begin{aligned} \hat{\pi}_{t, l} & = \frac{n_{d_t, l} + m_{d_t, l} + \rho(n_{d_t} + m_{d_t}) \hat{\pi}_{t-1, l}}{(n_{d_t} + m_{d_t})(1 + \rho)} \\ & = \frac{n_{d_t, l} + m_{d_t, l}}{n_{d_t} + m_{d_t}} \cdot \frac{1}{1 + \rho} + \hat{\pi}_{t-1, l} \cdot \frac{\rho}{1 + \rho}, \quad l = 1, \dots, S, \end{aligned} \quad [13]$$

where a total number of n_{d_t} words and m_{d_t} ratings in the document d_t are equally used for the estimation of document-level sentiments. A general weighting mechanism between words and ratings has been explored in the Appendix. Moreover, the posterior distribution over topics under the sentiment label $l \in \{1, \dots, S\}$ in the t th document is given by

$$\begin{aligned} \hat{\theta}_{t, l, z} & = \frac{n_{d_t, l, z} + \rho n_{d_t} \hat{\pi}_{t-1, l} \hat{\theta}_{t-1, l, z}}{n_{d_t, l} + \rho n_{d_t} \hat{\pi}_{t-1, l}} \\ & \approx \frac{n_{d_t, l, z}}{n_{d_t, l}} \cdot \frac{1}{1 + \rho} + \hat{\theta}_{t-1, l, z} \cdot \frac{\rho}{1 + \rho}, \quad z = 1, \dots, K. \end{aligned} \quad [14]$$

Eq. [13] and Eq. [14] present the posterior estimations of π and θ with the exponentially weighted moving average (EWMA)-like version, where $1/(1 + \rho)$ plays the role of the smoothing parameter λ in EWMA (empirically set at 0.3 in this study). Changes in document-level sentiments and topics are accumulated over time in this way. Moreover, the distributions of sentiments are estimated by words and ratings jointly, while only words are considered in the estimation of topic distributions as ratings are not assigned with topic labels. The Gibbs sampling procedure of the sequential JST-RMR is shown in Algorithm 2.

Algorithm 2 Gibbs sampling procedure of sequential JST-RMR in the online stage

Input: Word distribution ϕ , rating distribution μ , target sentiment distribution $\pi^{(0)}$, and topic distribution $\theta^{(0)}$.

Output: Document-level sentiment distribution π_t and topic distribution θ_t .

- 1: Assign initial topic/sentiment labels to all words/ratings at random.
- 2: **for** each document d_t **do**
- 3: **for** each Gibbs sampling iteration **do**
- 4: **for** each word w in d_t **do**
- 5: Exclude w associated with its sentiment label l and topic label z from variables $n_{d_t}, n_{d_t, l}, n_{d_t, l, z}$.
- 6: Sample a new sentiment-topic combination for w based on Eq. [11].
- 7: Update variables $n_{d_t}, n_{d_t, l}, n_{d_t, l, z}$ by incorporating the new sentiment/topic label of w .
- 8: **end for**
- 9: **for** each rating r in d_t **do**
- 10: Exclude r associated with its sentiment label l from variables $m_{d_t}, m_{d_t, l}$.
- 11: Sample a new sentiment assignment for r based on Eq. [12].
- 12: Update variables $m_{d_t}, m_{d_t, l}$ by incorporating the new sentiment label of r .
- 13: **end for**
- 14: **end for**
- 15: Estimate π_t and θ_t based on Eq. [13] and Eq. [14].
- 16: **end for**

3.2.2. Control charts

Considering that the state of customer responses is reflected by the document-level sentiment and topic distributions estimated above, we conduct the following hypothesis test to check the shifts in user sentiments and topics simultaneously for the t th document:

$$\begin{aligned} H_0 : \boldsymbol{\pi}_t &= \boldsymbol{\pi}^{(0)} \text{ and } \boldsymbol{\theta}_t = \boldsymbol{\theta}^{(0)}, \\ H_1 : \boldsymbol{\pi}_t &\neq \boldsymbol{\pi}^{(0)} \text{ or } \boldsymbol{\theta}_t \neq \boldsymbol{\theta}^{(0)}. \end{aligned} \quad [15]$$

The Kullback–Leibler (KL) divergence (Kullback 1959) is used for measuring the distance between the estimated document-level sentiment/topic distributions and their target values. For example, the KL divergence (also known as relative entropy) between the t th estimated joint sentiment-topic distribution $\hat{P}_t(l, z)$ and the in-control (IC) joint distribution $P_0(l, z)$ is defined as

$$D_{\text{KL}}(\hat{P}_t(l, z), P_0(l, z)) = \sum_{l=1}^S \sum_{z=1}^K \hat{P}_t(l, z) \log \frac{\hat{P}_t(l, z)}{P_0(l, z)}. \quad [16]$$

This distance measure is nonnegative, and it is equal to zero if and only if $\hat{P}_t(l, z) = P_0(l, z)$ for every sentiment category l and topic category z . In addition, this divergence can be further decomposed into the following independent terms:

$$\begin{aligned} & D_{\text{KL}}(\hat{P}_t(l, z), P_0(l, z)) \\ &= \sum_{l=1}^S \sum_{z=1}^K \hat{P}_t(l, z) \log \frac{\hat{P}_t(l, z)}{P_0(l, z)} \\ &= \sum_{l=1}^S \hat{P}_t(l) \log \frac{\hat{P}_t(l)}{P_0(l)} + \sum_{l=1}^S \hat{P}_t(l) \sum_{z=1}^K \hat{P}_t(z|l) \log \frac{\hat{P}_t(z|l)}{P_0(z|l)} \\ &= D_{\text{KL}}(\hat{P}_t(l), P_0(l)) + \sum_{l=1}^S \hat{P}_t(l) \cdot D_{\text{KL}}(\hat{P}_t(z|l), P_0(z|l)) \\ &= D_{\text{KL}}(\hat{\boldsymbol{\pi}}_t, \boldsymbol{\pi}^{(0)}) + \sum_{l=1}^S \hat{\pi}_{t,l} \cdot D_{\text{KL}}(\hat{\boldsymbol{\theta}}_{t,l}, \boldsymbol{\theta}_l^{(0)}), \end{aligned} \quad [17]$$

where the first term represents the distance between the conditioning variables (i.e., sentiments) and the second term represents the distance between the conditioned variables (i.e., topics).

Kullback (1959) has shown that the KL divergence (multiplied by a constant) between a c -category multinomial distribution $P(x)$ and its estimated distribution $\hat{P}(x)$ is asymptotically chi-squared distributed with $c - 1$ degrees of freedom:

$$2N \cdot D_{\text{KL}}(\hat{P}(x), P(x)) \rightarrow \sum_{x \in X} \frac{(n(x) - NP(x))^2}{NP(x)} \sim \chi_{c-1}^2, \quad [18]$$

where N is the size of a sample from the population distributed by $P(x)$, $n(x)$ is the count of observations that is assigned to category x in the sample, and the estimation of $P(x)$ is given by $\hat{P}(x) = n(x)/N$.

In this study, we propose a two-chart control scheme by implementing the KL divergence in the dimensions of sentiments and topics, respectively. Based on the decomposed terms in Eq. [17], the proposed charting statistics of individual charts are defined as

$$\begin{aligned} Q_t^{\text{sentiment}} &= 2(N_t + M_t) \cdot D_{\text{KL}}(\hat{\boldsymbol{\pi}}_t, \boldsymbol{\pi}^{(0)}) \\ &= 2(N_t + M_t) \cdot \sum_{l=1}^S \hat{\pi}_{t,l} \log \frac{\hat{\pi}_{t,l}}{\pi_l^{(0)}}, \end{aligned} \quad [19]$$

$$\begin{aligned} Q_t^{\text{topic}} &= 2N_t \cdot \sum_{l=1}^S \hat{\pi}_{t,l} D_{\text{KL}}(\hat{\boldsymbol{\theta}}_{t,l}, \boldsymbol{\theta}_l^{(0)}) \\ &= 2N_t \cdot \sum_{l=1}^S \hat{\pi}_{t,l} \sum_{z=1}^K \hat{\theta}_{t,l,z} \log \frac{\hat{\theta}_{t,l,z}}{\theta_{l,z}^{(0)}}, \end{aligned} \quad [20]$$

where $N_t = (1 + \rho)n_{d_t}$ is the total sample size of review words composed of n_{d_t} actual assignment counts and ρn_{d_t} prior virtual counts in the t th document, and similarly, $M_t = (1 + \rho)m_{d_t}$ is the total sample size of ratings. Variables in Eq. [19] and Eq. [20] are independent of each other, and they measure the shifts on document-level sentiments and sentiment-conditioned topics, respectively. Considering that the above two variables are in different scales, we do not sum them up for joint monitoring as in the previous work (Liang and Wang 2019); instead, a two-chart scheme is constructed for monitoring $Q_t^{\text{sentiment}}$ and Q_t^{topic} separately. An out-of-control (OC) signal is triggered when either $Q_t^{\text{sentiment}} > L^{\text{sentiment}}$ or $Q_t^{\text{topic}} > L^{\text{topic}}$, where $L^{\text{sentiment}}$ and L^{topic} are the control limits chosen for individual charts based on a specific reject region. We name it the sequential JST-RMR scheme.

Based on the property in Eq. [18], $Q_t^{\text{sentiment}}$ and Q_t^{topic} are asymptotically chi-squared distributed in the IC state with $S - 1$ and $S(K - 1)$ degrees of freedom, respectively. However, as the assignments of sentiments and topics for multinomial estimation in practice are not obtained from actual i.i.d. observations but from the results of Gibbs sampling, the accurate distributions of $Q_t^{\text{sentiment}}$ and Q_t^{topic} are more complicated than the chi-square distribution. Thus, the control limits for individual charts in this study are obtained through simulations such that a specified overall IC average run length (IC-ARL) of the two-chart scheme is achieved.

3.2.3. Diagnosis

Besides detecting shifts in the process, it is important to identify the root cause of the observed shift. For example, a targeted quality improvement requires figuring out the topic or quality aspect in which customers show their sentiment shifts. The previous model in Liang and Wang (2019) achieved diagnosis through tracing the decomposed terms of the KL divergence. We follow this procedure of diagnosis for identifying the truly OC terms in this study. Specifically, the KL divergence in Eq. [16] that measures the distance between the estimated and the target joint sentiment-topic distributions is decomposed into

$$\begin{aligned}
 & D_{\text{KL}}(\hat{P}_t(l, z), P_0(l, z)) \\
 &= \sum_{l=1}^S \sum_{z=1}^K \hat{P}_t(l, z) \log \frac{\hat{P}_t(l, z)}{P_0(l, z)} \\
 &= \sum_{z=1}^K \hat{P}_t(z) \log \frac{\hat{P}_t(z)}{P_0(z)} + \sum_{z=1}^K \hat{P}_t(z) \sum_{l=1}^S \hat{P}_t(l|z) \log \frac{\hat{P}_t(l|z)}{P_0(l|z)} \\
 &= D_{\text{KL}}(\hat{P}_t(z), P_0(z)) + \sum_{z=1}^K \hat{P}_t(z) D_{\text{KL}}(\hat{P}_t(l|z), P_0(l|z)),
 \end{aligned} \tag{21}$$

where the decomposed terms measure the shifts on topics and topic-specific sentiments, respectively. According to the decomposition above, the signal of topic shifts in the t th document is defined as

$$G_t = 2N_t \cdot D_{\text{KL}}(\hat{P}_t(z), P_0(z)). \tag{22}$$

And the signal of sentiment shifts conditioned on the topic $z \in \{1, \dots, K\}$ in the t th document is represented by

$$U_{t,z} = 2N_t \hat{P}_t(z) \cdot D_{\text{KL}}(\hat{P}_t(l|z), P_0(l|z)). \tag{23}$$

In general, the signal that corresponds to the truly OC term is supposed to grow significantly from its IC state. As we do not have a direct per-document topic distribution $\hat{P}_t(z)$ and topic-specific sentiment distributions $\hat{P}_t(l|z)$, a transformation is conducted prior to the diagnosis. For example, the marginal distribution over topic labels for the t th document is obtained by

$$\begin{aligned}
 \hat{P}_t(z) &= \sum_{l=1}^S \hat{P}_t(l) \hat{P}_t(z|l) \\
 &= \sum_{l=1}^S \hat{\pi}_{t,l} \hat{\theta}_{t,l,z}, \quad z = 1, \dots, K.
 \end{aligned} \tag{24}$$

And the sentiment distribution conditioned on the topic label $z \in \{1, \dots, K\}$ is given by

$$\begin{aligned}
 \hat{P}_t(l|z) &= \frac{\hat{P}_t(l) \hat{P}_t(z|l)}{\hat{P}_t(z)} \\
 &= \frac{\hat{\pi}_{t,l} \hat{\theta}_{t,l,z}}{\sum_{l=1}^S \hat{\pi}_{t,l} \hat{\theta}_{t,l,z}}, \quad l = 1, \dots, S.
 \end{aligned} \tag{25}$$

3.3. Alternative approach: SRJST scheme

The sequential JST-RMR scheme incorporates the user ratings that serve as an important part in the joint monitoring. With an intention of seeing how much the rating part of the mixed data would improve monitoring, the previously proposed SRJST model (Liang and Wang 2019) that considers only review texts is regarded as a benchmark.

For a fair comparison, a two-chart scheme is built based on the SRJST model to balance the shift detection in topics and sentiments. For each document $d_t, t = 1, \dots, T$, the SRJST model produces an estimation of the document-level topic distribution $\hat{P}'_t(z)$ and the topic-specific sentiment distribution $\hat{P}'_t(l|z)$ (differentiated by superscript ' hereafter), which are easily transformed into the document-level sentiment marginal distribution $\hat{P}'_t(l)$ and the conditional topic distribution $\hat{P}'_t(z|l)$. Similarly, the independent variables to be monitored for the t th document in the two-chart scheme of the SRJST are defined as

$$Q_t^{\text{sentiment}} = 2N_t \cdot D_{\text{KL}}(\hat{P}'_t(l), P_0(l)), \tag{26}$$

$$Q_t^{\text{topic}} = 2N_t \cdot \sum_{l=1}^S \hat{P}'_t(l) D_{\text{KL}}(\hat{P}'_t(z|l), P_0(z|l)). \tag{27}$$

The charting statistics above are only dependent on $N_t = (1 + \rho)n_{d_t}$ review words that are composed of n_{d_t} actual assignment counts and ρn_{d_t} prior virtual counts in the t th document.

4. Case study

In this section, an implementation of the proposed scheme is demonstrated on the publicly available Amazon data sets (McAuley et al. 2015). We modified Phan's Gibbs LDA++ package¹ for the model implementation. Specifically, we select the customer reviews related to Dell computers in 2013 and 2014 from the Amazon data sets. Data preprocessing is performed on reviews by removing nonwords and stop words. In addition, we stem words to their roots and delete the infrequent words to reduce the vocabulary size. The

¹<http://gibbslda.sourceforge.net/>

final corpus includes a total of 63,630 ratings and 446,219 words.

4.1. Offline training

First, in the offline training stage, the proposed JST-RMR model is applied on the collection of individual reviews for training word distributions ϕ and rating distributions μ (see Section 3.1). For the Dell computer corpus, we set the number of sentiment classifications $S=2$ (i.e., positive and negative) and the number of topics $K=10$ such that a high likelihood (or low perplexity) on the held-out test set is achieved (Blei, Ng, and Jordan 2003). We follow the symmetric hyperparameter setting of α in Lin et al. (2012): $\alpha_{i,z} = 0.05n/(S \times K)$, where n is the average number of words in a document, and the value of 0.05 makes the prior represented by α weighted as around 0.05 of the actual observation counts in the document. Similarly, we have $\delta_{l,r} = 0.05M/(S \times R)$, where M is the total number of ratings in the corpus, and R is the total number of rating scales. Asymmetric γ is used in this study to capture different correlations among sentiment labels. For the experimental corpus, γ_l is set to 1.6 for the positive sentiment l and 0.5 for the negative sentiment l by seeking the maximum likelihood of the observed data (Huang 2005).

Considering that many words are commonly treated as positive (e.g., “excellent”) or negative (e.g., “terrible”) regardless of the topics or domains involved, we use the method in Lin et al. (2012) that the sentiment classification of words is weakly supervised by incorporating a subjectivity lexicon into the prior setting of hyperparameter β . We select 1,051 positive words and 2,145 negative words from the sentiment lexicon MPQA² whose polarity orientations are domain independent. For the positive sentiment l , we set elements in β_l to be 0 for the words in the 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. This setting enables that the words in sentiment lexicons can only be drawn from the word distributions conditioned on their corresponding sentiment labels.

The estimated word distributions conditioned on combinations of topics and sentiments define the probability of words arising from a specific topic-sentiment pair and explain the quality aspects embedded in topics. The training results of word distributions are presented in Table 1, which shows the most

frequent words under each topic-sentiment combination. Different performance aspects of Dell computers are discovered, such as the battery (topic 1), speed (topic 2), price (topic 9), and product appearance (topic 10). There are also some aspects regarding service quality, including shipping and return (topic 4) and warranty (topic 5).

Compared with word distributions, rating distributions are decided only by their latent sentiment labels. Figure 2 shows the estimated probability density of 1 to 5 rating scales under positive and negative sentiment labels. The training result is natural and expected that the positive sentiment corresponds to higher user ratings and vice versa.

4.2. Control charts

We focus on the phase-II monitoring of daily collected review documents including both words and ratings in this case. Specifically, a total number of 150 documents from 6/20/2013 to 11/16/2013 are collected as IC samples. In the phase-II monitoring, a set of 60 documents from 11/17/2013 to 1/15/2014 are collected as phase-II samples and demonstrated in the proposed control charts.

We set the smoothing parameter $1/(1 + \rho) = 0.3$ for both schemes of sequential JST-RMR and SRJST. Simulations with bootstrap resampling from IC documents are used for the approximation of control limits such that the overall IC-ARL of 200 is achieved. We assume that the individual charts that detect sentiment and topic shifts, respectively, have the same IC-ARL. Because the individual charting statistics are independent, it is easy to show that the overall IC-ARL of 200 can be achieved once an IC-ARL of $(1 - \sqrt{1 - 200^{-1}})^{-1} = 400$ is selected for each individual chart (Mukherjee, McCracken, and Chakraborti 2015).

Figure 3 presents the control charts of sequential JST-RMR and SRJST in 210 days from 6/20/2013 to 1/15/2014, with the first 150 documents as IC samples and the subsequent 60 documents as phase-II samples. It could be seen that the proposed method is effective in detecting changes in phase-II samples. Both schemes of sequential JST-RMR and SRJST trigger their OC signals at the 160th sample. When looking from individual charts of both schemes, shifts are observed in dimensions of both sentiments and topics. The topic charts of both schemes show similar performance in topic-shift detection, while the sentiment chart of sequential JST-RMR enables a quicker detection of sentiment shifts.

²<http://www.cs.pitt.edu/mpqa/>

Table 1. Top words under 20 topic-sentiment pairs.

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Batteri	time	gb	drive	window	window	work	amazon	support	review
Power	hour	processor	hard	xp	instal	new	return	servic	problem
Life	work	ram	dvd	system	system	arriv	ship	call	product
Hour	problem	memori	cd	use	softwar	great	sent	warranti	custom
Charg	start	intel	disk	upgrad	updat	amazon	day	custom	issu
Onli	day	core	gb	like	upgrad	well	receiv	tech	servic
Star	month	ghz	replac	new	xp	receiv	back	fix	mani
Use	tri	drive	onli	oper	microsoft	good	seller	new	bad
Last	use	hard	instal	want	driver	box	order	help	peopl
Suppli	issu	card	extern	work	load	ship	replac	phone	say
Good	week	graphic	ssd	run	run	product	send	replac	model
Work	first	fast	optic	learn	oper	time	refund	send	never
Cord	boot	hd	player	still	offic	packag	week	repair	seem
Plug	turn	cpu	usb	bit	work	fast	took	contact	thing
New	minut	speed	work	machin	viru	happi	repair	ask	think
Like	back	ssd	ram	better	problem	like	time	person	compani
Be caus	shut	dual	doe	love	mcafe	order	call	even	experi
Come	got	pentium	use	desktop	use	item	anoth	work	fix
Doe	sever	gig	space	os	version	condit	item	technic	updat
Adapt	power	perform	tb	win	boot	got	compani	tri	see

Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Internet	usb	game	fan	screen	screen	purchas	purchas	look	look
Use	port	video	littl	keyboard	key	think	year	great	design
Web	wireless	play	doe	touch	keyboard	best	old	like	inspiron
Offic	connect	card	run	like	button	anoth	buy	good	littl
Work	card	graphic	onli	mous	touch	price	use	size	model
Program	hdmi	watch	thing	use	mous	want	month	work	size
Surf	failur	movi	heat	nice	use	better	replac	recommend	qualiti
Open	wifi	work	sound	good	pad	time	ago	light	color
Load	bluetooth	great	hot	great	click	good	last	nice	macbook
Fast	slot	good	time	realli	type	brand	time	easi	feel
Download	vga	music	bit	light	touchpad	refurbish	inspiron	small	howev
App	onli	high	nois	pad	back	product	price	design	inch
Like	internet	well	long	fast	press	buy	never	love	compar
Microsoft	network	stream	speaker	love	bottom	worth	problem	perfect	less
Want	monitor	handl	loud	easi	plastic	recommend	hp	pretti	bright
Run	cabl	fast	bad	better	finger	spend	still	weight	pro
File	problem	perform	lap	backlit	hand	money	befor	machin	heavi
Process	speaker	abl	cpu	resolut	case	like	school	want	cheap
Page	routr	enough	put	feel	side	new	first	carri	notebook
Start	plug	email	use	well	left	happi	work	solid	display

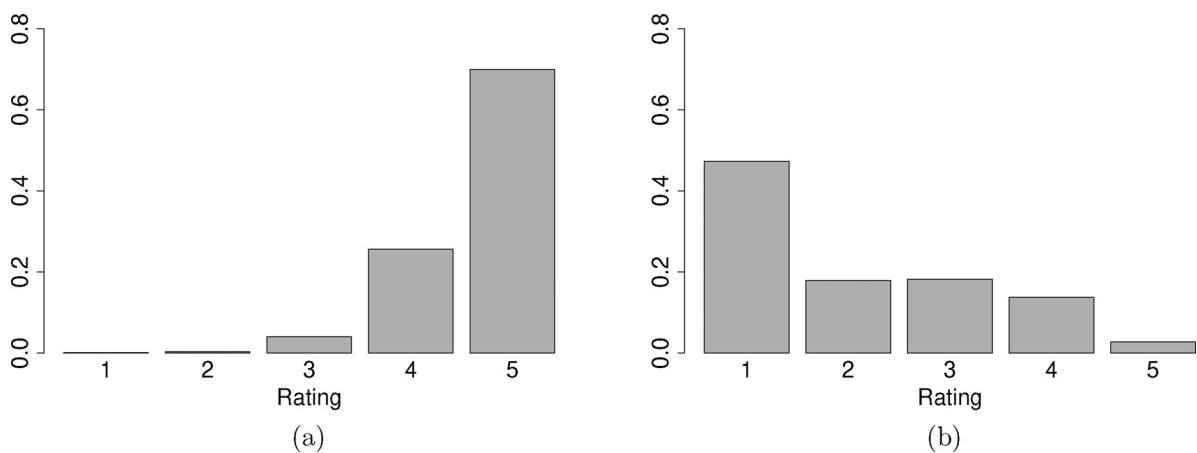


Figure 2. The distribution over ratings under (a) positive and (b) negative sentiment labels.

We can see from Figure 3 that major shifts occur during the time period from the 160th to the 175th sample (11/26/2013–12/11/2013). The average daily rating score in this period is 3.578, while the average

daily rating score in IC samples is 3.806. Moreover, the average KL divergence between the joint sentiment-topic distribution in this period and its IC value is 0.0386. By tracing the decomposed variables

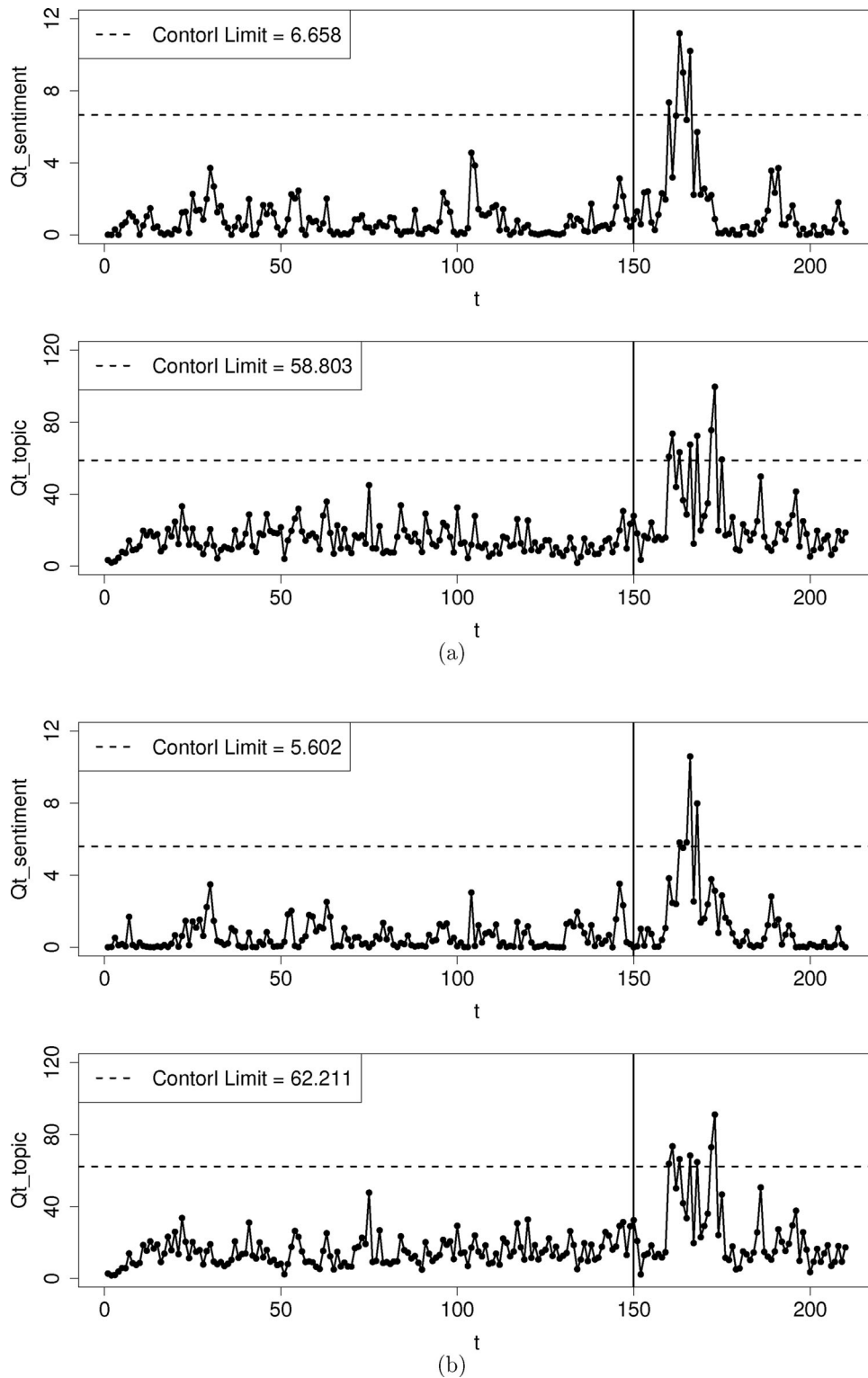


Figure 3. Control charts of (a) sequential JST-RMR scheme and (b) SRJST scheme.

measuring shifts of various dimensions (see Section 3.2.3), we find that such changes are mainly due to increased negative sentiment under topics shipping and return (topic 4), warranty (topic 5), and product appearance (topic 10). Although an accurate identification of the root cause for such changes needs more

careful studies from different perspectives, it is probably reasonable to link the traditional shopping festival at the end of November with this observed phenomenon; factors such as soaring sales volume and impulsive spending behavior may lead to deteriorated performance in the quality dimensions above.

5. Simulation

In this section, we evaluate the performance of the proposed sequential JST-RMR scheme and the alternative SRJST scheme in detecting the shifts of document-level sentiments and topics for simulated documents through ARL comparisons. Given the same IC-ARL (e.g., 370 in this study), the control scheme with shorter OC-ARLs is assumed to outperform the other scheme.

5.1. Simulated documents

We have introduced the Dell computer data set in Section 4. Based on the word distributions ϕ and rating distributions μ of the Dell computer corpus estimated in Section 4.1, the simulated IC and OC documents are generated by sampling words and ratings from their respective distributions under various sentiment and topic mixtures. For example, with the IC joint sentiment-topic distribution $P_0(l, z)$, the simulated IC documents d_1, \dots, d_T are generated as follows:

- For each document $d_t, t = 1, \dots, T$:
 - Draw the number of words $n_{d_t} \sim \text{Poisson}(n)$.
 - Draw the number of ratings $m_{d_t} \sim \text{Poisson}(m)$.
 - For the i th word $w_i, i = 1, \dots, n_{d_t}$:
 - Draw the sentiment and topic assignments $l_i, z_i \sim P_0(l, z)$.
 - Draw a specific word $w_i \sim \text{Multinomial}(\phi_{l_i, z_i})$ conditioned on the sentiment l_i and topic z_i .
 - For the i th rating $r_i, i = 1, \dots, m_{d_t}$:
 - Draw the sentiment assignment $l_i \sim P_0(l) = \sum_z P_0(l, z)$.
 - Draw a specific rating $r_i \sim \text{Multinomial}(\mu_{l_i})$ conditioned on the sentiment l_i .

5.2. Comparison results for shift detection

Different schemes are applied to the simulated documents and compared according to their OC-ARLs in detecting shifts of document-level sentiments and topics. The control limits of individual charts in each scheme are decided through simulations so that each of the two individual charts has the same IC-ARL of 740 and the resulting two-chart combo scheme has an overall IC-ARL of 370.

First, the shift is assumed to occur only on the document-level sentiment distribution $P(l|z)$ under a specific topic label z , while the marginal topic distribution $P(z)$ remains unchanged. Table 2 shows the comparison results (i.e., the OC-ARLs together with

their standard errors) of the proposed and alternative methods under different word/rating ratios, in which each ARL is approximated based on more than 10,000 replicates for the stability of results. The degree of shifts is measured by the KL divergence between the OC and IC joint sentiment-topic distributions.

The SRJST control scheme with an average number of words $n = 1,000$ is regarded as the benchmark for our comparison, which means that only the word part in the mixed data is used. According to the results in Table 2, the proposed sequential JST-RMR scheme performs better than the benchmark, showing decreasing OC-ARLs with an increasing number of ratings besides words (i.e., $m = 100, 200, 500, 1,000$). This proves the improvement resulting from the incorporation of rating data. With an additional intention of comparing words and ratings in improving sentiment shift detection, we also present the results of SRJST scheme with different average number of words (i.e., $n = 1, 100, 1,200, 1,500, 2,000$). The one-to-one comparisons (e.g., $n = 1,100$ to $n = 1,000m = 100$, $n = 1,200$ to $n = 1,000m = 200$, $n = 1,500$ to $n = 1,000m = 500$, and $n = 2,000$ to $n = 1,000m = 1,000$) show that bringing in ratings leads to higher improvement than adding the same amount of words. The reason why ratings speak louder than words in sentiment shift detection is that ratings are more informative for the task of sentiment classification and estimation, which will be discussed in Section 6.

Moreover, we compare the results of detecting topic shifts in Table 3 by assuming that only the document-level topic distribution $P(z)$ deviates from its target value. The proposed sequential JST-RMR scheme still performs better than the benchmark scenario with only the word part (i.e., $n = 1,000$), and it shows improving performances with an increasing number of ratings. However, Table 3 presents the opposite results when conducting the one-to-one comparisons between the sequential JST-RMR scheme and the SRJST scheme with the same total amount of words and ratings. It shows that adding words instead of incorporating the same amount of ratings results in quicker detection of topic shifts. Words speak louder than ratings in the case of topic shift detection because ratings are assumed to have only sentiment assignments, and they do not provide any information regarding topics. The shift in topics could also lead to the shift in marginal distribution over sentiments and, through the detection of the latter one, ratings can still help to improve the topic shift detection. However, this detection is less sensitive compared

Table 2. ARL comparison under sentiment shifts.

No.	KL divergence	SRJST					Sequential JST-RMR				
		$n = 1,000$ (base)	$n = 1,100$	$n = 1,200$	$n = 1,500$	$n = 2,000$	$n = 100$ $m = 100$	$n = 1,000$ $m = 200$	$n = 1,000$ $m = 500$	$n = 1,000$ $m = 1,000$	
1	0	373.3 (3.682)	372.5 (3.514)	370.8 (3.529)	370.8 (3.321)	370.0 (3.476)	371.5 (3.398)	373.0 (3.468)	369.2 (3.516)	372.3 (3.530)	
2	0.000475	141.5 (1.237)	134.6 (1.187)	126.0 (1.078)	107.1 (0.905)	87.15 (0.726)	127.2 (1.137)	120.1 (1.072)	95.95 (0.829)	71.33 (0.615)	
3	0.001266	66.08 (0.524)	59.55 (0.456)	54.31 (0.410)	45.29 (0.334)	35.31 (0.242)	55.91 (0.447)	52.41 (0.409)	38.44 (0.294)	26.32 (0.200)	
4	0.004629	19.10 (0.108)	17.36 (0.094)	16.38 (0.086)	13.82 (0.070)	11.09 (0.050)	15.92 (0.089)	14.61 (0.083)	10.67 (0.058)	7.387 (0.035)	
5	0.009680	9.955 (0.043)	9.264 (0.040)	8.778 (0.036)	7.471 (0.029)	6.237 (0.022)	8.345 (0.037)	7.559 (0.033)	5.617 (0.022)	4.107 (0.014)	

with the detection of topic shifts directly based on the latent topic labels of words.

5.3. Diagnosis

Once an OC signal is triggered, diagnosis is conducted by tracing the decomposed variables (i.e., G_t and $U_{t,z}, z = 1, \dots, 10$ in this case) that measure the topic shifts and topic-specific sentiment shifts, respectively. The variable showing the largest relative increase from its IC value is supposed to note the truly OC term. Table 4 presents the probability of identifying the truly OC variables in different methods, with shifts only on the topic distribution $P(z)$ or only on one of the topic-specific sentiment distributions $P(l|z)$. It is shown that the sequential JST-RMR method keeps the desirable properties of diagnosis and produces similar diagnostic accuracy with the benchmark method (i.e., $n = 1,000$). The incorporation of ratings (i.e., $n = 1,000m = 200$) does not lead to better performance on diagnosis compared with the benchmark scenario, while an increase in number of words (i.e., $n = 1,200$) does. Unlike words, ratings do not help in diagnosis improvement because they provide only the general tendencies of customer sentiments without any topic-related explanations.

6. Informative comparison between words and ratings

Both review words and ratings provide important information in sentiment discovery, while ratings are believed to be more informative in this task. In this section, we plan to explore the comparison between words and ratings and have an insight into their properties.

6.1. Comparison of information gain

First, we can measure the information gain achieved by both words and ratings in sentiment classifications based on the results of Gibbs sampling. According to the information theory by Shannon (1948), information gain is defined as the amount of information that is gained by knowing the value of a specific attribute, which is the entropy difference between the distributions before and after the attribute is included. For example, the information gain in the task of sentiment classification (denoted by the sentiment distribution $P(l)$) achieved by the i th word w_i in the data set is defined as

Table 3. ARL comparison under topic shifts.

No.	KL divergence	SRJST					Sequential JST-RMR				
		$n = 1,000$ (base)	$n = 1,100$	$n = 1,200$	$n = 1,500$	$n = 2,000$	$n = 1,000$ $m = 100$	$n = 1,000$ $m = 200$	$n = 1,000$ $m = 500$	$n = 1,000$ $m = 1,000$	
1	0	373.3 (3.682)	372.5 (3.514)	370.8 (3.529)	370.8 (3.321)	370.0 (3.476)	371.5 (3.398)	373.0 (3.468)	369.2 (3.516)	372.3 (3.530)	
2	0.000371	202.7 (1.834)	188.7 (1.674)	187.4 (1.688)	169.1 (1.494)	142.5 (1.237)	190.9 (1.784)	185.8 (1.689)	168.0 (1.242)	140.8 (1.286)	
3	0.001442	72.35 (0.576)	66.37 (0.528)	60.20 (0.464)	48.05 (0.354)	33.46 (0.216)	66.77 (0.548)	63.85 (0.504)	54.68 (0.433)	41.84 (0.336)	
4	0.003166	28.48 (0.175)	24.99 (0.143)	22.65 (0.122)	17.93 (0.084)	13.33 (0.051)	25.66 (0.162)	25.13 (0.153)	21.86 (0.133)	17.41 (0.106)	
5	0.008429	9.710 (0.030)	8.932 (0.026)	8.457 (0.023)	7.096 (0.017)	5.840 (0.012)	9.013 (0.030)	9.072 (0.030)	8.134 (0.028)	6.807 (0.026)	

$$\begin{aligned}
 IG(P(l), w_i) &= \text{Entropy}(P(l)) - \text{Entropy}(P(l|w_i)) \\
 &= - \sum_l P(l) \log(P(l)) + \sum_l P(l|w_i) \log(P(l|w_i)) \\
 &= - \sum_l \frac{\gamma_l}{\sum_k \gamma_k} \log\left(\frac{\gamma_l}{\sum_k \gamma_k}\right) + \sum_l \frac{n_{l,w_i}}{n_{w_i}} \log\left(\frac{n_{l,w_i}}{n_{w_i}}\right).
 \end{aligned} \tag{28}$$

And the information gain achieved by the i th rating r_i in the data set is

$$\begin{aligned}
 IG(P(l), r_i) &= \text{Entropy}(P(l)) - \text{Entropy}(P(l|r_i)) \\
 &= - \sum_l P(l) \log(P(l)) + \sum_l P(l|r_i) \log(P(l|r_i)) \\
 &= - \sum_l \frac{\gamma_l}{\sum_k \gamma_k} \log\left(\frac{\gamma_l}{\sum_k \gamma_k}\right) + \sum_l \frac{m_{l,r_i}}{m_{r_i}} \log\left(\frac{m_{l,r_i}}{m_{r_i}}\right).
 \end{aligned} \tag{29}$$

The average information gain in sentiment classification resulting from a word/rating is obtained by

$$\begin{aligned}
 IG(P(l), w) &= \sum_i P(w_i) \cdot IG(P(l), w_i), \\
 IG(P(l), r) &= \sum_i P(r_i) \cdot IG(P(l), r_i),
 \end{aligned} \tag{30}$$

where $P(w_i)$ and $P(r_i)$ measure the frequency of a specific word/rating in the entire data set.

To measure the contribution of words and ratings in the Dell computer data set on a comparable scale, we compute their average information gain in the sentiment classification. According to the training results in the offline stage, the average information gain of words and ratings in the experimental data set is 0.226 and 0.378, respectively. It proves that ratings, providing higher information gain, are more informative than words in inferring the latent sentiment polarity of documents.

6.2. Comparison of sentiment estimation

Moreover, we compare the words and ratings in the Dell computer data set by discussing their performance in estimating the document-level sentiment distributions under different sample size. The sentiment distribution π_t is estimated based on n_{d_t} words and m_{d_t} ratings according to Eq. [13], which can be further decomposed into

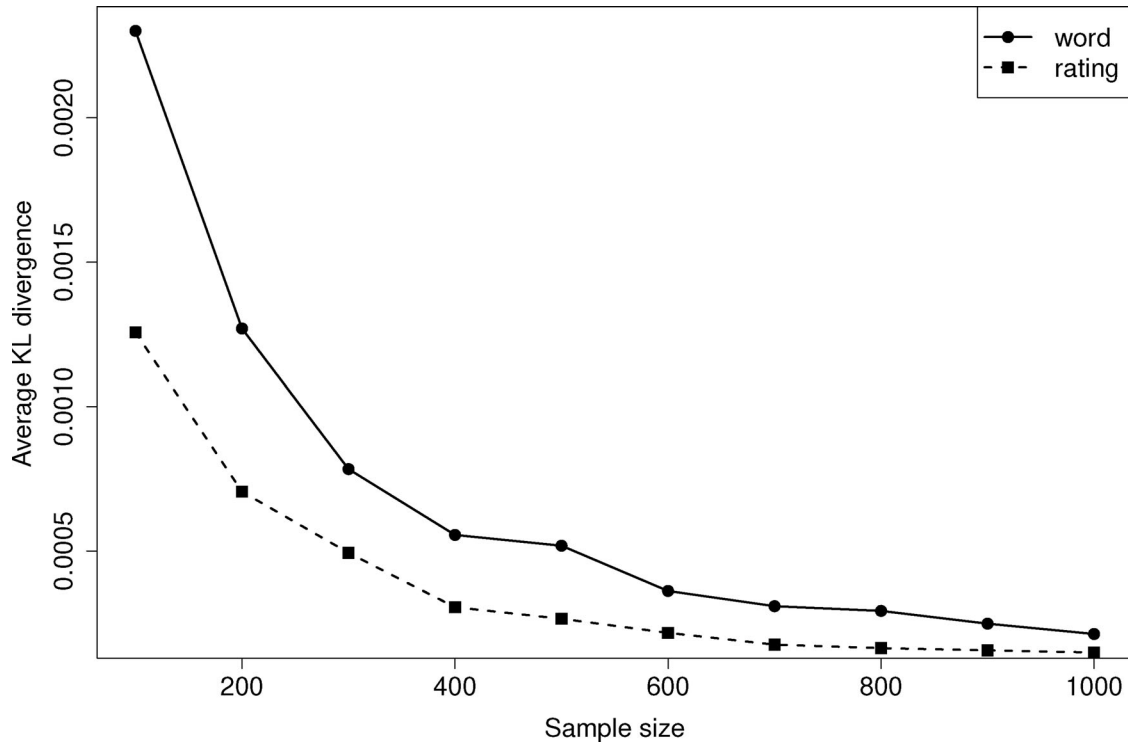
$$\hat{\pi}_{t,l} = \frac{m_{d_t}}{n_{d_t} + m_{d_t}} \cdot \hat{\pi}_{t,l}^{rating} + \frac{n_{d_t}}{n_{d_t} + m_{d_t}} \cdot \hat{\pi}_{t,l}^{word}, \tag{31}$$

$$l = 1, \dots, S,$$

where $\hat{\pi}_{t,l}^{rating}$ and $\hat{\pi}_{t,l}^{word}$ are represented as the sentiment distributions individually derived from ratings and words:

Table 4. Diagnostic accuracy of OC variables.

	No.	KL divergence	SRJST		Sequential JST-RMR	
			n = 1,000	n = 1,200	n = 1,000	m = 200
Topic shift	1	0.000371	0.153	0.157	0.149	
	2	0.001442	0.359	0.450	0.393	
	3	0.003166	0.619	0.670	0.625	
	4	0.008429	0.900	0.926	0.891	
Sentiment shift	1	0.000475	0.405	0.434	0.427	
	2	0.001266	0.620	0.709	0.616	
	3	0.004629	0.832	0.862	0.825	
	4	0.009680	0.872	0.904	0.869	

**Figure 4.** Average KL divergence between the target and the estimated sentiment distributions.

$$\hat{\pi}_{t,l}^{rating} = \frac{m_{d_t,l}}{m_{d_t}} \cdot \frac{1}{1+\rho} + \hat{\pi}_{t-1,l}^{rating} \cdot \frac{\rho}{1+\rho}, \quad l = 1, \dots, S,$$

$$\hat{\pi}_{t,l}^{word} = \frac{n_{d_t,l}}{n_{d_t}} \cdot \frac{1}{1+\rho} + \hat{\pi}_{t-1,l}^{word} \cdot \frac{\rho}{1+\rho}, \quad l = 1, \dots, S.$$

[32]

Figure 4 shows the average KL divergence between the target sentiment distribution and the estimated sentiment distributions that are individually derived from words or ratings. With the increase of sample size, the estimated sentiment distributions become more accurate and approach the target one. According to the comparison results in Figure 4, ratings show lower noise and higher effectiveness in the estimation of latent sentiment distributions compared with words of the same sample size. It explains why

the incorporation of ratings would result in a significant improvement in detecting sentiment shifts.

7. Conclusion

This article focuses on the modeling and monitoring of online customer reviews including text words and rating scores, which provide significant information of customer opinions and quality concerns for online products and services. Our method can fully incorporate the mixed-type data for monitoring, with ratings indicating the latent sentiment polarities and review texts interpreting the related topics. Specifically, the daily collected review texts and user ratings to be monitored are connected through a joint generative sentiment-topic model (sequential JST-RMR) and approximated by their latent sentiment/topic distributions. A two-chart control scheme is constructed for shift

detection in both user sentiments and related topics based on the results of sequential JST-RMR modeling. And a diagnostic procedure is developed for identifying the truly OC terms after a process change is detected.

We demonstrate the implementation of the proposed method on a real-world data set. The proposed monitoring scheme is compared with a benchmark method that considers only text words in the monitoring. Through simulation study, we have shown that the proposed sequential JST-RMR scheme outperforms the benchmark scheme of SRJST in detecting both sentiment and topic shifts when both schemes are implemented on the same amount of words. It is noted that the incorporation of user ratings in the sequential JST-RMR results in higher improvement of sentiment shift detection than adding the same amount of words in the SRJST, as ratings are proven to be more informative than words in the inference of sentiments. However, the incorporation of ratings in the sequential JST-RMR is less efficient for the topic shift detection than adding the same amount of words in the SRJST, as ratings do not directly provide information about document-level topics.

Future research can be motivated by the following points: (1) As review words in one day are aggregated for daily monitoring, the longer reviews, mostly negative ones, are naturally given higher weights, which may lead to underestimation of the overall quality. We believe the weight allocation strategy (or more generally speaking, information aggregation strategy) among review collections is a topic that is worth more in-depth study. (2) Besides the overall ratings that accompany the observed review texts, there are an increasing number of websites providing user ratings on specific aspects. Future research on the monitoring of customer responses can be extended by investigating the aspect ratings, through which the relationship between topics and sentiments can be constructed more appropriately.

About the authors

Qiao Liang is a Ph.D. student in the Department of Industrial Engineering, Tsinghua University, Beijing, China. She received her B.S. degree in Industrial Engineering from Tsinghua University in 2016. Her research interests include statistical modeling and data analytics for manufacturing and service processes, with a focus on statistical process control based on text analytics.

Kaibo Wang is a Professor in the Department of Industrial Engineering, Tsinghua University, Beijing, China. He received his B.S. and M.S. degrees in Mechatronics from Xi'an Jiaotong University, Xi'an, China, and his Ph.D. in Industrial Engineering and Engineering Management from

the Hong Kong University of Science and Technology, Hong Kong. His research focuses on statistical quality control and data-driven system modeling, monitoring, diagnosis, and control, with a special emphasis on the integration of engineering knowledge and statistical theories for solving problems from the real industry.

Funding

This study is funded by the Key Program of the National Natural Science Foundation of China under grants No. 71932006 and No. 71731008.

ORCID

Kaibo Wang  <http://orcid.org/0000-0001-9888-4323>

References

- Ashton, T., N. Evangelopoulos, and V. Prybutok. 2014. Extending monitoring methods to textual data: A research agenda. *Quality & Quantity* 48 (4):2277–94. doi: [10.1007/s11135-013-9891-8](https://doi.org/10.1007/s11135-013-9891-8).
- Ashton, T., N. Evangelopoulos, and V. Prybutok. 2015. Quantitative quality control from qualitative data: Control charts with latent semantic analysis. *Quality & Quantity* 49 (3):1081–99.
- Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.
- Dermouche, M., J. Velcin, L. Khouas, and S. Loudcher. 2014. A joint model for topic-sentiment evolution over time. In Proceedings of the 2014 IEEE International Conference on Data Mining, 773–778. IEEE.
- Diao, Q., M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang. 2014. 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, 193–202. ACM.
- Duan, W., Q. Cao, Y. Yu, and S. Levy. 2013. Mining online user-generated content: Using sentiment analysis technique to study hotel service quality. In 2013 46th Hawaii International Conference on System Sciences, 3119–3128. IEEE.
- Griffiths, T. L., and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101 (Suppl 1):5228–35. doi: [10.1073/pnas.0307752101](https://doi.org/10.1073/pnas.0307752101).
- Hofmann, T. 1999. Probabilistic latent semantic indexing. In Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, 50–57. ACM.
- Huang, J. 2005. Maximum likelihood estimation of Dirichlet distribution parameters. *CMU Technique Report*. <http://jonathan-huang.org/research/dirichlet/dirichlet.pdf>.
- Kullback, S. 1959. *Information theory and statistics*. New York, NY: John Wiley & Sons.
- Li, H., R. Lin, R. Hong, and Y. Ge. 2015. Generative models for mining latent aspects and their ratings from short reviews. In 2015 IEEE International Conference on Data Mining, 241–250. IEEE.

- Liang, Q., and K. Wang. 2019. Monitoring of user-generated reviews via a sequential reverse joint sentiment-topic model. *Quality and Reliability Engineering International* 35 (4):1180–99. doi: [10.1002/qre.2452](https://doi.org/10.1002/qre.2452).
- Lin, C., Y. He, R. Everson, and S. Ruger. 2012. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering* 24 (6):1134–45. doi: [10.1109/TKDE.2011.48](https://doi.org/10.1109/TKDE.2011.48).
- Lin, C., and Y. He. 2009. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, 375–384. ACM.
- Ling, G., M. R. Lyu, and I. King. 2014. Ratings meet reviews, a combined approach to recommend. In Proceedings of the 8th ACM Conference on Recommender Systems, 105–112. ACM.
- Lo, S. 2008. Web service quality control based on text mining using support vector machine. *Expert Systems with Applications* 34 (1):603–10. doi: [10.1016/j.eswa.2006.09.026](https://doi.org/10.1016/j.eswa.2006.09.026).
- Lu, Y., P. Tsaparas, A. Ntoulas, and L. Polanyi. 2010. Exploiting social context for review quality prediction. In Proceedings of the 19th International Conference on World Wide Web, 691–700. ACM.
- McAuley, J., and J. Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In Proceedings of the 7th ACM Conference on Recommender Systems, 165–172. ACM.
- McAuley, J., C. Targett, Q. Shi, and A. Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 43–52. ACM.
- Mei, Q., X. Ling, M. Wondra, H. Su, and C. Zhai. 2007. Topic sentiment mixture: Modeling facets and opinions in weblogs. In Proceedings of the 16th International Conference on World Wide Web, 171–180. ACM.
- Montgomery, D. C. 2012. *Introduction to statistical quality control*. New York, NY: John Wiley & Sons.
- Mukherjee, A., A. McCracken, and S. Chakraborti. 2015. Control charts for simultaneous monitoring of parameters of a shifted exponential distribution. *Journal of Quality Technology* 47 (2):176–92. doi: [10.1080/00224065.2015.11918123](https://doi.org/10.1080/00224065.2015.11918123).
- Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27 (3):379–423. doi: [10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x).
- Sperkova, L., F. Vencovsky, and T. Bruckner. 2015. How to measure quality of service using unstructured data analysis: A general method design. *Journal of Systems Integration* 6 (4):3–16.
- Titov, I., and R. McDonald. 2008a. A joint model of text and aspect ratings for sentiment summarization. In Proceedings of ACL-08: HLT, 308–316. Association for Computational Linguistics.
- Titov, I., and R. McDonald. 2008b. Modeling online reviews with multi-grain topic models. In Proceedings of the 17th International Conference on World Wide Web, 111–120. ACM.
- Wang, C., and D. M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 448–456. ACM.

- Wang, H., Y. Lu, and C. Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 618–626. ACM.
- Woodall, W. H., and D. C. Montgomery. 2014. Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology* 46 (1): 78–94. doi: [10.1080/00224065.2014.11917955](https://doi.org/10.1080/00224065.2014.11917955).
- Xu, Y., W. Lam, and T. Lin. 2014. Collaborative filtering incorporating review text and co-clusters of hidden user communities and item groups. In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, 251–260. ACM.

Appendix A. Weighting mechanism between words and ratings

Both words and ratings are used for the estimation of document-level sentiments as shown in Eq. [13]. To explore the weighting mechanism between words and ratings, Eq. [13] can be represented in a more general form by incorporating a weight parameter σ :

$$\hat{\pi}_{t,l} = \frac{n_{d_t,l} + \sigma m_{d_t,l}}{n_{d_t} + \sigma m_{d_t}} \cdot \frac{1}{1 + \rho} + \hat{\pi}_{t-1,l} \cdot \frac{\rho}{1 + \rho}, \quad l = 1, \dots, S,$$

where σ measures the relative weight of a rating to a word in the document-level sentiment prediction.

The choice of σ varies among studied data sets, and greedy algorithm is used to obtain its best setting. For example, Figure A1 shows the model performance in sentiment prediction under different values of σ , where the performance is measured by the average KL divergence between the estimated sentiment distribution $\hat{\pi}_t$ and its target value based on the simulated review documents in Section 5.1. A lower KL divergence indicates higher accuracy of sentiment prediction. We can see it from Figure A1 that the best performance is achieved under values of $\sigma \in [1, 2]$ among various scenarios of simulation, and the results are robust in this range. Specifically, we set $\sigma = 1$ for the experimental data set in this study such that words and ratings are equally treated.

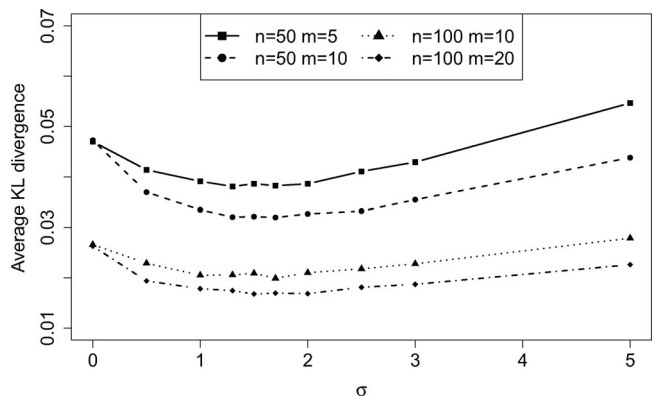


Figure A1. Average KL divergence between the target and the estimated sentiment distributions under different values of weight parameter σ .