Article

AMERICAN MARKETING ASSOCIATION

Journal of Marketing Research
2025, Vol. 62(6) 1045-1062

DOI: 10.1177/00222437251315745
journals.sagepub.com/home/mrj
Sage

Accounting for Formative and Reflective Topics in Product Review Data for Better Consumer Insights

Joachim Büschken, Thomas Otter, and Greg M. Allenby

Abstract

Observations of product and service reviews suggest that textual product reviews may contain statements about the overall experience ("We had a great time") or, similarly, about whether to recommend a particular product. The authors argue that such statements encapsulate an overall assessment and hence are not independently informative about, but rather reflect, overall ratings. The authors propose a model that allows for the distinction between topics that contribute to and topics that merely reflect an overall evaluation and apply the model to a dataset consisting of luxury hotel reviews. The findings show that, compared with a standard supervised latent Dirichlet allocation, the proposed model better fits the data and improves customer insights by resulting in more semantically coherent topics that point at specific aspects with positive and negative relationships to customers' evaluation of their experience.

Keywords

statistical text analysis, supervised LDA, rating analysis, structural modeling, Bayesian modeling

Online supplement https://doi.org/10.1177/00222437251315745

Submitted July 7, 2023

Product reviews and product ratings provide marketers the opportunity to understand the reasons for product success and failure. Written reviews give customers the opportunity to describe their experience with a product and to rationalize, argue for, and explain their overall rating in the same way they explain their opinions in real life, by stating what is important to them and providing an overall assessment. The analysis of product reviews among customers identifies commonalities in the relationship between what people say is important and their evaluation of the product that are useful for identifying features that could be improved.

A challenge in analyzing the relationship between product reviews and product ratings comes when some statements in a review rationalize the corresponding overall rating while other statements merely replicate the overall evaluation in the rating (e.g., "We had a great time"). The distinction between contributing and reflective statements, or topics, becomes important when quantifying the predictive relationship between reviews and ratings. The inclusion of statements that are merely reflective of the overall evaluations artificially inflates estimates of predictive accuracy and biases downward estimates of the importance of statements rationalizing the rating. At the limit, the estimated importance of statements that rationalize the rating is driven to zero when reflective statements are included in an analysis. This is a problem of model specification that will persist even at the limit of an infinite amount of data. Moreover, supervised models of—in practice necessarily finite—text data that ignore the possibility of reflective statements may fail to uncover a meaningful structure of distinct topics.

The purpose of this article is to develop a model that separates topics that contribute to overall ratings from topics that reflect an overall rating verbally, but without specifying what contributes to the rating. Our model addresses a key challenge in supervised text analysis, where document labels provide overall evaluations. Specifically, it deals with the tension between achieving accurate predictions and finding the most concise representation of text within the analyzed documents. This challenge arises in all analyses of finite datasets that aim to simultaneously reduce dimensionality and make predictions.

Joachim Büschken is Professor of Marketing, Catholic University Eichstätt-Ingolstadt, Germany, and Affiliated Professor, The Ohio State University, USA (email: joachim.bueschken@ku.de). Thomas Otter is Professor of Marketing, Goethe University, Germany, and Nova School of Business and Economics, Portugal (email: otter@marketing.uni-frankfurt.de). Greg M. Allenby is Helen C. Kurtz Chair in Marketing, Fisher College of Business, The Ohio State University, USA (email: allenby.1@osu.edu).

There is a trade-off between reducing the complexity of data X while preserving essential information and creating a simplified representation that effectively predicts labels y. In finite datasets, resolving this conflict based solely on overall fit often leads to dimensions that lack clear individual significance and are difficult to interpret beyond their strong predictive power.[1]

Contributing or formative topics are those that provide a rationale for the rating, whereas reflective topics verbally reexpress the evaluative label. Separation of these is needed to capture the rationale customers provide for their product ratings. Reflective topics are not informative about the rationale for a review, that is, more detailed and concrete evaluations of and experiences with features of a product or service. The goal of our model is to improve topic inference to obtain better insights into the origin of consumer ratings. By definition, reflective topics cannot provide such insights. In our empirical analysis, we find that our data contain multiple reflective topics that talk about consumers' experience in general with little to no reference to particular aspects of the object rated (hotel stay). We also find that our model identifies more coherent formative topics with a stronger relationship to actual product attributes.

Models of text data do not generally allow for restrictions on what the words in a set of documents will describe (Boyd-Graber, Blei, and Zhu 2007; Yang, Boyd-Graber, and Resnik 2017). The problem with specifying a model relating the overall rating to the topics in a text analysis is that it is not clear how to incorporate statements that merely reflect on an overall evaluation in this context. Topics derived from such statements fail to explain a numerical overall evaluation because they may, for example, only replicate the overall evaluation verbally ("This is a great restaurant"). Our integrated model simultaneously identifies reflective topics and conducts inference about separate formative topics and their relationship to the overall rating. In our empirical analysis we show that our dataset contains multiple reflective topics and that accounting for reflective content in textual customer satisfaction data changes inference regarding topics that contribute to the rating. Only contributing (formative) topics offer potential for successful managerial intervention.

Our analysis proceeds as follows. We start by reviewing extant literature that explores the relationship between features of text and consumer evaluations. In the following section, we develop our structural rating-topic model (henceforth SRTM) and outline our estimation approach. The essence of our approach is integration of model selection into the Markov chain Monte Carlo (MCMC) for simultaneous dimension reduction and allocation of topics to be either formative or reflective with respect to the rating, as detailed subsequently. The advantage of our approach is that it adaptively identifies formative and reflective topics by allowing for a theory-based structure between these two topic classes while preserving the key feature of topic models, that is, a reduction in dimensionality when moving from the word space to the topic space. We then provide an overview of the dataset used in our empirical analysis and present results from the empirical application of our model. A discussion of results and future research questions concludes the article.

Literature

Our article builds on research in statistical text analysis and marketing that explores the relationship between textual features of text data and consumer evaluations. An important contribution to the literature is the use of topic models to predict consumer ratings. This analysis is facilitated by latent Dirichlet allocation (LDA) models, which provide accessible high-level summaries of documents (McAuliffe and Blei 2007) that identify sets of frequently co-occurring terms (topics) and the topical composition of documents by way of topic counts or shares. This enables researchers to link topics directly to ratings by using document-topic counts as predictors of the rating in a regression model (Agarwal and Chen 2010; McAuliffe and Blei 2007), lending topics sentiment by way of their (conditionally) positive or negative (or lack of) relationship to the rating. A variant of this idea is to employ (normalized) topic shares as regressors (Büschken and Allenby 2016, 2020) given that topic counts often correlate highly with total word counts, confounding the two quantities in an unconditional analysis (Nguyen, Ying, and Resnik 2013). Li, Wu, and Mai (2019) propose a variant of supervised LDA by assuming that topics are a priori associated with positive or negative sentiment. Table 1 summarizes contributions to this literature relevant to the research at hand. Ansari, Li, and Zhang (2018) augment their topic-rating model with product attributes and user characteristics. Our article departs from this research by considering that topics are a reflection of the rating.

The advantage of LDA models compared with direct approaches (e.g., word counts) lies in the reduction of dimensionality of the model. Using observable word counts on the document level as predictors requires solutions in which the rating regression incorporates design matrices where the number of columns may exceed the number of cases by orders of magnitude. Taddy (2013) proposes to use ratings as priors to unigrams and uses forward regression to predict ratings. His goal was to reduce the (observed) word-document counts to a more manageable number of predictors of the rating. LDA solves this problem by reducing the number of predictors from the size of the vocabulary to the number of topics in a corpus.

An empirical analysis of the relationship between ratings provided by reviewers and topics often reveals that ratings are closely linked to topical features of consumer reviews. For example, from the regression models linking topics to ratings (Table 1), analysts often find strong relationships between topic occurrence and evaluations (Mankad et al. 2016). In other words, ratings provide a priori knowledge on topical composition (labels) of reviews. The standard method in the literature is to use topics as predictors in a regression (supervised

[1] We thank an anonymous reviewer for making us emphasize this point.

Table 1. Literature Linking Features of Text Data to Consumer Evaluations.

	Text Modeling Method	Additional Data	Dependent Variable	Independent Variables	Structural Approach	Role of Rating
McAuliffe and Blei (2007)	Topic modeling	None	Consumer rating	DTC	Regression	Dependent (predicted by topics)
Agarwal and Chen (2010)	Topic modeling	None	Consumer rating	DTC	Regression	Dependent
Nguyen, Ying, and Resnik (2013)	Topic modeling	None	Consumer rating	DTC, WDC	Regression	Dependent
Taddy (2013)	Unigram (conditional on rating)	None	Consumer rating	(Reduction of) WDC	Inverse regression (combined with forward regression predicting ratings)	Conditioning argument for unigrams, predictive target in forward regression
Rabinovich and Blei (2014)	Topic modeling	None	WTP	Rating	Not applicable	Prior information to topic shares
Büschken and Allenby (2016, 2020)	Topic modeling	None	Consumer rating	DTP	Regression	Dependent
Ansari, Li, and Zhang (2018)	Topic modeling	Product attributes, user characteristics	Consumer rating	WTP, product attributes, user characteristics	Regression	Dependent
Li, Wu, and Mai (2019)	Topic modeling	None	Consumer rating	Topic sentiment	Regression	Dependent
Mankad et al. (2016)	Topic modeling	None	Consumer rating	Topic occurrence, document sentiment	Regression	Dependent
Yang, Zhang, and Fan (2023)	Neural networks	None	Consumer rating	Latent node representation of DTC	Regression	Dependent
Chakraborty, Kim, and Sudhir (2022)	Neural networks	Attribute sentiment labels (humans)	Consumer rating	Attribute sentiment (predicted from text)	Latent class regression	Dependent
This article	Topic modeling	none	Topics reflective of consumer rating	Formative topics	Model search (flexible)	Facilitates distinction between formative and reflective topics

Notes: WTP = word-topic probabilities, WDC = word-document counts, DTC = document-topic counts, DTP = document-topic probabilities.

LDA); an alternative approach is to specify prior distributions to word-topic probabilities by way of the rating (Rabinovich and Blei 2014). In both cases, the goal is to exploit a priori knowledge about ratings and topics. When ratings are used, topics simply differ between reviews that give a bad rating and those that give a top-box rating.

More recently, advances in neural network modeling have been proposed to link latent features of text to consumer ratings. Yang, Zhang, and Fan (2023) propose a variational autoencoder model with (total) word counts as inputs and outputs. Nodes in this neural network model represent document-level topic distributions and word-topic distributions, as in LDA. Document-topic distributions are linked to the rating by way of (additional) latent nodes. Thus, their model can be viewed as a supervised LDA with more flexible links between word, topic, and document labels. It integrates the idea of (supervised) topic modeling into the hierarchical latent variable approach of neural networks (Dieng, Ruiz, and Blei 2020).

Chakraborty, Kim, and Sudhir (2022) use such a neural-networking approach to predict unobserved attribute evaluations from text data. This addresses the issue of topics from LDA not necessarily being related to product attributes that managers can adjust in order to improve customer satisfaction. In LDA, topics are only revealed a posteriori by (empirically) co-occurring terms, whatever these may be. It can therefore be difficult to map topics to actual areas of action. Chakraborty, Kim, and Sudhir obtain unobserved attribute evaluations by way of a neural network model that is trained on a sample of (observed) attribute-sentiment scores provided by human readers. Attributes are defined a priori, making the approach more similar to obtaining attribute-level customer satisfaction scores by way of surveys. The model then predicts evaluations of (all) predefined attributes, which are linked to the rating by way of regression, revealing their relative importance. Similarly, Timoshenko and Hauser (2019) propose a neural-network model to (simultaneously) identify consumer needs in a particular product category (e.g., toothbrush) and sentences in a textual dataset that talk about these needs. As in Chakraborty, Kim, and Sudhir, their model is trained on a sample of sentences, some of which talk about needs, as labeled by human readers. The model is then used to predict which sentences outside the training sample are informative about needs.

These approaches are different from ours in several ways. First, and in accordance with LDA-type models, we specify our model as a generative model of (all) text. Neither Chakraborty, Kim, and Sudhir (2022) nor Timoshenko and Hauser (2019) specify a likelihood of the words in a dataset. Second, they avoid the issue of reflective topics by limiting the analysis to sentences that talk about predefined product attributes or needs. In comparison, we consider all (text) data, including topics that merely reflect the rating. We note, however, that the proposed distinction of formative and reflective content in product review data could be integrated into their modeling approach.

To summarize, our research departs from the literature in several ways. First, and most importantly, we allow for unobserved topics to be a mere reflection of the rating, not predictors of the rating in the way of a standard supervised approach to text modeling. We call these topics reflective. Anecdotal, qualitative analyses of textual product and service review data strongly suggest the existence of such topics (e.g., in reviewer statements such as "we had a wonderful time" or "this is a great movie"). As a consequence, our model does not treat all topics a priori as predictors of the rating. We use the rating to facilitate the distinction between reflective topics and (with respect to the rating) predictive topics, which we call formative. To the best of our knowledge, this has not been attempted before. Second, our approach to exploring the origin of consumers' ratings builds on the distinction of formative versus reflective topics: We use only the formative topics to predict the rating. No set of (predefined) attributes (or needs) is necessary to achieve this. Third, we document that this approach results in more coherent topics both in terms of statistical fit to the text and substantively.

Model Development

Overview

Formative versus reflective topics in product review data. Figure 1 provides a stylized overview of our model, explaining the basic motivation behind our modeling approach. For this presentation, we use actual results from our model, using reviews of luxury hotels in Manhattan, New York (see the "Data and Preprocessing" section). In accordance with the LDA model, we assume that topics are unobserved and are to be inferred from data. Each topic is characterized by a vector of word-topic probabilities given the vocabulary in a corpus, and a vector of topic-document probabilities that are specific to the review.

Let us assume that the topics on the left side of the figure (e.g., "room," "noise") are formative of the overall rating, while the topics on the right (e.g., "hotel great," "great local attractions") are reflective, in a stylized preview of our empirical results. The configuration of formative topics on the left side of the figure predicts an overall evaluation in a generalized regression model. So, the arrows on the left originate from the topics and terminate in the rating. The reflective topic "hotel great," in this case, is a mere reexpression of a high rating, and so the arrows begin with the rating and end with the reflective topics listed on the right. When our model identifies a topic as reflective but ex post consideration of topical terms shows that the topic does not qualify as a reexpression of a high rating, our model picks up that reviewers will only muse about this topic conditional on a good hotel experience (assuming a positive connection between the ratings and the prevalence of this topic), as with the topic "great local attractions" in our example. Our model assumes that product reviews may contain a mixture of formative and reflective topics. The goal of our analysis is to jointly identify topics and their role as contributing to the rating (formative) versus reflecting the rating without making an independent contribution as to why a rating is high versus low.

The stylized model displayed in Figure 1 bears resemblance to the classic multiple indicator–multiple causes (MIMIC) model in marketing (Jöreskog and Goldberger 1975). As in the MIMIC model, we assume that relationships between formative and reflective indicators are mediated by a single variable (Bollen 1989), which results in conditional independence of reflective and formative indicators (Fornell and Bookstein 1982). In the MIMIC model, formative and reflective indicators are observed and the mediating variable is assumed to be latent. The distinction between formative and reflective indicators stems from the observation that, under certain conditions, causal indicators may better correspond to the relation of indicators to a (latent) construct than reflective measurement (Bollen and Lennox 1991). More recent contributions to this literature critically focus on these conditions (Bollen and Diamantopoulos 2017; Jarvis, MacKenzie, and Podsakoff 2003; Lee, Cadogan, and Chamberlain 2013; Podsakoff, Shen, and Podsakoff 2006).

Our model is different from the MIMIC approach as it assumes all indicators (i.e., topics) to be latent, only indirectly observed by way of words in reviews, and the mediating variable (rating) to be observed. In a sense, compared with the MIMIC model, our model flips the role of latent and observed variables. Formative topics, in our model, are unobserved and assumed to contribute to the (observed) rating. Reflective topics are equally unobserved and assumed to result from a rating, not contribute to it. More importantly, the split of any set of topics into formative versus reflective is assumed to be a priori unknown and part of our inference approach.

Why does our approach matter?. We use a simulation to illustrate why our approach matters to marketers interested in

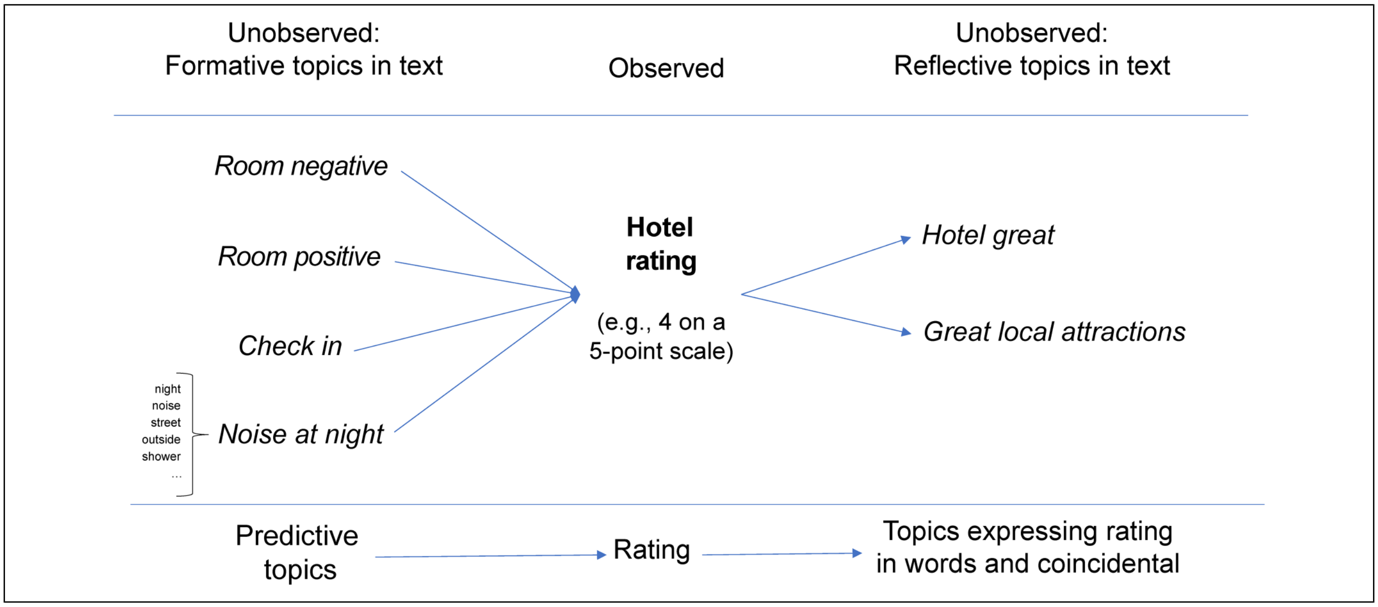

Figure 1. Overview of Proposed Model.
Notes: Top terms for the topic "noise" are shown for illustration.

(conditional) dependencies between topics and evaluative labels. We first abstract away from the dimension reduction inherent to text analysis and proceed as if latent variables corresponding to the prevalence of topics in documents were given, focusing on how reflective topics may bias inference. We note that this bias cannot be overcome by collecting more data. We then revisit the need to jointly conduct dimension reduction and infer (conditional) dependencies when given text data with labels.

Consider the following example:

$$y_1 = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + \epsilon,$$
$$y_2 = \gamma_0 + y_1\gamma_1 + \zeta,$$

where $X \sim N(0, 4I)$ with I being the 4×4 identity matrix, $\varepsilon \sim N(0, .33)$, $\zeta \sim N(0, \psi)$, $\beta' = (-1, 2, -2, 3)$, and $\gamma' = (0, 1)$. Here, the X variables are formative to y_1 while y_2 is reflective. Let y_1 be consumers' ratings. In other words, y_1 originates from X (formed by X), not from y_2.

We generate 100 observations from this model and run regressions using y_1 as the dependent variable and (X, y_2) as covariates, given different values of ψ. We explore the effect of using a reflective indicator of the dependent variable as regressor. Figure 2 shows how inference with respect to β changes as ψ changes. It reveals that all estimates of β go to 0 as ψ goes to 0. In other words, the inclusion of a perfectly reflective indicator of y_1 in the model effectively leads to excluding the actual predictors. It also leads to inflation of the predictive relationship of y_1, because y_2 explains nearly all of its variance. Thus, β estimates of formative indicators will be biased toward zero when the model contains reflective indicators. In this case, estimates of β and σ are inconsistent—adding data to the analysis will not change the bias introduced by including endogenous y_2 among the regressors.

The preceding illustrative simulation assumes away the need to identify variables, except of the label y, by dimension reduction in the applications we study here. In supervised LDA and related models that aim at predicting document labels from text subject to dimension reduction, the likelihood of the label y (the loss associated with predicting the labels y) will contribute information about how to reduce the dimensionality of the text data. In principle, this is helpful in necessarily finite datasets, and especially in situations where individual documents are short. However, a misspecified predictive model can disadvantage the dimension reduction in the sense that latent dimensions (topics) that result in minimal loss in predicting y are both statistically and substantively worse representations of the underlying text.

In an extended simulation study reported in Web Appendix A in which we fit a standard supervised LDA to a corpus containing reflective and formative topics, we demonstrate that ignoring the possibility of reflective topics leads to biased estimates of the model connecting the rating to topics. As in the preceding simulation, the coefficients linking formative topics to the rating go to zero, and the error variance is greatly deflated. We also find that estimates of the topics themselves are biased. In other words, a misspecified model changes how the text is represented in a lower-dimensional space. In our empirical analysis using textual review data (see the "Model-Based Results" section), we find that allowing for reflective topics improves the statistical representation of the text as measured by fit and results in a more coherent topic structure.[2]

[2] Hence, an ex post correction for the presence of reflective topics is impossible in finite datasets consisting of shorter documents where the model for the label y

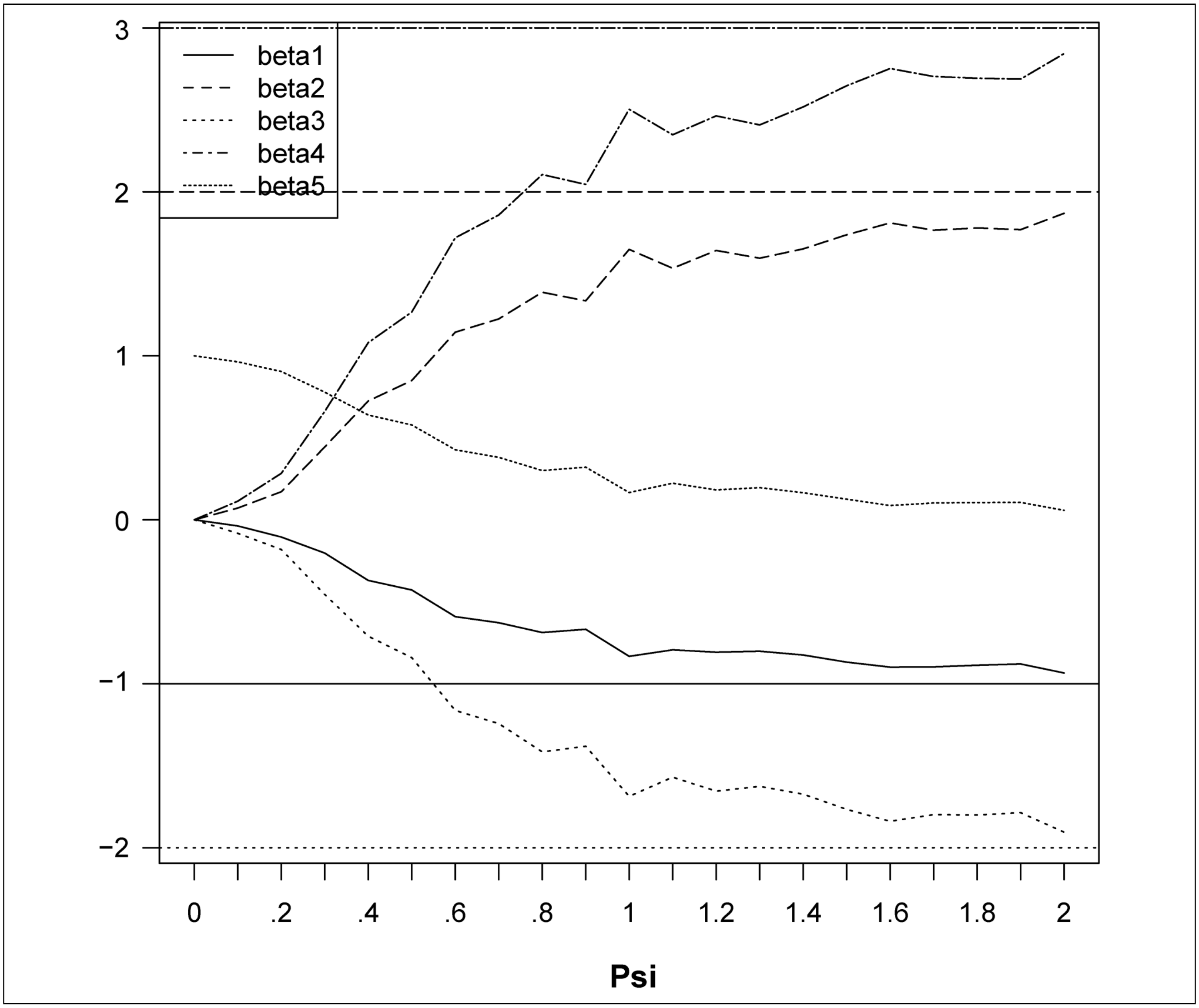

Figure 2. Results from Regression That Includes Reflective Indicator of Dependent Variable as Covariate, Given Different Values of ψ.
Notes: β_5 is the estimated slope of reflective indicator y_2. Horizontal lines indicate true values of β.

Data-Generating Process of the Proposed SRTM

Our model builds on the standard LDA topic model (Blei, Ng, and Jordan 2003; Hofmann 1999), which assumes that the data originate from a mixture of latent topics. The probability of observing word n in document d is

$$p(w_{n,d}|\phi, \theta_d) = \sum_{t=1}^{T} p(w_{n,d}|z_{n,d} = t, \phi_t)p(z_{n,d} = t|\theta_d) = \sum_{t=1}^{T} \phi_t(w_{n,d})\theta_{t,d}, \quad (1)$$

where $z_{n,d}$ indicates the topic assignment of word $w_{n,d}$, ϕ_t is a topic-specific vector of word probabilities over a (fixed) vocabulary of size V, and θ_d are document-specific prior probabilities of $z_{n,d}$, defined over a (fixed) number of topics T. The set of topic-specific word probabilities $\{\phi_t\}_{t=1}^{T}$ is collected in ϕ, a matrix of dimensionality V × T.

Our model assumes the following data-generating process, given an a priori fixed number of topics T and size of vocabulary V:

1. For each topic, draw word probabilities $\phi_t \sim \text{Dirichlet}(\alpha)$, a vector of length V. The vector α is a fixed prior vector of (homogeneous, positive) scalar entries of length V, resulting in a symmetric Dirichlet prior.
2. For each topic, draw (binary) indicator $I_t \sim \text{Bernoulli}(\delta)$ with δ being the (a priori fixed) prior probability of $I_t =$

materially influences the lower-dimensional representation of the text (i.e., the topic structure).

1. Let $I_t = 1$ indicate a reflective topic and $I_t = 0$ a contributing topic with respect to the rating.

3. Draw (scalar) error variance $\sigma^2 \sim IG(a_\sigma, b_\sigma)$, with a_σ and b_σ as fixed prior quantities. Then draw $\beta \sim MVN(0, \sigma^2\Sigma_\beta)$, a vector of regression coefficients of length equal to the number of formative topics plus an intercept. The mean and the covariance Σ_β are fixed prior quantities.
4. For each reflective topic, draw a scalar error variance $\psi_t \sim IG(a_\psi, b_\psi)$ with a_ψ and b_ψ as fixed prior values. Then draw $\lambda_t \sim MVN(0, \psi_t\Sigma_\lambda)$, where the mean and the covariance Σ_λ are again fixed prior quantities. Each λ_t is a vector of length two containing an intercept ($\lambda_{0,t}$) and a slope coefficient ($\lambda_{1,t}$). Collect all λ_t in matrix Λ, a matrix of coefficients with dimensionality two times the number of reflective topics.
5. For each document (i.i.d.):
   a. Draw vector $\theta_d^{*^{(form)}} \sim MVN(\mu_{*(form)}, \Sigma_{*(form)})$ of length equal to number of formative topics with $\mu_{*(form)}$ and $\Sigma_{*(form)}$ as (a priori fixed) parameters of dimensionality given by the number of formative topics.
   b. Draw scalar $\varepsilon_d \sim N(0, \sigma^2)$.
   c. Compute (continuous) scalar rating $r_d = \beta^T[1, (\theta_d^{*^{(form)}})]^T + \epsilon_d$; potentially combine with cut points (from an ordered uniform prior) to produce ordinal ratings.
   d. For each reflective topic, draw (i.i.d.) scalar-valued $\rho_{t,d} \sim N(0, \psi_t)$; collect in vector ρ_d.
   e. Compute $\theta_d^{*^{(refl)}} = \Lambda^T[1, r_d]^T + \rho_d$.
   f. Concatenate $\theta_d^* = (\theta_d^{*^{(form)}}, \theta_d^{*^{(refl)}})$ and, for each topic, compute $\theta_{d,t} = \frac{\exp(\theta_{d,t}^*)}{\sum_{texp}(\theta_{d,t}^*)}$.
   g. For each word n in document d (and continue as in LDA):
      i. Draw topic indicator $z_{n,d} \sim Multinomial(\theta_d)$.
      ii. Draw word $w_{d,n} \sim Multinomial(\phi_{z_{n,d}})$.

Next we discuss the two distinguishing components of our model in detail. The first component relates topic probabilities to overall ratings, using a regression model:

$$r_d = \beta_0 + \sum_{t:I_t=0} \theta_{d,t}^{*^{(form)}} \beta_t + \epsilon_d. \tag{2}$$

Here, r_d is the (scalar) rating associated with document d, ε_d is a (scalar) error component, and β weighs document-specific θ_d^* with respect to their contribution to the overall rating r_d. The superscript "(form)" indicates that this part of the model treats all formative topics symmetrically, consistent with (a priori) nonredundant contributions of individual topics to the overall rating (as in supervised LDA, with the notable difference that supervised LDA assumes that all topics are formative by default). When customers provide overall evaluations in the form of ordered categories, such as star ratings or similar, as is typical, an ordered probit model recovers underlying continuous evaluations r_d (Johnson and Albert 2006):

$$y_d = k \qquad \text{if} \qquad c_{k-1} \le r_d \le c_k. \tag{3}$$

Here y_d is the observed rating and c a vector of cut points mapping underlying continuous evaluations r_d to ordinal scale categories $k \in (1, \ldots, K)$.

The second component of our model allows for some topics to be reflective of the overall rating using a factor model where the topic probability is an outcome of the rating:

$$\theta_d^{*^{(refl)}} = \lambda_0 + \lambda_1 r_d + \rho_d. \tag{4}$$

Here, ρ_d is a vector of residuals, λ_1 is a vector of loadings, and λ_0 is a vector of intercepts. Topics and document-specific topic distributions are latent. As a result, topic–word relations and partitioning of topics into those that contribute to an overall rating (form) versus those that verbally reexpress this overall rating (refl) are subject to inference that is described subsequently. Figure 3 displays the directed acyclic graph of our model. Note the deterministic relationships between I and the splitting of θ^* into formative and reflective topics as well as between θ^* and θ, with the latter generated from the former by way of the softmax transformation.

Note that neither I nor the resulting split into formative and reflective topics are known a priori. Next, we explain how we approach the problem of inferring I from data. The essence of our inference strategy is to respecify our model as a super-model that allows for all topics to enter both Equations 2 and 4 and to constrain coefficients in this super-model to 0, given estimates of I, similar to variable selection in a regression model (George and McCulloch 1995). In other words, when a topic is identified as reflective, its β becomes 0, and when a topic is identified as formative, its λ becomes 0. This facilitates model selection (inference regarding I) by way of stochastic search (George and McCulloch 1995).

A Super-Model Framework for Topic Partitioning

Our estimation framework builds on an oversaturated super-model that simultaneously includes all topics in Equation 2, that is, on the right-side of the equation, for observed numeric document labels, and also those in Equation 4, that is, on the left-hand side. Thus, we associate each of the t = 1, …, T topics with an element in the indicator vector $I = (I_1, \ldots, I_T)$. Elements in this vector determine whether topic t reflects the overall assessment ($I_t = 1$), on the left-hand side, or combines into the overall assessment ($I_t = 0$), on the right-hand side. The indicator vector I is unobserved, and thus the parameter in our model that maps from the oversaturated super-model into the space of identifiable models with extreme points $(T^{(refl)} = \sum_{t=1}^T I_t = T, T^{(form)} = \sum_{t=1}^T (1 - I_t) = 0)$ and $(T^{(refl)} = 0, T^{(form)} = T)$. Note that $(T^{(refl)} = T, T^{(form)} = 0)$ corresponds to the most constrained, least informative situation, in which inferred topics only replicate overall assessments. The other extreme, $(T^{(refl)} = 0, T^{(form)} = 1)$, corresponds to the implicit assumption made in supervised LDA and other supervised models (Büschken and

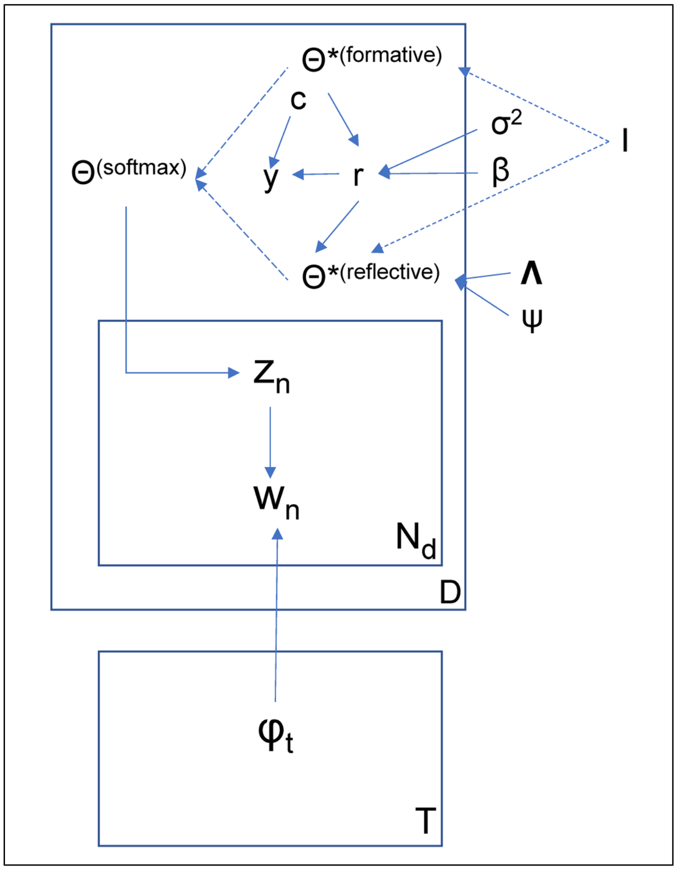

Figure 3. Directed Acyclic Graph of Proposed SRTM.
Notes: Dashed arrows indicate deterministic dependence. Plates indicate replications for D documents, N_d words per document, and T topics. Fixed prior distributions are omitted to reduce clutter.

Allenby 2020; McAuliffe and Blei 2007; Yang, Zhang, and Fan 2023).

The 2^T models implied by different configurations of I correspond to the following constraints of parameters in the underlying super-model expressed in the form of prior distributions for regression coefficients β and loadings λ in Equations 2 and 4:

$$p(\beta|\sigma^2, V_\beta, I) = \prod_{t=1}^{T} p(\beta_t|\mu_{\beta_t} = 0, \sigma^2 V_{\beta_t}, I_t),$$

where

$$V_{\beta_t} = \begin{cases} V_{\beta_t} > 0 \quad \text{if} \quad I_t = 0, \\ V_{\beta_t} \to 0 \quad \text{if} \quad I_t = 1, \end{cases}$$

and where V_{β_t} is the prior variance of β_t. This prior shrinks β_t to its prior expectation 0 when $I_t = 1$, that is, when the t^th^ topic only verbally reexpresses an overall sentiment or evaluation. Similarly, the prior for λ shrinks $\lambda_{0,t}$ and $\lambda_{1,t}$ to prior expectations 0 when $I_t = 0$, that is, when the t^{th} topic combines into an overall evaluation.

$$p(\lambda|\psi, V_\lambda, I) = \prod_{t=1}^{T} p(\lambda_t|\mu_{\lambda_t} = 0, \psi_t V_{\lambda_t}, I_t),$$

where

$$V_{\lambda_t} = \begin{cases} V_{\lambda_t} > 0 \quad \text{if} \quad I_t = 1. \\ V_{\lambda_t} \to 0 \quad \text{if} \quad I_t = 0. \end{cases}$$

Since each indicator I_t can only take values 0 or 1, our approach effectively allocates each topic either to Equation 2, the "right-hand side" of the model for the label, or Equation 4, the "left-hand side" of that model. When $I_t = 1$, β_t is zero, and the t^{th} topic does not independently contribute to the overall sentiment or evaluation in r. Similarly, when $I_t = 0$, λ_t is zero, and the t^{th} topic cannot be (another) expression of overall sentiment or evaluation. An extension of our model could explicitly account for topics that neither independently contribute to observed ratings nor reflect these ratings. However, we cover this possibility a posteriori upon finding combinations of $I_t = 0$ and β_t concentrated around zero, or $I_t = 1$ and λ_t concentrated around zero, or a mixture of both in the posterior. Note that the standard approach to linking topics to ratings is a special case of our model: $I_t = 0, \forall t$.

In our analysis, we use $V_{\beta_t} = a$ and $V_{\lambda_t} = abI_2$ if $I_t = 1$ (reflective), and $V_{\beta_t} = ab$ and $V_{\lambda_t} = aI_2$ if $I_t = 0$ (formative), where $a = 10^{-5}$ and $b = 10/a = 10^6$ and I_2 indicates the two-dimensional identity matrix (George and McCulloch 1995). A slab-and-spike prior can be obtained by setting a to a value arbitrarily close to 0. The effect of this specification of prior distributions of β and λ given I is that formative topics are effectively removed from the reflective model and reflective topics are effectively removed from the formative model. Finally, note that by shrinking λ_t to zero when a topic is classified as formative ($I_t = 0$), what remains of the reflective model for this topic, that is, $\theta^*_t \sim N(0, \psi_t)$, implicitly defines a prior distribution for this topic in the formative model. Web Appendix B illustrates the empirical identification of our model using simulated data.

The partitioning of topics implied by our model contributes to structuring the dependence between topics across documents. In turn, this structure can aid in the identification of more meaningful topics from finite amounts of data at the document level. Different from the specification of a more flexible dependence structure as proposed by Blei and Lafferty (2007), the dependence structure implied by our model derives from linguistic arguments. For example, upon recognizing that two different topics both express overall sentiments, one positive and one negative, our model implies that these two topics cannot be jointly present in documents with a very good or a very bad observed rating.

In Web Appendix C, we demonstrate by way of an additional simulation study that topic inference is biased when all topics are falsely assumed to be formative. In this analysis, we generate data from formative topics contributing to an observed numerical rating and reflective topics arising from that rating. We show that, when a standard supervised approach is applied, topic inference and inference regarding the contribution of topics to the rating are biased.

Joint Distribution and Identification

Joint Distribution and Factorization into Sampling Steps

We arrive at the following joint distribution of knowns and unknowns (suppressing indices n, d, and t):

$$p(y, w, r, z, \phi, \theta^*, \theta, \beta, \lambda, \sigma^2, \psi, I, c) =$$

$$p(y|r, c)\, p(w|z, \phi)\, p(z|\theta)\, p(\phi|\alpha)\, p(\theta^*|r, \lambda, \psi, I)\, p(r|\theta^*, \beta, \sigma^2, I)$$

$$\times\, p(\lambda|V_\lambda, \psi, I)\, p(\beta|V_\beta, \sigma^2, I)\, p(\psi, \sigma^2, I, c),$$

where ψ and σ^2 are variances of the error components and V_β and V_λ are fixed quantities, given I. As usual, we assume independent prior distributions: $p(\psi, \sigma^2, I, c) = p(\psi)p(\sigma^2)p(I)p(c)$.

The essence of our model is to impose a (cross-sectional) structure on the topics based on the rating. Topic allocation is achieved by way of a mixture prior for β and λ. As shown previously, θ_d can be obtained from $\theta_d^{*(form)}$ and $\theta_d^{*(refl)}$ in a deterministic fashion. Inference regarding I is key to estimating our model. Given I, θ_d^* is separated into formative and reflective topics, and updates of coefficients of the structural rating model (β, λ, σ, ψ) can proceed i`ndependently, conditional on r.

The conditional independencies characterizing the joint distribution give rise to the following complete set of conditional distributions:

1. $p(I|else) \propto p(\theta^*|r, \lambda, \psi, I)p(r|\theta^*, \beta, \sigma^2, I)$

   $p(\lambda, \psi|V_\lambda, I, a_\psi, b_\psi)p(\beta, \sigma^2|V_\beta, I, a_\sigma, b_\sigma)p(I),$

2. $p(\beta, \sigma^2|else) \propto p(r|\theta^*, \beta, \sigma^2, I)p(\beta, \sigma^2|V_\beta, I, a_\sigma, b_\sigma),$

3. $p(\lambda, \psi|else) \propto p(\theta^*|r, \lambda, \psi, I)p(\lambda, \psi|V_\lambda, I, a_\psi, b_\psi),$

4. $p(z|else) \propto p(w|z, \phi)p(z|\theta),$

5. $p(\theta|else) \propto p(z|\theta)p(\theta^*|r, \lambda, \psi, I)p(r|\theta^*, \beta, \sigma^2, I),$

6. $p(\phi|else) \propto p(w|z, \phi)p(\phi),$

7. $p(r|y, c, else) \propto p(y|r, c)p(\theta^*|r, \lambda, \psi, I)p(r|\theta^*, \beta, \sigma^2, I),$

8. $p(c|y, r, else) \propto p(y|r, c)p(c).$

We detail our MCMC estimation procedure for all steps in Web Appendix A. To generate I, that is, to draw from the posterior of all possible configurations of indicators identifying topics as formative or reflective, we combine Steps 1 to 3 into a single joint move. This way we overcome the (deterministic in the limit) link between I and "essentially zero" elements of β and λ preventing conditional sampling from these steps. Our approach is facilitated by conditional independence between the formative part in Equation 2 (Step 2) and the reflective part of our model in Equation 4 (Step 3). Conjugate priors in the form of $p(\beta, \sigma^2 \mid V_\beta, I, a_\sigma, b_\sigma) = IG(\sigma^2 \mid a_\sigma, b_\sigma)N(\beta \mid 0, \sigma^2 V\beta(I))$ (and analogously for $p(\lambda, \psi \mid V_\lambda, I, a_\psi, b_\psi)$) allow for a direct evaluation of the normalized posterior density. We then use the candidate's formula to obtain the normalizing constants of the formative and the reflective model as ratios of the respective nonnormalized densities over their normalized counterparts (evaluated at direct draws from the corresponding conditional posteriors). The normalizing constant required for sampling from the posterior distribution of I is then the product of these two normalizing constants, again as a consequence of conditional independence.

Identification

Empirical identification of the model structure comes from two sources: first, the word co-occurrences in the corpus to be rationalized by the topic model, and, second, the relationship among the topics and between the rating and the topics across documents. Different from LDA and supervised LDA, the proposed model implies that reflective topics are independent conditional on the rating (see Equation 4). In contrast, conditioning on the rating introduces dependence between formative topics, as in an indifference curve or surface that traces out combinations of formative topics while keeping the rating constant.[3]

More specifically, the proposed model implies that reflective topics are independent from formative topics conditional on ratings as in a model of full mediation (e.g., Laghaie and Otter 2023) and hence that the covariance of θ* can be fully rationalized by as many factors as there are formative topics (plus one from the error term in the formative regression model). Importantly, the rank reduction (in the covariance structure) implied by the presence of reflective topics as well as the conditional independence between formative and reflective topics can be assessed before uniquely identifying the associated model parameters. This is because conditional independence and the rank reduction (implied by a particular classification of topics into reflective and formative) occur for any draw of parameter values from a continuous prior.

The rank reduction in the covariance structure is another consequence of conditional independence given the overall rating and will contribute additional overidentifying restrictions once two or more topics are classified as reflective. The covariances between the t^th reflective topic score in θ* and *all formative* topic scores in θ* are equal to the product of formative regression coefficients and this reflective topic's loading $\lambda_{1,t}$. The covariances between the t^th reflective topic score in θ* and *all other reflective* topic scores in θ* are proportional to the product of this topic's loading $\lambda_{1,t}$ and those of all other reflective topics.[4] Together, this implies that covariances between any two reflective topic scores and all other topic scores will be perfectly linearly dependent, resulting in a rank reduction (factor structure) of covariances from classifying two or more topics as reflective. However, even with only

[3] As already mentioned, topic shares θ are dependently distributed across documents because of the adding-up constraint. All statements about dependence and independence here apply to latent θ*, before imposing the simplex normalization.

[4] The covariance matrix of reflective topic scores is proportional to the outer product of the corresponding vector of loadings and hence of rank 1 for any number of reflective topics.

one reflective topic, conditional independence between reflective and formative topics already overidentifies (relative to supervised LDA or standard, unsupervised LDA). The error term associated with the rating captures unmentioned formative topics in this context.

The covariance rank reduction from reflective topics and the implied conditional independence relations also distinguish the model qualitatively from a model that treats all topics as formative (as in supervised LDA) and from standard, unsupervised LDA, which has implications for what substantive topics can be identified from the data. As we will demonstrate subsequently, the implied prior over the joint distribution of topics has important implications for what topics can be identified from the data. Intuitively, a topic that only verbally reflects the overall evaluation in the rating should be independent from all other topics conditional on that rating. And the model's ability to isolate statements that merely reflect the overall evaluation and associate them with a topic is uniquely facilitated by allowing for the possibility of reflective topics.

However, conditional on a model structure, the implied distribution of θ* is only partially identified. To see this, consider that the structural model is specified on the level of unconstrained θ*, which are (deterministically) related to document-level topic proportions θ by way of the softmax transformation. Identification of θ is provided by the LDA part of the model. However, θ resides on the simplex, implying that each θ_d contains only T − 1 independent topic shares. More formally:

$$\theta_{d,t} = 1 - \sum_{-t} \theta_{d,t} \quad \forall t.$$

The softmax mapping from θ* to θ implies that we can only identify θ* up to an additive constant. In addition, this deterministic relationship implies that what can be empirically identified about the covariance structure of θ* is limited by the rank deficiency of the θs' covariance from the adding-up constraint. Hence, we need constraints to uniquely identify parameters in the mean and covariance structure of θ*.

We point identify model parameters, fixing the intercept ($\lambda_{0,t}$) and the slope ($\lambda_{1,t}$) of one reflective topic to zero and one, respectively. Here, the first constraint identifies the mean of θ*, and the second constraint is due to the rank deficiency of the covariance of the identified θs (which are obtained as the softmax mapping from θ*). In addition, we apply the usual constraints to identify the scale and location of latent r. Specifically, we fix the smallest and the largest (finite) cut point relating latent continuous ratings to observed ordinal ratings, implicitly defining proper uniform priors for the remaining cut points in between. Together, these constraints point identify all remaining parameters in our model.

As already mentioned, the classification of topics into formative and reflective can proceed in the absence of point identification. This is useful because it allows for imposing the constraints $\lambda_{0,t} = 0$ and $\lambda_{1,t} = 1$ for some topic t classified as reflective, which obviously requires knowledge about the classification of topics, only after learning this classification from the data. Empirically, we first learn the topics jointly with the conditional independence structure between topics given the ratings in an unidentified space with respect to the distribution of θ* and hence without point identifying intercepts and slope parameters. Upon identifying topics and their interrelationship, we apply the constraints necessary for point identifying parameters to a topic reliably classified as reflective, as follows:

1. Run the MCMC until the sampler converges to reliably identified topics allocated to be formative or reflective. Then select one reflective topic.
2. Set $\lambda_{t,0} = 0$ and $\lambda_{t,1} = 1$ for this topic for point identification of parameters in all possible models where the topic chosen for identification is classified as reflective. The choice of t is arbitrary as long as the corresponding topic is reliably classified as reflective.
3. Run MCMC with λ_t fixed for this topic until convergence of all remaining estimable parameters is achieved.

Empirical Analysis

Data and Preprocessing

In our empirical analysis, we use a dataset of (initially) 3,481 reviews of luxury hotels in Manhattan, New York. The dataset was obtained from Expedia. Hotels were selected because of their similarity in terms of hotel category rating (five stars, luxury category), location (all in downtown Manhattan, near Times Square/Broadway), and price point. We view this dataset as an example for (targeted) analysis at the intersection of product category, location, and price point. A priori, it seems reasonable to assume that a selection of reviews based on such criteria would facilitate an analysis centered around few, but focused topics. Table 2 shows summary statistics of the reviews after preprocessing. In preprocessing, we first eliminated stopwords, changed capital letters to small letters, omitted punctuation, and eliminated words appearing ten or fewer times in the corpus. Words were not stemmed (e.g., note the difference in meaning of "accommodating" and "accommodation"). In our model-based analysis, we exclude reviews containing fewer than ten tokens. This leads to 2,264 reviews remaining in the analysis.[5] In analyses not reported here, we find that short reviews are typically made up of a single topic. When the share of single-topic reviews in a corpus becomes too large, the empirical distinction between formative and reflective topics becomes tenuous.

After preprocessing, the hotel reviews initially contained 1,078 unique terms. Prior to model-based analysis, we investigated the frequency of terms and found the typical exponential distribution of (marginal) word counts and n-grams. Bigrams (e.g., "front desk") and trigrams (e.g., "within walking distance") specifically exhibited extreme distributions: Relatively few bigrams and trigrams accounted for nearly 100% of all

[5] We (initially) use the same set of luxury hotel reviews as in Büschken and Allenby (2020) but only consider reviews with more than ten word-tokens after preprocessing.

occurrences. For better topic interpretability, we identified the 200 and 100 most frequent bigrams and trigrams, respectively, in our data and added these to our vocabulary for 1,378 unique terms in total (Table 2). We then modified the reviews accordingly. For example, if the sequence "within walking distance" (which is a top trigram) appeared in a review, we concatenated these words ("within-walking-distance") and treated the sequence as a single (unique) token (Timoshenko and Hauser 2019). We did not do this if the observed sequence was, for example, "within easy walking distance" because it is not a (top) trigram. Adding n-grams enlarges the vocabulary while reducing the total count per token. By design, the n-gram tokens in our data contain the context in which words are used, lending lists of top topic terms more contextual information (see Web Appendix D). After our preprocessing, reviews in our dataset contained 29 tokens on average, with a corpus size of 66,241 tokens (Table 2). The reviews range in length from 11 to 228 words, indicating large differences in linguistic complexity and heterogeneity in topic composition. The distribution of ratings exhibits the typical skew toward top-box ratings.

Table 2. Descriptive Statistics of Product Review Dataset (After Preprocessing and Eliminating Short Reviews).

	Luxury Hotel Reviews
Number of reviews (ten words or more)	2,264
Total number of words (corpus size)	66,241
Number of unique terms	1,378
Number of words per review	
Mean	29.3
SD	19.0
Max	228
Number of sentences per review	
Mean	5.3
SD	3.2
Consumer rating (five-point scale)	
Mean	4.38
SD	.90

Model-Based Results

We apply our model to the luxury hotel reviews, using a fixed number of topics. We compare different topic solutions based on predictive fit, using a standard supervised LDA, and find $T = \{10\}$ to be optimal. Note that a plain-vanilla LDA yields similar results. We find that a larger number of topics only improves in-sample fit, indicating that such models overfit the data. We start by comparing our model to several benchmark models in terms of fit and in terms of its capacity to generate unique and coherent topics.

Comparing model fit. In this section, we compare the fit of the proposed SRTM to benchmark models. A natural benchmark for our model is the (standard) supervised LDA, which assumes all topics to be formative with respect to the rating. In our implementation, the supervised LDA is a constrained version of the SRTM with $I_t = 0 \ \forall t$. We also compute fit for a plain-vanilla unsupervised LDA and an LDA with a single topic only, both of which do not use the rating for topic inference. Comparing fit results from these two models reveals the contribution of introducing a mixture of topics. Comparing LDA and supervised LDA reveals the contribution of the rating as likelihood information to topic identification and topic shares. Comparing the supervised LDA and SRTM reveals the contribution of allowing for topics to be reflective of the rating. Our set of benchmark models is limited to LDA-type models of text. However, we believe that our model for the joint distribution of topics and ratings across documents will be useful regardless of what lower-level model of text is assumed. We revisit this point in our discussion.

Table 3 shows fit results for all models. Reported is the log marginal likelihood (LML) of observing all words in the hotel review corpus. LML is a measure of how well a model fits the data after integrating over its parameter space. Integration is achieved by computing the likelihood of all words in the corpus, given draws from the posterior distribution of unobserved quantities (Rossi, Allenby, and McCulloch 2005). LML allows for a direct comparison of models fitted to the same data that differ with respect to their parameter space.[6] To assess predictive fit, we report perplexity, a measure of how surprised a model is after seeing new data. We follow the approach in Blei, Ng, and Jordan (2003) to compute (per-word) perplexity and use 1,247 independently obtained customer reviews of luxury hotels in Manhattan (same source, location, rating, and brands) as holdout data. A lower value of perplexity indicates better generalization performance. Table 3 reveals that the proposed SRTM fits our data best with respect to both in-sample and out-of-sample fit. In the following section, we present results from the SRTM and show that the corpus contains multiple reflective topics that are not predictors of the rating. From Table 3, we also find that a standard supervised LDA performs about on par with a plain-vanilla, unsupervised LDA. This implies that assuming that all topics contribute to the rating does not improve representation of the corpus by the model. The slightly worse perplexity measure from supervised LDA relative to LDA is consistent with a tension between dimension reduction and prediction of ratings in this model.

Comparing topic coherence. A key feature of topic models is that they identify sets of frequently co-occurring terms describing the themes that authors use in their reviews. Ideally, topics identify unique and coherent sets of terms that describe the context in which words are being used. In the following, we compare the proposed SRTM's capacity to identify unique and coherent topics with that of benchmark models. We apply the following measures to evaluate topics:

[6] See Otter, Gilbride, and Allenby (2011) for the derivation and the discussion of what is effectively an LML conditional on the different *models* for the observed ratings in case of supervised LDA and the SRTM, including the special case of the absence of such a model in case of the LDA.

Table 3. Model Fit.

	LDA (T = 1)	LDA	Supervised LDA	SRTM
LML	−6.014	−5.337	−5.332	−5.234
Perplexity	124.07	47.75	48.39	46.66

Notes: Displayed is the (per-word) log marginal likelihood (LML) of the calibration data and perplexity of words in the holdout data, given the model.

- **Topic coherence.** Topics are viewed as semantically coherent when their top words appear jointly in the documents analyzed. To assess local co-occurrence, we use word embeddings from a word-to-vec model (Mikolov et al. 2013a, 2013b). Similar embeddings of top topic terms indicate that terms actually appear in a local context within reviews (Fang et al. 2016).
- **Topic uniqueness.** Topics are unique when their characteristic terms are not redundant or shared among topics (Pergola, Gui, and He 2020). We use counts of top term overlap among topics to assess topic uniqueness with respect to top terms. We also compute correlation of ϕ, which assesses similarity (i.e., nonuniqueness) of word-topic probabilities across the whole vocabulary. In general, less overlap of top terms and lower correlation of (all) word-topic probabilities indicate a more unique topic solution.

We assess semantic topic coherence by way of comparing embeddings of top topic terms. Word embeddings are high-level representations of the local context in which words appear. Higher similarity of embeddings of topic terms indicates that the top words from a topic generate similar contextual terms (Aletras and Stevenson 2013; Schütze 1998). We obtain embeddings using the same vocabulary extended by the most frequent bigrams and trigrams. We employ a skip-gram model with a 2 + 2 window surrounding focal words and 100 latent nodes linking focal tokens to context tokens. Given topic results (i.e., posterior means of $\phi_t \mid$ Model), we then compute cosine similarity of the top topic term pairs. If desired, topics can then be evaluated by averaging their top terms' pairwise similarity scores (Newman et al. 2010). Alternatives to using embeddings are using observed (normalized) counts of term pair co-occurrences within documents or using rolling windows of contextual terms (pointwise mutual information). Coherence measures such as UCI and UMass coherence scores are based on that approach (Mimno et al. 2011). The advantage of using model-based embeddings instead of counts lies in the projection of focal terms onto context terms through a flexible model structure that projects focal terms onto the whole vocabulary.

Figure 4 displays the distribution of all $(20 \times (20 - 1)/2 \times T)$ similarity scores of the top 20 words over topics, given a model. Figure 4 reveals that the SRTM shifts the distribution of the embedding of similarity across the range of scores, compared with a standard supervised approach.[7] It also suggests that topic coherence does not improve when the rating is used as information on document-topic probabilities in the standard way, compared with a simple LDA.

Figure 5 shows the distribution of topic correlation (Panel A) and topic overlap scores (Panel B) among topics for the SRTM and benchmark models. A topic solution given T topics gives rise to $T(T - 1)/2$ correlations and overlap scores. A set of unique topics from a model is characterized by a distribution of topic correlations (or overlap scores) skewed to the right and with a mode close to 0, indicating that topics are uncorrelated (do not share top terms). From Figure 5, we find that the proposed SRTM generates topics that exhibit both less overlap of top terms and lower correlation across the whole vocabulary. In fact, it pushes topic correlation practically to 0. In summary, we find the SRTM to outperform benchmark models with respect to all measures of topic coherence and uniqueness considered here.

Results from SRTM

Model structure. In the following, we investigate the a posteriori structure of the SRTM, given our hotel data. For point identification of parameters, we apply the strategy outlined in the "Identification" section. Table 4 displays the posterior mean of I_t, that is, the assignment of topics as reflective for the hotel reviews. A posterior mean of $I_t > .5$ implies a posterior mode allocation of a topic as reflective ($I_t < .5$ implies formative).

From Table 4 we find that the hotel reviews contain two reflective topics, both of which are reliably identified as indicators of the rating (I = 1). The reliability of identification of formative topics ranges from 1 (T = 5) to .96 (T = 10). For all topics, the posterior mode of I_t is well away from .5. The presence of reflective topics in our data calls into question the standard way of connecting the rating to topics in textual customer satisfaction data (Büschken and Allenby 2020; McAuliffe and Blei 2007; Yang, Zhang, and Fan 2023). The implication for a supervised LDA model is that inference regarding origin of the rating may be biased (see the "Model Development" section). Across all reviews in the hotel data, we find reflective topics to account for 38% of the corpus, implying that, in general, hotel reviews are rich in content that does not independently contribute to explaining the rating, including verbal representations of ratings. We also find that the number of words originating from reflective topics correlates highly (.45) with the total number of words in a review.

Model estimates. Table 5 displays posterior estimates of the coefficients of the structural rating model, conditional on the posterior mode of I. Here, λ_0 denotes the intercepts of the reflective Equation 4, and β_0 denotes the intercept of the

[7] In a comparison of semantic coherence of top topic terms from our hotel reviews without adding bigrams and trigrams to the vocabulary, we find that the advantage of the SRTM is even larger. Apparently, adding contextual information by way of (top) term concatenation reduces the relative performance of our model.

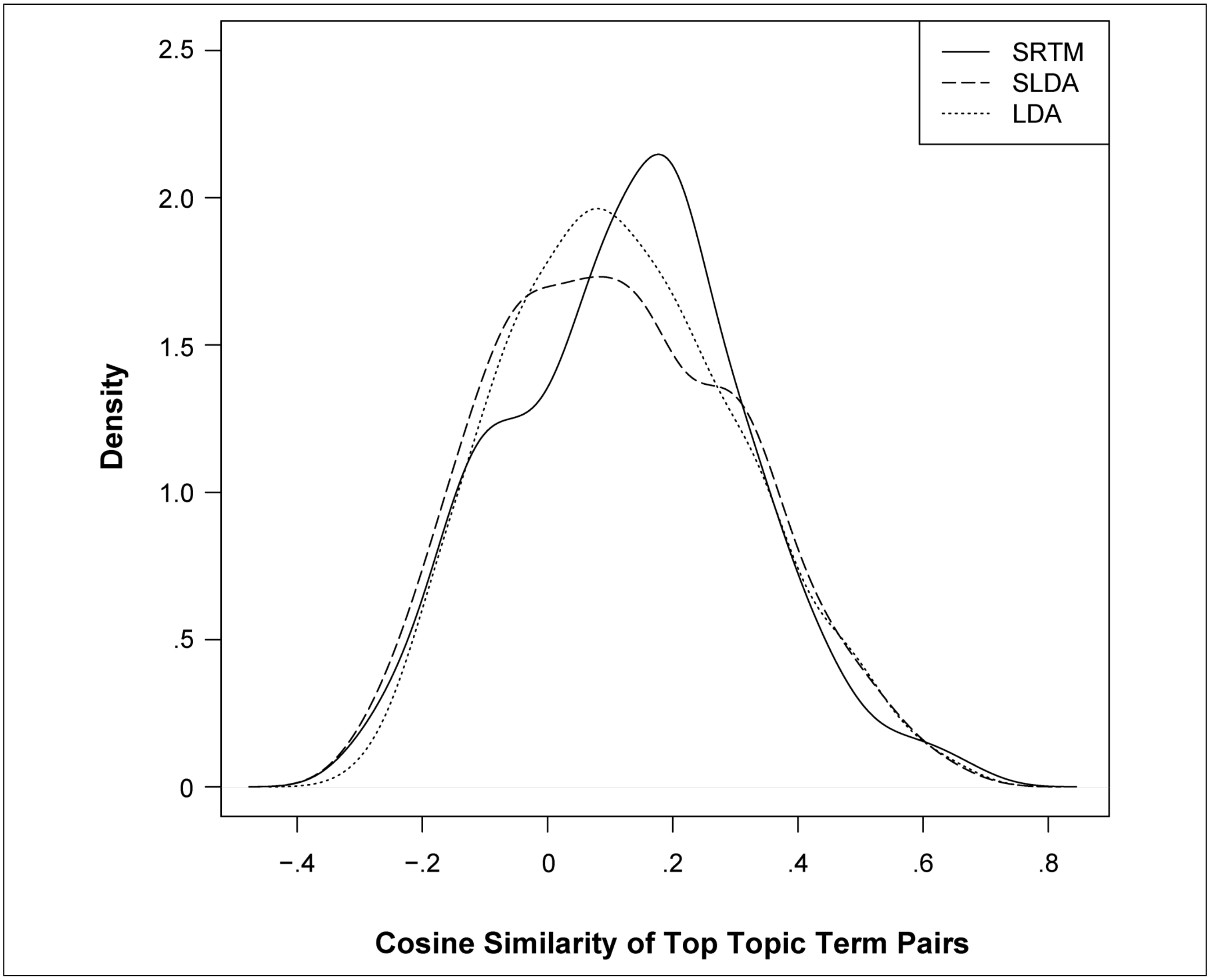

Figure 4. Distribution of Semantic Similarity of Top 20 Topic Words.
Notes: The figure shows similarity scores of words obtained from skip-gram word embeddings. SLDA = supervised LDA.

formative Equation 2. Note that when a topic is classified as reflective ($I_t = 1$), the corresponding coefficient β_t shrinks to 0. When a topic is classified as formative ($I_t = 0$), both $\lambda_{0,t}$ and $\lambda_{1,t}$ shrink to zero. We find reflective topics to be associated with positive ratings only (Table 5). This implies that reviewers present content associated with negative ratings using a formative, argumentative approach, combining positive and negative aspects of their experience into an overall assessment. In other words, reviewers presenting a critical account of their experience typically formulate a structured argument for its defense and not merely a (short, bad) overall verbal assessment.

Equation 2 of our model allows us to investigate the relationship between ratings and (formative) topics (Table 5). A first result from our data is the presence of multiple formative topics with coefficients credibly different from zero and exhibiting positive and negative signs.

Topic most predictive of consumer rating. In Figure 6, we show word clouds of the topic most predictive of the rating from the SRTM and the supervised LDA. Interest in analyzing this topic results from the model identifying the potentially most potent managerial intervention to improve customers' evaluations. From the SRTM, this is the formative topic ($I = 0$) with the largest (absolute) β. From the supervised LDA, this is the topic with the largest (absolute) β across all topics.

The standard supervised LDA (Figure 6, Panel B) identifies as the top predictor of the rating a topic that talks about a ("wonderful," "fabulous") visit to New York and stay at the hotel. Note that this is a priori known, given our data selection. The topic also indicates that reviewers would return ("go-back," "definitely") to the hotel, mention its brand ("Marriott," "Hilton"), and recommend their choice ("I-highly-recommend"). However, the topic provides little to no information as to why the stay was great or why the brand is recommended. In contrast, the most predictive

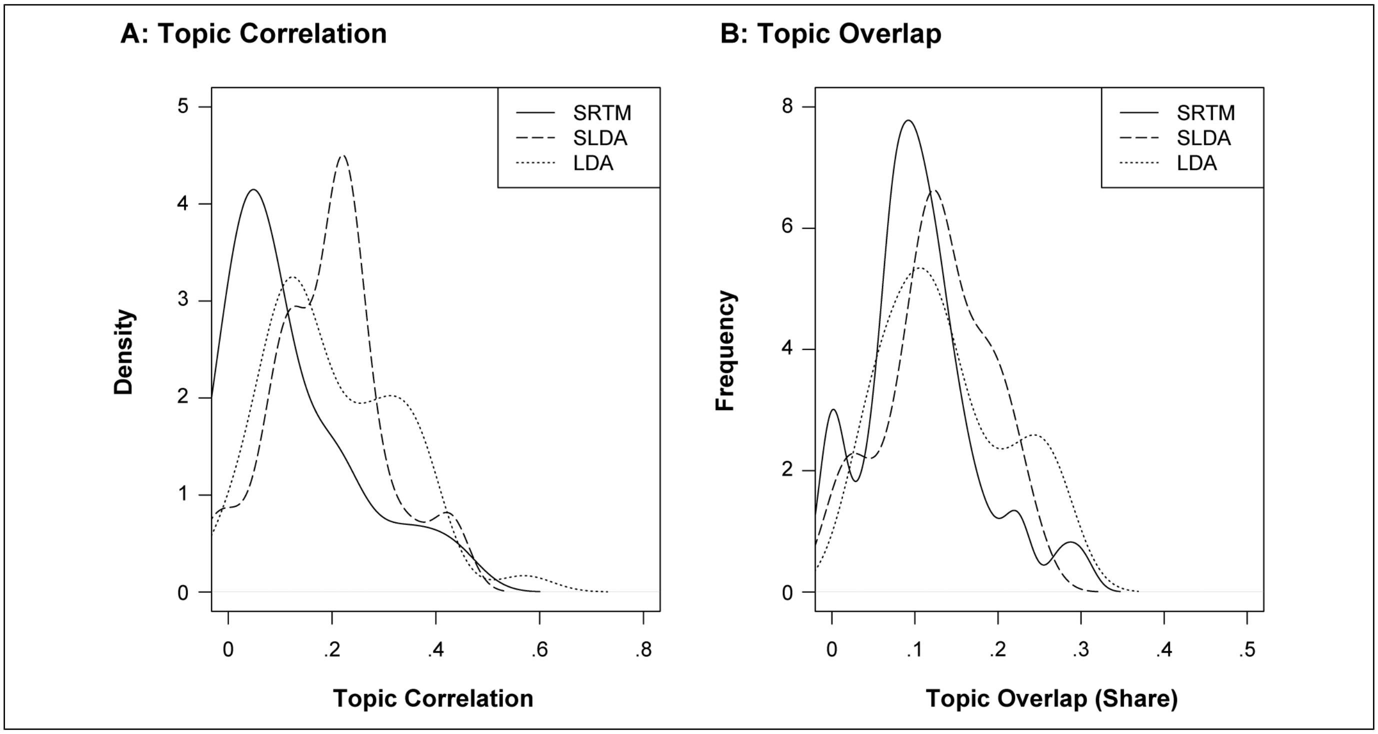

Figure 5. Distribution of Topic Correlation and Topic Overlap, Given the Model.
Notes: Displayed are the distributions of the T(T − 1)/2 uniqueness scores given the model and method. SLDA = supervised LDA.

Table 4. Rating Model Structure for Hotel Data.

	Posterior Mean of I
I_1: Hotel great	1.000
I_2: Room positive	.001
I_3: Local attractions	1.000
I_4: Elevator	.009
I_5: Room negative	.000
I_6: Check-in	.035
I_7: Amenities	.020
I_8: Noise	.002
I_9: Hotel	.002
I_{10}: Staff	.038

Notes: Displayed are posterior means of I for each topic. Labels for topics were assigned by the authors (see Web Appendix A).

(formative) topic from the SRTM (Figure 6, Panel A) describes (positively evaluated) features of the hotel room (it was "clean"/ "room-clean," "large," "nice," with a "comfortable"/"comfy" "bed"). In summary, the most predictive topic from the supervised LDA describes an overall positive experience (which is largely redundant given the observed rating), whereas the SRTM identifies as most important predictor a product attribute (i.e., the room and how it was appointed). We note that the topic describing a positive overall evaluation is identified as reflective by the SRTM (see next section) and, therefore, does not enter the formative model. In Web Appendix D, we show the top 20 words for all topics from our dataset for both the supervised LDA and the SRTM.

Reflective topics from SRTM. An important feature of our model is the identification of topics that merely reflect the overall rating. Figure 7 shows world clouds of the two topics identified as reflective by our SRTM. Several observations from Figure 7 are noteworthy. Topic 1 (Panel A) talks about a hotel stay in general. The topic is rich in evaluative terms such as "perfect," "excellent," "wonderful," "fantastic," "best," "fabulous," and "amazing." Reviewers also indicate their intention to return ("go-back") and recommend ("highly-recommend"). In summary, this topic talks about a very positive hotel experience in general.

In comparison, Topic 3 (Panel B) consists mainly of terms describing the hotel's closeness to attractions (e.g., restaurants, tourist spots) and convenient access (e.g., "subway," "short-walk") to them. This topic is not simply a reexpression of the rating in words. Instead, it expresses an element of consumers' experience (nightlife, entertainment). While closeness to attractions (relative to another hotel with less convenient access) could contribute to a positive rating as a formative topic, all hotels in this sample are only a few blocks away from Broadway and Times Square. Hence, closeness to local attractions is not a differentiating feature in this dataset. Nevertheless, we find that reviewers with better ratings spend more of their review on local attractions (see the credibly positive slope coefficient $\lambda_{1,3}$ linking ratings to the prevalence of this topic in a review in Table 5). Our ex post interpretation of this result is that an overall positive experience with the *hotel* leads reviewers to expand on other positive aspects of their *New York City trip/experience* such as their use of world-class local attractions

Table 5. Estimates of β and λ from the Rating Model.

	Rating Model Coefficients	
	Posterior Mean	Posterior SD
Formative Model		
β_0: Intercept	**.58**	.07
β_1: Hotel great	.02	.00
β_2: Room positive	**1.51**	.40
β_3: Local attractions	.01	.00
β_4: Elevator	.16	.15
β_5: Room negative	**−1.39**	.38
β_6: Check-in	**−.22**	.08
β_7: Amenities	.07	.27
β_8: Noise	**−.97**	.23
β_9: Hotel	.55	.54
β_{10}: Staff	.18	.51
σ^2	.89	.05
Reflective Model		
$\lambda_{0,1}$: Intercept: Hotel great	.00*	.00*
$\lambda_{1,1}$: Hotel great	1.00*	.00*
ψ_1	1.84	.10
$\lambda_{0,3}$: Intercept: Local attractions	−.40	.86
$\lambda_{1,3}$: Local attractions	**.63**	.19
ψ_3	2.98	.14

Notes: Displayed are posterior means and standard deviations, conditional on posterior mode of I. Coefficients credibly different from 0 on 95% level as indicated by posterior mass right or left of 0 are shown in boldface. To reduce clutter, reflective coefficients not credibly different from 0 are not shown. Asterisk (*) indicates parameters fixed for identification. Labels for topics were assigned by the authors (see Web Appendix A).

and how these contributed to their (positive) New York City experience.

Discussion

In this article, we introduce a structural model relating topics in textual product reviews to the rating provided by consumers in a novel way. Our model builds on supervised text models that use observed labels for inference regarding latent model variables (McAuliffe and Blei 2007; Yang, Boyd-Graber, and Resnik 2017). We propose that the standard way of relating the label (i.e., rating) to document-level latent variables (document-topic probabilities) ignores the possibility of reflective topics. The innovation introduced by our model allows for topics to be formative or reflective with respect to the rating. In simulation, we show that inference based on the (standard) supervised approach is biased and inconsistent in the presence of reflective topics. In our empirical application, we show that allowing for reflective topics in reviews changes topic inference in general.

Allocation as reflective in our model stems from a topic being an indicator and not a driver of the (overall) rating. Technically, identification results from the conditional independence relationships implied by classifying topics as reflective and the corresponding reduction in the dimensionality of the covariance among topics. From our empirical analysis, we find that reflective topics exist, consistent with the idea that reviewers build an argument that culminates in verbalized overall assessments. We also find that topics are statistically identified as either formative or reflective with high reliability. Reflective topics in our data are indicative of positive ratings of the hotel or a great overall experience in New York City. They are rich in evaluative terms (“great,” “recommend,” “perfect,” “very-convenient”), which is consistent with their classification as reflective of an evaluation. In comparison, formative topics identified by our model tend to be rich in nouns that describe a particular aspect or attribute of a product or service.[8]

The finding that one of our reflective topics, great local attractions, does not qualify as a simple reexpression of a high rating as per the topical terms suggests that reviews may well go beyond immediate aspects and attributes of the product/service in focus and their evaluation. The mechanism behind the empirical observation that reviewers are more ready to digress into other positive things experienced as part of their trip, conditional on a positive experience with the hotel, requires further study that we leave for future research.

Our model results in semantically more coherent and also more unique topics than a standard supervised LDA. In fact, we find that the topics from our model differ greatly from those identified by benchmark models. This, in conjunction with improved fit to the data, suggests that accounting for topics reflective of an overall rating may improve topic inference from review data in general, which, after all, is the key idea of inference from text models, in which meaning of words is inferred from term co-occurrence. Improved topic inference, in particular with respect to formative topics, aids managers in identifying potential interventions to improve customers’ experience.[9]

An important feature of our model is that it constrains topic discovery with respect to topics’ relationship with the rating. There is, however, a priori no constraint in our model on relating topics to attributes (Chakraborty, Kim, and Sudhir 2022). Topic-attribute relationships are only revealed a posteriori (if they exist). Marketers may find this insufficient when their interest lies in specific attributes (e.g., a restaurant owner interested in the impact of recent menu changes on customer satisfaction). Such a targeted analysis can be conducted by introducing domain knowledge through lexical priors to LDA’s word-topic probabilities (Jagarlamudi, Daumé, and Udupa 2012) or word-based lexicon approaches (Chakraborty, Kim, and Sudhir 2022). Our SRTM can potentially accommodate such prior knowledge. We leave this extension of our model to future research.

Our set of benchmark models is limited to LDA-type models of text. However, we believe that our model for the joint

[8] In an earlier version of this article, the model choice, that is, (conditional) classification of topics as formative or reflective, was mistakenly based on posterior predictive densities computed for the same data used to fit the model instead of prior predictive densities (see the Web Appendix for details), which led to qualitatively different results. We thank an anonymous reviewer for pointing out the problem.

[9] We thank an anonymous reviewer for emphasizing this point.

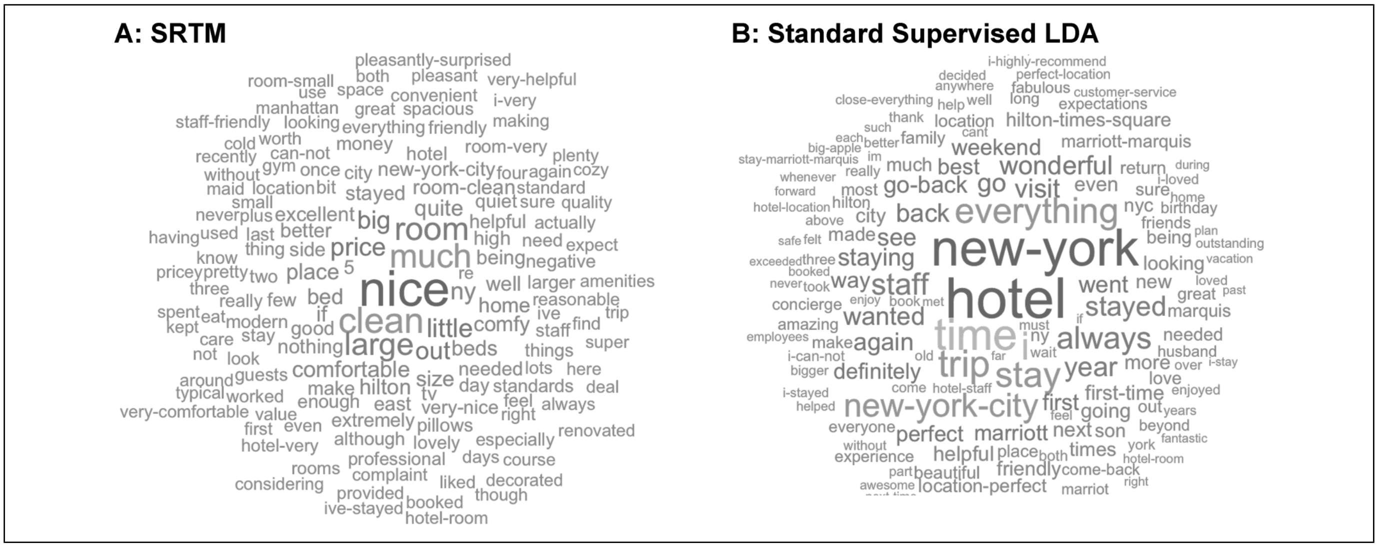

Figure 6. Word Clouds of Topic Most Predictive of the Rating.

Notes: The figure shows the top topic obtained by ordering posterior mean of β (in absolute terms) by size. For supervised LDA, all topics are ranked; for SRTM, only topics reliably identified as formative are ranked. Font size of terms in clouds is proportional to word-topic probability.

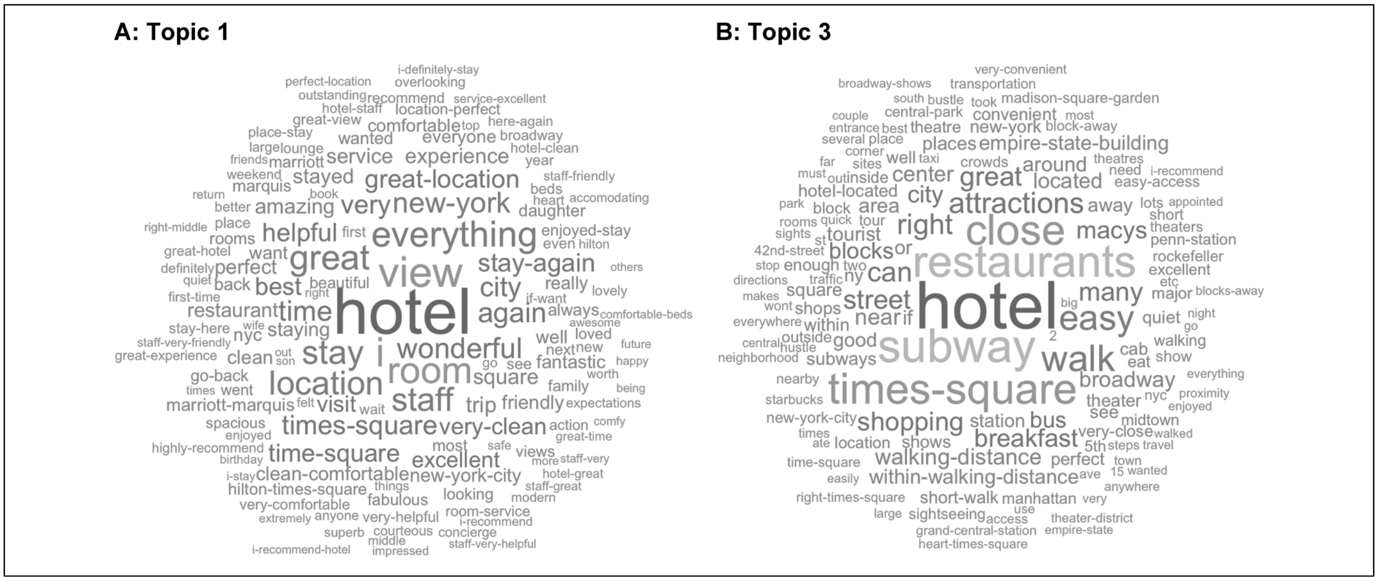

Figure 7. Word Clouds from Topics Identified as Reflective by SRTM.

Notes: Displayed are the top terms as indicated by posterior mean of ϕ_t. Font size of terms in clouds is proportional to word-topic probability.

distribution of topics and ratings across documents will be useful regardless of what lower-level model of text is assumed. Substantively, topic models might suffice for the type of analysis conducted here because relatively targeted datasets consisting of shorter documents exhibit characteristics that do not require the capacity of deep-learning models (Goodfellow, Bengio, and Courville 2016). In a sense, our data from the intersection of hotel category (five-star luxury rating), brands, location (Manhattan near Times Square/Broadway), and price point that results in a relatively smaller vocabulary is typical of text data analyzed in practice, which often accounts for a focal brand or company, a particular set of competing brands, locations, quality levels, prices, and so forth. It is, of course, possible that a different selection of data would result in a large and highly diverse vocabulary that may benefit from a deep-learning approach as proposed by Dieng, Ruiz, and Blei (2020). We leave the integration of our structural rating model that distinguishes between formative and reflective topics with deep-learning models of text to future research.

Acknowledgments

The authors thank participants at the Bayesian Marketing and Machine Learning Workshop at the Cambridge Judge Business School and seminar participants at Nova School of Business and Economics, Purdue University, and Hong Kong University for their contributions to this article.

Coeditor

Brett R. Gordon

Associate Editor

Christophe Van den Bulte

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Joachim Büschken https://orcid.org/0000-0001-5133-8122
Greg M. Allenby https://orcid.org/0000-0001-9759-0067

References

Agarwal, Deepak and Bee-Chung Chen (2010), "FLDA: Matrix Factorization Through Latent Dirichlet Allocation," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, 91–100.

Aletras, Nikolaos and Mark Stevenson (2013), "Evaluating Topic Coherence Using Distributional Semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Association for Computational Linguistics, 13–22.

Ansari, Asim, Yang Li, and Jonathan Z. Zhang (2018), "Probabilistic Topic Model for Hybrid Recommender Systems: A Stochastic Variational Bayesian Approach," *Marketing Science*, 37 (6), 987–1008.

Blei, David M. and John D. Lafferty (2007), "A Correlated Topic Model of Science," *Annals of Applied Statistics*, 1 (1), 17–35.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Bollen, Kenneth A. (1989), *Structural Equations with Latent Variables*. John Wiley & Sons.

Bollen, Kenneth A. and Adamantios Diamantopoulos (2017), "In Defense of Causal-Formative Indicators: A Minority Report," *Psychological Methods*, 22 (3), 581–96.

Bollen, Kenneth and Richard Lennox (1991), "Conventional Wisdom on Measurement: A Structural Equation Perspective," *Psychological Bulletin*, 110 (2), 305–14.

Boyd-Graber, Jordan, David Blei, and Xiaojin Zhu (2007), "A Topic Model for Word Sense Disambiguation," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, 1024–33.

Büschken, Joachim and Greg M. Allenby (2016), "Sentence-Based Text Analysis for Customer Reviews," *Marketing Science*, 35 (6), 953–75.

Büschken, Joachim and Greg M. Allenby (2020), "Improving Text Analysis Using Sentence Conjunctions and Punctuation," *Marketing Science*, 39 (4), 727–42.

Chakraborty, Ishita, Minkyung Kim, and K. Sudhir (2022), "Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Missing Attributes," *Journal of Marketing Research*, 59 (3), 600–622.

Dieng, Adji B., Francisco J.R. Ruiz, and David M. Blei (2020), "Topic Modeling in Embedding Spaces," *Transactions of the Association for Computational Linguistics*, 8, 439–53.

Fang, Anjie, Craig Macdonald, Iadh Ounis, and Philip Habel (2016), "Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 1057–60.

Fornell, Claes and Fred L Bookstein (1982), "Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory," *Journal of Marketing Research*, 19 (4), 440–52.

George, Edward I. and Robert E. McCulloch (1995), "Stochastic Search Variable Selection," *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds. Chapman & Hall/CRC, 203–14.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016), *Deep Learning*. MIT Press.

Hofmann, Thomas (1999), "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 50–57.

Jagarlamudi, Jagadeesh, Hal Daumé III, and Raghavendra Udupa (2012), "Incorporating Lexical Priors into Topic Models," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 204–13.

Jarvis, Cheryl Burke, Scott B. MacKenzie, and Philip M. Podsakoff (2003), "A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research," *Journal of Consumer Research*, 30 (2), 199–218.

Johnson, Valen E. and James H. Albert (2006), *Ordinal Data Modeling*. Springer Science and Business Media.

Jöreskog, Karl G. and Arthur S. Goldberger (1975), "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable," *Journal of the American Statistical Association*, 70 (351a), 631–39.

Laghaie, Arash and Thomas Otter (2023), "Measuring Evidence for Mediation in the Presence of Measurement Error," *Journal of Marketing Research*, 60 (5), 847–69.

Lee, Nick, John W. Cadogan, and Laura Chamberlain (2013), "The MIMIC Model and Formative Variables: Problems and Solutions," *AMS Review*, 3 (1), 3–17.

Li, Xiaolin, Chaojiang Wu, and Feng Mai (2019), "The Effect of Online Reviews on Product Sales: A Joint Sentiment-Topic Analysis," *Information & Management*, 56 (2), 172–84.

Mankad, Shawn, Hyunjeong "Spring" Han, Joel Goh, and Srinagesh Gavirneni (2016), "Understanding Online Hotel Reviews Through Automated Text Analysis," *Service Science*, 8 (2), 124–38.

McAuliffe, Jon and David Blei (2007), "Supervised Topic Models," *Advances in Neural Information Processing Systems*, 20.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a), "Efficient Estimation of Word Representations in Vector Space," arXiv, https://doi.org/10.48550/arXiv.1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean (2013b), "Distributed Representations of Words and Phrases and Their Compositionality," *Advances in Neural Information Processing Systems*, 26.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011), "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 262–72.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin (2010), "Automatic Evaluation of Topic Coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 100–108.

Nguyen, Viet-An, Jordan L Ying, and Philip Resnik (2013), "Lexical and Hierarchical Topic Regression," *Advances in Neural Information Processing Systems*, 26.

Otter, Thomas, Timothy J. Gilbride, and Greg M. Allenby (2011), "Testing Models of Strategic Behavior Characterized by Conditional Likelihoods," *Marketing Science*, 30 (4), 686–701.

Pergola, Gabriele, Lin Gui, and Yulan He (2020), "A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews," arXiv, https://doi.org/10.48550/arXiv.2010.11384.

Podsakoff, Nathan P., Wei Shen, and Philip M. Podsakoff (2006), "The Role of Formative Measurement Models in Strategic Management Research: Review, Critique, and Implications for Future Research," *Research Methodology in Strategy and Management*, Vol. 3, David J. Ketchen and Donald D. Bergh, eds. Emerald Group Publishing, 197–252.

Rabinovich, Maxim and David Blei (2014), "The Inverse Regression Topic Model," in *Proceedings of the 31st International Conference on Machine Learning*. PMLR, 32 (1), 199–207.

Rossi, Peter E., Greg M. Allenby, and Robert E. McCulloch (2005), *Bayesian Statistics and Marketing*. John Wiley & Sons.

Schütze, Hinrich (1998), "Automatic Word Sense Discrimination," *Computational Linguistics*, 24 (1), 97–123.

Taddy, Matt (2013), "Multinomial Inverse Regression for Text Analysis," *Journal of the American Statistical Association*, 108 (503), 755–70.

Timoshenko, Artem and John R Hauser (2019), "Identifying Customer Needs from User-Generated Content," *Marketing Science*, 38 (1), 1–20.

Yang, Weiwei, Jordan Boyd-Graber, and Philip Resnik (2017), "Adapting Topic Models Using Lexical Associations with Tree Priors," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1901–06.

Yang, Yi, Kunpeng Zhang, and Yangyang Fan (2023), "sDTM: A Supervised Bayesian Deep Topic Model for Text Analytics," *Information Systems Research*, 34 (1), 137–56.