Discussion Papers In Economics And Business

The Text-Score Allocation Model: Finding Latent Topics of Online Review Documents and Multi-Item Ratings

Sotaro Katsumata
Seungjin Kim

Discussion Paper 20-01

January 2020

Graduate School of Economics
Osaka University, Toyonaka, Osaka 560-0043, JAPAN

The Text-Score Allocation Model: Finding Latent Topics of Online Review Documents and Multi-Item Ratings

Sotaro Katsumata*
Seungjin Kim†

Abstract

This study focuses on online review data in which comments are written in natural languages and evaluations are attached as integers. This study develops a topic model incorporating both natural languages and evaluation scores, expanding latent Dirichlet allocation (LDA). The model consists of two components: LDA and a Dirichlet-binomial clustering model. The latter assumes binomial distributions for the review scores. Since the model assumes conjugate distributions, we can apply a fast and stable estimator based on collapsed Gibbs sampling to estimate the parameters. Further, the model enables us to examine the relationship between vocabulary words and review scores based on the topic allocation results.

Keywords: Topic Modeling, Customer Reviews, Forecasting
JEL Classifications Code: M33

*Graduate School of Economics, Osaka University, Address: 1-7 Machikaneyama, Toyonaka, Osaka 5650824, Japan, Email: katsumata@econ.osaka-u.ac.jp, Tel: +81-6-6850-5246
†Graduate School of Economics, Osaka University

1 Introduction

Consumer-generated media (CGM), which involves consumers transmitting and sharing information themselves, is an important source of information for marketing activities. Kannan and Li (2017) indicate that one of the characteristics of digital marketing is that the consumers ' purchasing process can be observed. This is expected to be used in consumer behavior research as a valuable source of information, which can be obtained without conducting a questionnaire survey. (Lee and Bradlow, 2011; Humphreys and Wang, 2017).

Among CGM, online reviews that evaluate products used and purchased by the reviewers serve not only as a source of information on consumers but also on firms. Furthermore, some studies indicate the relevance of online reviews to sales and corporate performance. The empirical study by Chevalier and Mayzlin (2006) show that online book reviews affect online book sales, and some studies show that online reviews affect the box office of motion pictures (Liu, 2006; Onishi and Manchanda, 2012; Dellarocas, Zhang, and Awad, 2007). These results imply that online word-of-mouth is closely related to sales. Additionally, this content is viewed by many potential consumers and influences their decision-making. In addition, Luo (2009); Tirunillai and Tellis (2012) show that CGM are related to the evaluation of a company as a whole, such as its stock price. Borah and Tellis (2016) show that CGM affect the market share of not only the focal company but also of its competitors and of companies that share its image. These studies suggest that online reviews have a substantial impact on company performance.

However, researchers should pay attention not only to the quantity but also to the qualitative aspects of online reviews. CGM is not a medium that companies can control. Following the three categories of Stephen and Galak (2012), the first, "paid media," requires paying regular advertising fees; the second, "owned media," such as company homepages owned by the company or its parent organization, can be easily used to control the company's message. However, the third category, "earned media," is managed by others or by consumers and regarded as an environmental factor. Since paid media and owned media can be controlled by companies themselves, it can be assumed that practically no negative information about the companies or their products is conveyed through these media. On the other hand, no financial expenditure is needed for earned media compared to paid media, and this point has been indicated as an advantage by several studies (Berger and Milkman, 2012; Tirunillai and Tellis, 2014). However, negative reviews may be sent out by consumers (Anderson and Simester, 2014; Chen and Lurie, 2013). Furthermore, several studies have indicated that negative reviews have a significant impact on consumer attitudes and market outcomes (Chevalier and Mayzlin, 2006; Floyd et al., 2014; Tirunillai and Tellis, 2012). Therefore, when analyzing CGM, it is important to consider the sentiment or valence of the transmitted information, not simply the transmission volume. Heretofore, there have been studies using dictionaries to assess these sentiments but which words express a negative impression and which words express a positive one may vary depending on the product category. This limits the analysis that can be performed using a general dictionary.

Therefore, the rating attached to a review is important. Ratings are often given as discrete integer scores, and considering these ratings and the words that appear in the review text, one can obtain information that will greatly contribute to marketing research and product development. However, few models exist that can be easily used to quantify a natural language such as the text of a review text and to analyze its relationship with an integer score such as a review score. Therefore, in this study, we present a classification model that considers both online review ratings and text.

2 Literature Review

2.1 Natural Language Analysis

Research on how to utilize information written in natural language, especially that posted by consumers, has generated many studies and is one of the most popular issues in marketing research (Berger et al., 2019; Humphreys and Wang, 2017; Wedel and Kannan, 2016). As already mentioned, online review analysis requires not only counting how many posts there are but also examining their valence and

sentiments, so it is necessary to not only perform a simple aggregation but also to analyze posts' contents.

Since CGM posts are written in natural language, their contents need to be quantified by an appropriate method. Text data are called Unstructured data (Balducci and Marinova, 2018). Recently, Balducci and Marinova (2018) have offered several quantification methods based on previous studies. Furthermore, Berger et al. (2019) present a procedure for analyzing text data. They show three tools for classifying text data. The first tool is for estimating moods and sentiments from the words themselves. For example, the sentiment expressed by a word can be determined by LIWC (Linguistic Inquiry and Word Count), which is used in Berger and Milkman (2012) and Hewett et al. (2016). The polarity (sentiment) dictionary makes it possible to separate negative reviews from positive ones. The second tool is for classifying words and extracting topics. By exploring and classifying topics that do not have external criteria, it is possible to interpret the object and obtain information. The third tool is for examining the relationship between words in a sentence. In the present study, a model for topic extraction is constructed and used for the analysis. In the next section, we will review the research on latent Dirichlet allocation (LDA), a representative topic model.

2.2 Expansion of the LDA

LDA and its extended model are widely used in marketing as a method for classifying sentences. LDA is a model proposed by Blei, Ng, and Jordan (2003), which assumes a high-dimensional categorical distribution for the words appearing in sentences and assumes a Dirichlet distribution that is its conjugate distribution as a prior distribution. In machine learning, there are many models that are extensions of LDA. For example, the author-topic model (Rosen-Zvi et al., 2004) considers author factors, the dynamic topic Model (Blei and Lafferty, 2006) considers time series factors, and the Pachinko allocation model (Li and MaCallum, 2006) incorporates hierarchical classification parameters. In addition, LDA can use collapsed Gibbs sampling, which is a stable and fast estimator (Griffiths and Steyvers, 2004). Collapsed Gibbs sampling is a kind of Markov-Chain Monte Carlo (MCMC) estimator.

In the marketing literature, Liu, Burns, and Hou (2017) present an analysis model for SNS text and apply it to Twitter data. In addition, Tirunillai and Tellis (2014) apply an extended LDA model to online reviews, and Büschken and Allenby (2016) present the results of an extension of the author-topic model. The author-topic model is also based on the model of Nam, Joshi, and Kannan (2017). Further, Toubia et al. (2019) apply an extended topic model to CGM. LDA itself can be applied to any data which can be assumed to follow a multinomial distribution and is also applicable to areas other than natural language. Trusov, Ma, and Jamal (2016) have developed a model that expands the correlated topic model (Blei and Lafferty, 2007) and conducts an empirical analysis of website visit behavior.

Using collapsed Gibbs sampling as an estimator, the parameters of an LDA model can be obtained from the conditional posterior distribution in the same way as in other Bayesian models. This implies that LDA can be easily expanded and combined with other models and its parameters can be estimated by the Metropolis-Hastings (M-H) method. Trusov, Ma, and Jamal (2016) and Büschken and Allenby (2016) use the M-H method as part of the parameter estimation. However, since the M-H method has a high computational load and it is difficult to stabilize the estimation result, ideally, it is desirable to create a model using only combinations of conjugate distributions. Therefore, in this study, we present a model that assumes a binomial distribution, taking advantage of the property that a review score is a continuous integer with a maximum value.

3 Model

3.1 Settings and Notation

In this section, we define the model. In the first part of this section, we define LDA, and then the score allocation model (SAM), which assumes a binomial distribution for the input data. Subsequently, the text-score allocation model (TSAM), our proposed model, is defined by combining these two models.

Assume that there are D pairs of review comments and ratings. We call each review a *document*. Document d includes both text and score data. The text is decomposed into *words*. Let the total number of words be N; w_i is the i-th $(i = 1, \cdots, N)$ observed word. There are V *vocabulary words* observed in the data. We assume that for the i-th word, any one of the V vocabulary words is actually observed; therefore, w_i is a V-dimensional vector, of which each element takes the value 1 if the corresponding vocabulary word is observed and the value 0 otherwise. Therefore, $w_{iv} = 1$ if vocabulary word $v(v = 1, \cdots, V)$ is observed and then $w_{iv'} = 0, \forall v' \neq v$. Note that each of the N words may belong to any one of the D documents. We introduce word-document indicator x_i, a D-dimensional vector, as follows: if the i-th word is observed in document d, $x_{id} = 1$; otherwise, $x_{id} = 0$.

Next, we define the notation for ratings. Assume that there are J item ratings in review d and the observed scores are denoted by y_{dj}. For item j, a score is given by reviewers on a $Q_j + 1$-point scale. Let the lowest score be 0, so $y_{dj} = \{0, 1, \cdots, Q_j\}$.

3.2 Text Allocation Model: Latent Dirichlet Allocation (LDA)

LDA was proposed by Blei, Ng, and Jordan (2003) and assumes that word w_i follows a V-dimensional categorical distribution (a single-trial multinomial distribution):

$$w_i \sim Categorical_V\left(\tilde{\phi}_i\right) \tag{1}$$

which has K latent parameters $\tilde{\phi}_i$ called *topics*. Let us introduce the K-dimensional parameter z_i. If w_i belongs to topic k, $z_{ik} = 1$ and $z_{ik'} = 0, \forall k \neq k'$. The elements of the latent parameters are defined as follows:

$$\tilde{\phi}_{iv} = \prod_{k=1}^{K} \phi_{kv}^{z_{ik}} \tag{2}$$

In addition, assume that z_{ik} follows a categorical distribution where x_i is the word-document indicator defined above.

$$z_i \sim Categorical_K\left(\tilde{\theta}_i\right), \tilde{\theta}_{ik} = \prod_{d=1}^{D} \theta_{dk}^{x_{id}} \tag{3}$$

As prior distributions, ϕ_k and θ_d follow Dirichlet distributions that are conjugates of the categorical distribution in equation (1). Let $\phi_k \sim Dirichlet(\beta)$ and $\theta_d \sim Dirichlet(\alpha)$; the full conditional posterior distribution is obtained as follows:

$$\pi(z, \Phi, \Theta|w) \propto \pi(w|z, \Phi)\pi(z|\Theta)\pi(\Phi)\pi(\Theta) \tag{4}$$

We can obtain posterior sample of the unknown parameter z by using the collapsed Gibbs sampling method proposed by Griffiths and Steyvers (2004).

$$\pi(z|w) \propto \int\int \pi(z, \Phi, \Theta|w) d\Phi d\Theta \tag{5}$$

The conditional posterior distribution of z_i is easily obtained as follows:

$$\pi(z_{ik}|z_{-ik}, w) \propto \frac{n_{kv^*,-i} + \beta}{n_{k,-i} + V\beta}(n_{d^*k} + \alpha) \tag{6}$$

where v^* is observed vocabulary at i-th word and, therefore, $x_{iv^*} = 1$. d^* is the document the i-th word belongs to; therefore, $x_{id^*} = 1$. $n_{kv^*} = \sum_{i'=1, i' \neq i}^{N} x_{i'v^*} z_{i'k}$, $n_{k,-i} = \sum_{v=1}^{N} n_{kv}$, and $n_{d^*k} = \sum_{i'=1, i' \neq i}^{N} x_{i'd^*} z_{dk}$. Griffiths and Steyvers (2004) have shown that collapsed Gibbs sampling is a fast, stable estimator and steeply converges to the stationary distribution.

3.3 Score Allocation Model (SAM)

LDA only focuses on the classification of natural language. On the other hand, the SAM defined in this section is a model for the classification of multi-dimensional scores observed in a review. First, let us assume that the rating of item j in review d follows a Binominal distribution:

$$y_{dj} \sim Binomial(Q_j, \tilde{\psi}_{dj}) \tag{7}$$

where we assume the same K latent topics for $\tilde{\psi}_{dj}$ as in LDA. Let the topic indicator parameter be c_d, which is a K-dimensional vector. If document d belongs to topic k, $c_{dk} = 1$, and $c_{dk'} = 0$ for $k' \neq k$. Then,

$$\tilde{\psi}_{dj} = \prod_{k=1}^{K} \psi_{kj}^{c_{dk}} \tag{8}$$

Similar to the LDA, we assume the following categorical distribution for c_d:

$$c_d \sim Categorical_K(\theta_d) \tag{9}$$

For ψ_{kj}, we assume a Beta distribution, a conjugate of the binomial distribution. For c_d, we assume a Dirichlet distribution. Let $\psi_{kj} \sim Beta(\gamma)$ and $\theta_d \sim Dirichlet(\alpha)$; the full conditional posterior distribution is obtained as follows:

$$\pi(c, \Psi, \Theta|y) \propto \pi(y|c, \Psi)\pi(c|\Theta)\pi(\Psi)\pi(\Theta) \tag{10}$$

Since this model consists of the conjugate distributions, the conditional posterior samples of c_d are generated by collapsed Gibbs sampling.

$$\pi(c|y) \propto \int\int \pi(c, \Psi, \Theta|y) d\Psi d\Theta \tag{11}$$

The conditional posterior samples for c_d are generated from the following categorical distribution:

$$\pi(c_{dk}|c_{-dk}, y) \propto \prod_{j=1}^{J} \frac{\prod_{q=0}^{y_{dj}}(n_{kj1,-d} + q + \gamma) \prod_{q=0}^{Q_j - y_{dj}}(n_{kj0,-d} + q + \gamma)}{\prod_{q=0}^{Q_j}(n_{kj,-d} + q + 2\gamma)} \tag{12}$$

where $n_{kj1,-d} = \sum_{d'=1,d'\neq d}^{D} y_{d'j} c_{d'k}$, $n_{kj0,-d} = \sum_{d'=1,d'\neq d}^{D} (Q_j - y_{d'j}) c_{d'k}$, and $n_{kj,-d} = \sum_{d'=1,d'\neq d}^{D} Q_j c_{d'k}$.

If c_d is given, parameter ψ_{kj} can be easily obtained:

$$\psi_{kj} = \frac{n_{kj1} + \gamma}{n_{kj} + 2\gamma} \tag{13}$$

where $n_{kj1} = \sum_{d=1}^{D} y_{dj} c_{dk}$ and $n_{kj} = \sum_{d=1}^{D} Q_j c_{dk}$. ψ_{kj} is a useful parameter for examining the score of item j when document d belongs to topic k. Since the expected value of the score is $Q_j \psi_{kj}$, a larger ψ_{kj} implies a better rating. The model is regarded as a clustering technique and ψ_k is regarded as the mean vector of cluster k. In the appendix, we compare the results of the SAM with those of k-means using sample data. In addition, if $K < J$, a $K \times J$ matrix $\Psi = \{\psi_{kj}\}$ can be interpreted as a dimension reduction matrix from the K-th to the J-th dimension.

3.4 Text-Score Allocation Model (TSAM)

By combining LDA and the SAM, we define the TSAM, which assumes a common topic. Mimno et al. (2009) have proposed the polylingual topic model (also known as the joint topic model), which merges two text data sets sharing a common topic. However, this study also involves two distinct data sets: reviews and scores.

The model has two input data sets. First, we assume that the score of review d follows the binomial distribution. At this time, the parameter depends on a latent topic, and the parameter c_{dk} indicates the affiliations for topic k.

$$y_{dj} \sim Binomial(Q_j, \tilde{\psi}_{dj}), \tilde{\psi}_{dj} = \prod_{k=1}^{K} \psi_{kj}^{c_{dk}} \tag{14}$$

Second, as in LDA, the observed word w_i follows the categorical distribution below.

$$w_i \sim Categorical_V \left(\tilde{\phi}_i\right), \tilde{\phi}_{iv} = \prod_{k=1}^{K} \phi_{kv}^{z_{ik}} \tag{15}$$

The model has two latent parameters: z_i and c_d. To connect these two latent parameters, we assume there exists a common prior parameter θ_d. If word w_i is affiliated with document d, the latent parameter z_i follows a categorical distribution with parameter θ_d. Further, c_d follows a categorical distribution with parameter θ_d:

$$z_i \sim Categorical_K \left(\tilde{\theta}_i\right), \tilde{\theta}_{ik} = \prod_{d=1}^{D} \theta_{dk}^{x_{id}} \tag{16}$$

$$c_d \sim Categorical_K \left(\theta_d\right) \tag{17}$$

For prior parameters, we assume a Dirichlet distribution for ϕ_i and θ_d; for ψ_{kj}, we assume a Beta distribution, which is the conjugate of the binomial distribution. Therefore, $\Phi_i \sim Dirichlet(\beta)$, $\theta_d \sim Dirichlet(\alpha)$, and $\psi_{kj} \sim Beta(\gamma)$. The full conditional posterior distribution is as follows:

$$\pi(z, c, \Phi, \Psi, \Theta | w, y) \propto \pi(w|z, \Phi)\pi(z|\Theta)\pi(\Phi)\pi(y|c, \Psi)\pi(c|\Theta)\pi(\Psi)\pi(\Theta) \tag{18}$$

As in LDA and the SAM defined above, this model also has an analytical solution for implementing collapsed Gibbs sampling.

$$\pi(z, c|w, y) \propto \int\int\int \pi(z, c, \Phi, \Psi, \Theta|w, y) d\Phi d\Psi d\Theta \tag{19}$$

Therefore, we need to estimate only two parameters, z and c, whose posterior samples are obtained as follows:

$$\pi(z_{ik}|z_{-ik}, c, w) \propto \frac{n_{kv^*,-i} + \beta}{n_{k,-i} + V\beta}(n_{d^*k} + c_{d^*k} + \alpha) \tag{20}$$

where $n_{kv^*,-i}, n_{k,-i}$, and n_{d^*k} are the same those defined above for LDA. A posterior sample for c_{dk} is obtained from the following distribution:

$$\pi(c_{dk}|c_{-dk}, z, y) \propto \left[\prod_{j=1}^{J} \frac{\prod_{q=0}^{y_{dj}}(n_{kj1,-d} + q + \gamma) \prod_{q=0}^{Q_j - y_{dj}}(n_{kj0,-d} + q + \gamma)}{\prod_{q=0}^{Q_j}(n_{kj,-d} + q + 2\gamma)}\right](n_{dk} + \alpha) \tag{21}$$

where $n_{kj1,-d}, n_{kj0,-d}$, and $n_{kj,-d}$ are the same as those defined for the SAM above.

If z and c are given, the prior parameters ϕ_{kv}, ψ_{kj}, and θ_{dk} can be obtained as follows:

$$\phi_{kv} = \frac{n_{kv} + \beta}{n_k + V\beta} \tag{22}$$

$$\psi_{kj} = \frac{n_{kj1} + \gamma}{n_{kj} + 2\gamma} \tag{23}$$

$$\theta_{dk} = \frac{n_{dk} + c_{dk} + \alpha}{n_d + 1 + K\alpha} \tag{24}$$

Figure 1 shows a graphical model of the three models above. TSAM is obtained by combining LDA and the SAM. The parameter θ is shared, and the potential topic allocation parameters c and z are generated by using this parameter.

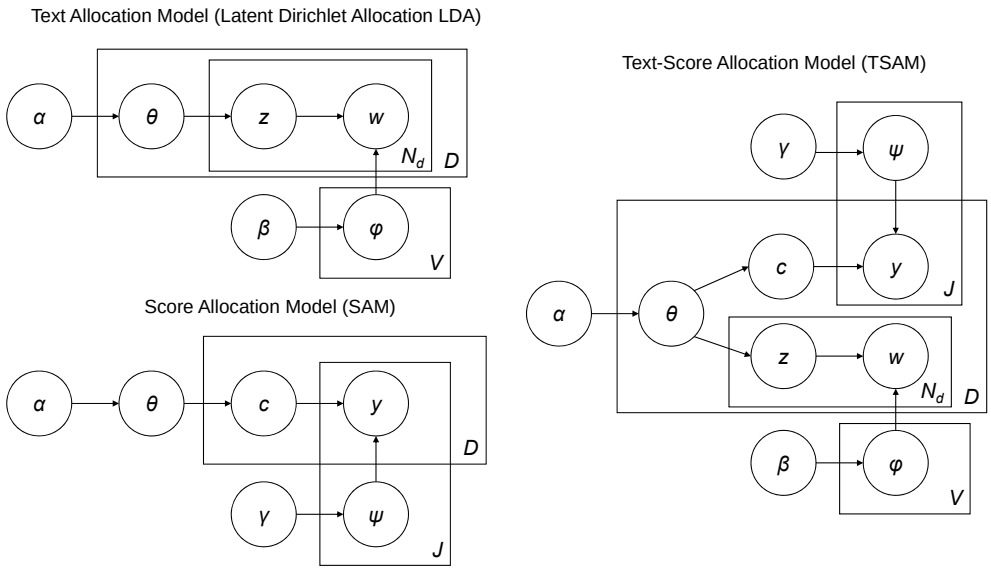


Figure 1: Graphical Expression of Three Models

3.4.1 Forecasting

Since, in the estimation, the samples are obtained from the conditional distribution by Gibbs sampling, predicted values can be obtained by using an algorithm for missing value augmentation. For document d, if the natural language w of a review is observed but the evaluation score y_{dj} is missing, the model can generate a predicted value for it from the following distribution:

$$\pi(y_{dj} = q | y_{-dj}, c, z) \quad = \quad \left(\begin{array}{c} Q_j \\ q \end{array} \right) \left(\frac{n_{k^* j 1,-d} + \gamma}{n_{k^* j,-d} + 2\gamma} \right)^q \left(\frac{n_{k^* j 0,-d} + \gamma}{n_{k^* j,-d} + 2\gamma} \right)^{Q_j - q} \tag{25}$$

where k^* is the topic document d belongs to; therefore, $c_{dk^*} = 1$.

One of the simplest ways to generate the random samples of y_{dj} following the distribution above, is repeated Bernoulli trials. Generate $r^{(h)}$ from $r^{(h)} \sim Bernoulli\left((n_{k^* j 1,-d} + \gamma)/(n_{k^* j,-d} + 2\gamma)\right)$ Q_j times and summarize the results. Then, $y_{dj} = \sum_{h=1}^{Q_j} r^{(h)}$ is a random sample from the binomial distribution above.

4 Data

4.1 Data Collection

The data are reviews of the Japanese price comparison site Kakaku.com (https://kakaku.com). Kakaku.com is a website operated by Kakaku.com, Inc., where visitors can view comparisons of actual sales prices of electrical appliances and other product/service categories, such as alcoholic beverages and types of insurance, and consumer reviews of the products/services. Reviewed and evaluated products/services are available mainly in Japan and reviews are mostly written in Japanese. However, according to SimilarWeb (https://www.similarweb.com), Kakaku.com is the most visited website in the "E-commerce and shopping: Price Comparison" category in the world.

In this study, we select a smartphone as the objective product category. In particular, we focus on the iPhone series, which are manufactured and sold by Apple, Inc. The analysis period is nine and a half years: from August 2008, when iPhone 3G was released in Japan, to February 2018. During this time, 13 models have been released, and reviews of these models are in the objective data set. For data collection, we use Python to scrape websites. The data include the posting date and time of the review, the name of the product (model) to be reviewed, the review title, the review text, the reviewer, and the scores. The scores comprise six items—the five part-wise evaluations and the overall satisfaction—and all are scored on a 5-point scale. Among these, *satisfaction* is a comprehensive evaluation, and the other five aspects are partial evaluations. Specifically, they are *design*, *mobility*, *response*, *screen*, and *battery*. For example, the score for design is the specific evaluation of a model's appearance. This includes evaluation of colors. The score for mobility evaluates a model's ease of portability as a mobile device. The score for screen evaluates the size and resolution of the screen. Finally, with regard to battery, the evaluated element is battery durability. In the following, scores for the five aspects other than satisfaction will be collectively called *subscores*.

Of the collected reviews, 5,375 reviews were finally included in the analysis, excluding those that did not have a full text or that lacked scores for the 6 items. Fig. 2 shows the total number of reviews used in the analysis, by month, and provides the release month of the models. As can be seen, the number of reviews has increased greatly in the months when new models were released. It can also be seen that reviews have been observed throughout the entire period.

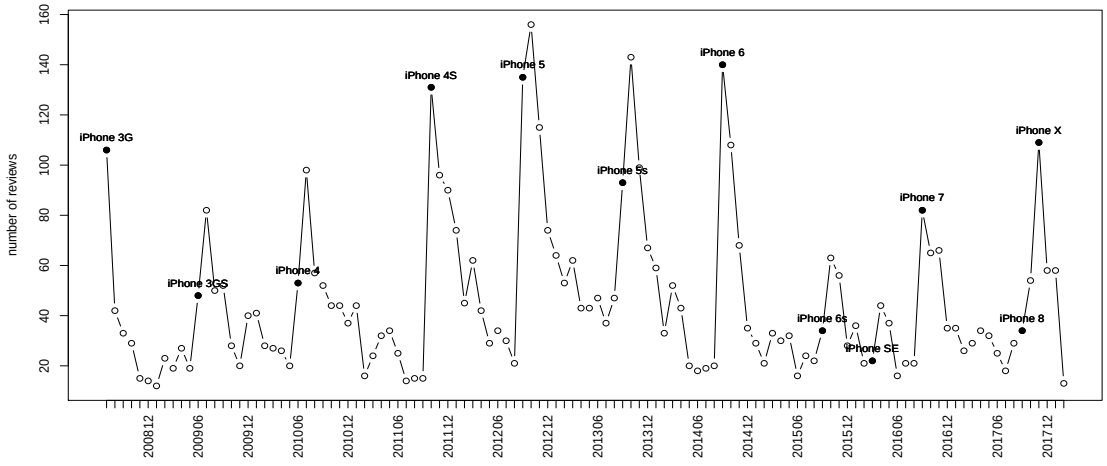


Figure 2: Number of Reviews and Release Date

4.2 Data Cleaning

Since the review text to be analyzed is written in Japanese, a Japanese morphological analyzer needs to be used. In this study, words are decomposed using MeCab, a general-purpose morphological analysis engine developed by Kudo, Yamamoto, and Matsumoto (2004). In addition, ipadic-Neologd, a new neologism dictionary developed and managed by Sato, Hashimoto, and Okumura (2017), is used as a dictionary to decompose words. Nouns, verbs, and adjectives are extracted and used in the analysis.

In review text, a negation tag may be used for verbs and adjectives. In Japanese sentences, an auxiliary verb for negation is added after the verb or adjective. Therefore, we check whether such an auxiliary verb appears after a verb or adjective is observed, and if it does, a negation tag is attached to the target verb or adjective. As a result, in the decomposition, a verb or an adjective with a meaning opposite to that of its regular form is counted as a separate word by adding (-1) after the word. For

example, if the word *Ugoka-nai* (does not move) is observed, with the negation expression *nai* (not) attached to the verb *Ugoku*(move), the word *move (-1)* is recorded in the data set.

Next, we need to clean the decomposed words. Cleaning consists of the following four steps: (1) exclude words observed only one time, (2) exclude words with length 1 (single-character words), (3) exclude words with only numerals, (4) exclude the product name (iPhone) and words related to the model name (e.g., 3G), and (5) exclude general indicating pronouns (e.g., this, that). As a result of the cleaning, 567,249 words composed of $12,815$ vocabulary words were obtained. There are 5,375 reviews, the shortest of which contains 5 words and the longest of which contains 506 words.

4.3 Distribution of Scores

The number of documents $D = 5375$, the number of evaluation items $J = 6$, the maximum score for each evaluation item is $Q-1 = Q_j-1 = 4, \forall j = 1, \cdots, J$, the number of vocabulary words $V = 12,815$, and the corpus length $N = 567,249$. The evaluation data y consist of a $D \times J$ matrix that takes $0, \cdots, Q-1$ values, and the natural language w is an $N \times V$ matrix that takes binomial values $\{0, 1\}$ for each element. Further, the document index x is a $N \times D$ matrix that takes the binomial values $\{0, 1\}$.

The observed score distribution is shown in 1. This table shows that all the items exhibit high average scores. For design and screen, over 70% of reviewers give the highest rating of 4, whereas very few reviewers give a rating of 0 or 1. For satisfaction, which is the comprehensive evaluation 60% of reviewers also give a rating of 4. This implies that the consumers reviewing the products tend to give a rating of 4 if they are generally satisfied, and a rating of 3 or less if the product has any clear defects.

Table 1: Summary of Scores

Score	Subscores										Satisfaction	
	Design		Mobility		Response		Screen		Battery			
0	59	1.1%	101	1.9%	62	1.2%	44	0.8%	216	4.0%	159	3.0%
1	78	1.5%	232	4.3%	128	2.4%	52	1.0%	494	9.2%	185	3.4%
2	344	6.4%	687	12.8%	405	7.5%	329	6.1%	1212	22.5%	374	7.0%
3	1080	20.1%	1485	27.6%	1091	20.3%	988	18.4%	1825	34.0%	1458	27.1%
4	3814	71.0%	2870	53.4%	3689	68.6%	3962	73.7%	1628	30.3%	3199	59.5%
Total	5375	100.0%	5375	100.0%	5375	100.0%	5375	100.0%	5375	100.0%	5375	100.0%

4.4 Forecasting and Model Comparison

To assess the forecasting performance of the model, some observed scores are masked. Hereafter, the reviews in which none of the information is masked are called *complete reviews*, and the prediction samples are not based on these reviews. We prepare two types of samples for prediction. The first type involves a subset of the reviews where text data and score data other than satisfaction are used; only the satisfaction score is not used. In this subset, satisfaction has actually been evaluated on a scale of 0 to 4, but these data are masked. On the other hand, the review text and the subscores of the five aspects are not masked. This set of reviews is called *text-subscore reviews*. In text-subscore reviews, the satisfaction scores are predicted by the text and subscores. Next, we prepare *text-only reviews*, from which lack satisfaction and subscores are omitted. We divide newer posts, those submitted after 2015, into three groups to prepare text-subscore reviews and text-only reviews. One-third of each is not masked and classified as complete reviews. We masked the satisfaction ratings of another one-third of the posts to prepare text-subscore reviews and masked the satisfactions and other five subscore ratings of the remaining one-third of the posts for text-only reviews. The size of each set of reviews is shown in Table 2, where NA implies the masked scores. In model estimation, all the observations in each set are included in the analysis and masked scores are augmented by MCMC simulations.

The model presented in this study needs the number of potential topics to be provided, as does LDA. Therefore, to determine the most appropriate number of topics, some models that assume different numbers of topics are used, to compare their fit and predictive performances. In this study, a

Table 2: Summary of Three Datasets

	Complete reviews	Text-subscore reviews	Text only reviews
Text	available	available	available
Subscore (5 aspects)	available	available	NA
Satisfaction	available	NA	NA
# of reviews			
2008-2014	3849	0	0
2015	138	138	138
2016	171	171	171
2017	174	174	174
2018	25	26	26
Total	4357	509	509
% of reviews	81.1%	9.5%	9.5%

total of 15 models where the number of topics is $K = 2, 5, 10, 20, \cdots, 100, 110, 120, 150, 200, 300$ are estimated. Since the model is estimated by the MCMC method, it is necessary to determine the number of iterations. According to Griffiths and Steyvers (2004), an LDA estimated by collapsed Gibbs sampling converges to the posterior distribution within 1,000 iterations; therefore, the first 1,000 iterations are withdrawn as burn-in and we collect the samples of subsequent 2,000 iterations to determine the posterior density. The parameters included in the model are z and c. In addition to this, the missing score data y also need to be estimated. The satisfaction scores missing from the text-subscore reviews and the subscores and satisfaction scores missing from the text-only dataset are obtained from the posterior distributions. In the simulation, samples for all parameters are generated sequentially based on the other samples.

To compare the models' forecasting accuracy, hit rate, AUC (Area Under Curve)—also known as ROC (Receiver Operating Characteristics)—score, and prediction MSE (Mean Squared Error) are used. For the hit rate, the expected value of the predicted score is rounded to an integer to obtain the predicted value. Since scores are five ordered integers, we consider a predicted value up to ± 1 of the actual value to be a correct prediction. Therefore, our hit rate is not a strict evaluation but a hit rate with nearest neighbors, which is also applied in Ansari, Essegaier, and Kohli (2000). Although the AUC score is only applied in binomial discrimination problems, the binary value of $\{satisfaction = 4, satisfaction \leq 3\}$ is used as the prediction target. For the prediction MSE, the expected values of the scores are compared with the observed integer.

To compare the prediction accuracy of our model with that of other models, we estimate some regression models. Specifically, an ordered logit model of 5-point dependent variables is estimated to compare the hit rate (with nearest neighbors) and MSE, and the binomial logit model ($\{1, 0\} = \{satisfaction = 4, satisfaction \leq 3\}$) is estimated to compare AUC scores. In the predictions for text-score and text-only reviews, a set of complete reviews is used to obtain the parameters. To forecast text-subscore review scores, the 5 subscores and the number of observations for the top 50 words (most frequently observed in the set of text-subscore reviews) are used as explanatory variables. Therefore, the number of explanatory variables in the model is $5 + 50 = 55$. To forecast text-only review scores, the number of observations for the top 50 words is also used as an explanatory variable. However, the model does not include the subscores; therefore, the number of explanatory variables is 50.

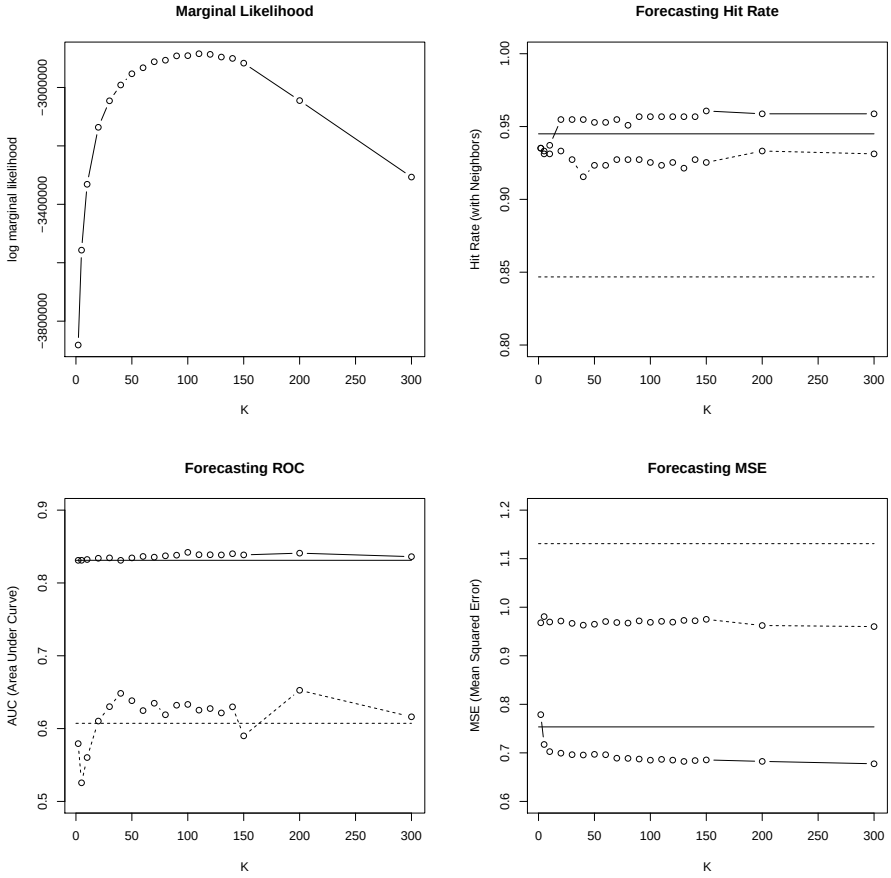
5 Results

5.1 Model Comparison

As defined above, several indices are calculated to assess model fit. First, the log of the marginal likelihood is calculated as an index for comparing the in-sample fit. The marginal likelihood is obtained from the harmonic mean, which is also used in Griffiths and Steyvers (2004) to determine the optimal

number of topics.

The upper left part of Figure 3 shows the relationship between the number of topics and the model's marginal likelihood. The marginal likelihood is an index whose larger values are assumed to be better. From here, it can be seen that the fit increases rapidly up to approximately $K = 80$, it is largest when $K = 110$, and it gradually decreases for larger values of K. Therefore, based on the marginal likelihood, it can be said that $K = 110$ corresponds to the most appropriate model. However, the marginal likelihood shows the in-sample fit of the model. Therefore, we also examine the accuracy using other indicators.



Upper left: log marginal likelihood. Upper right: Hit rate (with the nearest neighbors). Bottom left: AUC. Bottom right: Forecasting MSE. The solid line is the prediction results for the text-subscore reviews, and the dotted line those for the text-only reviews. The prediction target is the satisfaction scores. Hit rate and MSE indicate the prediction results in terms of the 5 ordered integers, and AUC indicates the prediction results of a binomial variables for which $\{1, 0\} = \{satisfaction = 4, satisfaction \leq 3\}$. The straight lines are the prediction results of the regression models (the ordered logit model and the binomial logit model).

Figure 3: Results of Forecasting

The upper right, lower left, and lower right of Figure 3 are the prediction results for the text-subscore reviews and text-only reviews. The prediction target is satisfaction score. The upper right part of the figure is the hit rate including the nearest neighbors, the lower left is the AUC score calculated from the ROC curve, and the lower right is the forecast MSE. The hit rate is determined for the five values as the ratio of the diagonal component ± 1 elements of the predicted/observed cross-tabulation. AUC predicts the binomial result of whether a score is 4 or not. The larger the hit rate and ROC are, the better. On the other hand, the smaller the MSE is, the better. In the figure, the solid line is the prediction results

for the text-subscore reviews, and the dotted line is the prediction results for the text-only reviews. The straight lines are the prediction results of the regression models the ordered logit model and binomial logit model), where the solid lines are the predictions for the text-subscore reviews, and the dotted line is the prediction for the text-subscore reviews.

Theoretically, AUC is expected to be 0.5 when predicted completely at random. Therefore, the prediction ability of all models is better than random prediction. In addition, compared with the results of the regression models, the results show that the proposed TSAMs with different numbers of topics have a generally better predictive ability. The regression models performed better than some TSAMs compared with small K. However, when K increases to some extent, the TSAM's accuracy is higher than that of the regression models, with a few exceptions (e.g., AUC for text-only reviews for the $K = 150$ model). In addition, the differences between the TSAM's results and the regression models' results for the text-only reviews are generally larger than for the text-subscore reviews. This is because the regression models include only the top 50 words as explanatory variables to forecast text-only review scores, whereas the TSAM classifies topics based on more than 10,000 words. This implies that the predictive accuracy increases by utilizing this massive number of words.

For each part of the figure and for all models, the prediction results for the text-subscore reviews are much better than the prediction results for the text-only reviews. This implies that the relationships between the five subscores and satisfaction are strong. However, even if the number of topics increases, the prediction ability does not greatly improve.

From the model comparisons, the model with the highest marginal likelihood was for $K = 110$, but it can be seen that there is almost no change in the marginal likelihood from $K = 100$ to $K = 130$. Regarding the hit rate, AUC, and MSE, TSAMs with small values of K do not exhibit good prediction accuracy. However, the figure shows that the predictive accuracy increases until K is approximately 20 to 30. After that, the predictive accuracy does not dramatically improve. Therefore, in the following, we examine the result of the model with $K = 110$. The sampling path of this model is shown in Figure 7 in the Appendix. It can be seen that this path converges to the posterior distribution in fewer than 1,000 iterations.

5.2 Topic Interpretation

In this section, we examine the results of the topic analysis obtained from the model. In an ordinary topic model such as LDA, researchers have to interpret the characteristics of topics based on the words that have high generating probabilities for each topic. In addition, researchers must also examine whether a topic is a *good* topic or a *bad* one. However, our TSAM incorporates review scores and we can use the expected scores as a guideline. Parameter ψ_{kj}, which affects the expected value of a review score, is estimated for each topic. By examining this ψ_{kj}, we can intuitively interpret whether the topic is a compilation of good or bad reviews. Based on these parameters, it is possible to assess whether a word with a high probability of belonging to each topic is related to a good review or a bad one. Furthermore, since the evaluation is divided into five subscores in addition to the satisfaction score, which is an overall evaluation, it is possible to examine how highly each of the five aspects was evaluated.

Table 3 shows the topics for which the value of ψ_{kj} in any of the six scores is in the top three among the 110 topics. Conversely, Table 4 shows the topics for which the value of ψ_{kj} in any of the six items is in the bottom three places among all the 110 topics. Topics for which any of the six scores are in the top three are called *good topics*, and topics for which any of the six scores are in the bottom three are called *bad topics*. The numbers shown in the table are the posterior expectations of ψ_{kj}, and the numbers in parentheses indicate the ranking across all topics. In addition, the top 10 words obtained based on ϕ_{kv} are shown in the right column. Note that, since the model analyzed Japanese words and the words in the tables are translated, some words consist of more than two English words.

First, from Table 3, we can see that positive words appear frequently in the top words. For example, in Topic No. 73, there are many words that can be interpreted as corresponding to a good rating, such as *highest* and *excellent*. For topics No. 73, No. 65, and No. 2, the most relevant words are *highest*, *love*, and *great*, respectively, and the topic with the highest rating is a positive word. However, it is difficult

to understand exactly what aspects are related to good evaluations by using only words with a positive meaning. We cannot examine which aspect is excellent or great without subscores. Our model allows us to easily examine the relationship between scores and frequent words. For example, in topic No. 21, words such as *thin*, *light*, and *fast* are frequently observed. This implies that a thin and lightweight shape and high-speed processing lead to a positive evaluation of mobility and response, which rank third among the 110 topics. These results indicate that some words are closely related to the subscores given by the reviewers and analysts can interpret these relationships to find the characteristics of each topic.

On the other hand, there are topics where high and low ratings are greatly divided among items. No. 53 in Table 3 exhibits very high ratings for design and screen, whereas its ratings for mobility and battery are in the bottom 10% of all topics. This suggests that a certain number of consumers were satisfied with the design and appearance of the product but experienced issues with battery life. Regarding No. 71, battery has a high rating, whereas design and mobility have low ratings. We find that words such as *large* and *huge* are frequently observed in the reviews of topic No. 71. These results suggest that the words *big* and *huge* were observed when reviewers were evaluating the difficulty of holding the product and are related to the lower scores in design and mobility.

No.	Design	Mobility	Response	Screen	Battery	Satisfaction	Frequent words
73	0.981 (2)	0.959 (2)	0.98 (2)	0.984 (2)	0.859 (1)	0.971 (1)	highest, excellent, response, design, have, mobile, relinquish (-1), luxury, incomparable, all
65	0.976 (3)	0.926 (8)	0.973 (6)	0.979 (5)	0.825 (6)	0.966 (2)	satisfaction, change-model, very, love, overall, concern, comprehensive, demerit-point, change-device, long-lasting
2	0.968 (11)	0.928 (7)	0.973 (8)	0.98 (4)	0.849 (2)	0.964 (3)	great, have, single-word, all, excitement, design, perfect, only-this, mistake (-1), sufficient
21	0.95 (34)	0.948 (3)	0.978 (3)	0.977 (7)	0.793 (14)	0.956 (4)	thin, light, fast, rather, map, long, speed, surprise, vertical, lightness
50	0.974 (4)	0.975 (1)	0.985 (1)	0.982 (3)	0.796 (11)	0.954 (6)	LTE, area, communication, tethering, correspondence, my-home, internet-speed, network, speed, Mbps
53	0.985 (1)	0.729 (100)	0.813 (95)	0.988 (1)	0.429 (105)	0.947 (11)	mobile, internet, PC, able, email, see, pleasant, site, review, convenient
71	0.763 (106)	0.552 (106)	0.942 (30)	0.959 (27)	0.846 (3)	0.85 (77)	large, screen, see, conspicuous, huge, roundness, have, design, thin, take-on

Note a) The numbers shown in the table are the posterior expectations of ψ_{kj}, and the numbers in parentheses indicate the ranking.
Note b) Since Japanese words were analyzed in the model and the words in the tables are their translation, some of them consist of more than one English word.
Note c) The sign (-1) added after some words indicates negation.

Table 3: Good Topics

Next, Table 4 shows topics with low ratings. Contrary to topics with high ratings, the words *bad*, *worst* and *disappointment* are found here. However, looking at the evaluations by the five aspects, there are topics where high evaluations and low evaluations are mixed. The most remarkable topic is No. 49, which shows that the overall satisfaction rating is very low, at the 109th place, even though mobility and screen ratings are ranked high, within the top 10%. Looking at the evaluations by the five aspects, response and battery ratings are also low, but the reason why satisfaction is lower than that can be understood by looking at the frequently observed words. Looking at the right column of Table 4, the most frequently occurring words are related to issues with telecommunication qualities; examples of these words include: *signal*, *out-of-service*, *antenna*, and *weak*, and *SB* (abbreviation for the telecommunications company SoftBank). In addition, *carrier* also indicates a communications company. These words imply that these low evaluations are not for the product model itself but for the communications environment. Issues with telecommunication qualities are strongly related to the low degree of satisfaction. Although this review site itself is not for a telecommunications company but an

online community for evaluating mobile terminal models, consumers often confuse the quality of service of the models provided by manufacturers with the quality of the communications environment provided by telecommunications companies. Therefore, since an evaluation as low as the one mentioned above depends on the communications environment, it is not necessary for the manufacturer to improve its own product in response. It can be said that, instead, it is the telecommunications companies that have to improve the environment for comfortable communication services.

In the following sections, we will further examine each topic. The next section focuses on iPhone generations and examines topics that frequently appear in each generation.

No.	Design	Mobility	Response	Screen	Battery	Satisfaction	Frequent words
85	0.402 (110)	0.412 (109)	0.624 (105)	0.55 (109)	0.332 (108)	0.252 (110)	bad, extremely, able (-1), disappointment, usable (-1), not-good, worst, can (-1), less-than, slow
49	0.95 (35)	0.921 (9)	0.871 (85)	0.976 (8)	0.562 (97)	0.276 (109)	signal, telephone-call, out-of-service, place, situation, SB, quality, carrier, antenna, weak
47	0.518 (109)	0.396 (110)	0.491 (109)	0.529 (110)	0.333 (107)	0.344 (108)	screen, operation, conversion, bright, telephone-call, sound-quality, hard-to-do, letter, menu, display
39	0.764 (105)	0.682 (103)	0.617 (107)	0.614 (108)	0.492 (104)	0.406 (107)	exchange, correspondence, get, model-body, failure, support, status, repaire, -able, mal-function
93	0.896 (83)	0.655 (104)	0.403 (110)	0.832 (103)	0.267 (109)	0.504 (106)	email, receive, incoming, -able, able (-1), con-firmation, setting, correspondence, MMS, no-tification
41	0.836 (99)	0.424 (108)	0.612 (108)	0.859 (99)	0.254 (110)	0.616 (105)	phone, mobile-phone, function, mobile, re-view, touch-panel, usual, dial, pleasant, nec-essary
9	0.609 (108)	0.505 (107)	0.881 (80)	0.892 (85)	0.728 (50)	0.705 (100)	button, home, push, power-supply, volume, screen, location, arrangement, hard-to-do, sleep

Note a) The numbers shown in the table are the posterior expectations of ψ_{kj}, and the numbers in parentheses indicate the ranking.
Note b) Since Japanese words were analyzed in the model and the words in the tables are their translation, some of them consist of more than one English word.
Note c) The sign (-1) added after some words indicates negation.

Table 4: Bad Topics

5.3 Generations and Topics

We focus on multiple generations of the iPhone series models released between 2008 and 2018. We can examine the relationship between model reviews, topics, words, and ratings. Assume that there are D reviews, and let the set of reviews written for generation g be denoted by D_g and the number of reviews included in the set be $n(D_g)$. The topic share (appearance probability of topic k) for each generation g is obtained from the average value of the appearance probability the topic in each document $\bar{\theta}_{dk}$.

$$\bar{\theta}_{gk} = \frac{1}{n(D_g)} \sum_{d \in D_g} \theta_{dk} \tag{26}$$

Table 5 summarizes the top three topics, having the largest values of $\bar{\theta}_g$ for all generations. In addition, Figure 4 shows the trend in the topic share $\bar{\theta}_{gk}$ for each generation and the evaluation for each of the five aspects for each topic. Table 5 shows the topic share, satisfaction value, ranking, and frequent words. For other subscore values, the relationships between topic shares and generations are shown in Figure 4.

Based on the frequently observed topics for each generation shown in Table , we find that many generations have at least one good topic and at least one bad topic. Further, the top topics gradually change from iPhone 3G, launched in the Japanese market in 2008, to iPhone X, launched in 2017. This

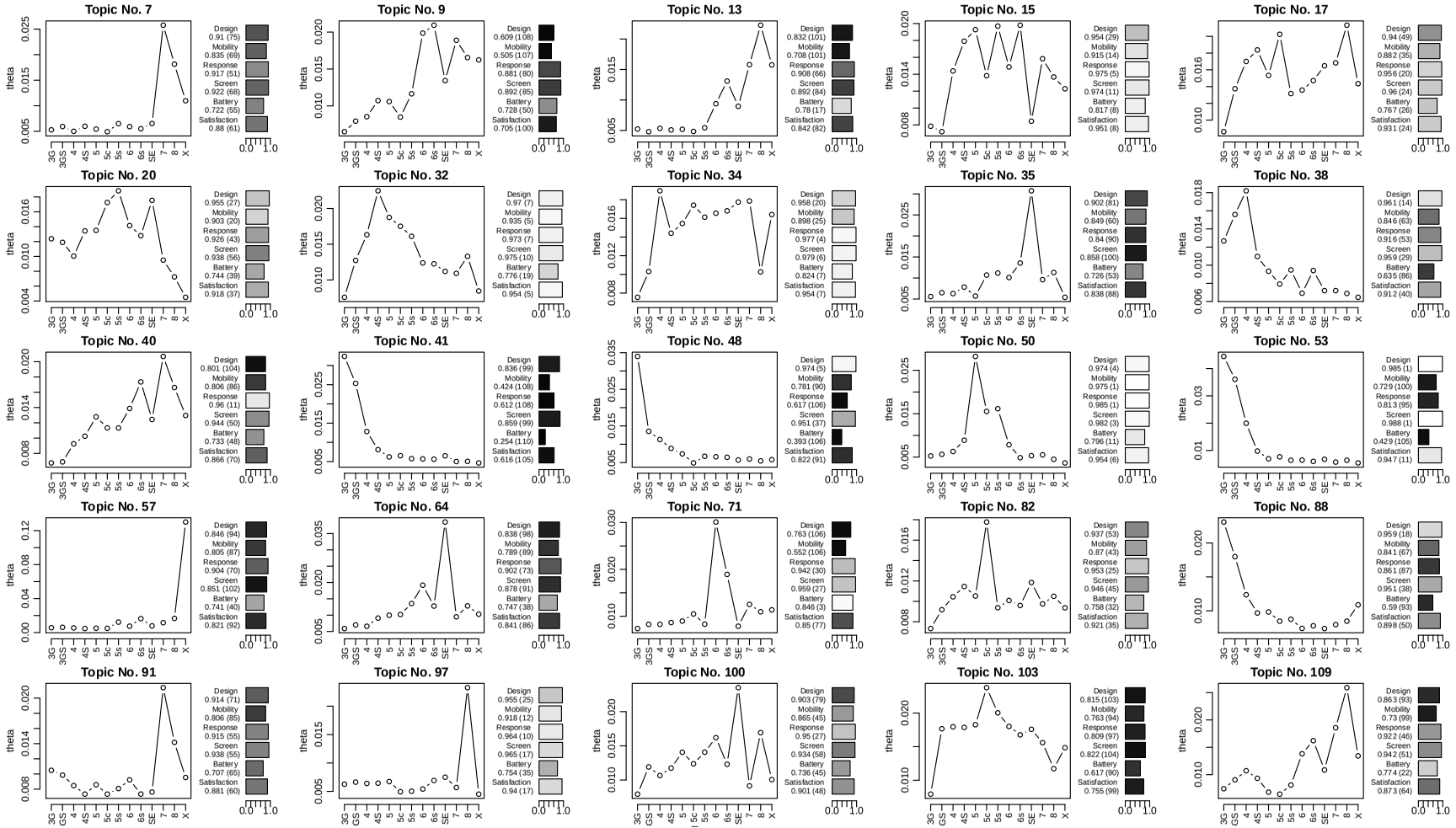
implies that the environmental factors have significantly changed during this tenyear period. Table 6 shows the correlation coefficient matrix of the topic shares of each generation. This shows that correlation coefficients tend to be higher the closer they are to the diagonal elements of the matrix. This implies that the same topic is likely to appear in close generations, and topics are different for distant generations.

For example, No.53 is a frequent topic in 3G, 3GS, and 4. In Figure 4, in terms of the topic share $\bar{\theta}_g$, 3G has the highest value, followed by those of 3GS and 4. Further, as the generation changes, the topic share decreases. Additionally, we find that design and screen are highly evaluated. However, the battery evaluation is lower than that of the other subscores. This implies that this topic was reviewed by comparing the feature phones popular at the time. Feature phones have a very long continuous standby time; on the other hand, the iPhone uses more electric power than its rivals. Therefore, its performance is relatively inferior in terms of battery. However, large screens and sophisticated designs have proved popular with consumers. Reflecting this, the overall evaluation was also higher. In addition, the reason this topic has not appeared much after the iPhone 4s is that the diffusion rate of smartphones sharply increased from 2011 to 2012. When the iPhone 4s was released, many competing smartphone models were released, and feature phones were no longer comparison targets at the time. According to the Ministry of Internal Affairs and Communications of Japan (2015), the diffusion rate of smartphones at the end of 2010 was 9.7%, but it rapidly increased to 29.3% by the end of 2011 and reached 49.5% by the end of 2012.

Gen/No.	Topic Share	Satisfaction	Frequent words
iPhone 3G			
53	0.044	0.947 (11)	*See* Table 3
48	0.034	0.822 (91)	function, cellphone, usable (-1), able (-1), 1seg, flaw, drop-down, Safari, review, ease-of-use
41	0.032	0.616 (105)	*See* Table 4
iPhone 3GS			
53	0.036	0.947 (11)	*See* Table 3
41	0.025	0.616 (105)	*See* Table 4
88	0.018	0.898 (50)	iTunes, computer, software, synchronization, able, data, Mac, cooperation, PC, convenient
iPhone 4			
53	0.02	0.947 (11)	*See* Table 3
34	0.019	0.954 (7)	problem (-1), mobile, design, response, conversion, sound-quality, display, telephone-call, operation, letter
38	0.018	0.912 (40)	able, setting, ringtone, simple, app, able (-1), folder, detailed, worry, dictionary
iPhone 4S			
32	0.022	0.954 (5)	conversion, design, letter, display, operation, mobile, response, menu, telephone-call, button
17	0.018	0.931 (24)	music, design, conversion, sound-quality, response, letter, display, telephone-call, mobile, menu
15	0.018	0.951 (8)	pretty, simple, telephone-call, design, music, conversion, letter, operation, response, sound-quality
iPhone 5			
50	0.028	0.954 (6)	*See* Table 3
15	0.019	0.951 (8)	*See above*
32	0.019	0.954 (5)	*See above*
iPhone 5c			
103	0.024	0.755 (99)	normal, design, mobile, response, display, letter, operation, telephone-call, sound-quality, conversion
17	0.02	0.931 (24)	*See above*
82	0.018	0.921 (35)	compare, feeling, have, like, slightly, personal, here, inferior, think, special
iPhone 5s			
103	0.02	0.755 (99)	*See above*
15	0.02	0.951 (8)	*See above*
20	0.019	0.918 (37)	MNP, cheap, contract, fee, lump-sum, plan, basis, monthly, cancellation, trade-in
iPhone 6			
71	0.03	0.85 (77)	*See* Table 3
9	0.02	0.705 (100)	*See* Table 4
64	0.019	0.841 (86)	size, operation, small, one-hand, screen, large, best, 4-inches, screen-size, 5-inches
iPhone 6s			
9	0.021	0.705 (100)	*See* Table 4
15	0.02	0.951 (8)	*See above*
71	0.019	0.85 (77)	*See* Table 3
iPhone SE			
64	0.039	0.841 (86)	*See above*
35	0.031	0.838 (88)	SIM-free, SIM, carrier, operate, able, usable, overseas, price, MVNO, MVNO-SIM
100	0.024	0.901 (48)	design, telephone-call, mobile, response, conversion, letter, sound-quality, music, display, menu
iPhone 7			
7	0.026	0.88 (61)	waterproof, after, other-than, usable, migration, correspondence, Suica, mobile-wallet, card, convenient
91	0.021	0.881 (60)	music, earphone, speaker, sound-quality, listen, playback, hear, attached, jack, as-such
40	0.021	0.866 (70)	change (-1), change, evolution, improvement, increase, change, this-time, feel, rise, up
iPhone 8			
109	0.026	0.873 (64)	camera, photo, video, image-quality, photograph, take-a-phonto, pretty, see, function, take
97	0.023	0.94 (17)	speed, processing, resolution, rise, experience, ability, strongest, current, current, street
13	0.022	0.842 (82)	Plus, model, correspondence, same, VoLTE, change, disappointment, rose-gold, next, personal
iPhone X			
57	0.13	0.821 (92)	screen, ID, touch, button, lock, home, certification, display, unlock, disappear
34	0.016	0.954 (7)	*See above*
9	0.016	0.705 (100)	*See* Table 4

Topic share is the value of $\bar{\theta}_g$, and satisfaction is the value of ψ_{kJ}. The numbers in parentheses after topic share and satisfaction indicate the rankings. *See* Table 3 and *See* Table 4 indicate that the frequent words of the topic are listed in the suggested table. *See above* indicates that the topic is already listed at the top of the table.

Table 5: Popular Topics for Each Segment



This Figure displays the topics shown in Table 5. The line chart is the topic share ($\bar{\theta}_g$) in each generation. The bar graph on the right of each topic shows the value of ψ. The lighter the colors are, the higher the rank is (it has a relatively large value), and the darker the colors are, the lower the rank is (it has a relatively small value).

Figure 4: Dynamics and Profiles of Popular Topics for Each Segment

	3G	3GS	4	4S	5	5c	5s	6	6s	SE	7	8	X
iPhone 3G	1.00												
iPhone 3GS	0.87	1.00											
iPhone 4	0.54	0.74	1.00										
iPhone 4S	0.21	0.41	0.76	1.00									
iPhone 5	-0.02	0.11	0.42	0.66	1.00								
iPhone 5c	-0.02	0.19	0.50	0.72	0.68	1.00							
iPhone 5s	-0.04	0.09	0.38	0.66	0.73	0.77	1.00						
iPhone 6	-0.07	0.05	0.28	0.41	0.43	0.55	0.58	1.00					
iPhone 6s	-0.11	0.00	0.28	0.45	0.36	0.52	0.61	0.81	1.00				
iPhone SE	-0.08	0.03	0.15	0.32	0.24	0.48	0.53	0.55	0.53	1.00			
iPhone 7	-0.15	-0.05	0.16	0.26	0.23	0.37	0.37	0.51	0.63	0.32	1.00		
iPhone 8	-0.13	-0.06	0.09	0.24	0.18	0.32	0.36	0.51	0.64	0.45	0.70	1.00	
iPhone X	-0.08	-0.07	-0.03	-0.05	-0.04	0.01	0.17	0.13	0.37	0.07	0.26	0.31	1.00

Table 6: Correlation Matrix of Topic Share

5.4 Share of Salient Topics

In this section, we examine the factors that increase or decrease the topics from the basic information obtained from each review on salient topics mentioned in the previous section. The parameter θ_d is obtained for each document. We set $K = 110$; therefore, θ_d is a 110-dimensional vector and has the property that $\sum_{k=1}^{K} \theta_{dk} = 1$. Following Puranam, Narayan, and Kadiyali (2017), we conduct a regression analysis to assess the precedence of θ_d. There are some models that assume objective variables such as this. For example, Cooper and Nakanishi (1988) and Berry (1994) proposed models that analyze market share. We apply a model developed by Berry (1994), which assumes a logit model for the consumers' choice probabilities, which are components of market share. Berry (1994) assumes an external good and subtracts it from the market share of each alternative to obtain a linear model. Therefore, in this study, the share of the 85 topics excluding the 25 salient topics shown in Table 5 and Figure 4 in the previous section is regarded as the share of external goods. We examine the relative impact of each salient topic compared with other topics with a high share. Let the total share of 85 topics in review d be denoted by θ_{d0} and the basic information attached to each review denoted by X_d. We estimate the following linear regression model:

$$\ln(\theta_{dk}) - \ln(\theta_{d0}) = X_d \lambda_k + \varepsilon_{dk},\ \varepsilon_{dk} \sim N(0, \sigma_k^2) \tag{27}$$

In X_d, the following explanatory variables and control variables are used:

- *Latest Model*: As many new smartphone models are released during the analysis period, existing models become relatively obsolete. Therefore, a dummy variable that takes the value 1 if the reviewed model was the latest model when the review was posted, and 0 otherwise, is included in the explanatory variables. If the estimated coefficient is positive, the topic is likely to be a review of the latest model, and if the parameter is negative, it is likely to be a review of an outdated model. Out of 5,375 reviews, 4,193 (78 %) reviews were for the latest models at the time of posting, whereas 1,182 (22 %) reviews were for outdated models.

- *Reviewer's Experience*: Each review includes the author's name; therefore, we can track if the author has posted reviews in the past. Based on the name, for each review, a dummy variable that takes the value 1 if the author has posted a review of the iPhone in the past, and 0 otherwise, is included in the explanatory variables. If the estimated coefficient is positive, the topic is likely to be mentioned by consumers who have posted reviews in the past, and if the parameter is negative, it is likely to be mentioned by consumers who have posted reviews more recently.

- *Year Control*: The regression model includes the year a review was posted as a control variable. Since time trends may not change linearly, a separate dummy variable is included for each year. In the estimation, we introduce the constraint that the parameter for 2008 (the first year) is 0.

- *Generation Control*: We also include a dummy variable corresponding to the smartphone model to be reviewed as a control variable. In the estimation, we introduce the constraint that the parameter for 3G (the first model) is 0.

Figure 5 shows the obtained results, specifically, the estimated coefficients of latest model and reviewer's experience. The horizontal axis shows the estimated coefficients of latest model, and the vertical axis those of reviewer's experience. The numbers in the figure are topic numbers, and the symbols attached as subscripts of the numbers indicate the significance level of latest model/reviewer's experience. If the symbol is "+" or "−", the estimated coefficient is significant at the 5% level; "o" indicates that it is not significant.

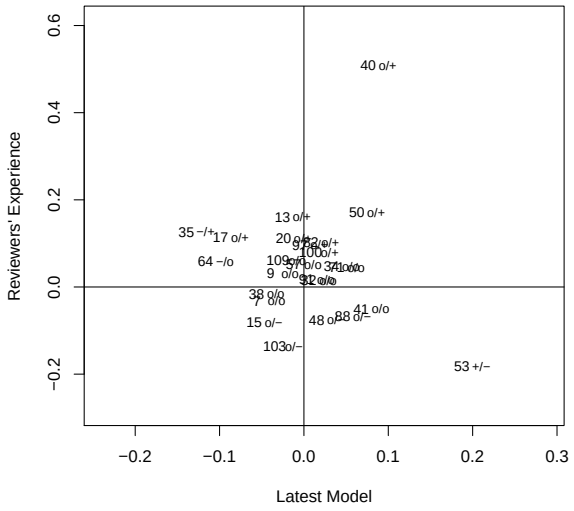


Figure 5: Result of Regression Analyisis

The results in Figure 5 can be better interpreted by comparing them with those in Table 5 and Figure 4. For example, if more reviews of outdated models were posted, the appearance probability of topic No.35 would increase. Furthermore, consumers who have previously posted reviews of the iPhone would also increase the likelihood of this topic. Looking at Table 5, No.35 is a topic about SIM-free and MVNO (Mobile Virtual Network Operator, a new low-priced carrier), and it seems to include reviews by consumers relatively familiar with telecommunication policies and technologies. This implies that the reviewers have obtained the iPhone at a low price by purchasing a model when it has become outdated and the price has dropped. Looking at Figure 4, the probability that this topic appears would increase for reviews of the iPhone SE, which was released as a low-priced version. This suggests that the reviews are by consumers who want to acquire smartphones at a low price. In addition, the results show that No.40 is more likely to consist of reviews by consumers who have posted reviews in the past. Looking at Figure 4, the relationship of this topic with the iPhone 7, 8, and 6s is particularly strong. Therefore, this topic seems to have been created by consumers who had been using an older model before the time of their review and then bought a new model and reviewed it.

6 Conclusion

In this study, we developed the TSAM, expanding the commonly used LDA method and applied it to online reviews that have both document information written in natural language and evaluation

information in the form of continuous scores, to classify natural language and score information simultaneously. In the empirical analysis, the TSAM was applied to 10 years of review data on the iPhone 3G through the iPhone X in the Japanese market. The following results were obtained:

First, we found that the TSAM can be used as a prediction model. We cannot use LDA for prediction because LDA is an unsupervised model. However, using the TSAM proposed in this study, continuous scores are predicted from document data written in natural language. In particular, as a result of a comparison with regression models (an ordered logit model and a binomial logit model), the TSAM's prediction performance when only document information was obtained was very high. The TSAM can be used as a prediction model using a large amount of document information as input data. Further, in terms of computational efficiency, since collapsed Gibbs sampling is used for parameter estimation, fast and stable results can be obtained. Second, by interpreting the model, it is possible to obtain more useful information for marketing decision-making compared to that obtained from LDA. Although LDA cannot directly interpret the sentiment and mood of a document, the TSAM provides an average rating score for each topic. Therefore, we can directly determine whether the topic itself corresponds to a good rating or a bad one. Furthermore, our model provides parameter estimates and expectations for five aspects of a product. Although the data adapted in this study include multiple evaluation aspects and there may be large differences in the scores across the different aspects, we can visually interpret the relationship between topics and aspects based on the output of the model. In addition, by checking words that are strongly related to the topic, researchers can find factors affecting each score that cannot be discovered by simply looking at the evaluation score of each aspect. Furthermore, the factors affecting topic shares were also examined using a market share model.

Here, we would like to identify some issues with this study and propose a direction for future research. The first issue is that the examination of the factors affecting each topic is limited in this study. Some previous studies have proposed a model that can examine precedence by assuming a linear combination regression term for the parameters of the Dirichlet distribution. However, it is difficult to assume conjugate prior distributions for the prior parameters of this linear combination. This implies that the estimation result will be inefficient and unstable. In this study, since the model is developed assuming only conjugate priors, we cannot incorporate linear combination priors. The second issue is the number of topics. In marketing decision-making, it is better to discuss a number of topics that can be interpreted by analysts. In the $K = 110$ model, it is not easy to interpret all the topics. As a future research topic, we could try to use a hierarchical model such as Pachinko allocation model (Li and MaCallum, 2006). As regards the practical range, it is necessary to analyze not only the online reviews of the smartphones covered in this study but also those of other smartphones and to verify the properties of the TSAM.

A SAM Simulation Results

In this section, we examine the SAM, comparing it with another well-known method. As defined above, SAM requires J-a dimensional integer score vector as input data. Each score needs to be evaluated on the Q_j point scale, but the range does not have to be the same for all items, and the binomial value $\{0, 1\}$ scale is also acceptable. In this section, we aim to assess the results by comparing them with those from another method. Therefore, we simulate input data and compare the results. The method we compare the SAM with is *k-means*, which is frequently used for clustering. Two-dimensional data with $D = 400$ are generated as sample data. All the data are integers from 0 to 99. We assume there are four groups, with the following average values: $(70, 70), (30, 70), (70, 30), (30, 30)$. There are 100 observations in each group. Therefore, the settings for SAM are $J = 2$, $(Q_1, Q_2) = (100, 100)$.

To compare the two methods, we focus on the SAM parameter $\{\psi_{kj}\}$. This parameter corresponds to the cluster mean in the k-means method. For SAM, the first 1,000 iterations are omitted as burn-in and the subsequent 1,000 iterations are collected as the sample. $\psi_{kj}^{(h)}$ is obtained from $c_d^{(h)}$ in each iteration, and the median of $\psi_{kj}^{(h)}$ is used for the analysis. In addition, we classify each observation into one of four topics by using the topic allocation parameter $c_d^{(h)}$. Figure 6 shows the results of the k-means

clustering (left) and the expected values $Q_j\psi_{kj}$ (right). It can be seen that almost the same values are obtained from both methods. This implies that the proposed TSAM combines LDA and k-means.

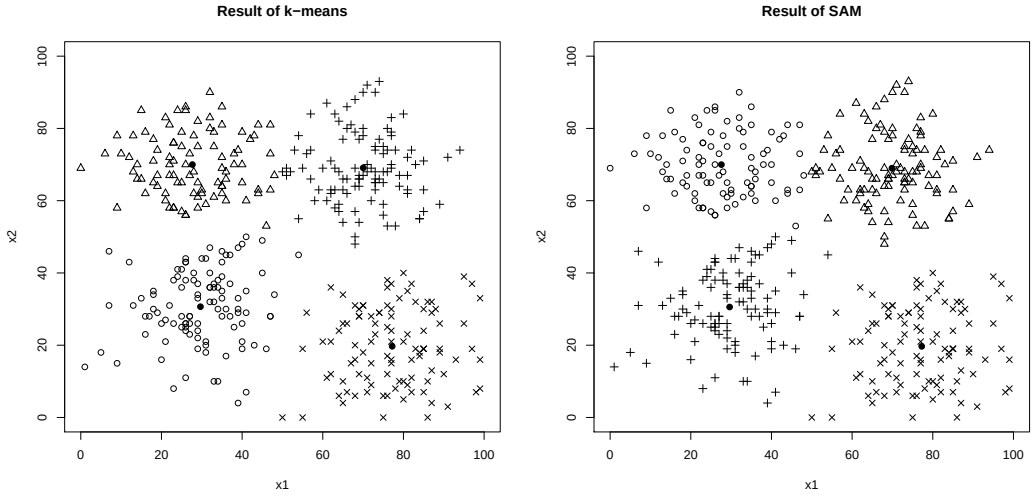


Figure 6: Result of k-means and SAM

B Sampling Path of TSAM

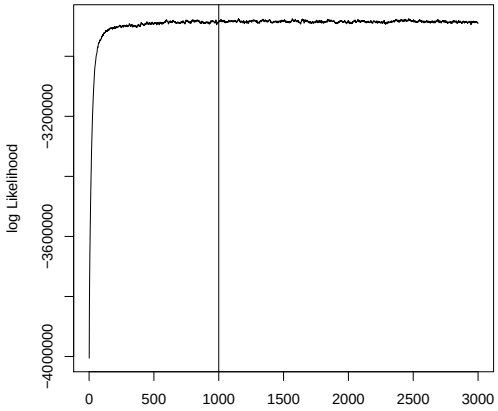


Figure 7: Sampling Path of TSAM $K = 110$

References

Anderson, E. T., & Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research, 51*(3), 249-269.

Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing Research, 37*(3), 363-375.

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science, 46*(4), 557-590.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2019). Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing.*

Berger, J., & Milkman, K. L. (2012). What makes online content viral?. *Journal of marketing research, 49*(2), 192-205.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics, 25*(2), 242-262.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).* ACM.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17-35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Borah, A., & Tellis, G. J. (2016). Halo (spillover) effects in social media: do product recalls of one brand hurt or help rival brands?. *Journal of Marketing Research, 53*(2), 143-160.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science, 35*(6), 953-975.

Chen, Z., & Lurie, N. H. (2013). Temporal contiguity and negativity bias in the impact of online word of mouth. *Journal of Marketing Research, 50*(4), 463-476.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research, 43*(3), 345-354.

Cooper, L. G., & Nakanishi, M. (1988). *Market Share Analysis, International Series in Quantitative Marketing*, Boston: Kluwer Academic Publishers.

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing, 21*(4), 23-45.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing, 90*(2), 217-232.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences, 101*(suppl 1), 5228-5235.

Hewett, K., Rand, W., Rust, R. T., & Van Heerde, H. J. (2016). Brand buzz in the echoverse. *Journal of Marketing, 80*(3), 1-24.

Humphreys, A., & Wang, R. J. H. (2017). Automated text analysis for consumer research. *Journal of Consumer Research, 44*(6), 1274-1306.

Kannan, P. K., & Li. H. A. (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing, 34*(1), 22-45.

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. *In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 230-237).*

Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research, 48*(5), 881-894.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *In Proceedings of the 23rd international conference on Machine learning (pp. 577-584).* ACM.

Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on Twitter. *Journal of Advertising, 46*(2), 236-247.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing, 70*(3), 74-89.

Luo, X. (2009). Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. *Marketing Science, 28*(1), 148-165.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2 (pp. 880-889).* Association for Computational Linguistics.

Nam, H., Joshi, Y. V., & Kannan, P. K. (2017). Harvesting brand information from social tags. *Journal of Marketing, 81*(4), 88-108.

Onishi, H., & Manchanda, P. (2012). Marketing activity, blogging and sales. *International Journal of Research in Marketing, 29*(3), 221-234.

Puranam, D., Narayan, V., & Kadiyali, V. (2017). The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors. *Marketing Science, 36*(5), 726-746.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *In Proceedings of the 20th conference on Uncertainty in artificial intelligence (pp. 487-494).* AUAI Press.

Sato, T., Hashimoto, T., & Okumura, M. (2017). Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval. *In Proceedings of the twenty-three annual meeting of the association for natural language processing.* The Association for Natural Language Processing.

Stephen, A. T., & Galak, J. (2012). The effects of traditional and social earned media on sales: A study of a microlending marketplace. *Journal of marketing research, 49*(5), 624-639.

Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science, 31*(2), 198-215.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research, 51*(4), 463-479.

Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption. *Journal of Marketing Research, 56*(1), 18-36.

Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science, 35*(3), 405-426.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing, 80*(6), 97-121.

A Online Appendix: Estimation of TSAM

A.1 Model Definition

Assume that there are D documents and N words ($D < N$). There are V vocabulary words observed in the text data; therefore, word w_i is a V-dimensional vector whose element $w_{iv} = 1$ if the i-th word is vocabulary word v; otherwise, $w_{iv'} = 0, \forall v' \neq v$. w_i is affiliated with any one of the documents; therefore, we introduce a document affiliation index x_i. If w_i if affiliated with document d, $x_{id} = 1$, and $x_{id'} = 0, \forall d' \neq d$. In addition, for each document, J item ratings are observed. L let y_{dj} be the ordered integer score for aspect j in document d, where the minimum score is 0 and the maximum score is $Q_j - 1$. In addition, assume there are K latent topics and denote the topic affiliation index for text as z_i and the score affiliation index as c_d.

$$w_i \sim Categorical_V(\tilde{\phi}_i), \tilde{\phi}_{iv} = \prod_{k=1}^{K} \phi_{kv}^{z_{ik}} \tag{28}$$

$$z_i \sim Categorical_K(\tilde{\theta}_i), \tilde{\theta}_{ik} = \prod_{d=1}^{D} \theta_{dk}^{x_{id}} \tag{29}$$

$$y_{dj} \sim Binomial(Q_j, \tilde{\psi}_{dj}), \tilde{\psi}_{dj} = \prod_{k=1}^{K} \psi_{kj}^{c_{dj}} \tag{30}$$

$$c_d \sim Categorical_K(\theta_d) \tag{31}$$

The prior distributions of ϕ_k, ψ_d and θ_d are defined as follows:

$$\begin{aligned} \phi_k &\sim Dirichlet_V(\beta) && (32)\\ \theta_d &\sim Dirichlet_K(\alpha) && (33)\\ \psi_{kj} &\sim Beta(\gamma) && (34)\\ & && (35) \end{aligned}$$

A.2 Densities

$$\begin{aligned} \pi(W|\Phi, Z) &= \prod_{i=1}^{N} \pi(w_i|\Phi) && (36)\\ &= \prod_{i=1}^{N} \prod_{v=1}^{V} \left(\prod_{k=1}^{K} \phi_{kv}^{z_{ik}} \right)^{w_{iv}} && (37)\\ &= \prod_{i=1}^{N} \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{kv}^{z_{ik} \times w_{iv}} && (38) \end{aligned}$$

$$\begin{aligned} \pi(Z|\Theta) &= \prod_{i=1}^{N} \pi(z_i|\Theta) && (39)\\ &= \prod_{i=1}^{N} \prod_{k=1}^{K} \left(\prod_{d=1}^{D} \theta_{dk}^{x_{id}} \right)^{z_{ik}} && (40)\\ &= \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{x_{id} \times z_{ik}} && (41) \end{aligned}$$

To define the density of Y, we introduce the sub-observation y_{djq}. For example, if $Q_j = 4$ and $y_{dj} = 2$, we divide y_{dj} into the Q_j-dimensional observation vector $\{y_{dj1}, y_{dj2}, y_{dj3}, y_{dj4}\} = \{1, 1, 0, 0\}$. y_{djq} is the q-th element of the vector.

$$\begin{aligned}
\pi(Y|\Psi, C) &= \prod_{d=1}^{D}\prod_{j=1}^{J}\prod_{q=1}^{Q_j}\pi(y_{djq}|\Psi, C) && (42)\\
&= \prod_{d=1}^{D}\prod_{j=1}^{J}\prod_{q=1}^{Q_j}\left(\prod_{k=1}^{K}\psi_{kj}^{c_{dk}}\right)^{y_{djq}}\left(\prod_{k=1}^{K}[1-\psi_{kj}]^{c_{dk}}\right)^{1-x_{djq}} && (43)\\
&= \prod_{d=1}^{D}\prod_{j=1}^{J}\prod_{q=1}^{Q_j}\prod_{k=1}^{K}\left((\psi_{kj})^{y_{djq}}(1-\psi_{kj})^{1-y_{djq}}\right)^{c_{dk}} && (44)\\
& && (45)
\end{aligned}$$

$$\begin{aligned}
\pi(C|\Theta) &= \prod_{d=1}^{D}\pi(c_d|\Theta) && (46)\\
&= \prod_{d=1}^{D}\prod_{k=1}^{K}(\theta_{dk})^{c_{dk}} && (47)\\
&= \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{dk}^{c_{dk}} && (48)
\end{aligned}$$

$$\begin{aligned}
\pi(\Phi) &= \prod_{k=1}^{K}\pi(\phi_k) && (49)\\
&= \prod_{k=1}^{K}\frac{1}{B(\beta)}\prod_{v=1}^{V}\phi_{kv}^{\beta-1} && (50)\\
&\propto \prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{kv}^{\beta-1} && (51)
\end{aligned}$$

$$\begin{aligned}
\pi(\Psi) &= \prod_{k=1}^{K}\prod_{j=1}^{J}\pi(\psi_{kj}) && (52)\\
&= \prod_{k=1}^{K}\prod_{j=1}^{J}\frac{1}{B(\gamma)}\psi_{kj}^{\gamma-1}(1-\psi_{kj})^{\gamma-1} && (53)\\
&\propto \prod_{k=1}^{K}\prod_{j=1}^{J}\psi_{kj}^{\gamma-1}(1-\psi_{kj})^{\gamma-1} && (54)
\end{aligned}$$

$$\pi(\Theta) = \prod_{d=1}^{D} \pi(\theta_d) \tag{55}$$

$$= \prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{dk}^{\alpha-1} \tag{56}$$

$$\propto \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha-1} \tag{57}$$

A.3 Likelihood Function

$$L(\mathcal{D}|Z, C, \Phi, \Psi) = \pi(W, Y|\Phi, \Psi, Z, C) \tag{58}$$

$$= \pi(W|\Phi, Z)\pi(Y|\Psi, C) \tag{59}$$

$$= \prod_{i=1}^{N} \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{kv}^{z_{ik} \times w_{iv}} \times \prod_{d=1}^{D} \prod_{j=1}^{J} \prod_{q=1}^{Q_j} \prod_{k=1}^{K} \left((\psi_{kj})^{y_{djq}} (1 - \psi_{kj})^{1-y_{djq}} \right)^{c_{dk}} \tag{60}$$

A.4 Full Conditional Posterior Distribution

$$\pi(Z, C, \Phi, \Psi, \Theta|\mathcal{D}) \propto L(\mathcal{D}|Z, C, \Phi, \Psi)\pi(Z|\Theta)\pi(C|\Theta)\pi(\Phi)\pi(\Psi)\pi(\Theta) \tag{61}$$

$$\propto \prod_{i=1}^{N} \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{kv}^{z_{ik} \times w_{iv}} \times \prod_{d=1}^{D} \prod_{j=1}^{J} \prod_{q=1}^{Q_j} \prod_{k=1}^{K} \left((\psi_{kj})^{y_{djq}} (1 - \psi_{kj})^{1-y_{djq}} \right)^{c_{dk}} \tag{62}$$

$$\times \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{x_{id} \times z_{ik}} \times \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{c_{dk}} \tag{63}$$

$$\times \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{kv}^{\beta-1} \times \prod_{k=1}^{K} \prod_{j=1}^{J} \psi_{kj}^{\gamma-1} (1 - \psi_{kj})^{\gamma-1} \times \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha-1} \tag{64}$$

$$\propto \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{kv}^{\sum_{i=1}^{N} z_{ik} w_{iv}} \tag{65}$$

$$\times \prod_{j=1}^{J} \prod_{k=1}^{K} (\psi_{kj})^{\sum_{d=1}^{D} \sum_{q=1}^{Q_j} y_{djq} c_{dk}} (1 - \psi_{kj})^{\sum_{d=1}^{D} \sum_{q=1}^{Q_j} (1 - y_{djq}) c_{dk}} \tag{66}$$

$$\times \prod_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{\sum_{i=1}^{N} x_{id} z_{ik}} \times \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{c_{dk}} \tag{67}$$

$$\times \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{kv}^{\beta-1} \times \prod_{k=1}^{K} \prod_{j=1}^{J} \psi_{kj}^{\gamma-1} (1 - \psi_{kj})^{\gamma-1} \times \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha-1} \tag{68}$$

To simplify the expression, we define the following variables:

$$n_{kv} = \sum_{i=1}^{N} z_{ik} w_{iv} \tag{69}$$

$$n_{kj1} = \sum_{d=1}^{D} \sum_{q=1}^{Q_j} y_{djq} c_{dk} \tag{70}$$

$$n_{kj0} = \sum_{d=1}^{D} \sum_{q=1}^{Q_j} (1 - y_{djq}) c_{dk} \tag{71}$$

$$n_{dk} = \sum_{i=1}^{N} x_{id} z_{ik} \tag{72}$$

$$\pi(Z, C, \Phi, \Psi, \Theta | \mathcal{D}) \propto \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{kv}^{n_{kv}} \times \prod_{j=1}^{J} \prod_{k=1}^{K} (\psi_{kj})^{n_{kj1}} (1 - \psi_{kj})^{n_{kj0}} \tag{73}$$

$$\times \prod_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{n_{dk}} \times \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{c_{dk}} \tag{74}$$

$$\times \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{kv}^{\beta-1} \times \prod_{k=1}^{K} \prod_{j=1}^{J} \psi_{kj}^{\alpha-1} (1 - \psi_{kj})^{\alpha-1} \times \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\gamma-1} \tag{75}$$

$$\propto \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{kv}^{n_{kv}+\beta-1} \tag{76}$$

$$\times \prod_{j=1}^{J} \prod_{k=1}^{K} (\psi_{kj})^{n_{kj1}+\gamma-1} (1 - \psi_{kj})^{n_{kj0}+\gamma-1} \tag{77}$$

$$\times \prod_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{n_{dk}+c_{dk}+\alpha-1} \tag{78}$$

The TSAM parameters Z and C are obtained by collapsed Gibbs sampling.

$$\begin{aligned}
\pi(Z,C|\mathcal{D}) &\propto \int\int\int \pi(Z,C,\Phi,\Psi,\Theta|\mathcal{D})d\Phi d\Psi d\Theta && (79)\\
&\propto \int \prod_{v=1}^{V}\prod_{k=1}^{K} \phi_{kv}^{n_{kv}+\beta-1} d\Phi && (80)\\
&\quad \times \int \prod_{j=1}^{J}\prod_{k=1}^{K} (\psi_{kj})^{n_{kj1}+\gamma-1}(1-\psi_{kj})^{n_{kj0}+\gamma-1} d\Psi && (81)\\
&\quad \times \int \prod_{k=1}^{K}\prod_{d=1}^{D} \theta_{dk}^{n_{dk}+c_{dk}+\alpha-1} d\Theta && (82)\\
&\propto \prod_{k=1}^{K} \frac{\prod_{v=1}^{V}\Gamma(n_{kv}+\beta)}{\Gamma(\sum_{v=1}^{V} n_{kv}+V\beta)} && (83)\\
&\quad \times \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\Gamma(n_{kj1}+\alpha)\Gamma(n_{kj0}+\gamma)}{\Gamma(n_{kj1}+n_{kj0}+2\gamma)} && (84)\\
&\quad \times \prod_{d=1}^{D} \frac{\prod_{k=1}^{K}\Gamma(n_{dk}+c_{dk}+\alpha)}{\Gamma(\sum_{k=1}^{K}(n_{dk}+c_{dk})+K\alpha)} && (85)
\end{aligned}$$

Let us introduce the following variables:

$$\begin{aligned}
n_k &= \sum_{v=1}^{V} n_{kv} && (86)\\
n_{kj} &= n_{kj1}+n_{kj0} && (87)\\
n_d &= \sum_{k=1}^{K} n_{dk} && (88)
\end{aligned}$$

Note that $\sum_{k=1}^{K} c_{dk} = 1$.

A.5 Posterior Distributions

A.5.1 Conditional Posterior Distribution of z_{ik}

The conditional posterior distribution of z_{ik} is similar to that of the original LDA.

$$\begin{aligned}
\pi(z_{ik}=1|Z_{-i}C,\mathcal{D}) &\propto \left(\frac{n_{kv^*,-i}+\beta}{n_{k,-i}+V\beta}\right)\left(\frac{n_{d^*k}+c_{dk}+\gamma}{n_{d^*,-i}+1+K\gamma}\right) && (89)\\
&\propto \left(\frac{n_{kv^*,-i}+\beta}{n_{k,-i}+V\beta}\right)(n_{d^*k,-i}+c_{dk}+\gamma) && (90)
\end{aligned}$$

A.5.2 Conditional Posterior Distribution of c_{dk}

$$\pi(c_{dk}=1|Z,C_{-d},\mathcal{D}) \propto \prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\Gamma(n_{kj1}+\alpha)\Gamma(n_{kj0}+\alpha)}{\Gamma(n_{kj1}+n_{kj0}+2\alpha)} \tag{91}$$

$$\times\prod_{d=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(n_{dk}+c_{dk}+\gamma)}{\Gamma(\sum_{k=1}^{K}(n_{dk}+c_{dk})+K\gamma)} \tag{92}$$

$$\propto \prod_{j=1}^{J}\frac{\Gamma(n_{kj1}+\alpha)\Gamma(n_{kj0}+\alpha)}{\Gamma(n_{kj1}+n_{kj0}+2\alpha)} \tag{93}$$

$$\times\frac{\Gamma(n_{dk}+c_{dk}+\gamma)}{\Gamma(\sum_{k=1}^{K}(n_{dk}+c_{dk})+K\gamma)} \tag{94}$$

where $n_{kj1}=\sum_{d=1}^{D}\sum_{q=1}^{Q_j}y_{djq}c_{dk}=\sum_{d=1}^{D}c_{dk}y_{dj}$ and $n_{kj0}=\sum_{d=1}^{D}\sum_{q=1}^{Q_j}c_{dk}(1-y_{djq})=\sum_{d=1}^{D}c_{dk}(Q_j-y_{dj})$. Therefore, $n_{kj1}+n_{kj0}=\sum_{d=1}^{D}c_{dk}Q_j$ and we can obtain the following equations:

$$\Gamma(n_{kj1}+\alpha) = \Gamma\left(\sum_{d'=1,d'\neq d}^{D}c_{d'k}y_{d'j}+c_{dk}y_{dj}+\alpha\right) \tag{95}$$

$$= \Gamma\left(\sum_{d'=1,d'\neq d}^{D}c_{d'k}y_{d'j}+y_{dj}+\alpha\right) \tag{96}$$

$$= \Gamma\left(n_{kj1,-d}+y_{dj}+\alpha\right) \tag{97}$$

$$= \prod_{q=1}^{y_{dj}}\left(n_{kj1,-i}+q-1+\alpha\right)\Gamma\left(n_{kj1,-d}+\alpha\right) \tag{98}$$

$$\tag{99}$$

where $n_{kj1,-d}=\sum_{d'=1,d'\neq d}^{D}c_{d'k}y_{d'j}$.

$$\Gamma(n_{kj0}+\alpha) = \Gamma\left(\sum_{d'=1,d'\neq d}^{D}c_{d'k}(Q_j-y_{d'j})+c_{dk}(Q_j-y_{dj})+\alpha\right) \tag{100}$$

$$= \Gamma\left(\sum_{d'=1,d'\neq d}^{D}c_{d'k}y_{d'j}+(Q_j-y_{dj})+\alpha\right) \tag{101}$$

$$= \Gamma\left(n_{kj0,-d}+(Q-y_{dj})+\alpha\right) \tag{102}$$

$$= \prod_{q=1}^{(Q_j-y_{dj})}\left(n_{kj0,-d}+q-1+\alpha\right)\Gamma\left(n_{kj0,-d}+\alpha\right) \tag{103}$$

$$\tag{104}$$

where $n_{kj0,-d}=\sum_{d'=1,d'\neq d}^{D}z_{d'k}(Q_j-x_{d'j})$.

$$\Gamma(n_{kj1}+n_{kj0}+2\alpha) = \Gamma(\sum_{d=1}^{D}c_{dk}Q_j+2\alpha) \tag{105}$$

$$= \Gamma\left(n_{kj,-d}+Q_j+2\alpha\right) \tag{106}$$

$$= \prod_{q=1}^{Q_j}\left(n_{kj,-d}+q-1+2\alpha\right)\Gamma\left(n_{dk,-d}+2\alpha\right) \tag{107}$$

where $n_{k,-d} = \sum_{d'=1,d'\neq d}^{D} z_{d'k} Q$.

From the equations above, we can obtain the following relationship:

$$\prod_{j=1}^{J} \frac{\Gamma(n_{kj1}+\alpha)\Gamma(n_{kj0}+\alpha)}{\Gamma(n_{kj1}+n_{kj0}+2\alpha)} \tag{108}$$

$$\propto \frac{\prod_{q=1}^{y_{dj}} (n_{kj1,-d}+q-1+\alpha) \prod_{q=1}^{(Q_j-y_{dj})} (n_{kj0,-d}+q-1+\alpha)}{\prod_{q=1}^{Q_j} (n_{kj,-d}+q-1+2\alpha)} \tag{109}$$

In addition, we can transform the component as follows:

$$\frac{\Gamma(n_{dk}+c_{dk}+\gamma)}{\Gamma(\sum_{k=1}^{K}(n_{dk}+c_{dk})+K\gamma)} \propto \frac{\Gamma(n_{dk}+c_{dk}+\gamma)}{\Gamma(n_d+1+K\gamma)} \tag{110}$$

$$\propto \Gamma(n_{dk}+1+\gamma) \tag{111}$$

$$\propto (n_{dk}+\gamma)\Gamma(n_{dk}+\gamma) \tag{112}$$

The posterior distribution of c_{dk} is obtained as follows:

$$\pi(c_{dk}=1|Z,C_{-d},\mathcal{D}) \tag{113}$$

$$\propto \frac{\prod_{q=1}^{y_{dj}} (n_{kj1,-d}+q-1+\alpha) \prod_{q=1}^{(Q_j-y_{dj})} (n_{kj0,-d}+q-1+\alpha)}{\prod_{q=1}^{Q_j} (n_{kj,-d}+q-1+2\alpha)} (n_{dk}+\gamma) \tag{114}$$

A.6 Parameters Ψ, Φ, and Θ

Given Z and C, the conditional posterior distribution of ψ_{kj} is a beta distribution; therefore, the expectation is obtained as follows:

$$\pi(\psi_{kj}|Z,C,\Phi,\Psi_{-kj},\Theta,\mathcal{D}) \propto \prod_{v=1}^{V}\prod_{k=1}^{K} \phi_{kv}^{n_{kv}+\beta-1} \tag{115}$$

$$\times \prod_{j=1}^{J}\prod_{k=1}^{K} (\psi_{kj})^{n_{kj1}+\gamma-1} (1-\psi_{kj})^{n_{kj0}+\gamma-1} \tag{116}$$

$$\times \prod_{k=1}^{K}\prod_{d=1}^{D} \theta_{dk}^{n_{dk}+c_{dk}+\alpha-1} \tag{117}$$

$$\propto (\psi_{kj})^{n_{kj1}+\gamma-1} (1-\psi_{kj})^{n_{kj0}+\gamma-1} \tag{118}$$

$$E[\psi_{kj}] = \frac{n_{kj1}+\gamma}{n_{kj1}+n_{kj0}+2\gamma} \tag{119}$$

Similarly, given Z and C, the conditional posterior distributions of ϕ_{kv} and θ_{dk} are Dirichlet distributions. Therefore, their respective expectations are obtained as follows:

$$\pi(\phi_{kv}|Z,C,\Phi_{-kv},\Psi,\Theta,\mathcal{D}) \propto \phi_{kv}^{n_{kv}+\beta-1} \tag{120}$$

$$E[\phi_{kv}] = \frac{n_{kv}+\beta}{n_k+V\beta} \tag{121}$$

$$\pi(\theta_{dk}|Z,C,\Phi,\Psi,\Theta_{-dk},\mathcal{D}) \propto \theta_{dk}^{n_{dk}+c_{dk}+\alpha-1} \tag{122}$$

$$E[\theta_{dk}] = \frac{n_{dk}+\alpha}{n_d+K\alpha} \tag{123}$$

B Online Appendix: Pseudocode

B.1 Pseudocode for LDA

Algorithm 1 Algorithm for LDA

Input: K (Number of topics): Integer
Input: BN (Number of burn-in iteratoins): Integer
Input: NN (Number of sampling iteratoins): Integer
Input: X (Document Index): A N-dimensional vector, $X[i] \in \{1, \cdots, D\}, \forall i = 1, \cdots, N$
Input: W (Word Index): A N-dimensional vector, $W[i] \in \{1, \cdots, V\}, \forall i = 1, \cdots, N$
Input: Z (Topic Allocation Initial Value): A N-dimensional vector, $Z[i] \in \{1, \cdots, K\}, \forall i = 1, \cdots, N$
Input: α, β: Hyperparameters
Output: ZZ: A $N \times K$ matrix
Initialisation:
define: ZZ is a $NN \times N$ matrix.
define: Nkv is a $K \times V$ matrix, Nk is a K-dimensional vector, Ndk is a $D \times K$ matrix.
For $i \leftarrow 1$ **to** N **do**:
    $v \leftarrow W[i]$, $k \leftarrow Z[i]$, $d \leftarrow X[i]$
    $Nkv[k, v] \leftarrow Nkv[k, v] + 1$, $Nk[k] \leftarrow Nk[k] + 1$, $Ndk[d, k] \leftarrow Ndk[d, k] + 1$
End for
Collapsed Gibbs Sampling Loop:
define: p is a K-dimensional vector.
For $nn \leftarrow 1$ **to** $BN + NN$ **do**:
    For $i \leftarrow 1$ **to** N **do**:
        $v \leftarrow W[i]$, $k \leftarrow Z[i]$, $d \leftarrow X[i]$
        $Nkv[k, v] \leftarrow Nkv[k, v] - 1$, $Nk[k] \leftarrow Nk[k] - 1$, $Ndk[d, k] \leftarrow Ndk[d, k] - 1$
        For $k \leftarrow 1$ **to** K **do**:
            $p[k] \leftarrow \frac{Nkv[k,v]+\beta}{Nk[k]+V*\beta}(Ndk[d, k] + \alpha)$
        End for
        $k \sim Cat\,(p)$
        $Nkv[k, v] \leftarrow Nkv[k, v] + 1$, $Nk[k] \leftarrow Nk[k] + 1$, $Ndk[d, k] \leftarrow Ndk[d, k] + 1$
        If $nn > BN$:
            $ZZ[i, k] \leftarrow ZZ[i, k] + 1$
        End for
    End for
End for

B.2 Online Appendix: Pseudocode for SAM (Score Allocation Model)

Algorithm 2 Algorithm for SAM (Score Allocation Model)

Input: K (Number of topics): Integer
Input: BN (Number of burn-in iteratoins): Integer
Input: NN (Number of sampling iteratoins): Integer
Input: Y (Score Index): A $D \times J$ matrix, $Y[d, j] \in \{1, \cdots, Q_j\}, \forall d = 1, \cdots, D$
Input: C (Topic Allocation Initial Value): A D-dimensional vector, $C[d] \in \{1, \cdots, K\}, \forall d = 1, \cdots, D$
Input: γ: Hyperparameter
Output: CC: A $D \times K$ matrix
Initialisation:
Define: CC is a $D \times K$ matrix.
Define: Nkj is a $K \times J$ matrix, $Nkj1$ is a $K \times J$ matrix, $Nkj0$ is a $K \times J$ matrix
For $d \leftarrow 1$ **to** D **do**:
    For $j \leftarrow 1$ **to** J **do**:
        $q \leftarrow Y[d, j]$, $k \leftarrow C[d]$
        $Nkj[k, j] \leftarrow Nkv[k, j] + Q_j$,
        $Nkj1[k, j] \leftarrow Nkj1[k, j] + q$, $Nkj0[k, j] \leftarrow Nkj0[d, k] + Q_j - q$

End for
End for
Collapsed Gibbs Sampling Loop:
define: $p1, p0, pA, p$ is a K-dimensional vector.
For $nn \leftarrow 1$ **to** $BN + NN$ **do**:
For $d \leftarrow 1$ **to** D **do**:
For $j \leftarrow 1$ **to** J **do**:
$q \leftarrow Y[d, j]$, $k \leftarrow C[d]$
$Nkj[k, j] \leftarrow Nkv[k, j] - Q_j$,
$Nkj1[k, j] \leftarrow Nkj1[k, j] - q$, $Nkj0[k, j] \leftarrow Nkj0[d, k] - (Q_j - q)$
End for

For $k \leftarrow 1$ **to** K **do**:
For $j \leftarrow 1$ **to** J **do**:
For $q \leftarrow 0$ **to** Q_j **do**:
If $q \leq Y[d, j]$ **do**:
$p1[k] \leftarrow p1[k] * (Nkj1[k, j] + q + \gamma)$
End If
If $q > Y[d, j]$ **do**:
$p0[k] \leftarrow p2[k] * (Nkj0[k, j] + Q_j - q + \gamma)$
End If
$pA[k] \leftarrow pA[k] * (Nkj[k, j] + q + 2 * \gamma)$
End for
End for
$p[k] \leftarrow \frac{p1[k] * p2[k]}{pA[k]}$
End for

$k \sim Cat(p)$
For $j \leftarrow 1$ **to** J **do**:
$q \leftarrow Y[d, j]$
$Nkj[k, j] \leftarrow Nkv[k, j] - Q_j$,
$Nkj1[k, j] \leftarrow Nkj1[k, j] - q$, $Nkj0[k, j] \leftarrow Nkj0[d, k] - (Q_j - q)$
End for
If $nn > BN$:
$CC[d, k] \leftarrow CC[d, k] + 1$
End for
End for
End for

B.3 Pseudocode for TSAM (Text-Score Allocation Model)

Algorithm 3 Algorithm for SAM (Score Allocation Model)

Input: K (Number of topics): Integer
Input: BN (Number of burn-in iteratoins): Integer
Input: NN (Number of sampling iteratoins): Integer
Input: X (Document Index): A N-dimensional vector, $X[i] \in \{1, \cdots, D\}, \forall i = 1, \cdots, N$
Input: Y (Score Index): A $D \times J$ matrix, $Y[d, j] \in \{1, \cdots, Q_j\}, \forall d = 1, \cdots, D$
Input: W (Word Index): A N-dimensional vector, $W[i] \in \{1, \cdots, V\}, \forall i = 1, \cdots, N$
Input: Z (Topic Allocation Initial Value): A N-dimensional vector, $Z[i] \in \{1, \cdots, K\}, \forall i = 1, \cdots, N$
Input: C (Topic Allocation Initial Value): A D-dimensional vector, $C[d] \in \{1, \cdots, K\}, \forall d = 1, \cdots, D$
Input: α, β, γ: Hyperparameters
Output: ZZ: A $N \times K$ matrix
Output: CC: A $D \times K$ matrix
Initialisation:
define: ZZ is a $NN \times N$ matrix.

Define: CC is a $D \times K$ matrix.
define: Nkv is a $K \times V$ matrix, Nk is a K-dimensional vector, Ndk is a $D \times K$ matrix.
Define: Nkj is a $K \times J$ matrix, $Nkj1$ is a $K \times J$ matrix, $Nkj0$ is a $K \times J$ matrix
For $i \leftarrow 1$ **to** N **do**:
    $v \leftarrow W[i]$, $k \leftarrow Z[i]$, $d \leftarrow X[i]$
    $Nkv[k,v] \leftarrow Nkv[k,v]+1$, $Nk[k] \leftarrow Nk[k]+1$, $Ndk[d,k] \leftarrow Ndk[d,k]+1$
End for
For $d \leftarrow 1$ **to** D **do**:
    For $j \leftarrow 1$ **to** J **do**:
        $q \leftarrow Y[d,j]$, $k \leftarrow C[d]$
        $Nkj[k,j] \leftarrow Nkv[k,j]+Q_j$,
        $Nkj1[k,j] \leftarrow Nkj1[k,j]+q$, $Nkj0[k,j] \leftarrow Nkj0[d,k]+Q_j-q$
    End for
End for
Collapsed Gibbs Sampling Loop:
define: $p1, p0, pA, p$ is a K-dimensional vector.
For $nn \leftarrow 1$ **to** $BN+NN$ **do**:

    Sampling Z:
    For $i \leftarrow 1$ **to** N **do**:
        $v \leftarrow W[i]$, $k \leftarrow Z[i]$, $d \leftarrow X[i]$
        $Nkv[k,v] \leftarrow Nkv[k,v]-1$, $Nk[k] \leftarrow Nk[k]-1$, $Ndk[d,k] \leftarrow Ndk[d,k]-1$
        For $k \leftarrow 1$ **to** K **do**:
            $p[k] \leftarrow \frac{Nkv[k,v]+\beta}{Nk[k]+V*\beta}(Ndk[d,k]+C[d,k]+\alpha)$
        End for
        $k \sim Cat\,(p)$
        $Nkv[k,v] \leftarrow Nkv[k,v]+1$, $Nk[k] \leftarrow Nk[k]+1$, $Ndk[d,k] \leftarrow Ndk[d,k]+1$
        If $nn > BN$:
            $ZZ[i,k] \leftarrow ZZ[i,k]+1$
        End for
    End for

    Sampling C:
    $p1 \leftarrow \mathbf{0}_K, p0 \leftarrow \mathbf{0}_K, pA \leftarrow \mathbf{0}_K$
    For $d \leftarrow 1$ **to** D **do**:
        For $j \leftarrow 1$ **to** J **do**:
            $q \leftarrow Y[d,j]$, $k \leftarrow C[d]$
            $Nkj[k,j] \leftarrow Nkv[k,j]-Q_j$,
            $Nkj1[k,j] \leftarrow Nkj1[k,j]-q$, $Nkj0[k,j] \leftarrow Nkj0[d,k]-(Q_j-q)$
        End for

        For $k \leftarrow 1$ **to** K **do**:
            For $j \leftarrow 1$ **to** J **do**:
                For $q \leftarrow 0$ **to** Q_j **do**:
                    If $q \leq Y[d,j]$ **do**:
                        $p1[k] \leftarrow p1[k]*(Nkj1[k,j]+q+\gamma)$
                    End If
                    If $q > Y[d,j]$ **do**:
                        $p0[k] \leftarrow p2[k]*(Nkj0[k,j]+Q_j-q+\gamma)$
                    End If
                    $pA[k] \leftarrow pA[k]*(Nkj[k,j]+q+2*\gamma)$
                End for
            End for
            $p[k] \leftarrow \frac{p1[k]*p2[k]}{pA[k]}(Ndk[k]+\alpha)$
        End for

$k \sim Cat(p)$
For $j \leftarrow 1$ **to** J **do**:
$q \leftarrow Y[d, j]$
$Nkj[k, j] \leftarrow Nkv[k, j] - Q_j$,
$Nkj1[k, j] \leftarrow Nkj1[k, j] - q$, $Nkj0[k, j] \leftarrow Nkj0[d, k] - (Q_j - q)$
End for
If $nn > BN$:
$CC[d, k] \leftarrow CC[d, k] + 1$
End for
End for
End for
