Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Sentence-Based Text Analysis for Customer Reviews

Joachim Büschken, Greg M. Allenby

To cite this article:
Joachim Büschken, Greg M. Allenby (2016) Sentence-Based Text Analysis for Customer Reviews. Marketing Science 35(6):953-975. https://doi.org/10.1287/mksc.2016.0993

MARKETING SCIENCE
Vol. 35, No. 6, November–December 2016, pp. 953–975
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

https://doi.org/10.1287/mksc.2016.0993

Sentence-Based Text Analysis for Customer Reviews

Joachim Büschken
Catholic University of Eichstätt–Ingolstadt, 85049 Ingolstadt, Germany, joachim.bueschken@ku.de

Greg M. Allenby
Fisher College of Business, Ohio State University, Columbus, Ohio 43210, allenby.1@osu.edu

Firms collect an increasing amount of consumer feedback in the form of unstructured consumer reviews. These reviews contain text about consumer experiences with products and services that are different from surveys that query consumers for specific information. A challenge in analyzing unstructured consumer reviews is in making sense of the topics that are expressed in the words used to describe these experiences. We propose a new model for text analysis that makes use of the sentence structure contained in the reviews and show that it leads to improved inference and prediction of consumer ratings relative to existing models using data from www.expedia.com and www.we8there.com. Sentence-based topics are found to be more distinguished and coherent than those identified from a word-based analysis.

Data, as supplemental material, are available at https://doi.org/10.1287/mksc.2016.0993.

Keywords: extended LDA model; user-generated content; text data; unstructured data; Bayesian analysis; big data

History: Received: August 26, 2013; accepted: October 8, 2015; Preyas Desai served as the editor-in-chief and Peter Fader served as associate editor for this article. Published online in *Articles in Advance* July 18, 2016.

1. Introduction

One of the challenges in understanding consumers is comprehending the language they use to express themselves. Words are difficult to understand because of their varied meaning among people. The word "data" may mean one thing to an analyst and something else to a teenager. Marketing has a long history of devising ways of cutting through the ambiguous use of words by designing questionnaires and experiments in such a way that questions are widely understood and expressed in simple terms. Qualitative interviews and other forms of pretesting are routinely used to identify the best way to query respondents for useful information.

Despite attempts to make things clear, the analysis of consumer response data continues to be challenged in providing useful insight for marketing analysis. Data collected on fixed-point rating scales, for example, are known to suffer from a multitude of problems such as yea-saying, nay-saying, and scale use tendencies that challenge inference. Moreover, some respondents have the expertise to provide meaningful feedback while others do not, and some provide somewhat independent evaluations about aspects of a product or service, while others tend to halo their responses (Büschken et al. 2013). Respondents are also known to substitute answers to questions different than the one being posed (Gal and Rucker 2011) and exhibit state-dependent responses where item responses carry forward and influence later responses (de Jong et al. 2012). Conjoint analysis is similarly challenged in getting respondents to make choices that mimic marketplace sensitivities (Ding et al. 2005), i.e., to obtain coherent and valid answers to the questions posed.

The growing availability of text data in the form of unstructured consumer reviews provides the opportunity for consumers to express themselves naturally while not being restricted to the design of a survey in the form of preselected items, available response items, and the forced use of rating scales. They simply say whatever they want to say in a manner and order that seems appropriate to them. The challenge in analyzing text data, as mentioned earlier, is in understanding what the words mean. The use of the word "hot" has a different meaning if it is paired with the word "kettle" as opposed to the word "car." As a result, a simple summary of word counts in text data will likely be confusing unless the analysis relates it to the other words that also appear without assuming an independent process of word choice.

The model and analysis presented in this paper is based on a class of models that are generally known as "topic" models (Blei et al. 2003, Rosen-Zvi et al. 2004), where the words contained in a consumer review reflect a latent set of ideas or sentiments,

each of which is expressed with its own vocabulary. A consumer review may provide opinions on different aspects of a product or service, such as its technical features and ease of use, and also on aspects of service and training. The goal of these models is to understand the prevalence of the topics present in the text and to make inferences about the likelihood of the appearance of different words. Words that are likely to appear more often command greater weight in drawing inferences about the latent topic, while the co-occurring words add depth to interpretation.

Topic models provide a simple, yet powerful way to model high-level interaction of words in speech. The meaning of speech arises from the words jointly used in a sentence or paragraph of a document. Meaning can often not be derived from looking at singular words. This is very much evident in consumer reviews where consumers may use the adjective "great" in conjunction with the noun "experience" or "disappointment." When doing so, they may refer to different attributes of a particular product or service.

Empirical analysis of high level interaction of variables present unique challenges. Consider the hotel review data that we use in our empirical analysis (see Section 4). These data consist of 1,011 unique terms. An analysis of all two-level interactions, using this data set, implies to consider up to $1{,}011^2$ or 1.02 million variables. It is immediately clear that an analysis of high-level interaction effects using traditional methods such as regression or factor analysis is very hard to conduct. By comparison, topic models do not require prior specification of interaction effects and are capable of capturing the pertinent co-occurring words up to the dimensionality of the whole vocabulary.

We propose a new variant of the topic model and compare it to existing models using data from online user-generated reviews of hotels and restaurants posted on the Internet. We find that, through a simple model-free analysis of the data, the sentences used in online reviews often pertain to one topic; that is, while a review may be comprised of multiple topics such as location and service, any particular sentence tends to deal with just one. We derive a restricted version of a topic model for predicting consumer reviews that constrains analysis so that each sentence is associated with just one topic, while allowing for the possibility that other sentences can also pertain to the same topic. We find this restriction is statistically supported in the data and leads to more coherent inferences about the hotel and restaurant reviews.

The remainder of this paper is organized as follows. We review alternative topic models and our proposed extension in the next section. In the appendix, we report on a simulation study that demonstrates the ability of our model to uncover the true data generating mechanism. We then present data on hotel reviews taken from www.expedia.com and restaurant reviews from www.we8there.com and examine the ability of our model to predict customer reviews. A comparison to alternative models is provided in Section 5, followed by concluding comments.

2. Topic Models for Customer Reviews

Text-based analysis of user-generated content (UGC) and consumer reviews has attracted considerable attention in the recent marketing literature. Textual consumer reviews have been used for a variety of purposes in marketing research:

- Predicting the impact of consumer reviews on sales using the valence of sentences (Berger et al. 2010)
- Determining the relative importance of reviews in comparison to own experience in the learning process of consumers about products (Zhao et al. 2013)
- Analyzing the change in conversion rates as a result of changes in affective content and linguistic style of online reviews (Ludwig et al. 2013)
- Predicting the sales of a product based on review content and sentiment (Godes and Mayzlin 2004, Dellarocas et al. 2007, Ghose et al. 2012)
- Eliciting product attributes and consumers preferences for attributes (Lee and Bradlow 2011, Archak et al. 2011)
- Deriving market structure (Netzer et al. 2012, Lee and Bradlow 2011)

These papers assume that informative aspects of text data are readily observed and can directly serve as covariates and inputs to other analyses. Typically, word counts and frequencies are used as explanatory variables to identify words that are influential in determining customer behavior or in discriminating among outcomes (e.g., satisfied versus unsatisfied experiences).

Alternatively, one may assume that specific words in UGC are only indicators of latent topics and that these topics are a priori unknown (Tirunillai and Tellis 2014). Latent topics are defined by a collection of words with a relatively high probability of usage and not from the prevalence or significance of single words. This is the key idea of latent topic modeling in latent Dirichlet allocation (LDA) (Blei et al. 2003) and the author–topic model (Rosen-Zvi et al. 2004) and the idea we are following here. Tirunillai and Tellis (2014) apply a variant of the LDA model to UGC to capture latent topics and valence in UGC, to analyze topic importance for various industries over time and utilize the emerging topics for brand positioning and market segmentation.

The goals of our analysis of customer review data are to (i) identify latent topics in customer reviews

and assess their predictive performance of satisfaction and (ii) contrast alternative methods of creating inferences about the latent topics. Issues present in both questions are whether simple word choice probabilities are sufficient for establishing meaning in the evaluations and the degree to which topics provide richer insights through the co-occurrence of words in a review.

2.1. Latent Dirichlet Allocation Model

A simple model for the analysis of latent topics in text data is the LDA model (Blei et al. 2003). The LDA model assumes the existence of a fixed number of latent topics that appear across multiple documents, or reviews. Each document is characterized by its own mixture of topics (θ_d), and each topic is characterized by a discrete probability distribution over words; that is, the probability that a specific word is present in a text document depends on the presence of a latent topic. It is convenient to think of a dictionary of words that pertain to all reviews, with each topic defined by a unique probability vector of potential word use. Words with high probability are used to characterize the latent topics.

The nth word appearing in review d, w_{dn}, is thought to be generated by the following process in the LDA model:

1. Choose a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$.
2. Choose a word $w_{dn} \sim$ from $p(w_{dn} \mid z_{dn}, \Phi)$.

In the model, θ_d is a document-specific probability vector associated with the topics z_{dn}, and Φ is a matrix of word-topic probabilities $\{\phi_{m,t}\}$ for word m and topic t, with $p(w_{dn} = m \mid z_{dn} = t, \Phi) = p(w_{dn} = m \mid \phi_t)$. The vector of word probabilities for topic t is thus ϕ_t.

Topics $\{z_{dn}\}$ and words $\{w_{dn}\}$ are viewed as discrete random variables in the LDA model, and both are modeled using a multinomial, or discrete, distribution. The objects of inference are the parameters $\{\theta_d\}$ and Φ that indicate the probabilities of the topics for each document d and associated words for each topic t. A model involving T topics has $\dim(\theta_d) = T$, and Φ is an $M \times T$ matrix of probabilities for the M unique words that appear in the collection of customer reviews. The first element of θ_d is the probability of the first topic in document d, and the first column of Φ is the word probability vector ϕ_1 of length M for this first topic.

The potential advantage of the LDA model is its ability to collect words together that reflect topics of potential interest to marketers. Co-occurring words appearing within a document indicate the presence of a latent topic. These topics introduce a set of word interactions into an analysis so that words with high topic probabilities (ϕ_t) are jointly predicted to be present. Since different topics are associated with different word probabilities, the topics offer a parsimonious way of introducing interaction terms into text analysis. Moreover, the LDA model is not overly restrictive in that it allows each document, or customer review, to be characterized by its own set of topic probabilities (θ_d).

We complete the specification of the standard LDA model by assuming a homogeneous Dirichlet prior for θ_d and ϕ_t

$$p(\theta_d) \sim \text{Dirichlet}(\alpha),$$
$$p(\phi_t) \sim \text{Dirichlet}(\beta).$$

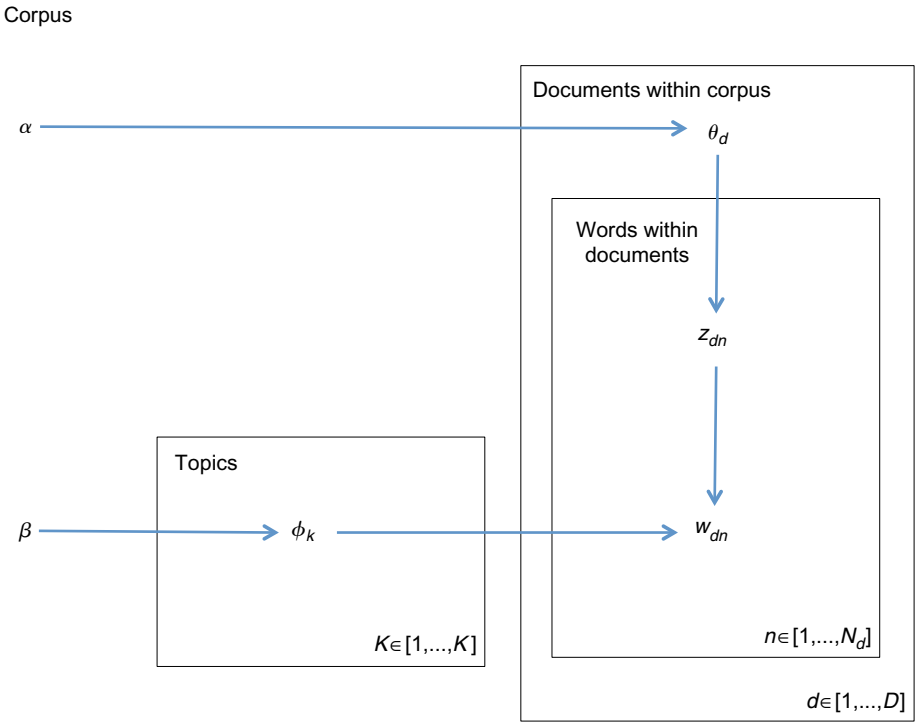
Figure 1 displays a plate diagram for the LDA model. The plates indicate replications of documents ($d = 1, \ldots, D$), words ($n = 1, \ldots, N_d$), and topics ($t = 1, \ldots, T$). We note that the LDA model does not impose any structure on the data related to the plates; i.e., it assumes that the latent topics z_{dn} can vary from word to word, sometimes referred to as a "bag-of-words" assumption in the text analysis literature. This assumption differs from the traditional marketing assumption of heterogeneity that exploits the panel structure often found in marketing data where multiple observations are known to be associated with the same unit of analysis. There is a marketing literature on what is known as context-dependent or structural heterogeneity (Kamakura et al. 1996, Yang and Allenby 2000, Yang et al. 2002) that allows the model likelihood to vary across observations. Restricted versions of the assumption made by the LDA model for observational heterogeneity include models of change points (DeSarbo et al. 2004) and latent Markov models (Fader et al. 2004, Netzer et al. 2008, Montoya et al. 2010).

2.2. Sentence-Constrained LDA Model

We find in the analysis of our data presented below that it is beneficial to constrain the LDA model so that words within a sentence pertain to the same topic. People tend to change topics across sentences, but not within a sentence. The LDA model assumes that the words within a document provide exchangeable information regarding the latent topics of interest, and we note that the data index (n) is simply an index for the word; i.e., it is not related to the authors or the reviews themselves. Our sentence-constrained model moves away from this bag-of-words assumption.

Figure 2 displays a plate diagram for our proposed sentence-constrained LDA (SC-LDA) model. A replication plate is added to distinguish the sentences within a review from the words within each sentence. Additional indexing notation is introduced into the model to keep track of the words (n) contained within the sentences (s) within each review (d), w_{dsn}. The latent topic variable z_{ds} is assumed to be the same for all words within the sentence and is displayed outside of the word plate in Figure 2. We assume that

Figure 1 **(Color online) Graphical Representation of the LDA Model**



the number of sentences in a document (S_d) and the number of words per sentence (N_{ds}) are determined independently from the topic probabilities (θ_d).

The probability of topic assignment changes because all words within a sentence are used to draw the latent topic assignment, z_{ds}. This requires the estimation algorithm to keep track of the topic assignments by sentence, C_{mt}^{SWT}, as well as the number of words in each sentence, n_{ds}. The appendix describes the estimation algorithm for the SC-LDA model.

The LDA model has been extended in a variety of ways in the statistics literature, by

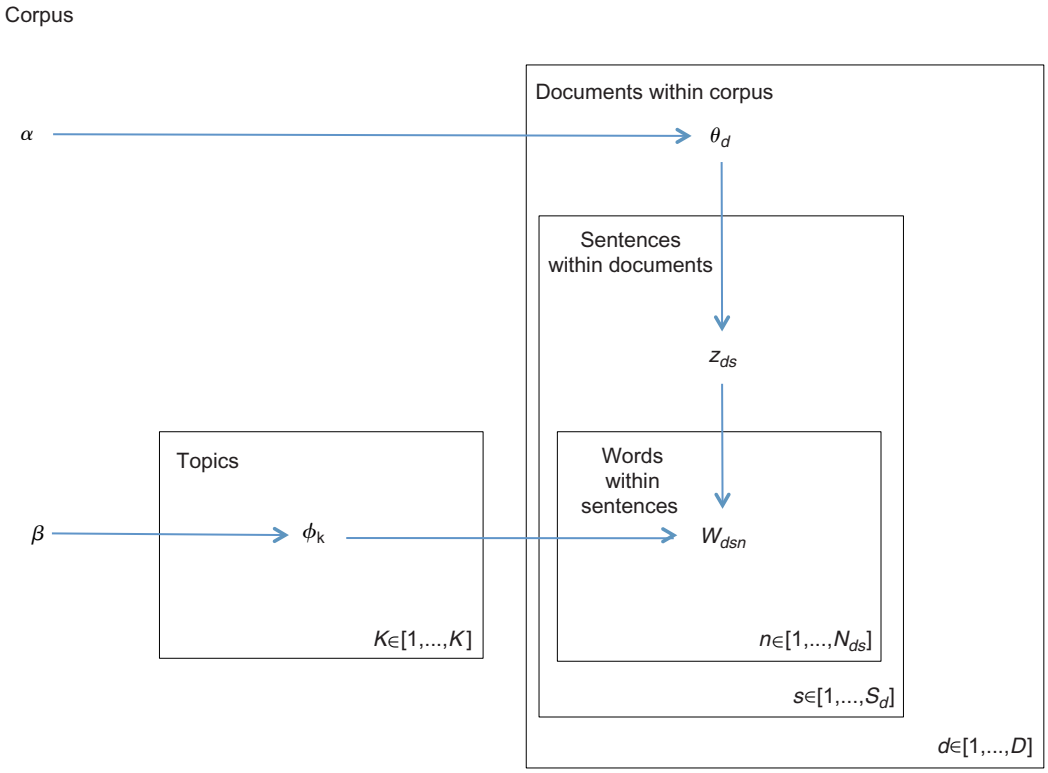
- introducing author information (Rosen-Zvi et al. 2004) that allows information to be shared across multiple documents by the same author,
- introducing latent labels for documents (Ramage et al. 2010) that allow for unobserved associations of documents,
- incorporating a dynamic topic structure by modeling documents from different periods (Blei and Lafferty 2006) or assuming that topic assignments are conditional on the previous word (Wallach 2006) or topic (Gruber et al. 2007),
- developing multiple topic layers (Titov and McDonald 2008) where words in a document may stem either from a document-specific global topic or from the content of the words in the vicinity of a focal word,
- incorporating the sender–recipient structure of written communication into topic models (McCallum et al. 2005; in the author–recipient topic model, both the sender and the recipient determine the topic assignment of a word), and
- incorporating informative word-topic probabilities consistent with domain knowledge through the prior distribution (Andrzejewski et al. 2009).

Our analysis of text data is designed to uncover latent topics associated with user-generated topics and relate them to product ratings. In marketing, the amount of text available for analysis per review is limited, often having less than 20 words, and multiple reviews for the same author are rare. We therefore do not attempt to develop the LDA model by making it dynamic, having multiple layers of topics, or constraining the prior to reflect prior notions of topics. Instead, we relate user reviews to the topic probabilities with a latent regression model.

2.3. Sentence-Constrained LDA Model with Ratings Data

We extend the SC-LDA model with a cut-point model (Rossi et al. 2001, Büschken et al. 2013) to relate the topic probabilities to the ratings data. The advantage of employing a topic model is the ability to collect co-occurring words together as topics, which improves the interpretation of text data. Relating the latent topic probabilities to ratings data is similar to traditional driver analysis, but with UGC that is not constrained to a set of prespecified drivers. Our model offers an alternative to models of ratings data that

Figure 2 **(Color online) Graphical Representation of the Sentence-Constrained LDA Model**



use subscales or single words represented by dummy variables.

A cut-point model relates responses on a fixed-point rating scale to a continuous latent variable and a set of cut points,

$$r_d = k \quad \text{if } c_{k-1} \le \tau_d \le c_k$$

and

$$\tau_d \sim N(\theta_d'\beta, \sigma),$$

where the cut points $\{c_k\}$ provide a mechanism for viewing the discrete rating as a censored realization of the latent continuous variable (τ_d) that is related to the topic probabilities (θ_d) through a regression model. Our regression model is similar to a factor model where β are the factor loadings and θ_d are the factor scores.

The plate diagram for the SC-LDA model with ratings data (SC-LDA-Rating) is provided in Figure 3. Our cut-point model is a simplified (i.e., homogenous) version of the model used by Ying et al. (2006)

$$\begin{aligned} c &= (c_1, c_2, \ldots, c_{K-1}) \\ &= \left(c_1 c_1, +\delta_1, c_1 + \sum_{k=1}^{2} \delta_k, \ldots, c_1 + \sum_{k=1}^{K-2} \delta_k\right), \end{aligned}$$

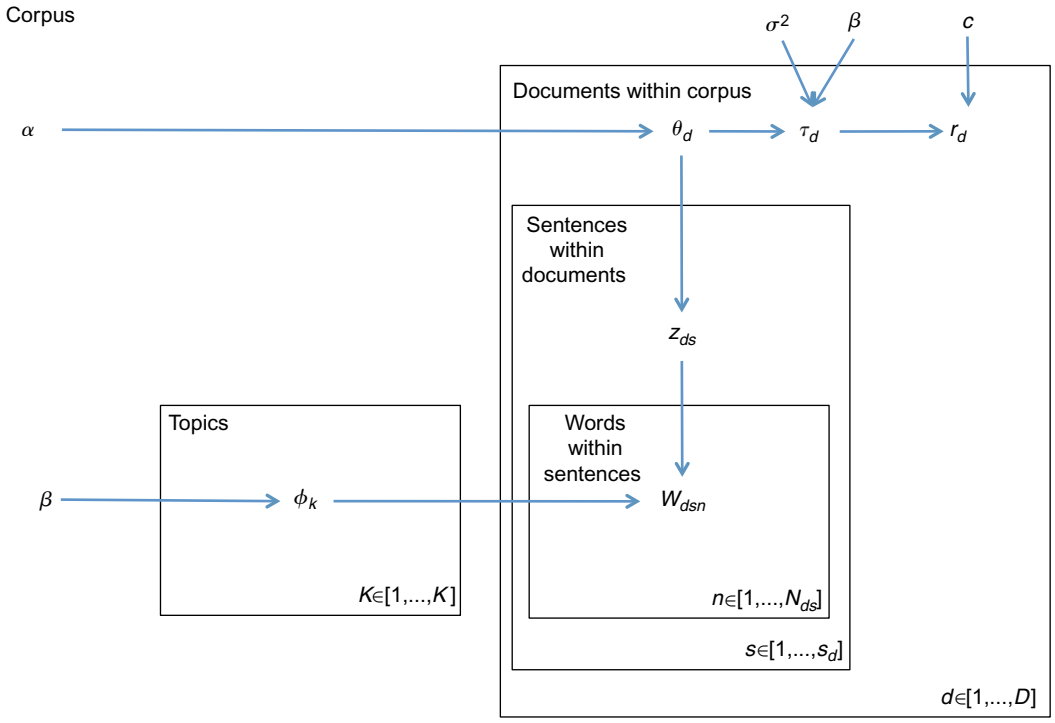
where cut points c_0 and c_K are $-\infty$ and ∞, respectively, and the δ are strictly positive cut-point increments. Constraints are needed to identify the SC-LDA-Rating model. For K points in the rating scale, there are traditionally $K-1$ free cutoffs if we set $c_0 = -\infty$ and $c_K = +\infty$. Two additional cutoff constraints are needed in our analysis because the regression model is specified with an intercept and error scale, and shifting all of the cutoffs by a constant or scaling all of the cutoffs is redundant with these parameters.

We also note that the topic probabilities for each document, θ_d, are constrained to sum to one, and as a result the likelihood for the latent regression model is not statistically identified without additional constraints. As discussed in the appendix, we postprocess the draws from the Markov Chain Monte Carlo (MCMC) chain, arbitrarily picking one of the topics to form a contrast for the remaining topics (see Rossi et al. 2005, Chapter 4). Postprocessing the draws results in inferences based on a statistically identified likelihood. Our proposed model and estimation strategy is discussed in more detail in the appendix.

2.4. SC-LDA Model with Sticky Topics

Figure 4 displays a hotel review. The color coding in the display is present to identify different potential topics, which are seen to change across sentences but not within sentences. For example, sentences describing "breakfast" are coded green, and sentences describing the "general experience" are coded blue.

Figure 3 **(Color online) Graphical Representation of the SC-LDA-Rating Model**



We note that in this review topics exhibit stickiness in the sense that the reviewer repeatedly stays with one topic over a number of consecutive sentences. Topic stickiness presents a potential violation of the assumption of independent and identically distributed (i.i.d.) topic assignments in the LDA model and its variants.

To account for sticky topics, we consider an extension of the SC-LDA-Rating model in which the topic z_{n-1}, assigned to sentence s_{n-1}, can exhibit carryover to s_n. Stickiness for the purpose of this model is defined as $z_n = z_{n-1}$. To develop this model, we consider a latent binary variable ζ_n that indicates whether the topic assignment to sentence s_n is sticky

$$\begin{aligned} \zeta_n = 1\text{: } z_n &= z_{n-1}, \\ \zeta_n = 0\text{: } z_n &\sim \text{Multinomial}(\theta_d). \end{aligned} \tag{1}$$

In the SC-LDA model, $\zeta_n = 0\ \forall\, n$, which implies that the SC-LDA with sticky topics can be thought of as a general case of the SC-LDA. We assume ζ to be distributed Binomial with a topic-specific probability ψ_t

$$\zeta_n \sim \text{Binomial}(\psi_t). \tag{2}$$

Figure 5 displays an example of a DAG (Directed Acyclic Graph) for the sticky topic model, given five consecutive sentences in a review. In the upper panel of Figure 5, we consider the general case of ζ being unknown. In the lower panel of Figure 5, we consider the case of a particular ζ-sequence that reveals sticky and nonsticky topics. In both versions of the DAG, we omit all fixed priors and the assignment of words to the sentences for better readability. In the lower panel of Figure 5, for cases of $\zeta_n = 1$, relationships between z and ζ are omitted, and the resulting (deterministic) relationships between z_n and z_{n-1} are added to the graph, indicating first-order dependency of the topic assignments. As the DAG in the lower panel of Figure 5 shows, a value of $\zeta_n = 1$ shuts off the relationship between z_n and θ_d and establishes a relationship between z_n and z_{n-1} so that $z_n = z_{n-1}$. This also implies that "observed" topic switches ($z_n \neq z_{n-1}$) are indicative of a topic draw from θ_d. Note that in Figure 5 we omitted ζ_1 for the first sentence because topic assignments do not carry over between documents. Thus, we fix $\zeta_1 = 0$ and assume z_1 to be generated by θ_d, as no prior topic assignment exists.

We relate the stickiness of topics to the number of sentences in a review through a regression model

$$\psi_{d,t} = \frac{e^{X_d \gamma_t}}{1 + e^{X_d \gamma_t}}, \tag{3}$$

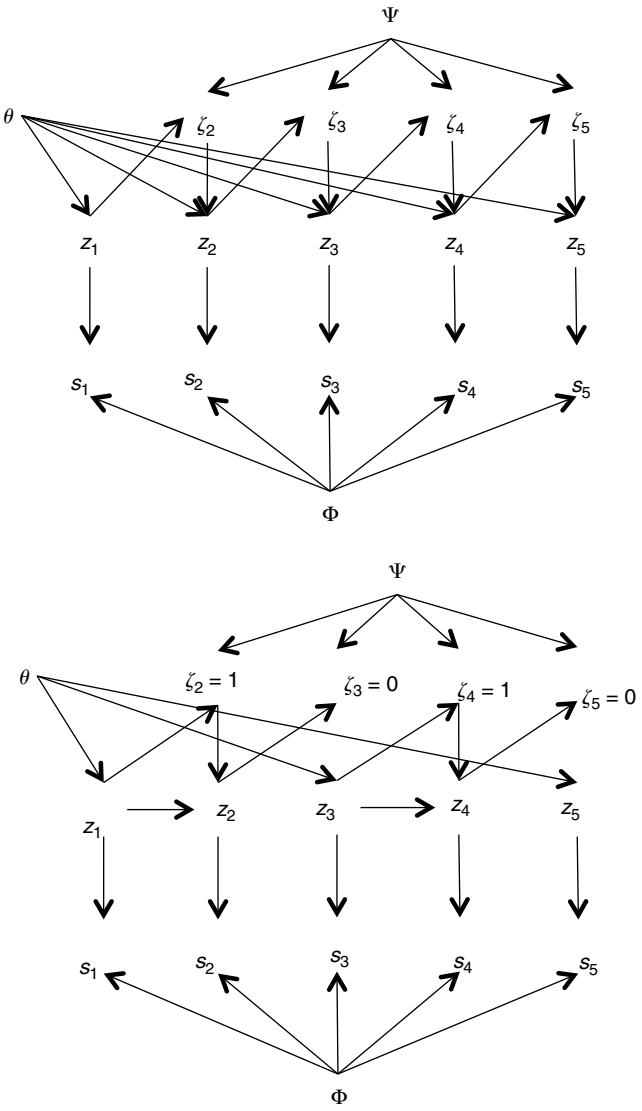
where covariate vector X_d consists of a baseline and the observed number of sentences in each review, and γ_t is a vector of topic-specific regression coefficients. A priori, it seems reasonable to assume that, as reviews contain more sentences, topics have a tendency to be carried over to the next sentence (see an example in Figure 4). The approach in Equation (3) allows for heterogeneity among reviews with respect to topic stickiness. In the appendix, we outline the estimation details for the SC-LDA model with sticky

Figure 4 **(Color online) A Hotel Review**

"The hotel was really nice and clean. It was also very quiet. There was a thermostat in each room so you can control the coolness. The bathroom was larger than in most hotels. The breakfast was sausage and scrambled eggs, or waffles you make yourself on a waffle iron. All types of juice, coffee, and cereal available. The breakfast was hot and very good at no extra charge. The only problem was the parking for the car. The parking garage is over a block away. It is $15.00 per day. You don't want to take the car out much because you can't find a place to park in the city, unless it is in a parking garage. The best form of travel is walking, bus, tour bus, or taxi for the traveler. The hotel is near most of the historic things you want to see anyway. I would return to this hotel and would recommend it highly."

Note. Potential sentence topics are highlighted in color.

Figure 5 **Graphical Representation of the SC-LDA Model with Sticky Topics**



topics and demonstrate statistical identification using a simulation study.

We also address in the appendix, through simulation, the question of whether a standard LDA model, which assumes that topics are assigned to words and not sentences, is able to recover topic probabilities that are sentence based. We find that the standard LDA model exhibits low recovery rates of the true topic assignments when topics are characterized by larger sets of co-occurring terms and longer sentences (i.e., more words). Only when topics are associated with a few unique terms and when sentences contain a few words will using the LDA model yield results similar to that of the SC-LDA model.

An assumption common to all LDA-based models analyzed in this research, including the sentence-constrained topic model and the model with sticky topics, is independence of the latent topics to the number of words and sentences in a review. In effect, we treat these observed quantities as independently determined and uninformative with respect to topics. Because the topic probabilities (and topic assignments) are latent in our model, this can only be ascertained by building a new model that allows for a dependency and comparing its fit relative to the fit of our proposed model. We note that while we postulate a priori that θ_d (or z_{ds}) is independent of S_d (or N_{ds}), this does not imply that they are a posteriori independent, given the data. We investigated this issue and found the average (absolute) correlation of the topic shares to the number of sentences across topics and data sets to be 0.08 (standard deviation (SD) = 0.08). The maximum (absolute) correlation of θ_d to S_d for one of the topics from any data set is 0.3. We find the same result for the correlation of θ_d to the number of words in a review. Additionally, conditional on the topic assignment of sentences (z_{ds}), we find that the topics are very similar with respect to the number of words generated for each sentence across all data sets. In conclusion, we do not believe our assumption of independency to be a significant issue for our data.

3. Empirical Analysis

This section presents results from applying the LDA, SC-LDA, and sticky SC-LDA models to consumer review data. Since ratings are available in all data sets, we only use topic models that incorporate the rating as a function of θ. We employ three data sets for comparison purposes: reviews of Italian restaurants from the website www.we8there.com and two sets of reviews from www.expedia.com pertaining to upscale hotels in Manhattan and hotels near John F. Kennedy

(JFK) International Airport. We find that our proposed SC-LDA-Rating model more accurately predicts consumer ratings than other models and is shown to lead to more coherent inferences about the latent topics. Characteristics of the data and preprocessing are discussed first, followed by a model comparison of in-sample and predictive fit.

Prior to data analysis, reviews were preprocessed using the following sequence of steps:

1. Splitting text into sentences identified through ".", ",", "!", or "?"; after the sentence split, all punctuation is removed
2. Substituting capital letters with lower-case letters
3. Removing all terms which appear in less than 1% of the reviews of a data set (i.e., "rare words")
4. Removing stop words using a standard vocabulary of stop words in the English language

The removal of rare words is motivated by the search for co-occurring terms ("topics"). Rare words make little to no contribution to the identification of such topics because of their rarity. Similarly, the removal of stop words is motivated by their lack of discriminatory power with respect to topics as all topics typically contain such words.

Stemming is absent from our preprocessing procedure. This is because words sharing the same stem may have different meaning. Consider, for example, the words "accommodating" and "accommodation," which share the stem "accommod." The word "accommodating" is mostly used to describe aspects of a service process or interaction with service personnel. The term "accommodation" is often used in the context of a hotel stay and typically refers to amenities of a hotel room and does not refer to interactions with service personnel. Thus, stemming may eliminate differences in meaning which, for identification and interpretation of latent topics, is not desirable.

3.1. Data

We obtained 696 reviews of Italian restaurants comprising a corpus of 43,685 words. The vocabulary of this data set consists of $W = 1{,}312$ unique terms (after preprocessing). For the analysis of hotels, we consider hotels located in downtown New York (Manhattan) and hotels within a two-mile radius of JFK airport. We obtained 3,212 reviews of Manhattan upscale hotels and 1,255 reviews of midscale hotels near JFK airport. The corpora of Manhattan hotel reviews and JFK hotel reviews comprise 73,314 and 25,970 words, respectively. Both hotel data sets are based on a working vocabulary of $W = 1{,}011$ words. The hotel and restaurant data sets contain an overall evaluation of the service experience on a five-point rating scale, where a higher rating indicates a better experience.

Table 1 provides numerical summary statistics of the preprocessed data based on word and sentence counts. On average, upscale hotel reviews contain 4.3 sentences with 5.3 words per sentence. The standard deviation of the number of sentences per review is 3.4, indicating significant heterogeneity among reviews with regard to the amount of information contained therein. Midscale hotel reviews contain a similar number of sentences (3.8) on average, with an average of 5.4 words per sentence. The Italian restaurant reviews contain an average of 12.2 sentences, each of which contain, on average, 5.2 words. The range of the number of sentences is 90, significantly higher than in the hotel data sets. Thus, restaurant reviews are longer and significantly more heterogeneous with respect to sentence count. It appears that restaurant reviewers feel the need to inform readers about restaurants in a more detailed fashion.

Reviews provided by both hotel and restaurant customers typically exhibit a sentence structure, although such a structure is not required; that is, reviewers voluntarily organize their reviews by using periods and capital letters to structure text. For example, Expedia accepts content in many forms, and some reviews exhibit a structure more compatible with a bag-of-words assumption. However, such a free structure is apparently not the norm. On average, hotel reviewers use about 4 sentences, which indicates their desire to differentiate statements within a review. The standard deviation of the number of sentences is about 3 across the segments, pointing at heterogeneity of structure. The Italian restaurant reviews in our data contain an average of 12 sentences, with a standard deviation of 11.

Table 1 reveals that the Manhattan hotels received an average rating of 4.4. The standard deviation of the rating of 0.9 indicates that many customers rated their experience at the top of the scale (share of 61.3%). Very few customers (4.5%) rated their experience toward the bottom of the scale (1 or 2). This is different for the airport hotels, which, on average, received a lower rating of 3.8 and where a larger share of customers (17.4%) rated their experience as bad (rating of 1 or 2). Italian restaurants received an average rating of 3.8. Thirty-two percent of the reviewers rated their experience as bad (1 or 2 stars). Forty-seven percent chose the best rating possible. Apparently, restaurant reviews are particularly useful to identify critical issues best avoided, and Manhattan hotel reviews are more informative about positive drivers of customers' experiences. Whereas restaurant reviews contain a lot of information (by word and sentence count), the challenge in hotel reviews is to extract managerially relevant information from less data per review.

We begin our analysis of the text by providing a simple summary of words appearing by rating for

Table 1 Summary Statistics

	Mean	Median	Standard deviation	Range
Number of sentences per review				
Midscale hotel	3.8	3	2.9	25
Upscale hotel	4.3	4	3.2	41
Italian restaurant	12.2	8	11.8	90
Number of words per sentence				
Midscale hotel	5.4	5	3.6	42
Upscale hotel	5.3	5	3.4	52
Italian restaurant	5.2	5	3.1	29
Rating				
Midscale hotel	3.5	4	1.1	4
Upscale hotel	4.4	5	0.9	4
Italian restaurant	3.8	4	1.4	4

Table 2 Most Frequently Used Words by Rating in Reviews

	Upscale hotel		Midscale hotel		Italian restaurant	
Rank	Rating 1 or 2	Rating 4 or 5	Rating 1 or 2	Rating 4 or 5	Rating 1 or 2	Rating 4 or 5
1	Room	Hotel	Hotel	Hotel	Pizza	Pizza
2	Hotel	Room	Room	Room	Food	Food
3	Location	Great	Airport	Airport	Good	Good
4	Rooms	Location	Stay	Shuttle	Restaurant	Great
5	Stay	Staff	Breakfast	Breakfast	Just	Restaurant
6	Good	Square	Shuttle	Good	One	One
7	Staff	Stay	Good	Clean	Us	Place
8	Service	Times	JFK	Stay	Back	Italian
9	Great	Clean	Staff	JFK	Like	Just
10	Times	New	One	Staff	Place	Like
11	Time	Time	Night	Service	Ordered	Best
12	One	Nice	Small	Free	Really	Cheese
13	Bed	Rooms	Clean	Nice	Got	Service
14	Nice	York	Rooms	Great	Came	Sauce
15	Square	Friendly	Get	Comfortable	Order	Time
16	Get	Good	Place	Night	Italian	Really
17	Us	Helpful	Free	Helpful	Cheese	Will
18	Breakfast	Comfortable	Flight	Flight	Get	Also
19	Floor	City	Close	Close	Menu	Us
20	Small	View	Service	Friendly	Minutes	Little
21	Desk	Service	Area	One	Time	Go
22	Night	Breakfast	Time	Rooms	Service	Get
23	Bathroom	Excellent	Like	Time	Go	Back
24	2	Right	Bed	Early	Said	Menu
25	Clean	Close	Next	Get	Will	Can
26	Like	Will	Us	Small	Sauce	Crust
27	Didn't	Everything	Desk	Hour	Two	Got
28	Front	One	Hour	Convenient	Salad	Two
29	New	Stayed	Location	Morning	Table	Order
30	Also	Us	Morning	Us	Eat	Made

the hotel and restaurant reviews, given the preprocessed data sets. A rating of four or five on overall satisfaction indicates satisfaction with the hotel stay or restaurant visit, whereas a rating of one or two indicates dissatisfaction. Table 2 provides a list of the top 30 words occurring in good and bad overall evaluations for the hotel and restaurant data described in Table 1.

Both good and bad upscale hotel evaluations are associated with adjectives "great," "good," "nice," and "clean." Frequent nouns in both categories are "location," "staff" "service," and "room(s)." Bad upscale reviews are uniquely associated with the adjective "small" and the nouns "bathroom" and "bed," indicating possible reasons for a bad experience. Good upscale reviews are uniquely associated with the terms "excellent" and "everything." Neither of these terms point at possible reasons for the good experience. Frequent words in midscale hotel reviews contain terms exclusive to the review selection; that is, terms such as "airport," "JFK," and "shuttle" are unique to the location of the hotels selected here. However, similar to upscale hotel reviews, we find that the vocabulary differs little with respect to ratings. The sets of the top 10 words in good and bad reviews are identical except for the term "one" in bad reviews (rank 21 in good reviews). Frequent words in both good and bad restaurant reviews include "pizza," "good," and "food," which indicates that these terms cannot discriminate ratings. In general, a simple listing of frequently observed words in good and bad reviews does not help much to discriminate good from bad ratings.

A problem with the simple analysis of word frequencies is that it is limited to the marginal analysis of predefined groups. The analysis of word counts or frequencies by rating or other observed variables is informative only of these individual classification variables. It does not identify the combinations of classification variables that lead to unique themes or topics for analysis. The reason for employing model-based analysis of the data is that it helps to reveal the combination of classification variables for which unique themes and points of differentiation are present.

3.2. Model Fit

Table 3 summarizes the in-sample fit and predictive fit of the topic models applied to the three data sets. In the empirical analysis, we only use topic models that incorporate a customer's rating information in the model estimation. Table 3 reports the log-marginal density (LMD) of the data for different models. The fit statistics are averaged over the number of topics to save space. For each model and data set, we estimate $T \in \{2{:}\ 20\}$ and find a consistent ordering of the fit statistic for the in-sample and predictive fit. We use 90% of the available data for calibration and the remaining 10% for out-of-sample prediction based on a random split of the reviews.

Table 3 reveals that, in terms of predictive fit, a topic model with a sentence constraint is generally preferred over a model that assigns topics to words. This is evidenced by the predictive fit of the LDA

Table 3 Model Fit of Topic Rating Models

Data	Model	In-sample fit	Predictive fit
Italian restaurant	LDA	−214,344.3	−27,001.7
	SC-LDA	−235,587.5	−26,458.3
	Sticky SC-LDA	−236,972.3	−26,515.3
Upscale hotel	LDA	−328,963.7	−42,675.6
	SC-LDA	−361,173.2	−41,236.6
	Sticky SC-LDA	−363,216.8	−41,649.9
Midscale hotel	LDA	−111,289.7	−17,563.9
	SC-LDA	−124,440.7	−16,970.7
	Sticky SC-LDA	−126,229.5	−17,147.8

rating model to be lower than the predictive fit of both SC-LDA-based topic models. Within the sample, however, the LDA fits better across all data sets, compared to both the SC-LDA model and the SC-LDA model with sticky topics. This result is due to the LDA model being more flexible, but this flexibility apparently does not help in predicting new data.

Table 3 also shows that the SC-LDA model with i.i.d. topic assignments performs consistently better than the SC-LDA model with sticky topics. This result is independent of using in-sample or out-of-sample fit as the fit measure. However, the difference in fit is relatively small for all data sets. For example, for the Italian restaurant data, the in-sample log marginal density of the data, using the SC-LDA model, is −235,586, compared to −236,972 for the SC-LDA model with sticky topics. The difference in out-of-sample fit is similarly small (LMD of −26,458 compared to −26,515) for this data set. We find that the SC-LDA model with sticky topics rarely points at topics being very "sticky." In fact, we very rarely observe values for ψ_t larger than 0.20 for any topic in all three data sets. The average ψ_t across data sets and topic numbers is less than 0.03, implying that the SC-LDA model with sticky topics becomes equivalent to the SC-LDA model in many cases. This also implies that stickiness of topics across consecutive sentences is not an important feature of the customer review data sets analyzed here.

Further analysis of the results indicates that the sentence constraint reduces the likelihood of observing frequent words and increases the likelihood of infrequently occurring words within topics. To illustrate, we consider results from the Expedia midscale hotel data. Figure 6 plots Φ_t for the sentence constrained and unconstrained LDA model and for each topic, ordered by their word choice probabilities. Note that for all topics and models, the area under the curve of Φ_t must integrate to one. The left panels in Figure 6 show Φ_t for the top 200 ranked words, and the right panels in Figure 6 show Φ_t for the lower ranked words (ranks 201 to 1,000) in the topics.

Figure 6 reveals that the sentence constraint leads to smaller probabilities for the most likely words than the unconstrained model, and higher probabilities for words that are less likely. This result is independent of the topics. It suggests that the SC-LDA model penalizes the most likely words compared with the LDA model by assigning relatively lower probabilities to these words. In comparison, the sentence constraint favors less frequent words. The reason for the penalty on frequent terms is that the sentence constraint assigns topics on the basis of context, where context is provided by the words appearing together in a sentence. The reductions in in-sample fit reported above are influenced by the tendency of the sentence constraint to assign less extreme word-choice probabilities to the terms compared to the unconstrained topic models.

Table 4 Pseudo-R^2 from Rating-Based Topic Models

Model	Midscale hotel	T	Upscale hotel	T	Italian restaurant	T
LDA-Rating	0.581	8	0.488	10	0.492	9
SC-LDA-Rating	0.719	8	0.663	10	0.649	9
Sticky SC-LDA	0.646	8	0.634	10	0.625	9

The fit of the rating-based topic models can also be evaluated on the basis of the explanatory power with respect to the satisfaction rating. Table 4 compares the share of variance of the latent continuous evaluation τ explained by the three topic models for the three data sets. The fit measure presented is the share of variance of τ explained by the covariates. We report the posterior mean and the posterior SD of this pseudo-R^2 and the number of topics (T) for the best-fitting model. The results in Table 4 imply that the sentence constraint leads to improved explanatory power of the latent topics with respect to the satisfaction rating in all three data sets. The improvement ranges from 10% (restaurant data) to 36% (upscale hotel).

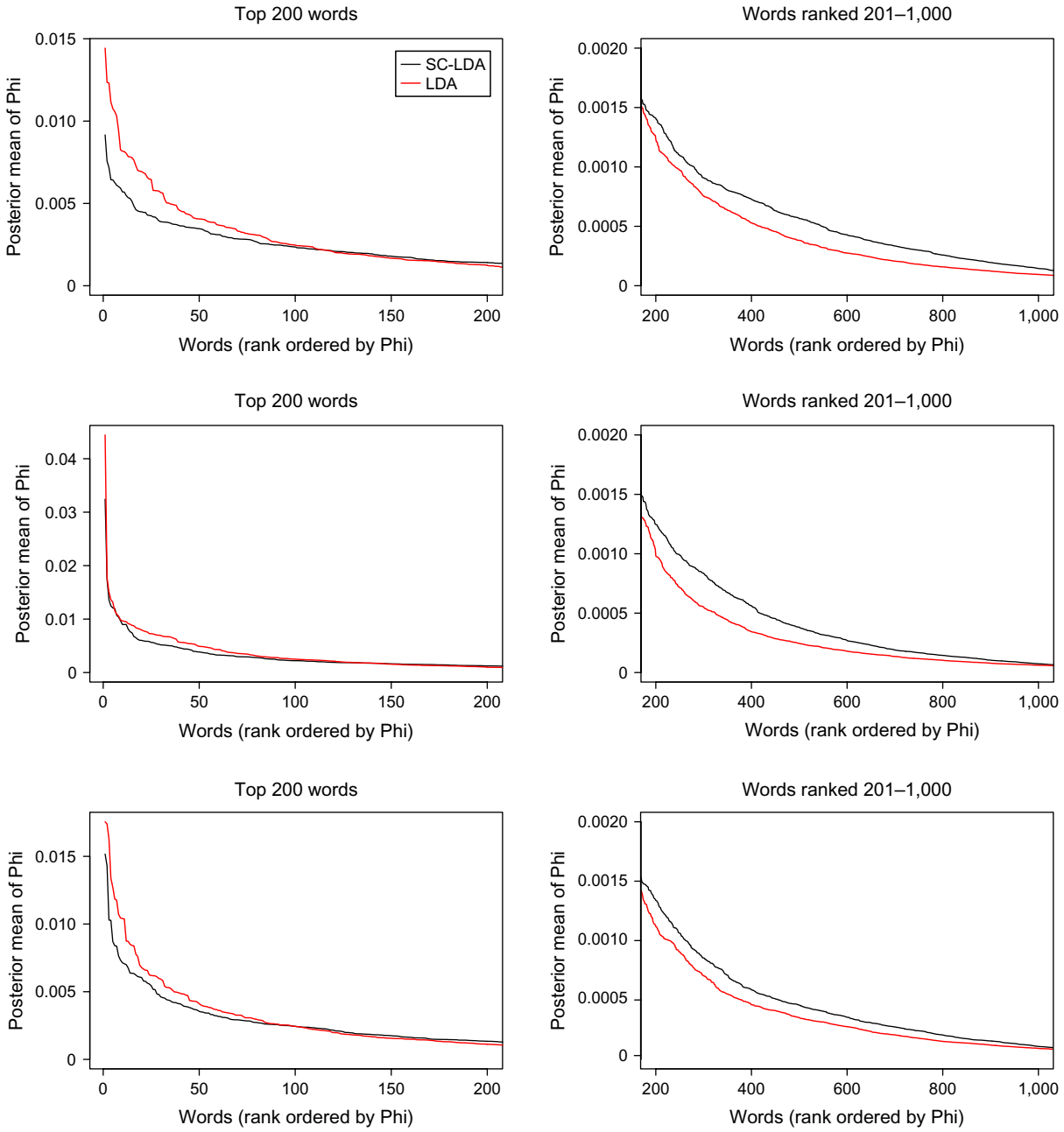
4. Predicting Customer Ratings

We investigate use of the latent topics to predict and explain consumer ratings of hotels and restaurants. The goal of our analysis is to identify themes associated with positive and negative reviews, comparing results from the model-free analysis reported in Table 2 to topics in the SC-LDA-Rating model. This information is useful for improving products and services by identifying potential drivers of customer satisfaction. For all subsequent analyses, we use the SC-LDA-Rating model with a number of topics that maximizes predictive fit.

4.1. Italian Restaurants

Table 5 displays the top 30 words associated with the best-fitting SC-LDA-Rating model for the Italian restaurant data set ($T = 9$). Summary descriptions of the topics are offered at the top of Table 5. We find

Figure 6 (Color online) Word Choice Probabilities (Φ) of LDA and SC-LDA Models for $T = 3$ (Midscale Hotel Data)



Note. From top to bottom, results show ϕ_t for $t = 1$, $t = 2$, and $t = 3$.

that in this data set, and in the other two data sets, the words for each topic provide a more coherent description of the product than that provided by the most frequently used words list in Table 2. Topic 1, for example, is a description of "real pizza," as evidenced by the use of words such as "crust," "thin," "Chicago," "style," "New," and "York." Topic 3 is a collection of words associated with customers' willingness to return to the restaurant ("will," "go," "back"). Topic 5 talks about service and staff in a positive fashion. Most adjectives in this topic have positive valence ("friendly," "attentive," "wonderful," "nice"). Topics 8 and 9 describe aspects of a negative service experience. Topic 8 is concerned with various issues with customers' orders. Topic 9 talks about issues regarding time ("minutes," "time," "wait," "never") and (frustrating) interaction with personnel ("asked," "came," "told"). Interestingly, topic 9 also contains the words "owner" and "manager," indicating that customers asked to talk to them. Ordinarily, restaurant patrons only do so as a last resort to resolve escalated conflicts with service personnel.

Table 5 We8There Italian Restaurant Data, Top 30 Words from the SC-LDA-Rating Model ($T = 9$)

Rank	Topic 1 "Real pizza"	Topic 2 "Menu"	Topic 3 "Return"	Topic 4 "Food ordered"	Topic 5 "Service and staff"	Topic 6 "Recommend"	Topic 7 "Layout"	Topic 8 "Issues with order"	Topic 9 "Conflict"
1	Pizza	Salad	Will	Sauce	Food	Food	Restaurant	Pizza	Us
2	Crust	Good	Restaurant	Cheese	Service	Italian	Dining	Just	Minutes
3	Really	Menu	Go	Pizza	Great	Restaurant	Room	Got	Food
4	Like	Ordered	Back	Ordered	Good	Best	Bar	Two	Order
5	Good	Also	Place	Fresh	Friendly	Place	Tables	Good	Came
6	Chicago	Pasta	Time	Bread	Staff	Recommend	Area	Order	Table
7	Thin	Bread	Food	Italian	Atmosphere	Pizza	Located	Really	Asked
8	Style	Food	Try	Sandwich	Excellent	Great	One	Cheese	Back
9	Best	Pizza	Pizza	Like	Place	One	Small	Get	Waitress
10	One	Wine	One	Good	Restaurant	Ever	Pizza	Back	Time
11	Just	Italian	Good	Came	Prices	Experience	Parking	One	Restaurant
12	New	Great	Never	Salad	Well	Anyone	Place	Us	Took
13	Pizzas	Salads	Visit	Just	Always	Good	Lot	Like	Said
14	Great	Delicious	Return	Tomato	Experience	Highly	Street	Pizzas	Just
15	Italian	Sauce	Great	Pasta	Wonderful	Restaurants	Building	Took	Get
16	Little	Dinner	Years	Mozzarella	Nice	Area	Just	Slice	Wait
17	York	Meal	Definitely	Sausage	Italian	Better	Kitchen	Little	Waiter
18	Cheese	Dishes	Many	Flavor	Wait	Worst	Table	Pretty	One
19	Place	Dessert	Just	Beef	Attentive	Family	Nice	Go	Bar
20	Get	One	Dinner	Garlic	Wine	Dining	Little	Time	Got
21	Know	Us	Going	Meat	Menu	Style	Good	Slices	Even
22	Much	House	Italian	Little	Family	New	Back	Said	Service
23	Beef	Special	Eat	Served	Reasonable	Favorite	There's	Came	Told
24	Lot	Large	Since	Crust	Dining	Will	Can	Much	Menu
25	Sauce	Veal	First	Delicious	Pleasant	Just	Front	Half	Never
26	Chain	Fresh	Family	Dish	Delicious	Far	Pretty	Home	Seated
27	Got	Selection	Like	One	Outstanding	Eaten	Like	Minutes	Owner
28	Flavor	Lasagna	Went	Really	Everything	Wonderful	Open	Enough	Manager
29	Dish	Shrimp	Experience	Marinara	Time	Many	Italian	Went	Bill
30	Find	Served	Times	Side	Just	Go	Two	Meal	Dinner

Table 6 We8There Italian Restaurant Data: Results from Topic Regression ($T = 9$, Topic 2 as Contrast)

Topic	Parameter	Posterior Mean	Posterior SD
Baseline	β_0	0.588	0.478
Real pizza	β_1	0.404	0.728
Menu	β_2	0[a]	0[a]
Return	β_3	−0.723	0.772
Food ordered	β_4	−1.149	0.704
Service and staff	β_5	2.549	0.798
Recommend	β_6	1.562	1.012
Layout	β_7	0.154	0.956
Issues with order	β_8	−2.175	0.672
Conflict	β_9	−5.568	0.687
Cut points	c_4	0.128	0[a]
	c_3	−0.507	0.057
	c_2	−1.099	0.071
	c_1	−1.643	0[a]
Fit	R^2	0.649	0.047

[a]Fixed parameter.

Table 6 displays the results of the regression analysis of overall satisfaction on the topic probabilities. We report the R^2 of the latent continuous evaluations τ as a measure of fit of this model. For the Italian restaurant data, the fit is high ($R^2 = 0.65$), indicating that topic probabilities are meaningful devices to explain customer ratings. The coefficient estimates (β^*) are the expected increase (given contrast topic) in the latent rating that is observed in censored form on the rating scale. The cut-point estimates (c_i) for the model indicate that a 0.50 increase in the latent rating is associated with a one-point increase in the observed rating. Since the coefficient estimates are multiplied by the topic probabilities (θ), a 0.10 increase in the topic probabilities are often associated with substantive increases of the ratings. For example, if the probability that a review is associated with the topic "conflict" increases by 0.10, the expected change in the latent rating is −0.56, translating to an almost one-point decline in overall satisfaction.

The regression analysis provides information on which of the coefficients have mass away from zero and which have mass near zero. The posterior standard deviations average about 0.75 (without the contrast topic), indicating that coefficients greater than 1.5 in absolute magnitude are "significant." Thus, topics 5 (service and staff), 8 (issues with order), and 9 (conflict) are worthy of special attention in our analysis, with the presence of topic 5 in a review associated with higher ratings, and that of topics 8 and 9 associated with lower ratings.

Traditional driver analysis in customer satisfaction analysis involves regressing an overall measure of

Table 7 Expedia Upscale Hotel Data, Top 30 Words from the SC-LDA-Rating Model ($T = 10$)

Rank	Topic 1 "Problems" at check-in"	Topic 2 "Nearby" attractions"	Topic 3 "Recommend and return"	Topic 4 "Noise and room negative"	Topic 5 "Room positive"	Topic 6 "Location"	Topic 7 "Amenities"	Topic 8 "Everything great"	Topic 9 "Friendly staff"	Topic 10 "New York experience"
1	Room	Hotel	Stay	Room	Room	Square	Breakfast	Location	Staff	Hotel
2	Hotel	Square	Hotel	Hotel	Clean	Times	Hotel	Hotel	Helpful	New
3	Us	Location	Will	Floor	Comfortable	Hotel	Room	Great	Friendly	York
4	Check	Times	Definitely	Street	Rooms	Location	Free	Staff	Hotel	Stayed
5	Desk	Close	Recommend	Bathroom	Hotel	Great	Good	Clean	Desk	City
6	Got	Subway	Back	One	Beds	View	Great	Good	Service	Stay
7	Day	Walking	Go	Night	Bed	Right	Restaurant	Service	Great	Times
8	Rooms	Distance	Great	Noise	Nice	Room	Food	Room	Nice	Time
9	Early	Walk	Time	Elevator	Large	Time	Service	Excellent	Front	Location
10	Time	Great	Highly	Elevators	New	Stay	Bar	Nice	Clean	Square
11	Front	Station	New	Get	Small	Heart	Internet	Rooms	Room	Great
12	Arrived	Central	Next	Little	Size	Middle	View	Comfortable	Us	Marriott
13	One	Within	York	Didn't	Spacious	Located	Expensive	Overall	Everyone	Hotels
14	Bed	Blocks	Place	Rooms	York	Perfect	Price	Stay	Concierge	Trip
15	Get	Everything	Enjoyed	Lobby	Great	Everything	Wi-Fi	Experience	Courteous	First
16	Staff	Broadway	Trip	Like	Good	Floor	Also	Friendly	Extremely	Best
17	Told	Away	NYC	Shower	City	Quiet	Nice	Price	Excellent	Hilton
18	2	Restaurants	Staying	Time	Bathroom	Nice	Included	Perfect	Pleasant	Marquis
19	Ready	Easy	City	Work	Well	Hilton	Worth	Wonderful	Always	Room
20	Even	Just	Marriott	Great	Quiet	Want	Rooms	Value	Polite	One
21	Called	Block	Visit	Day	Staff	Place	Get	Helpful	Professional	Perfect
22	Asked	Attractions	Return	Bit	Big	Excellent	Buffet	Fantastic	Location	Place
23	King	Right	Hilton	Quiet	Comfy	Building	Just	Everything	Check	Nights
24	Back	Located	Anyone	Small	View	Clean	Day	View	Good	Experience
25	Service	Shopping	Come	Problem	Standards	Good	Coffee	Loved	Every	Price
26	Also	Macy's	Can't	Outside	King	Close	Floor	Amazing	Accommodating	NY
27	Stay	Many	Definitely	Also	Pillows	Views	One	Better	Help	Year
28	Check-in	Convenient	Marquis	Even	Two	Fantastic	Little	Quality	Really	Night
29	Two	Grand	Overall	People	Modern	Staff	Staff	Quiet	Time	Weekend
30	Booked	Penn	Friends	Stay	Friendly	Wonderful	Lobby	Food	Attentive	Much

satisfaction with predefined subscales such as "food quality" or "service," where higher ratings on the subscales are associated with higher expected overall ratings. Such analysis typically only produces positive coefficient values, whereas the SC-LDA-Rating model produces both positive and negative regression coefficients. Moreover, traditional analysis is prone to haloing and other factors that express themselves as colinear regressors (Büschken et al. 2013). Such problems are not present in our topic-based analysis.

4.2. Upscale Hotels

Table 7 displays the top words for each topic in the upscale hotel data, and Table 8 displays the results of the associated regression analysis. Both results are based on the best-fitting SC-LDA-Rating model ($T = 10$). We start by noting that the topic proportions θ_d, obtained from the SC-LDA-Rating model, explain the rating very well ($R^2 = 0.66$). In the subsequent analysis of the topics, we find that most of the top 30 terms of the topics are unique to the topics. Thus, the R^2 of nearly 0.7 is not the result of topic overlap.

Similar to the topics emerging from the analysis of restaurant data, we find coherent topics in the upscale hotel data that center around a common theme. Descriptions of these themes are offered in Table 7. For example, topic 1 talks exclusively about problems for customers at check-in. Among the most frequent words in this topic are "one," "two," "bed," "asked," "got," "king," "ready," "told," and "booked." These words suggest that customers booked specific room types (e.g., room with a king-size bed or two separate beds), but apparently, at check-in, the front desk staff was unable to fulfill those requests. Topic 4 centers around noise problems during the night and its sources ("elevator(s)," "floor," "street," "people") and negative issues with the room ("bathroom," "small," "shower," "didn't," "work," "problem"). Topic 5, by contrast, reports aspects of a positive experience with the room (e.g., "clean," "comfortable," "nice," "spacious"). Topics 2, 6, and 10 cover various aspects of staying at a Manhattan hotel location. It seems that this location offers customers a potential for diverse experiences and that reviewers like to talk about the various aspects of that experience. Topic 3 centers around customers' willingness to return to the hotel ("will," "definitely," "go," "back") and recommend it to others.

From the regression analysis, we find that topics 4 (noise and problems with room), 7 (amenities), and

Table 8 Expedia Upscale Data: Results from Topic Regression ($T = 10$, Topic 2 as Contrast)

Topic	Parameter	Posterior Mean	Posterior SD
Baseline	β_0	0.558	0.583
Problems at check-in	β_1	−3.790	0.683
Nearby attractions	β_2	0[a]	0[a]
Recommend and return	β_3	3.289	1.286
Noise and room negative	β_4	−4.176	0.627
Room positive	β_5	−0.504	0.890
Location	β_6	0.696	1.025
Amenities	β_7	−2.563	0.777
Everything great	β_8	0.140	0.861
Friendly staff	β_9	1.365	0.832
New York experience	β_{10}	−0.091	0.914
Cut points	c_4	−0.259	0[a]
	c_3	−1.040	0.039
	c_2	−1.512	0.045
	c_1	−1.892	0[a]
Fit	R^2	0.663	0.032

[a]Fixed parameter.

1 (problems at check-in) are all significantly negative relative to topic 2 (nearby attractions). Most hotels in this data set charge additionally for Wi-Fi Internet access or breakfast. This is not appreciated much by customers who pay premium prices for these hotels and may expect such services to be included (or priced lower). The largest contributor to a negative rating is topic 4. A 10% increase of the proportion of this topic results in a change of the rating of nearly one rating scale point. This is determined from the regression results reported in Table 8. A 0.10 increase in the topic probability is multiplied by the regression coefficient for topic 4, −4.176, to yield a change in the latent overall rating by −0.42, or about a one-point difference in the rating scale as indicated by the cut-point estimates, c_i. The mention of aspects of hotel check-in (topic 1) is also associated with lower reviews. Apparently, if a customer cares enough to write about their stay and mentions early arrival or the correct room (not) being available, then they probably had a bad experience. One of the themes that emerges out of topic 1 is problems with the room configuration or beds, like a king-sized bed present when it should not be or vice versa. Similarly, the mention of an elevator (topic 4) is associated with lower satisfaction for upscale hotels, and is used in conjunction with words such as "floor," "people," and "noise." Thus, it is not the mechanical operation of the elevator that is problematic, but instead the noise it brings to the floors when it opens.

From Table 7, we find that the topics "friendly staff" (topic 9), "everything great" (topic 8), and "location" (topic 6) are all positively, but not significantly, associated with positive ratings. This result suggests that when booking upscale hotels in Manhattan, customers expect a positive experience characterized by these topics. To find expectations fulfilled seems to be worth mentioning in reviews, but it does not improve ratings. The only topic that significantly drives ratings up is topic 3, which talks about willingness to return.

4.3. Midscale Hotels

Table 9 displays the top words for each topic in the midscale hotel data, and Table 10 displays the associated regression coefficients. For midscale hotel data, we find results very similar to those for restaurant reviews and upscale hotel reviews; that is, we obtain coherent topics from applying the SC-LDA-Rating model that positively and negatively drive the overall rating. We report results from $T = 8$, which is the best-fitting SC-LDA-Rating model.

Several differences emerge from comparing the topics in the two hotel data sets. In the midscale JFK data (Table 9), we do not find a large variety in location-related topics compared to upscale data (Table 7). Topic 3 in the midscale data talks about food/dinner options in the vicinity of the hotels. Hotels in this price segment typically do not have restaurants, so patrons need other accessible food options ("deliver(y)," "nearby," "restaurant(s)," "walking"). Topic 7 is also concerned with location, but from the perspective of air travelers in need of a hotel close to JFK airport for ease of access. For these travelers, the shuttle service to and from the airport is a relevant feature of the hotel (topic 8). In the upscale hotel data, we find none of these aspects of location. By contrast to upscale hotels in Manhattan, midscale hotels offer several free amenities to guests (free Wi-Fi and breakfast, topic 5) that customers feel the need to report.

From the regression analysis (Table 10), two topics emerge as negative drivers of satisfaction for JFK midscale hotels—topic 1 (noise and smell) and topic 6 (front desk). Topic 1 reports significant problems with the room ("carpet") and the hotel ("floor") and talks about unpleasant odors, dirt, and noise. This topic exerts a strong significant negative effect on the rating, with a 10% increase associated with an approximate one-point decrease in rating. Interaction with front desk employees (topic 6) also effects ratings negatively. The top words in this topic suggest that issues arise from the front desk failing to organize transportation at the appropriate time ("time," "get," "check," "early" "morning," "flight," "shuttle," "late").

Service (topic 4) and room/free amenities (topic 5) emerge as positive drivers of satisfaction. The change in rating as a result of an increase of 10% of topic 4 is comparable to the effect of "noise and smell." This

Table 9 Expedia Midscale Hotel Data, Top 30 Words from the SC-LDA-Rating Model ($T = 8$)

Rank	Topic 1 "Noise and smell"	Topic 2 "Recommend"	Topic 3 "Food"	Topic 4 "Service"	Topic 5 "Room and free amenities"	Topic 6 "Front desk"	Topic 7 "JFK"	Topic 8 "Shuttle"
1	Room	Hotel	Breakfast	Staff	Clean	Room	Hotel	Shuttle
2	Small	Stay	Hotel	Helpful	Room	Hotel	Stay	Airport
3	Hotel	Will	Food	Friendly	Breakfast	Desk	JFK	Hotel
4	Rooms	New	Good	Hotel	Comfortable	Breakfast	Night	Hour
5	Little	Recommend	Restaurant	Clean	Good	Front	Flight	Free
6	Air	Inn	Restaurants	Good	Hotel	Us	Airport	JFK
7	Clean	York	Menus	Nice	Rooms	One	Good	Service
8	Floor	Hotels	Area	Service	Small	Time	Place	Every
9	Noisy	Area	Free	Breakfast	Free	Staff	Early	Get
10	Noise	Stayed	Eat	Room	Bed	Check	One	Bus
11	Like	Good	Delivery	Shuttle	Nice	Told	Overnight	Take
12	Smell	Express	Room	Desk	Great	Get	Close	Close
13	Night	Holiday	Nearby	Airport	Service	Early	Morning	Minutes
14	One	JFK	Also	Front	Airport	Morning	Near	Time
15	Window	Price	Get	Stay	Well	Got	Stayed	Subway
16	Bad	Best	Staff	Comfortable	Price	Flight	Next	Convenient
17	People	City	Take	Courteous	Shuttle	Arrived	Great	Us
18	Old	One	Can	Rooms	Bathroom	Back	Just	Breakfast
19	First	Time	Places	Overall	Beds	Smoking	Needed	Easy
20	Stay	Definitely	Great	Great	Quiet	Stay	Need	Good
21	Basement	Back	Dinner	Pleasant	Close	Didn't	Convenient	Train
22	Smoking	Next	Provided	Free	JFK	Went	Fine	Runs
23	Bathroom	Place	Coffee	Location	Excellent	Night	Flights	Flight
24	Just	Airport	Deliver	Excellent	Internet	Bed	Sleep	Station
25	Smelled	Around	Local	Convenient	Convenient	Left	Location	Cab
26	Didn't	Room	Shuttle	JFK	TV	Airport	Perfect	Ride
27	Area	Much	Us	Customer	Nothing	Shuttle	Day	City
28	Dirty	Better	Nice	Efficient	Fine	Late	Short	Great
29	Carpet	Don't	Place	Extremely	Comfy	Even	Flying	Can
30	Find	NY	Walking	Small	Wi-Fi	Asked	Late	Airtrain

Table 10 Expedia Midscale Data: Results from Topic Regression ($T = 8$, Topic 2 as Contrast)

Topic	Parameter	Posterior Mean	Posterior SD
Baseline	β_0	−0.637	0.954
Noise and smell	β_1	−4.846	1.423
Recommend	β_2	0[a]	0[a]
Food	β_3	1.558	1.310
Service	β_4	5.078	1.581
Room and free amenities	β_5	3.060	1.026
Front desk	β_6	−1.960	1.270
JFK	β_7	0.937	1.367
Shuttle	β_8	1.187	1.126
Cut points	c_4	0.890	0[a]
	c_3	−0.218	0.060
	c_2	−1.031	0.047
	c_1	−1.730	0[a]
Fit	R^2	0.719	0.045

[a]Fixed parameter.

suggests that front desk personnel reacting properly to complaints about noise and odors may be able to neutralize the negative effect. In the price segment studied here, free amenities (Wi-Fi, breakfast) are appreciated features and generate better ratings. The presence of topics 7 (JFK) and 8 (shuttle) in a review are associated with more positive review ratings using words such as "overnight" and "convenient" (JFK location) and free and frequent options to get to and from the airport (shuttle).

Finally, we note that the fit of the model with respect to the (latent) rating is the highest for the midscale data set ($R^2 = 0.72$). This is despite the fact that, for this data, the smallest number of topics is needed to maximize predictive fit compared to the other data sets. This suggests that it is not the number of topics that is important to explain ratings, but their coherence. Topic 8 from the midscale hotel data provides a good example of a set of low-probability words being gathered together by the model to provide an interpretable theme for describing variation in the satisfaction ratings.

5. Concluding Remarks

The advantage of using a latent topic model is the ability to uncover collections of words that co-occur in the customer reviews. In the analysis of our reviews, we find that many words are indiscriminately used in all evaluations of hotels and restaurants and therefore do not provide diagnostic value for interpreting

their use. Words like "great" and "not" are often mentioned in reviews and are not interpretable without knowing the object to which they refer. More generally, the analysis of consumer reviews is challenged by the lack of structure in the data. The words used in a bad review of a product can be different from those used in a good review, but not necessarily in a manner that is easy to detect. There may be words that are common to both bad and good reviews, as well as words that are infrequently used but which imply exceptionally good and bad aspects of the product. Simple summaries of words in the form of frequency tables may not be diagnostic of word combinations with good discriminating ability.

We introduce a sentenced-based topic model as a means of structuring the unstructured data. The idea behind topic analysis is that topics are defined by word groups that are used with relatively high probability, being distinct from the probabilities associated with other topics. The high probability word groups supports the presence of co-occurring words that provide added context to the analysis, allowing for a richer interpretation of the data. We extend the structure in topic models by restricting the topics to be the same within a sentence. In a variant of this model, we allow topics to be "sticky" and topic assignments to be non-i.i.d. From the three data sets used here, this variant is not favored over the SC-LDA-Rating model. We believe this is at least partly due to the small number of sentences present in most consumer reviews.

A casual inspection of this and other consumer reviews in our analysis, however, supports the use of the topic sentence restriction, and we find that it improves the predictive fit of the model. The effect of the sentence constraint smoothes out the author–topic probabilities because all words in the sentence are assumed to be part of the same topic. This increases the probability of infrequent words within a topic and decreases the probability of frequent words. We find that reviewers predominantly structure text by forming sentences, many of which express a single underlying topic. Our model naturally exploits this structure and correctly clusters words which are only jointly used in sentences ("front desk," "airport shuttle," "every hour," "walking distance," "comfy bed"), instead of assigning them to different topics.

We relate the topic probabilities to customers' overall satisfaction ratings using a latent cut-point model similar to that used in customer satisfaction driver analysis. We find many significant drivers in each of the data sets examined, with some drivers associated with positive ratings and others associated with negative ratings. We often find that an increase in a topic probability of 0.10 is associated with a unit increase in the rating, and that we consistently explain about 60%–70% of the variation in the latent evaluations. The regression coefficients are useful to identify significant drivers of positive and negative reviews.

Our model allows for the order of words to be changed freely within sentences, but not between sentences. This is because of the dependency of the topic assignment among words observed to part of the same sentence. Removing a word from a sentence implies that the topic assignment of the remaining words may change. The topic assignment of a sentence, however, is independent of the order of the sentences in a document. This introduces a "bag-of-sentences" property to our model in contrast to the standard bag-of-words assumption in stochastic modeling of text. We believe that the bag-of-sentence property more naturally reflects the use of speech in consumer reviews.

This paper demonstrates the usefulness of model-based analysis for unstructured data. The key in the analysis of unstructured data is to impose some type of structure on the analysis. Our analysis employs the structure of latent topics coupled with the assumption that topics change at the period. A challenge in the development of models for unstructured data is in knowing what structure to embed in models used for analysis. We believe that additional linguistic structure of the reviews, in the form of paragraphs and lists, may provide additional opportunities to extend the models used in our analysis.

Additional research is needed on a variety of topics connected to our model. First, we do not attempt to model the factors driving a respondent to post a review. In doing this, we are assuming that the objects of inference are the topics associated with good and bad reviews, and we avoid making statements of the intended consequences of any interventions the firm might undertake or the effects of incentives to get people to post reviews. In addition, we do not attempt to model the number of words per review. We assume that latent topic probabilities are independently determined and, thus, independent of the number of sentences (S_d) and the number of words per sentence (N_{ds}). With the data sets analyzed in this study, this assumption does not seem to be violated. With other data sets and longer reviews in particular (e.g., movie reviews often contain several hundred words), it might be inappropriate. An area of future research would therefore be to build a new model that allows for a dependency between a review's length and latent topics and compare its fit relative to the fit of our proposed model. We leave this model extension, and other generalizations of our model, to future research.

Supplemental Material

Supplemental material to this paper is available at https://doi.org/10.1287/mksc.2016.0993.

Appendix

A.1. Estimation of the LDA Model

The standard LDA model proposed by Blei et al. (2003) employs a Bayesian approach to augment the unobserved topic assignments z_w of the words w. To derive the expression necessary to sample the topic indicators, we start by considering the joint likelihood of observing the words (w) and topic indicators (z), integrated over the word choice probabilities given topics

$$p(\mathbf{w},\mathbf{z}\,|\,\cdot) \propto \prod_{t=1}^{T} \frac{\prod_{w=1}^{W}\Gamma(C_{mt}^{WT}+\beta)}{\Gamma(\sum_w (C_{mt}^{WT}+\beta))} \prod_{d=1}^{D} \frac{\prod_{t=1}^{T}\Gamma(C_{td}^{TD}+\alpha)}{\Gamma(\sum_w (C_{td}^{TD}+\alpha))}. \quad (4)$$

By Bayes' theorem, the full conditional posterior of z_{dn}, the nth word in document d in the corpus, is given by

$$\begin{aligned} p(\mathbf{z}_{dn}\,|\,\mathbf{w},\mathbf{z}_{-dn},\alpha,\beta,D) &= \frac{p(\mathbf{w},\mathbf{z}\,|\,\alpha,\beta,D)}{p(\{\mathbf{w}_{dn},\mathbf{w}_{-dn}\},\mathbf{z}_{-dn}\,|\,\alpha,\beta,D)} \\ &= \frac{p(\mathbf{w},\mathbf{z}\,|\,\alpha,\beta,D)}{p(\mathbf{w}_{dn}\,|\,\alpha,\beta,D)p(\mathbf{w}_{-dn},\mathbf{z}_{-dn}\,|\,\alpha,\beta,D)} \\ &\propto \frac{p(\mathbf{w},\mathbf{z}\,|\,\alpha,\beta,D)}{p(\mathbf{w}_{-dn},\mathbf{z}_{-dn}\,|\,\alpha,\beta,D)}, \end{aligned}$$

where dn denotes word n in document d. Solving this expression gives (Blei et al. 2003)

$$\begin{aligned} &p(z_{dn}=t\,|\,w_{dn}=m,z_{-dn},\alpha,\beta) \\ &\quad\propto \frac{C_{mt,-dn}^{WT}+\beta}{\sum_{m'} C_{m't,-dn}^{WT}+W\beta} \cdot \frac{C_{td,-dn}^{TD}+\alpha}{\sum_{t'} C_{t'd,-dn}^{TD}+T\alpha}, \end{aligned}$$

where $C_{mt,-dn}^{WT}$ and $C_{t,-dn}^{TD}$ are the count matrices with the topic assignment for the current word z_{dn} excluded. This expression can be used to obtain samples from z_{dn} conditional on the data (w) and the topic assignments of all other words.

A.2. Estimation of the Sentence-Constrained LDA Model

The LDA model can be modified as a sentence-based model (SC-LDA) in which topics are assigned to sentences instead of words. In our implementation of this model, periods in consumer reviews provided by consumers identify "sentences," which are assumed to have a unique topic (see Figure 4). Thus, the set of words between periods is assumed to originate from an unobserved topic from a fixed topic set T.

According to the DAG of our model presented in Figure 2, a topic $z_{d,s}$ for sentence s in document d is drawn from the observed set T. Conditional on θ_d, a topic t is drawn independently from a multinomial distribution for each sentence s in document d. Conditional on $\phi_{t=z}$, all words in a sentence for document d are drawn. It follows that $z_{dsi}=z_{dsj}\ \forall i,j\in s$.

By Bayes' theorem, the target distribution is given by

$$\begin{aligned} p(\mathbf{z}_s\,|\,\mathbf{w},\mathbf{z}_{-s},\alpha,\beta,D) &= \frac{p(\mathbf{w},\mathbf{z}\,|\,\alpha,\beta,D)}{p(\{\mathbf{w}_s,\mathbf{w}_{-s}\},\mathbf{z}_{-s}\,|\,\alpha,\beta,D)} \\ &= \frac{p(\mathbf{w},\mathbf{z}\,|\,\alpha,\beta,D)}{p(\mathbf{w}_s\,|\,\alpha,\beta,D)p(\mathbf{w}_{-s},\mathbf{z}_{-s}\,|\,\alpha,\beta,D)} \\ &\propto \frac{p(\mathbf{w},\mathbf{z}\,|\,\alpha,\beta,D)}{p(\mathbf{w}_{-s},\mathbf{z}_{-s}\,|\,\alpha,\beta,D)}, \end{aligned}$$

where s denotes sentence s.

To consider the effect of removing sentence s from the corpus, we introduce count matrix C^{SWT}, the count of words by topics for sentence s in the corpus. The matrix C^{SWT} has zero entries except in the topic column to which all words of sentence s are allocated. It also has zero entries in all rows referring to words from the vocabulary that do not appear in sentence s. We use $C_{mt,-s}^{WT}$ to denote the entries in the count matrix C_{-s}^{WT} obtained after removing sentence s. Note that $C_{mt}^{WT}=C_{mt,-s}^{WT}$ for all topics except the topic to which the words in sentence s were allocated and for all words that do not appear in sentence s.

We define matrix C_s^{TD} as the matrix indicating the allocation of sentence s to a certain topic t in document d. Following from (4), we can write down the likelihood of observing all words, except for sentence s

$$p(\mathbf{w}_{-s},\mathbf{z}_{-s}\,|\,\cdot) \propto \prod_{t=1}^{T} \frac{\prod_{w=1}^{W}\Gamma(C_{mt,-s}^{WT}+\beta)}{\Gamma(\sum_w (C_{mt,-s}^{WT}+\beta))} \prod_{d=1}^{D} \frac{\prod_{t=1}^{T}\Gamma(C_{td,-s}^{TD}+\alpha)}{\Gamma(\sum_w (C_{td,-s}^{TD}+\alpha))}. \quad (5)$$

To arrive at the target distribution, we divide (4) by (5). For this step, consider the first factor on the right-hand side (RHS) of both equations. This implies

$$\begin{aligned} &\frac{\prod_{t=1}^{T}(\prod_{w=1}^{W}\Gamma(C_{mt}^{WT}+\beta)/\Gamma(\sum_w (C_{mt}^{WT}+\beta)))}{\prod_{t=1}^{T}(\prod_{w=1}^{W}\Gamma(C_{mt,-s}^{WT}+\beta)/\Gamma(\sum_w (C_{mt,-s}^{WT}+\beta)))} \\ &\quad= \underbrace{\prod_{w\in s}\frac{\Gamma(C_{mt}^{WT}+\beta)}{\Gamma(C_{mt,-s}^{WT}+\beta)}}_{A} \underbrace{\frac{\Gamma(\sum_w (C_{mt,-s}^{WT}+\beta))}{\Gamma(\sum_w (C_{mt}^{WT}+\beta))}}_{B}. \end{aligned}$$

We consider part A first. By the recursive property of the gamma function

$$\begin{aligned} \frac{\Gamma(C_{mt}^{WT}+\beta)}{\Gamma(C_{mt,-s}^{WT}+\beta)} &= \frac{\Gamma(C_{mt,-s}^{WT}+\beta+C_{mt}^{SWT})}{\Gamma(C_{mt,-s}^{WT}+\beta)} \\ &= (C_{mt,-s}^{WT}+\beta)(C_{mt,-s}^{WT}+\beta+1)\cdot\ldots\cdot \\ &\quad (C_{mt,-s}^{WT}+\beta+(C_{mt}^{SWT}-1)), \end{aligned}$$

where C_{mt}^{SWT} denotes the number of times word w appears in sentence s, allocated to topic t. If $C_{mt}^{SWT}=1$

$$\frac{\Gamma(C_{mt}^{WT}+\beta)}{\Gamma(C_{mt,-s}^{WT}+\beta)} = (C_{mt,-s}^{WT}+\beta).$$

If $C_{mt}^{SWT}=2$

$$\frac{\Gamma(C_{mt}^{WT}+\beta)}{\Gamma(C_{mt,-s}^{WT}+\beta)} = (C_{mt,-s}^{WT}+\beta)(C_{mt,-s}^{WT}+\beta+1),$$

and so forth. It follows that

$$A=\prod_{w\in s}(C_{mt,-s}^{WT}+\beta)(C_{mt,-s}^{WT}+\beta+1)\cdot\ldots\cdot(C_{mt,-s}^{WT}+\beta+(C_{mt}^{SWT}-1)).$$

We next consider part B and denote the number of words in sentence s allocated to topic t by n_{wst}

$$B=\frac{\Gamma(\sum_w (C_{mt,-s}^{WT}+\beta))}{\Gamma(\sum_w (C_{mt}^{WT}+\beta))} = \frac{\Gamma(\sum_w (C_{mt,-s}^{WT}+\beta))}{\Gamma(\sum_w (C_{mt,-s}^{WT}+\beta)+n_{wst})}.$$

Again, by the recursive property of the gamma function

$$\begin{aligned} B = \Bigg[\Big(\sum_w (C_{mt,-s}^{WT}+\beta)\Big)\Big(\sum_w (C_{mt,-s}^{WT}+\beta)+1\Big)\cdot\ldots\cdot \\ \Big(\sum_w (C_{mt,-s}^{WT}+\beta)+(n_{wst}-1)\Big)\Bigg]^{-1}. \end{aligned}$$

The second factor on the RHS of (4) yields the same formal result as in Blei et al. (2003). However, the count matrix C^{TD} is obtained over the allocation of sentences to topics. We arrive at the following expression for the target distribution:

$$p(z_s \mid w_s, n_{ds}, w_{-s}, \alpha, \beta, D)$$
$$= \left(\prod_{w \in s} (C^{WT}_{mt,-s}+\beta)(C^{WT}_{mt,-s}+\beta+1)\cdot\ldots\cdot(C^{WT}_{mt,-s}+\beta+(C^{SWT}_{mt}-1))\right)$$
$$\cdot\left[\left(\sum_w (C^{WT}_{mt,-s}+\beta)\right)\left(\sum_w (C^{WT}_{mt,-s}+\beta)+1\right)\cdot\ldots\cdot\right.$$
$$\left.\left(\sum_w (C^{WT}_{mt,-s}+\beta)+(n_{ds}-1)\right)\right]^{-1}\cdot\frac{C^{TD}_{td}-C^{TD}_{s}+\alpha}{\sum_t (C^{TD}_{td}-C^{TD}_{s}+\alpha)}.$$

A.3. Estimation and Identification of the SC-LDA-Rating Model

We integrate customers' ratings into the SC-LDA model via an ordinal probit regression model. More specifically, we allow the latent, continuous rating τ_d to be a function of a reviews' topic proportions (θ_d)

$$r_d = k \quad \text{if } c_{k-1} \le \tau_d \le c_k,$$
$$\tau_d = \beta_0 + \beta' \theta_d + \epsilon_d,$$

where c is a vector of $K+1$ ordered cut points, β_0 is a baseline, β is a vector of coefficients of length T, and r_d is the observed rating. Cut-points c_0 and c_{K+1} have fixed values. We note that this model, even with cut points c_0, c_1, c_K, and c_{K+1} fixed, is not identified due to the nature of the covariates. We develop an identification strategy for this unidentified model later in this appendix.

The presence of a rating r_d as a function of θ_d implies that, after integrating out θ and ϕ, the rating in a document and the topic assignment of the sentences in that document are no longer independent. To account for this fact, we employ a "semicollapsed" Gibbs sampler where the ϕ are integrated out

$$p(z_s \mid w_s, n_{ds}, w_{-s}, \theta, \beta, D)$$
$$\propto \left(\prod_{w \in s} (C^{WT}_{mt,-s}+\beta)(C^{WT}_{mt,-s}+\beta+1)\cdots(C^{WT}_{mt,-s}+\beta+(C^{SWT}_{mt}-1))\right)$$
$$\cdot\left[\left(\sum_w (C^{WT}_{mt,-s}+\beta)\right)\left(\sum_w (C^{WT}_{mt,-s}+\beta)+1\right)\cdots\right.$$
$$\left.\left(\sum_w (C^{WT}_{mt,-s}+\beta)+(n_{ds}-1)\right)\right]^{-1}\cdot\theta_d.$$

Our regression model (see a DAG of this model in Figure 3) implies that the rating makes a likelihood contribution to the draw of θ_d. As a result, the draw of θ_d changes. We apply the following Metropolis–Hastings (MH) sampling scheme to the draw of θ_d:

1. Generate a candidate θ_d^{cand} from Dirichlet$(C^{TD}+\alpha)$.
2. Accept/reject θ_d^{cand} based on the Metropolis ratio

$$\alpha = \frac{p(y_d \mid \beta, \theta_d^{cand}, \sigma_\epsilon^2, c)}{p(y_d \mid \beta, \theta_d, \sigma_\epsilon^2, c)},$$

which are truncated univariate normal distributions. Note that we generate the candidate θ_d^{cand} from the posterior of the LDA model, which assures that the candidates for θ_d are always probabilities. As a result of this candidate-generating density, all elements in the Metropolis acceptance ratio α cancel out, except for the likelihood component of the regression model. For the draw of the parameters of the ordinal regression model (β, σ^2) and for the augmentation of the continuous ratings τ and the cut points c, we use standard results from the literature.

Regressing the rating on the topics requires an identification strategy. To see this, consider the case of $T=3$, i.e., a model with three topics. The regression equation is then

$$\tau_d = \beta_0 + \beta_1 \frac{t_{1,d}}{\sum_j t_{j,d}} + \beta_2 \frac{t_{2,d}}{\sum_j t_{j,d}} + \beta_3 \frac{t_{3,d}}{\sum_j t_{j,d}} + \epsilon_d, \qquad (6)$$

where

$t_{j,d}$: number of times a word in document d is allocated to topic j,
$\sum_j t_{j,d}$: number of words in document d,
τ: latent continuous rating,
β: regression coefficients,
ϵ: regression error.

The ratio $t_{j,d}/\sum_j t_{j,d}$ expresses the share of topic j in document d (e.g., θ_d from LDA).

Using $\sum_j t_{j,d} = t_{1,d} + t_{2,d} + t_{3,d}$, Equation (6) can be expressed as

$$\tau_d = \beta_0 + \beta_1 \frac{t_{1,d}}{\sum_j t_{j,d}} + \beta_2 \frac{t_{2,d}}{\sum_j t_{j,d}} + \beta_3 \left(1 - \frac{t_{1,d}+t_{2,d}}{\sum_j t_{j,d}}\right) + \epsilon_d. \qquad (7)$$

Simplifying (2) leads to

$$\tau_d = (\beta_0+\beta_3) + (\beta_1-\beta_3)\frac{t_{1,d}}{\sum_j t_{j,d}} + (\beta_2-\beta_3)\frac{t_{2,d}}{\sum_j t_{j,d}} + \epsilon_d,$$

which we rewrite as

$$\tau_d = \beta_0^* + \beta_1^* \frac{t_{1,d}}{\sum_j t_{j,d}} + \beta_2^* \frac{t_{2,d}}{\sum_j t_{j,d}} + \epsilon_d. \qquad (8)$$

Equation (8) demonstrates that the regression in Equation (6) is not identified. The reason for nonidentification is the redundancy of any one share of topics which can be expressed as the residual of the other shares. Equation (8), however, also shows that any slope coefficient in (6) can be omitted, and the resulting β^* are obtained as contrasts to this coefficient. In (8), the "new" baseline β_0^* is a baseline in relation to the "omitted" β_3, as are the slope coefficients β_1^* and β_2^*.

Table A.1 outlines the relationship between the nonidentified parameters of the model and the identified parameters (β^*). From Table A.1 it is clear that only the differences of the true parameters are identified. The choice of the contrast is arbitrary. Table A.1 also suggests a postprocessing

Table A.1 **Relationships** $(T=3)$

Contrast β_1	Contrast β_2	Contrast β_3
$\beta_0^* = \beta_0 + \beta_1$	$\beta_0^* = \beta_0 + \beta_2$	$\beta_0^* = \beta_0 + \beta_3$
—	$\beta_1^* = \beta_1 - \beta_2$	$\beta_1^* = \beta_1 - \beta_3$
$\beta_2^* = \beta_2 - \beta_1$	—	$\beta_2^* = \beta_2 - \beta_3$
$\beta_3^* = \beta_3 - \beta_1$	$\beta_3^* = \beta_3 - \beta_2$	—

strategy for the identified parameters when an MCMC procedure is applied to the estimation of Equation (6). We can use the MCMC procedure to sample from the nonidentified parameter space and postprocess down to the identified parameter space via results in Table A.1.

We demonstrate postprocessing for the following example from which we generate synthetic data:

$$\tau_d = 1 + -1\frac{t_{1,d}}{\sum_j t_{j,d}} + 1\frac{t_{2,d}}{\sum_j t_{j,d}} + 2\frac{t_{3,d}}{\sum_j t_{j,d}} + \epsilon_d,$$

and $\sigma_\epsilon^2 = 0.1$, $N = 2{,}000$, and the topic shares given $T = 3$ generated from a Dirichlet distribution with $\alpha_t = 0.5\ \forall t$. For the MCMC procedure, we use standard weakly informative priors and conjugate results for the conditional posterior distributions of the unknowns (β, σ).

Figure A.1 shows results from the MCMC for β_1. The left panel shows the direct results from the MCMC. It is obvious that the sampler does not recover the true value ($\beta_1 = -1$). The posterior mean obtained from the MCMC is -1.25, and the posterior SD is 1.59. The right panel of Figure A.1 shows the postprocessed parameter β_1^*, using β_3 as contrast. The posterior mean of β_1^* is -2.996, and the posterior SD is 0.09. Note that $\beta_1^* = \beta_1 - \beta_3 = -3$.

This demonstrates that we can use the samples from the MCMC, using Equation (6), and postprocess the results using the equations in Table A.1. An a priori choice of contrast to identify the model as in Equation (8) is not necessary.

A.4. Simulation Study: Efficiency of the LDA Model

In the following, we evaluate the efficiency of the LDA model when a topic sentence constraint is present in the data. Theoretically, an LDA model can assign the same topic to the words in a sentence. Also, the LDA model and the proposed SC-LDA model both operate on the same sufficient statistic, the word counts by document. This raises the issue of efficiency of the LDA model compared to the SC-LDA model. To explore this issue we conducted a simulation study. In this simulation study, we generated data from a SC-LDA-Rating model (i.e., *with* a sentence constraint) and then estimated an LDA rating model (i.e., *without* the sentence constraint). The question we tried to answer is under what conditions the LDA model without the sentence constraint is able to pick up the true data mechanism in which the words in a sentence originate from the same topic. The setup of the simulation is as follows:

- We set $T = 8$ and $V = 1{,}000$.
- We simulate θ_d from symmetric Dirichlet distributions using $\alpha = 2/T$ and ϕ_t from symmetric Dirichlet distributions using $\beta = 2{,}000/V$ or $\beta = 100/V$.
- We generate $D = 2{,}500$ documents with 4–10 or 18–36 sentences per document and 2–6 or 12–18 words per sentence (words and sentences uniformly distributed over indicated range).

A smaller value of β reduces the number of co-occurring terms under a topic, as the ϕ_t are then concentrated among relatively few terms. Assigning topics wordwise, as with the LDA model, should be less of a problem when the number of co-occurring terms is small. By contrast, a larger value of β increases the number of co-occurring terms. Topics can then only be identified correctly when all words in a sentence are considered. In summary, ignoring a sentence constraint present in the data should be less important when

- the number of words per sentence is small and
- the number of terms uniquely associated with a topic is small.

We evaluate the efficiency of the estimation procedure by the hit rate of the topic assignments of all words in the corpus. Recovering the true topic assignments of the words is essential for recovery of all other parameters of the model, including the parameters of the rating model. Figure A.2 displays the posterior means of the hit rates of the topic assignments for the eight simulation scenarios. The left panel of Figure A.2 shows the topic hit rates for $\beta = 2{,}000/V$, and the right panel of Figure A.2 shows the topic hit rates for $\beta = 100/V$.

Figure A.2 reveals that the topic hit rate of the LDA model is smaller than that of the SC-LDA model for all scenarios. For a high β (left panel of Figure A.2), the difference in topic hit rates is significant, especially when the number of words in the sentences is high. The advantage of the SC-LDA model is small when topics are characterized by few frequently occurring words ($\beta = 100/V$). In this situation, specific terms are highly indicative of a topic, and co-occurrence of such terms with less frequent terms within sentences is less likely. It is in this situation that ignoring the sentence constraint in the data introduces less bias in estimation.

A.5. MCMC for the SC-LDA-Rating Model with Sticky Topics

To develop an MCMC estimation procedure for the SC-LDA-Rating model with sticky topics, we start by defining the generate model of SC-LDA with first-order topic carryover. The generative model of the sticky topic model with fixed priors α, β, and ϵ is as follows:

1. Draw ψ_t from Beta(ϵ) $\forall t$ i.i.d.
2. Draw ϕ_t from Dirichlet(β) $\forall t$ i.i.d.
3. Draw θ_d from Dirichlet(α) $\forall d$ i.i.d.
4. For the first sentence in document d, s_1:
   (a) Draw z_1 from Multinomial(θ_d)
   (b) Draw set of words $\{w_1\}$ in sentence s_1 i.i.d. from Multinomial($\phi_{t=z_1}$)
   (c) Draw ζ_2 from Binomial($\psi_{t=z_1}$)
5. For sentences s_N, $N \in [2 : n_D]$:
   (a) if $\zeta_n = 0$: draw z_n from Multinomial(θ_d); if $\zeta_n = 1$: set $z_n = z_{n-1}$
   (b) Draw $\{w_n\}$ i.i.d. from Multinomial($\phi_{t=z_n}$)
   (c) Draw ζ_{n+1} from Binomial($\psi_{t=z_n}$)
6. Repeat steps 4 and 5 for all documents $d \in D$ (except for draw of ζ_{N_d})

Based on the DAG in Figure 5, we can factorize the joint distribution of the knowns and unknowns for a single document as follows:

$$\begin{aligned} &p(\{w\}_d, \{z\}_d, \theta_d, \phi, \{\zeta\}_d, \alpha, \beta, \psi, \epsilon) \\ &\propto p(w_1 \mid \phi, z_1) \times p(z_1 \mid \theta_d) \times \prod_{n=2}^{N_d} p(w_n \mid \phi, z_n, z_{n-1}, \zeta_n) \\ &\quad \times p(z_n \mid z_{n-1}, \theta_d, \zeta_n) \times p(\zeta_n \mid z_{n-1}, \psi) \times p(\phi \mid \beta) \times p(\theta_d \mid \alpha) \\ &\quad \times p(\psi \mid \epsilon) \times p(\beta) \times p(\alpha) \times p(\epsilon). \end{aligned} \tag{9}$$

Figure A.1 Raw and Postprocessed Results from the MCMC

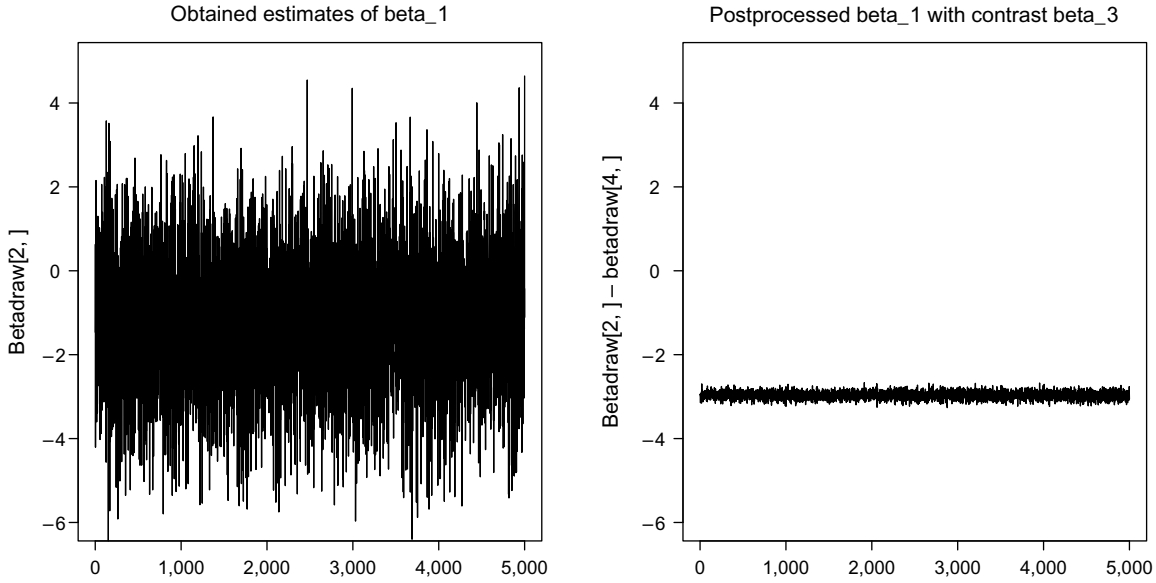
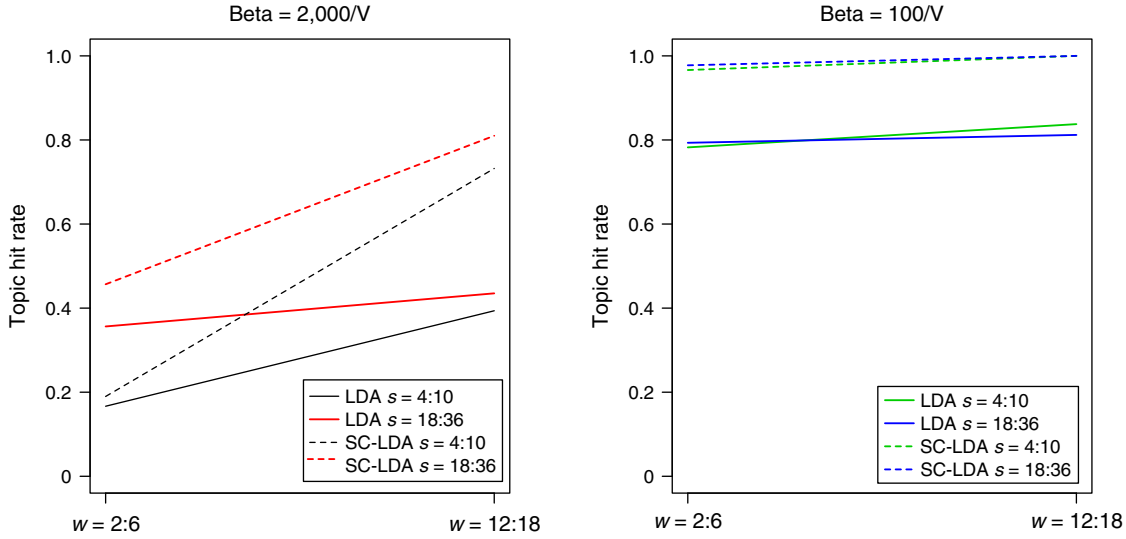


Figure A.2 (Color online) Topic Hit Rates



The likelihood of a word (or sentence), conditional on ζ_n, is

$$p(w_n \mid \phi, z_n, z_{n-1}, \zeta_n = 0) = p(w_n \mid \phi, z_n),$$
$$p(w_n \mid \phi, z_n, z_{n-1}, \zeta_n = 1) = p(w_n \mid \phi, z_{n-1}).$$

The likelihood of a topic assignment, conditional on ζ_n, is

$$p(z_n \mid z_{n-1}, \theta_d, \zeta_n = 0) = p(z_n \mid \theta_d),$$
$$p(z_n \mid z_{n-1}, \theta_d, \zeta_n = 1) = p(z_n = z_{n-1}) = 1.$$

Our model with sticky topics is a sentence-based model that constrains topic assignments to sentences in the same way as in the SC-LDA model

$$p(\{w\}_d, \{z\}_d, \theta_d, \phi, \{\zeta\}_d, \alpha, \beta, \psi, \epsilon)$$
$$\propto p(\{w\}_{s=1} \mid \phi, z_{s=1}) \times p(z_{s=1} \mid \theta_d) \times \prod_{s=2}^{N_{S_d}} p(\{w\}_s \mid \phi, z_s, z_{s-1}, \zeta_s)$$
$$\times p(z_s \mid z_{s-1}, \theta_d, \zeta_s) \times p(\zeta_s \mid z_{s-1}, \psi) \times p(\phi \mid \beta) \times p(\theta_d \mid \alpha)$$
$$\times p(\psi \mid \epsilon) \times p(\beta) \times p(\alpha) \times p(\epsilon). \quad (10)$$

In the following, we develop an MCMC sampling scheme for the sticky topic LDA model. The factorization of the joint posterior distribution of the parameters suggests the following sampling steps:

1. On the document level (omitting subscript d for z and w to improve readability):

(a) $p(z, \zeta \mid else) \propto p(w_1 \mid \phi, z_1) \times p(z_1 \mid \theta_d) \times \prod_{n=2}^{N_d} p(w_n \mid \phi, z_n, z_{n-1}, \zeta_n) \times p(z_n \mid z_{n-1}, \theta_d, \zeta_n) \times p(\zeta_n \mid z_{n-1}, \psi)$

(b) $p(\theta_d \mid else) \propto p(z_1 \mid \theta_d) \times \prod_{n=2}^{N_d} p(z_n \mid z_{n-1}, \theta_d, \zeta_n) \times p(\theta_d \mid \alpha)$

2. $p(\phi_t \mid w, z, \beta) \propto \prod_{d=1}^{d} \prod_{n=1}^{N_d} p(w_n \mid \phi_t, z_n) \times p(\phi_t \mid \beta)\ \forall t$
3. $p(\psi_t \mid else) \propto \prod_{d=1}^{d} \prod_{n=1}^{N_d} p(\zeta_n \mid \psi, z_{n-1}) \times p(\psi_t \mid \epsilon)\ \forall t$

Because of the first order carryover effect of the topics, it is useful to write down the joint probability of all quantities with respect to two subsequent sentences

$$\begin{aligned} &p(w_n, w_{n+1}, z_n, z_{n+1}, \zeta_n, \zeta_{n+1}, \phi, \psi, \theta_d) \\ &\quad \propto p(w_n, w_{n+1} \mid \phi, z_n, z_{n-1}, \zeta_n) \times p(z_n \mid z_{n-1}, \theta_d, \zeta_n) \\ &\qquad \times p(\zeta_n \mid z_{n-1}, \psi) \times p(\zeta_{n+1} \mid z_n, \psi). \end{aligned}$$

In the above, the expression $p(z_{n+1} \mid z_n, \theta_d, \zeta_{n+1})$ was omitted because it is a constant with respect to z_n. Note that

- $p(w_n, w_{n+1} \mid \phi, z_n, z_{n-1}, \zeta_n = 0, \zeta_{n+1} = 0) = p(w_n \mid \phi, z_n) \times p(w_{n+1} \mid \phi, z_{n+1})$,
- $p(w_n, w_{n+1} \mid \phi, z_n, z_{n-1}, \zeta_n = 0, \zeta_{n+1} = 1) = p(w_n \mid \phi, z_n) \times p(w_{n+1} \mid \phi, z_n)$,
- $p(w_n, w_{n+1} \mid \phi, z_n, z_{n-1}, \zeta_n = 1, \zeta_{n+1} = 0) = p(w_n \mid \phi, z_{n-1}) \times p(w_{n+1} \mid \phi, z_{n+1})$,
- $p(w_n, w_{n+1} \mid \phi, z_n, z_{n-1}, \zeta_n = 1, \zeta_{n+1} = 1) = p(w_n \mid \phi, z_{n-1}) \times p(w_{n+1} \mid \phi, z_{n-1})$.

where the last expression presents the case of a repeated topic carryover.

A.5.1. Draw of z_n and ζ_n. Analogous to Gibbs sampling for the Hidden Markov Model (Frühwirth-Schnatter 2006), we consider a joint "single-move" Gibbs sampler of the topic and the stickiness indicator. The joint posterior of z_n, ζ_n is obtained by dropping all elements independent of z_n and ζ_n from Equation (9) (sentence-based model, from (10)) and treating the latent variables z_{n-1}, ζ_{n-1}, z_{n+1}, and ζ_{n+1} as observed

$$\begin{aligned} p(z_n = t, \zeta_n \mid else) \propto\ & p(w_n, w_{n+1} \mid \phi, z_n = t, z_{n-1}, \zeta_n, \zeta_{n+1}) \\ & \times p(z_n = t \mid z_{n-1}, \theta_d, \zeta_n) \\ & \times p(\zeta_n \mid z_{n-1}, \psi) \times p(\zeta_{n+1} \mid z_n = t, \psi). \end{aligned}$$

Using results from the above:

$$\begin{aligned} p(z_n = t, \zeta_n = 1 \mid \zeta_{n+1} = 0, else) &\propto p(w_n \mid \phi, z_{n-1}) \\ &\quad \times p(w_{n+1} \mid \phi, z_{n+1}) \times p(z_n = t \mid z_{n-1}, \theta_d, \zeta_n = 1) \\ &\quad \times p(\zeta_n = 1 \mid z_{n-1}, \psi) \times p(\zeta_{n+1} \mid z_{n-1}, \psi) \\ &\propto \theta_{z_{n-1}}^{(w_n)} \times \psi_{z_{n+1}} \times (1 - \psi_{z_{n-1}}), \end{aligned}$$

$$\begin{aligned} p(z_n = t, \zeta_n = 0 \mid \zeta_{n+1} = 0, else) &\propto p(w_n \mid \phi, z_n = t) \\ &\quad \times p(w_{n+1} \mid \phi, z_{n+1}) \times p(\zeta_n = t \mid z_{n-1}, \theta_d, \zeta_n = 0) \\ &\quad \times p(\zeta_n = 0 \mid z_{n-1}, \psi) \times p(\zeta_{n+1} \mid z_n = t, \psi) \\ &\propto \psi_t^{(w_n)} \times \theta_{d,t} (1 - \psi_{z_{n-1}}) \times (1 - \psi_t), \end{aligned}$$

$$\begin{aligned} p(z_n = t, \zeta_n = 0 \mid \zeta_{n+1} = 1, else) &\propto p(w_n \mid \phi, z_n = t) \\ &\quad \times p(w_{n+1} \mid \phi, z_n = t) \times p(\zeta_n = t \mid z_{n-1}, \theta_d, \zeta_n = 0) \\ &\quad \times p(\zeta_n = 0 \mid z_{n-1}, \psi) \times p(\zeta_{n+1} \mid z_n = t, \psi) \\ &\propto \phi_t^{(w_n)} \times \phi_t^{(w_{n+1})} \times \theta_{d,t} \times (1 - \psi_{z_{n-1}}) \times \psi_t. \end{aligned}$$

In the case of $n = N_d$, $p(w_{n+1} \mid \cdot)$ and $p(\zeta_{n+1} \mid \cdot)$ can be dropped because these distributions do not exist. Note that, in the case of a topic carryover from word n to $n+1$ and $\zeta_n = 0$, the downstream likelihood of z_n consists of two words. If, however, $\zeta_n = 1$ the posterior does not depend on z_n because the topic is already determined. Essentially, the above expressions deal with the question whether to choose the "observed" previous topic assignment z_{n-1} for the current word w_n or to consider the case that z_n originates from θ_d. The above expressions give rise to $T+1$ multinomial probabilities from which we can jointly draw z_n, ζ_n. An alternative sampling scheme may consider $\{z_n\}_d$, $\{\zeta_n\}_d$ for a simultaneous update of all the latent topic and stickiness indicators in a document.

A.5.2. Draw of θ_d. In MCMC sampling for the standard LDA, the full conditional draw of θ_d is based on using the multinomial topic assignment of all sentences in a document as likelihood information. The multinomial likelihood of the topic assignments is combined with the Dirichlet prior $p(\theta \mid \alpha)$ for a conjugate update via a Dirichlet posterior in which the topic assignments are simple counts

$$p(\theta_d \mid else) \propto \text{Dirichlet}(C^{TD} + \alpha). \tag{11}$$

For the sticky LDA model, we have to keep track of the topic assignments that are downstream of θ_d and disregard topic assignments due to $\zeta = 1$

$$\begin{aligned} p(\theta_d \mid else) &\propto p(z_1 \mid \theta_d) \times \prod_{n=2}^{N_d} p(z_n \mid z_{n-1}, \theta_d, \zeta_n) \times p(\theta_d \mid \alpha) \\ &= \prod_{n:\, \zeta_n = 0} p(z_n \mid \theta_d) \times p(\theta_d \mid \alpha). \end{aligned}$$

We use the count matrix C^{TD} to collect topic assignments conditional on $\zeta_n = 0$ and then proceed as in the standard LDA.

A.5.3. Draw of ϕ. The draw of ϕ_t is not affected by the mixture prior for the topic assignments because of conditioning on z and can therefore be conducted in the usual way

$$p(\psi_t \mid else) \propto \text{Dirichlet}(C^{WT} + \beta). \tag{12}$$

A.5.4. Draw of ψ. For the model without covariates, the update of ψ is accomplished as follows:

$$\begin{aligned} &\prod_{d=1}^{D} \prod_{n=1}^{N_d} p(\zeta_{d,n} \mid \psi, z_{d,n} = t) \times p(\psi_t \mid \epsilon) \\ &\quad \propto \prod_{t=1}^{T} \psi_t^{S^t} \cdot (1 - \psi_t)^{C^t - S^t} \cdot \psi_t^{\epsilon_0 - 1} \cdot (1 - \psi_t)^{\epsilon_1 - 1} \\ &\quad = \prod_{t=1}^{T} \text{Beta}(S^t + \epsilon_0 - 1, C^t - S^t + \epsilon_1 - 1), \end{aligned} \tag{13}$$

where $S^t = \sum_d^D \sum_n^{N_d} \zeta_{d,n}^t$ or the number of times an assignment of topic t was "observed" to be sticky. C^t is the number or "trials," i.e., the total number of assignments of topic t to the sentences in the corpus except for $z_{d,1}$, the topic assignment of the first word (or sentence) in each document. For the model with covariates (Equation (3)) we use a binary probit regression model (Rossi et al. 2005).

Table A.2 Simulation Results: Parameter Recovery

	$\beta = 100/V$	$\beta = 1{,}000/V$	$\beta = 2{,}000/V$
$\alpha = 1/T$			
Φ	0.997 (0.001)	0.974 (0.001)	0.950 (0.001)
Θ	0.940 (0.002)	0.921 (0.002)	0.892 (0.003)
σ^2	0.209 (0.011)	0.205 (0.014)	0.199 (0.018)
Ψ	0.026 (0.011)	0.033 (0.018)	0.054 (0.041)
$\alpha = 2/T$			
Φ	0.998 (0.001)	0.975 (0.001)	0.949 (0.001)
Θ	0.886 (0.003)	0.860 (0.003)	0.803 (0.004)
σ^2	0.201 (0.013)	0.193 (0.017)	0.187 (0.024)
Ψ	0.016 (0.006)	0.027 (0.016)	0.063 (0.033)
$\alpha = 4/T$			
Φ	0.997 (0.001)	0.974 (0.001)	0.951 (0.001)
Θ	0.792 (0.004)	0.747 (0.005)	0.690 (0.007)
σ^2	0.199 (0.012)	0.213 (0.020)	0.195 (0.025)
Ψ	0.014 (0.003)	0.023 (0.020)	0.039 (0.024)

A.5.5. Prior Distributions. We use the following (fixed) prior distributions in our analysis:

$$\begin{aligned}
\theta_d &\sim \text{Dirichlet}(5/T),\\
\psi_t &\sim \text{Dirichlet}(100/V),\\
\sigma^2 &\sim \text{Inverse Gamma}(1,1),\\
\beta^{(reg)} &\sim N(0,10),\\
\gamma_t &\sim N(0,10),\\
\log(\delta) &\sim N(\mu_\delta, \Sigma_\delta).
\end{aligned}$$

All fixed prior distributions are weakly informative, conjugate prior distributions. To see this for θ_d, ψ_t, consider that fixing α, β is equivalent to fixing prior pseudocounts from an imaginary prior data set; that is, assuming $\beta = 100/V$, given $V = 1{,}000$, is equivalent to assuming 0.1 prior pseudocounts per unique term and topic. Similarly, given $T = 10$, $\alpha = 5/T$ is equivalent to 0.5 prior pseudocounts per topic and document. Larger values for α, β have a smoothing effect on estimates of θ, ψ, respectively. We tested prior setups for "smoothed" estimates, using larger values for α, β, and did not find that the results obtained from the three data sets differ significantly.

A.6. Simulation Study: Empirical Identification of the Sticky SC-LDA-Rating Model

In the following, we demonstrate statistical identification of the SC-LDA-Rating model with sticky topics using a simulation study. The study is based on a vocabulary of $V = 1{,}000$ unique terms and four topics ($T = 4$). The number of sentences per document is drawn from a uniform distribution across values 15 to 20, and we draw the number of words per sentence from a uniform distribution over values $\{3, 4, \ldots, 6\}$. We generate $M = 2{,}000$ documents and a corpus of about 150,000 words.

We generate the true word-topic probabilities (Φ) and the true document-topic probabilities (θ_d) from symmetric Dirichlet distributions. We allow β, the prior of Φ_t, to range from $100/V$ to $2{,}000/V$. We set α, the prior of θ_d, to values in $(1/T, 2/T, 4/T)$. We vary β and α independently, resulting in nine simulation scenarios (Table A.2). We set Ψ to 0.12, 0.02, 0.05, and 0.40 for topics 1 to 4, respectively. Note that, in the limit, homogenous ψ_t lead to marginal topic frequencies equal to θ_d.

In the SC-LDA-Rating model, the rating for each document is assumed to be generated via an ordinal probit regression model. For the simulation of data for this model, we use a baseline and slope coefficients with values $\beta_0^{\text{reg}} = -0.5$, $\beta_1^{\text{reg}} = 1$, $\beta_2^{\text{reg}} = -2$, and $\beta_3^{\text{reg}} = 1.8$. The error variance of the model is fixed at $\sigma^2 = 0.2$. To obtain ordinal ratings from the latent continuous evaluations τ generated by this model, we use cut points c fixed at values so that all rating categories are equally populated. In parameter estimation, we use data augmentation for the latent continuous evaluation τ and estimate all parameters of the ordinal probit model using the identification strategy outlined above.

In Table A.2, we report the correlation of the simulated and true parameters of Φ and Θ, σ^2, and the Mean Absolute Deviation (MAD) for Ψ from the nine scenarios. Recovery of $\sigma^2 = 0.2$ implies recovery of all parameters of the regression model, as this parameter is invariant to switches of the topic labels. For each scenario, we simulated data 100 times. For each of the 100 runs, we computed the correlation of the posterior means of Φ and Θ with true values, the posterior mean of σ^2, and the MAD of Ψ. We then computed the mean and SD of these quantities across the 100 simulation runs for purposes of reporting (Table A.2).

Table A.2 reveals that the parameters of our model can be recovered in all scenarios with high accuracy for $\beta \leq 1{,}000/V$ and $\alpha \leq 2/T$. In general, accuracy declines as β and α are increased. Higher values of β induce a more uniform distribution of words over the vocabulary. Higher values of α induce a more ambiguous relationship between documents and topics. We note that a more ambiguous relationship between documents and topics has a detrimental effect on the recovery of Ψ. This is because identification of Ψ depends on carryover of topics that are relatively rare, given θ_d.

A viable question to ask is whether our sampler identifies the true T, which must be fixed for an empirical application of topic models. Given a fixed simulated data set using $T_{\text{true}} = 4$ and an informative setup ($\alpha = 1/T$, $\beta = 100/V$, $V = 1{,}000$, $M_{\text{calib}} = 1{,}000$, $M_{\text{pred}} = 500$), we ran our model using alternative values for T. Table A.3 shows the in-sample fit and predictive fit of the model with T ranging from 2 to 12. Reported is the log marginal density of the data for the calibration and the holdout data. In Table A.3, results from uneven topic numbers are omitted for brevity. The results indicate that the model correctly identifies $T = 4$ as the true data generating process.

Table A.3 Model Fit for Simulated Data

	$T = 2$	$T = 4$	$T = 6$	$T = 8$	$T = 10$	$T = 12$
In-sample fit	−384,323.9	−380,492.2	−381,603.9	−382,060.7	−382,393.5	−383,815.1
Predictive fit	−196,132.3	−193,796.6	−194,510.8	−194,954.4	−194,889.7	−195,682.0

References

Andrzejewski D, Zhu X, Craven M (2009) Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *Proc. Internat. Conf. Machine Learn.* 382(26):25–32.

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Berger J, Sorensen AT, Rasmussen SJ (2010) Positive effects of negative publicity: When negative reviews increase sales. *Marketing Sci.* 29(5):815–827.

Blei DM, Lafferty JD (2006) Dynamic topic models. *Proc. 23rd Internat. Conf. Machine Learn.* (ACM, New York), 113–120.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3(January):993–1022.

Büschken J, Otter T, Allenby GM (2013) The dimensionality of customer satisfaction survey responses and implications for driver analysis. *Marketing Sci.* 32(4):533–553.

de Jong MG, Lehmann DR, Netzer O (2012) State-dependence effects in surveys. *Marketing Sci.* 31(5):838–853.

Dellarocas C, Zhang XM, Awad NF (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. Interactive Marketing* 21(4):23–45.

DeSarbo WS, Lehmann DR, Hollman FG (2004) Modeling dynamic effects in repeated-measures experiments involving preference/choice: An illustration involving stated preference analysis. *Appl. Psych. Measurement* 28(3):186–209.

Ding M, Grewal R, Liechty J (2005) Incentive-aligned conjoint analysis. *J. Marketing Res.* 42(1):67–82.

Fader PS, Hardie BGS, Huang CY (2004) A dynamic changepoint model for new product sales forecasting. *Marketing Sci.* 23(1): 50–65.

Frühwirth-Schnatter S (2006) *Finite Mixture and Markov Switching Models* (Springer Science+Business Media, New York).

Gal D, Rucker DD (2011) Answering the unasked question: Response substitution in consumer surveys. *J. Marketing Res.* 48(1):185–195.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4): 545–560.

Gruber A, Weiss Y, Rosen-Zvi M (2007) Hidden topic Markov models. *Internat. Conf. Artificial Intelligence Statist.*, 163–170.

Kamakura WA, Kim BD, Lee J (1996) Modeling preference and structural heterogeneity in consumer choice. *Marketing Sci.* 15(2):152–172.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(5):881–894.

Ludwig S, de Ruyter K, Friedman M, Brüggen EC, Wetzels M, Pfann G (2013) More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *J. Marketing* 77(1):87–103.

McCallum A, Corrada-Emmanuel A, Wang X (2005) The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and academic email. Technical report, University of Massachusetts, Amherst.

Montoya R, Netzer O, Jedidi K (2010) Dynamic allocation of pharmaceutical detailing and sampling for long-term profitability. *Marketing Sci.* 29(5):909–924.

Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. *Marketing Sci.* 27(2):185–204.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.

Ramage D, Dumais ST, Liebling DJ (2010) Characterizing microblogs with topic models. *Proc. Fourth Internat. AAAI Conf. Weblogs Social Media*.

Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. *Proc. 20th Conf. Uncertainty Artificial Intelligence* (AUAI Press, Arlington, VA), 487–494.

Rossi PE, Allenby GM, McCulloch RE (2005) *Bayesian Statistics and Marketing* (John Wiley & Sons, West Sussex, UK).

Rossi PE, Gilula Z, Allenby GM (2001) Overcoming scale usage heterogeneity. *J. Amer. Statist. Assoc.* 96(453):20–31.

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Marketing Res.* 51(4):463–479.

Titov I, McDonald R (2008) A joint model of text and aspect ratings for sentiment summarization. *Proc. ACL*, Vol. 8, 308–316.

Wallach HM (2006) Topic modeling: Beyond bag-of-words. *Proc. 23rd Internat. Conf. Machine Learn.* (ACM, New York), 977–984.

Yang S, Allenby GM (2000) A model for observation, structural, and household heterogeneity in panel data. *Marketing Lett.* 11(2):137–149.

Yang S, Allenby GM, Fennel G (2002) Modeling variation in brand preference: The roles of objective environment and motivating conditions. *Marketing Sci.* 21(1):14–31.

Ying Y, Feinberg F, Wedel M (2006) Leveraging missing ratings to improve online recommendation systems. *J. Marketing Res.* 43(3):355–365.

Zhao Y, Yang S, Narayan V (2013) Modeling consumer learning from online product reviews. *Marketing Sci.* 32(1):153–169.

CORRECTION

In this article, "Sentence-Based Text Analysis for Customer Reviews" by Joachim Büschken and Greg M. Allenby (first published in *Articles in Advance*, July 18, 2016, *Marketing Science*, DOI:10.1287/mksc.2016.0993), Appendix A.5.1 has been updated and Equation 13 has been corrected.