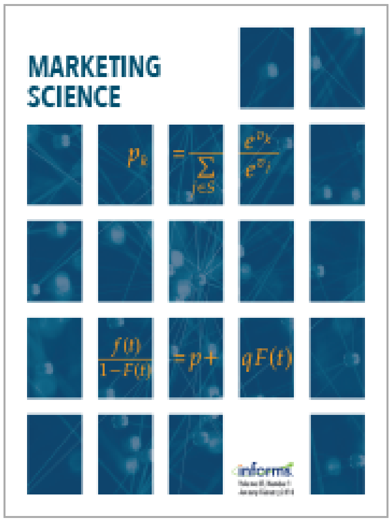


Marketing Science

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

Improving Text Analysis Using Sentence Conjunctions and Punctuation

Joachim Büschken, Greg M. Allenby

informs
http://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Vol. 39, No. 4, July–August 2020, pp. 727–742
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Improving Text Analysis Using Sentence Conjunctions and Punctuation

Joachim Büschken,[a] Greg M. Allenby[b]

[a] School of Management, Catholic University of Eichstätt-Ingolstadt, 85049 Ingolstadt, Germany; [b] Fisher College of Business, Ohio State University, Columbus, Ohio 43210

Contact: joachim.bueschken@ku.de, https://orcid.org/0000-0002-9673-8928 (JB); allenby.1@osu.edu, https://orcid.org/0000-0001-9759-0067 (GMA)

Received: January 31, 2017
Revised: April 4, 2018; January 30, 2019; July 16, 2019
Accepted: August 22, 2019
Published Online in Articles in Advance: July 7, 2020

https://doi.org/10.1287/mksc.2019.1214

Abstract. User-generated content in the form of customer reviews, blogs, and tweets is an emerging and rich source of data for marketers. Topic models have been successfully applied to such data, demonstrating that empirical text analysis benefits greatly from a latent variable approach that summarizes high-level interactions among words. We propose a new topic model that allows for serial dependency of topics in text. That is, topics may carry over from word to word in a document, violating the bag-of-words assumption in traditional topic models. In the proposed model, topic carryover is informed by sentence conjunctions and punctuation. Typically, such observed information is eliminated prior to analyzing text data (i.e., *preprocessing*) because words such as "and" and "but" do not differentiate topics. We find that these elements of grammar contain information relevant to topic changes. We examine the performance of our models using multiple data sets and establish boundary conditions for when our model leads to improved inference about customer evaluations. Implications and opportunities for future research are discussed.

History: Yuxin Chen served as the senior editor and Peter Fader served as associate editor for this article.
Supplemental Material: The data and online appendix are available at https://doi.org/10.1287/mksc.2019.1214.

Keywords: user-generated content • latent Dirichlet allocation (LDA) • topic dependency • syntactic covariates • Bayesian analysis • customer satisfaction analysis

1. Introduction

Text data in the form of customer reviews, blogs, and tweets is a fast-growing and rich source of data for marketing researchers. Websites such as Tripadvisor.com and Yelp.com offer a growing range of products and services for which customers can post reviews. An important and fruitful area of model-based empirical text analysis is the application of latent topic models (Blei et al. 2003, Tirunillai and Tellis 2014, Humphreys and Wang 2017) to such data. Topic models identify sets of words that frequently co-occur, giving rise to the ability to account for high-level interaction among words. Essentially, topic models are devices to detect latent clusters of co-occurring multinomial variables such as words from a given vocabulary. The clusters emerging from these models can be used to analyze the relationship between topics and variables of interest such as purchase intention and customer satisfaction (Büschken and Allenby 2016), providing insights into consumer preferences and behavior. Topic models compete with approaches in text analysis that are based on observed features of text such as dictionary-based approaches (Humphreys and Wang 2017). The advantage of topic models lies in the detection of latent variables that provide high-level summaries of text.

A challenge in applying topic models to customer reviews is the limited amount of data contained in any one review. The number of words in a review is typically fewer than 100, making it difficult to assess topic and word probabilities without imposing additional structure. An assumption typically present in topic models is that topics exhibit zero autocorrelation in that the probability of the topic assignment to word $t+1$ is independent of the topic assignment to word t. This assumption gives rise to the *bag-of-words property* and to word counts being sufficient statistics for the standard model. Recently, Büschken and Allenby (2016) proposed a model in which topics are constrained to not change within a sentence. They show that this restriction leads to better fit to the data and more interpretable topic word probabilities in customer review data. A similar approach was proposed by Nallapati and Allan (2002), who used sentence boundaries as structural information to a unigram-type probabilistic language model. The common element of both these models is the introduction of a sentence-based constraint to the model, imposing common topics within observed boundaries of text.

In this paper, we propose a new model that allows for topic assignments to carry over to the next word.

The model is an autocorrelated topic model in which the probability of a word-to-word topic carryover is parameterized as a binary logit model with covariates, and we find that punctuation (e.g., periods, exclamation marks, and commas) and conjunctions (e.g., "and," "but," and "because") are predictive of topic carryover. These syntactic elements are frequently discarded as part of data cleaning in text analysis and have not been previously analyzed for their value in predicting topic changes within sentences and inferences about the topics themselves. An alternative approach to modeling correlation among topic assignments to observed sequences of words involves the changepoint models that, in marketing, have been applied to time-series data to discover changes in the underlying data-generating process (Chib 1998, DeSarbo et al. 2004, Fader et al. 2004, Netzer et al. 2008, Gopalakrishnan et al. 2016).

A key element in our model is to relate the unobserved topic shares at the document level to the observed overall rating provided by reviewers in the way of a supervised latent Dirichlet allocation (LDA) model (Mcauliffe and Blei 2007). With this approach, the rating provides additional likelihood information to the latent topic shares, assuming a common mapping of topics to ratings across documents through a regression model. In our empirical analysis, we find this mapping to be highly discriminatory regarding the importance and valence of the topics.

We apply the proposed autocorrelated topic model to multiple data sets, and we find the autocorrelated topic model to fit the data better, so we establish boundary conditions for when it is preferred to a standard LDA model. For this analysis, we evaluate the performance of all models in terms of predictive fit, given holdout reviews, and the explained variance of customers' overall satisfaction rating. The autocorrelated topic model works best for reviews containing more words and longer sentences and in which we observe more incidents of conjunctions and punctuation, suggesting that a model of local topic dependency works better for linguistically more complex data. The LDA model, in comparison, works better when text is less complex. Interestingly, this includes the case where many relatively short reviews contain a richer vocabulary, suggesting that the number of unique terms in a corpus plays a minor role. We demonstrate managerial relevance of results from our topic model by identifying topics that disproportionately influence a customer's overall evaluation of a product or service experience.

The remainder of this paper is organized as follows. In Section 2, we develop the autocorrelated topic model. In Section 3, we present a summary of the data that we use in our empirical analysis. In Section 4, we present results from applying our model to the data and compare it with various alternative models. Section 5 presents results from using topics as predictors of customer ratings. And finally, in Section 6, we summarize our findings and offer concluding comments.

2. Model Development

The LDA model assumes that each word in a text document is generated from latent topics characterized by topic-specific word probabilities across a fixed vocabulary. Each document d is described by a vector of topic probabilities θ_d of (typically a priori fixed) dimension T, and each topic is characterized by vector ϕ_t, which specifies the word probabilities associated with that topic. Words in a document are generated by first drawing a topic indicator variable z from a discrete (multinomial) distribution with probability vector θ_d and then drawing a word from the vocabulary list with probability vector $\phi_{t=z}$. Büschken and Allenby (2016) propose a constrained version of this model by restricting all words within a sentence to be generated from the same topic. This is accomplished by drawing the latent variable z once for all words in a sentence and then "rolling out" this topic assignment across all words in a sentence.

The LDA and sentence-constrained LDA (SC-LDA) models represent two extremes in topic generation. Topics generated from the LDA model are assumed to be independent and identically distributed (i.i.d.) across all words, whereas the SC-LDA is a model of deterministic dependency within an observed locale and with only a sentence's period (e.g., full stop, exclamation point) allowing for topic variation within a document. In this section, we develop an approach to the LDA model that does away with the i.i.d. assumption to topic assignments. The model we propose is different in the way (local) dependence of topic assignments emerges. Our model proposes probabilistic topic carryover on the word level. This implies that topic assignments are locally correlated.

The issue of correlated topics has been examined previously in the literature. Griffiths et al. (2004) propose a model in which words differ in syntactic (i.e., related to placement or arrangement) and semantic (i.e., related to meaning) content. A hidden Markov model (HMM) is used to generate sentence syntax, and an LDA (topic) model is used to generate its content. The HMM introduces serial correlation between syntactic vocabularies and topics, but not among the topics themselves. Wallach (2006) proposes a model in which topics generate words conditional on the previous word, introducing first-order autocorrelation in word generation but not topic generation. That is, the topic indicator variable is still assumed to be i.i.d. across words. Blei and Lafferty (2007) employ a logistic Normal distribution instead of a Dirichlet distribution to allow for correlations in the prior to topics. Their model affects the topic probabilities θ_d but does not induce autocorrelation of topics within the document indicated by the latent indicator variables z. Trusov et al. (2016) propose a model with correlated topics for website visitation data to account for latent interests (or, as they are called by the authors, "roles"; p. 406), simultaneously driving the number of times different websites are visited up (or down). The common element of these approaches is that topics may exhibit a priori dependence (e.g., if reviewers talk extensively about service problems in restaurant reviews, they might also talk more about inflated prices). Finally, in a version of their SC-LDA model, Büschken and Allenby (2016) allow for a probabilistic carryover of topics from sentence to sentence but do not find empirical support for this model.

Topic dependency across sequences of words has important implications for inference regarding θ_d, the topic shares of individual documents. If topics exhibit word-to-word carryover, estimates of θ_d obtained via the LDA model are potentially biased. This is because, in the LDA model, every topic assignment on the word level is treated as a draw from θ_d. For example, if a document consists of two topic *runs* (consecutive topic carryover) of three words and six words, respectively, the LDA model treats these as three i.i.d. draws of $z = t_1$ from $p(z|\theta_d)$ and six i.i.d. draws of $z = t_2$ from $p(z|\theta_d)$. The expected value of θ_d suggested by the LDA model (ignoring the role of the prior) is then $(\frac{3}{9}, \frac{6}{9})$. A model with autocorrelated topics, however, suggests that $E(\theta_d) = (0.5, 0.5)$ because topic repeats as a result of carryover do not originate from $p(z|\theta_d)$. It is clear that when documents contain sequences of words exhibiting topic carryover, estimates from the LDA model are biased toward topics with high word-to-word carryover probability. As result, LDA model–based estimates may differ greatly from those using models that account for the serial structure of topic assignments. In our empirical analysis, we show that topics in customer reviews exhibit strong local dependency and that models accounting for this dependency result in better estimates of θ_d because they carry higher predictive power with respect to customers' satisfaction ratings. Next, we turn to the development of our model with autocorrelated topics.

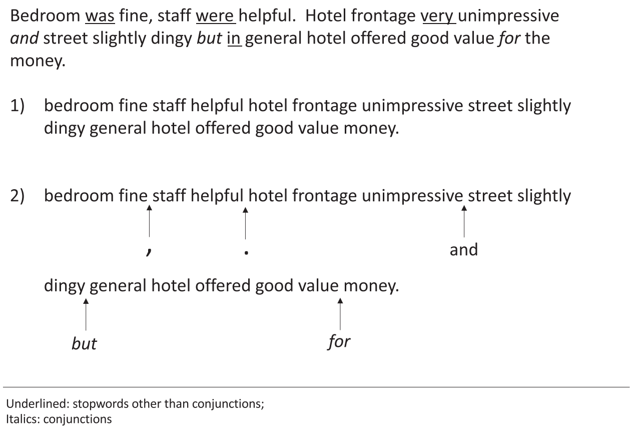
2.1. Autocorrelated Topic Model

Our autocorrelated topic model assumes topic carryover at the word level, with the degree of carryover affected by observed structural information in the text. An important structural feature of written Romanic and Germanic languages is the use of conjunctions such as "and" and "but" that play a syntactic role by joining parts of a sentence. As such, they do not represent topics or semantic content. A typical example of this role is present in this hotel review, "Comfy beds but not your usual large American beds," in which the conjunction "but" links two different perspectives of evaluation, one personal and the other more general ("not American"). In comparison, punctuation typically does not join, but separates, parts of speech. A full stop, for example, indicates the end of a sentence and introduces a pause to the flow of thoughts. Such a pause is a natural candidate for a topic change. Other examples of punctuation are exclamation points or question marks, both of which introduce structure to text in a similar way but also add weight or an interrogative notion to a statement. Structural punctuation for the purpose of this analysis consists of marks that act on parts of documents typically not larger than a sentence and not smaller than a word (e.g., hyphens; Meyer 1987, Say and Akman 1996). The central idea of our model is to use the observed structural information in text presented by punctuation and conjunctions for inference regarding the dynamics of topics in text.

It is interesting to note that in empirical applications of topics models, conjunctions as well as incidents of punctuation are typically removed from the data prior to analysis (i.e., *preprocessing*). Conjunctions are removed because they are stopwords. Stopwords typically carry very little power to discriminate topics. This is evident in the nearly uniform probabilities of stopwords, if included in the data, to appear under any topic. However, we propose that conjunctions and punctuation, as carriers of structural information, present information to topic change and introduce this information to our model. The challenge is in retaining the structural information without compromising inference about topics. In Figure 1, we present a stylized way of using these data in our model.

In the review at the top of Figure 1, we highlight conjunctions, punctuation, and stopwords other than conjunctions. In two versions of this review, numbered 1 and 2, we present different ways of exploiting structural information. Version 1 results from removal of all stopwords, including conjunctions, and all punctuation. This preprocessing of data is consistent with the bag-of-words assumption and uses no structural information other than the remaining words. For version 2, which is applied here, we view conjunctions (green) and punctuation (red) as prior information to a topic carryover between consecutive words. For example, the conjunction "but" is used as a covariate to the probability of a topic carryover from the word (street slightly) "dingy" to the word "general" (hotel offered good value). In a similar fashion, the full stop preceding the word "hotel" is

Figure 1. Using Syntactic Covariates in Topic Analysis for Actual Example of a Hotel Review


Bedroom was fine, staff were helpful. Hotel frontage very unimpressive and street slightly dingy but in general hotel offered good value for the money.
1) bedroom fine staff helpful hotel frontage unimpressive street slightly dingy general hotel offered good value money.
2) bedroom fine staff helpful hotel frontage unimpressive street slightly
,
.
and
dingy general hotel offered good value money.
but
for
Underlined: stopwords other than conjunctions;
Italics: conjunctions

Notes. Top: Original review. 1: Review after removing stopwords, including all conjunctions and all punctuation. 2: Review after removal of stopwords other than conjunctions and positional assignment of conjunctions and punctuation as covariates. Arrows indicate presence of covariates to topic carryover.

covariate to the probability of a carryover from (staff) "helpful" to "hotel" (frontage). The difference between the two approaches lies in the use of otherwise ignored data as observed covariates to topic change. In the empirical application of our model, we find topic carryover to be heavily driven by structural elements of text.

2.1.1. Joint Distribution of Model Parameters and Data.

The model with autocorrelated topics is based on the LDA topic model (Blei et al. 2003), which proposes the following joint distribution of knowns and unknowns:

$$p(w_n, z_n, \theta_d, \phi, \alpha, \beta) = p\left(w_n|\phi_{z_n}\right) \times p(z_n|\theta_d) \times p(\theta_d|\alpha) \times p(\phi|\beta) \times p(\alpha) \times p(\beta), \tag{1}$$

where w_n is the nth word of document d, z_n is the topic assignment of word w_n, θ_d is a vector of prior topic probabilities for document d, ϕ_t is a vector of word probabilities given topic t, and α, β are fixed priors of θ_d and ϕ_t, respectively.

We extend this model so that the topic z_{n-1} assigned to word w_{n-1} may carry over to word w_n independent of θ_d so that topics are autocorrelated. We define carryover of a topic as $z_n = z_{n-1}$ and introduce the latent binary variable ζ_n to indicate whether the topic assignment to word w_n is the result of carryover:

$$\begin{aligned} \zeta_n = 1 &: z_n = z_{n-1}, \\ \zeta_n = 0 &: z_n \sim \text{Multinomial}(\theta_d). \end{aligned} \tag{2}$$

In the LDA model, $\zeta_n = 0\ \forall n$, implying that this model is a special case of the autocorrelated-topics LDA (AT-LDA) model. The same holds for the SC-LDA, which imposes $\zeta_{n,s} = 1\ \forall n_s > 1$, where s is a sentence. We assume ζ_n to be a distributed binomial with probability ψ_n:

$$\begin{aligned} \zeta_n &\sim \text{Binomial}\left(\psi_n|z_{n-1}\right) \\ \psi_n|z_{n-1} &= \frac{\exp\left[\delta_{0,z_{n-1}} + \tilde{x}_n'\delta\right]}{1+\exp\left[\delta_{0,z_{n-1}} + \tilde{x}_n'\delta\right]} \end{aligned} \tag{3}$$

where $\tilde{x}_n$ is a vector of dummy variables that indicates conjunctions and punctuation to the current word (Figure 1), and δ are estimated coefficients that affect the probability of topic change. Negative values of δ increase the likelihood of an i.i.d. topic draw, whereas positive values indicate that a topic carryover is more likely. We allow for intercepts $\delta_{0,z_{n-1}}$ that depend on the previous word's topic z_{n-1}. Equation (3) specifies common coefficients δ for the conjunctions and punctuation in $\tilde{x}_{n-1}$. In our empirical analysis, we also consider interaction effects of the latent topics and covariates giving rise to topic-specific effects of conjunctions and punctuation.

2.1.2. Data-Generating Process.

The generative model of the AT-LDA with covariates to topic change, the regression of the customer rating on topic shares of documents, and fixed priors $\alpha, \beta, \mu_\delta, \Sigma_\delta, \mu_\beta, \Sigma_\beta, a, b, \lambda$ is as follows:

1. Draw independently
   a. δ from MV Normal$(\mu_\delta, \Sigma_\delta)$.
   b. ϕ_t from Dirichlet(β) $\forall t$ i.i.d.
   c. θ_d from Dirichlet(α) $\forall d$ i.i.d.
   d. $\beta^{(reg)}$ from MV Normal$(\mu_\beta, \Sigma_\beta)$.
   e. σ^2 from $IG(a, b)$.
   f. c from $U(\lambda)$.
2. Draw (latent) rating τ_d from $N(\theta_d^T\beta^{(reg)}, \sigma^2)$; compute r_d given τ_d via (5).
3. For the first word in document d, w_1
   a. Draw z_1 from Multinomial(θ_d).
   b. Draw w_1 from Multinomial$(\phi_{t=z_1})$.
   c. Compute $p(\psi_2|x_2, \delta, z_1)$ using (3); draw ζ_2 given ψ_2.
4. For words w_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 from Multinomial$(\phi_{t=z_n})$.
   c. Compute $p(\psi_{n+1}|x_{n+1}, \delta, z_n)$, draw ζ_{n+1} given ψ_{n+1}.
5. Repeat steps 2–4 for all documents $d \in D$ (except for draw of ζ_{N_d}).

Note that the draw of the words and the rating are independent given θ_d. The joint distribution of the knowns and unknowns of the AT-LDA model with covariates, given document d, factorizes as follows:

$$\begin{aligned} &p\left(r_d, \tau_d, \{w\}_d, \{z\}_d, \theta_d, \phi, \{\zeta\}_d, \beta^{(reg)}, \sigma^2, \delta, \alpha, \beta, x, a, b, c\right) \propto \\ &\quad p(r_d|\tau_d, c) \times p\left(\tau_d|\theta_d^T\beta^{(reg)}, \sigma^2\right) \times p\left(w_1|\phi, z_1\right) \times p(z_1|\theta_d) \times \\ &\prod_{n=2}^{N_d}\left[p\left(w_n|\phi, z_n, z_{n-1}, \zeta_n\right) \times p(z_n|z_{n-1}, \theta_d, \zeta_n)\right. \\ &\quad \left.\times\, p(\zeta_n|, x_n, \delta, z_{n-1})\right] \times p\left(\phi|\beta\right) \times p\left(\theta_d|\alpha\right) \times p(\sigma|a, b) \\ &\quad \times p(c|\lambda) \times p(\beta) \times p(\alpha) \times p(\delta) \times p(a) \times p(b) \times p(\lambda), \end{aligned} \tag{4}$$

where we, as usual, assume independent prior distributions. The likelihood of a word, conditional on ζ_n, is

$$\begin{aligned} p\left(w_n|\phi, z_n, z_{n-1}, \zeta_n = 0\right) &= p\left(w_n|\phi, z_n\right), \\ p\left(w_n|\phi, z_n, z_{n-1}, \zeta_n = 1\right) &= p\left(w_n|\phi, z_{n-1}\right). \end{aligned}$$

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

$$\begin{aligned} p(z_n|z_{n-1}, \theta_d, \zeta_n = 0) &= p(z_n|\theta_d), \\ p(z_n|z_{n-1}, \theta_d, \zeta_n = 1) &= p(z_n = z_{n-1}) = 1. \end{aligned}$$

In Online Appendix A.2, we outline our Markov chain Monte Carlo (MCMC) approach to estimating the AT-LDA model, and in Online Appendix A.2.7, we use simulation to show that the model is empirically identified.

2.2. Incorporating Rating Information

We relate topic probabilities θ_d to the overall rating in the form of a supervised topic model (Mcauliffe and Blei 2007). The "supervised" aspect of this feature of our model results from relating the observed rating provided by each reviewer to the latent topic shares of each review. More specifically, we assume the rating for a review r_d to be related to the latent topic probabilities using a standard cut-point model:

$$r_d = e, \quad \text{if} \quad c_{e-1} \le \tau_d \le c_e, \tag{5}$$

and

$$\tau_d \sim N\left(\theta_d'\beta^{(reg)}, \sigma^2\right). \tag{6}$$

The cut-point model implies E cut points for $E-1$ categories, where cut points c_0 and c_E are $-\infty$ and ∞, respectively, and c_1 and c_{E-1} are fixed values for identification of location and scale of the latent response variable τ, which we augment the usual way. The cut-point model applied here is a standard model from the literature (Johnson and Albert 2006; see Online Appendix A.2). Note that Equations (5) and (6) result in additional likelihood information to θ_d because the topic shares must also adhere to the cross-sectional mapping of shares to the rating through $\beta^{(reg)}$. In a standard LDA model, topic shares are informed by the topic assignments of the words in d, z_d, only. Also note that the fit of the regression model in (6) provides a way of assessing the predictive plausibility of competing topic models. In Section 5, we explore the mapping of topic shares to the rating in more detail.

3. Data

We examine four customer review data sets that differ with respect to size and lexical complexity. All data sets were obtained from publicly available sources and were selected based on length and richness. Prior to running the models, we preprocessed the data in the following way:

1. Changing capital letters to lowercase letters
2. Removing rare words and signs (appearing fewer than 10 times in the corpus)
3. Removing stopwords other than conjunctions used for structural analysis (e.g., "there" and "rather").

Note that we do not remove very frequently occurring terms (e.g., "restaurant" or "tent"), nor do we remove punctuation and conjunctions. The latter are used as covariates to word-to-word topic carryover. Stemming is absent from our preprocessing because terms carrying different meanings may share the same stem. Overall, our preprocessing strategy is targeted at using (nearly) all of the observed text data as information either to topics or to topic carryover. From our experience, more aggressive preprocessing (performing stemming, removing all words with low frequency, removing common terms other than stopwords, etc.) leads to shorter documents and smaller corpora, which, in turn, lead to models with hard constraints (e.g., the SC-LDA model) fitting the data very well. In other words, text data can be preprocessed in a way that favors particular models. It is also important to note that aggressive pruning results in destruction of the local context of conjunctions or punctuation and subsequent word(s). This is another reason why we preprocess raw data as little as possible prior to a model-based analysis.

Table 1 presents descriptive statistics of the data sets used in our study. Descriptive statistics were obtained after applying the same preprocessing procedure. The restaurant data set, obtained from We8there.com, contains 2,351 textual reviews of restaurants. This corpus contains a total of 171,385 words, 1,531 of which are unique terms. On average, each restaurant review consists of 72.9 words. The standard deviation of the number of words per review of 84 words indicates the long right tail of the words per review distribution. The 72.9 words are spread, on average, over 13.4 sentences with 5.6 words per sentence.

Table 1. Descriptive Statistics of Data Sets (After Preprocessing)

Statistic	Restaurants	Camping tents	Luxury hotels	Dog food
Number of reviews	2,351	7,973	3,481	6,018
Corpus size	171,385	364,761	79,377	94,165
Number of unique terms	1,531	3,664	1,060	1,980
Number of words per review				
Mean	72.9	45.7	24.7	15.7
Standard deviation	83.5	58.1	19.8	22.7
Maximum	606	792	205	536
Number of sentences per review				
Mean	13.4	6.2	4.9	3.0
Standard deviation	14.1	6.8	3.0	3.1
Consumer rating (five-point scale)				
Mean	3.75	4.19	4.42	4.38
Standard deviation	1.41	1.23	0.88	1.21

Sentence breaks present natural breaks in the narrative and opportunities for topic change. The camping tents data are made up of 7,973 reviews of three-season, multiple-person camping tents in the price range of $100–$200, obtained from Amazon.com. Its vocabulary consists of 3,664 unique terms. This implies that despite applying the same preprocessing rules, the vocabulary used by tent reviewers is much more diverse, suggesting higher lexical complexity of the data. On average, the camping tent reviews contain 6.2 sentences with 7.6 words per sentence, the highest number of words per sentence across all data sets. The third data set in our analysis is a set of reviews of luxury hotels (five stars) located in downtown New York (Manhattan), obtained from Expedia.com. This data set consists of 3,481 reviews with a vocabulary of 1,060 terms. The average number of words per review is 24.7, and the average number of sentences is 4.9 (five words/sentence). The dog food data, also obtained from Amazon.com, are comprised of 6,018 reviews with a total of 94,165 words. On average, dog food reviews contain 15.7 words over three sentences (5.3 words/sentence). Thus, these reviews are lexically less complex than the other data sets. All data sets exhibit significant heterogeneity in terms of review length, as indicated by the standard deviation and range of the number of words in each review. The coefficient of variation of the number of words exceeds one in all data sets, and the distributions of the number of words specifically in reviews of restaurants and camping tents exhibit very long tails, as indicated by the maximum number of words, suggesting that many customers feel the need to report about their experience in great detail. For all data sets, we find that customer ratings are skewed toward the right, as indicated by a mean rating close to the upper end of the scale.

Table 2 reports counts of the use of conjunctions and various forms of punctuation in our data sets. From Table 2, it is clear that all data sets in our analysis are rich in syntactic content. The conjunction “and” appears, on average, 5.8 times (13,683/2,351) in the restaurant review and 2.7 times in each tent review. Similarly, on average, full stops mark boundaries between sentences 6.5 times in restaurant reviews and 2.6 times in hotel reviews. A special case of punctuation is presented by the use of (round) parentheses. We record the use of parentheses more than 1,000 times in the corpus of restaurant reviews and about 2,400 times in the camping tent reviews. Apparently, reviewers feel the need to structure some part of their narrative by placing words within parentheses. Typically, parentheses are used to clarify preceding text or, when combined with a full stop, as a side remark. Parentheses provide an observable signal of words belonging together, suggesting some form of topical dependency. On average, a review in the restaurant data set contains 38 structural elements in the form of conjunctions or punctuation, and a review of a camping tent contains 17. The frequency at which structural elements appear in our data raises the question of why this information can be ignored.

4. Empirical Analysis

We evaluate the performance of the proposed model by examining model fit, the prediction performance for customer ratings, topic carryover, and the direction of effects of the various conjunctions and punctuation.

4.1. Model Fit

Performance of the proposed autocorrelated topic model is compared with the following models:

1. Latent Dirichlet allocation (LDA).
2. Models of (local) topic chunking:
   a. SC-LDA,
   b. A conjunction- and punctuation-constrained LDA (CPC-LDA) model with sections in the reviews

Table 2. Incidents of Conjunctions and Punctuation in the Data Sets

Conjunctions	Restaurants	Camping tents	Luxury hotels	Dog food
for	4,439	9,274	1,796	2,929
and	13,683	21,683	5,912	6,929
but	3,258	5,374	1,083	1,629
or	933	1,868	278	529
so	1,878	3,819	550	1,321
after	737	1,090	165	445
as	1,725	3,292	550	1,114
because	512	1,138	148	482
before	472	737	102	208
even	620	1,155	199	299
if	1,260	2,509	430	484
now	264	696	23	425
once	226	652	41	101
since	341	441	60	385
than	615	1,244	212	453
that	3,867	6,042	904	1,772
though	217	483	71	93
when	1,303	1,912	370	515
where	326	541	114	68
which	1,091	983	310	331
while	407	586	83	134
who	384	308	83	223
what	909	732	155	372
Punctuation				
,	15,316	20,660	5,762	6,130
.	27,674	38,463	12,710	10.794
;	855	1,175	318	249
!	1,774	3,201	0	1,123
?	309	317	48	100
&	330	335	2	210
(	1,168	2,391	532	716
)	1,151	2,445	538	718
Total occurrences	88,651	136,756	33,853	41,606
Number of documents	2,351	7,927	3,215	6,018
Covariates per document	37.7	17.3	10.5	6.9
Covariates per word	0.52	0.37	0.43	0.44

Notes. Conjunctions appearing fewer than 200 times (e.g., "provided," "until") are omitted to reduce clutter. Total occurrences include omitted covariates.

defined by observed conjunctions and punctuation, and

c. Variants of both, allowing for topic carry-over across observed word sequences ("sticky" SC/CPC-LDA).

3. Topic models with autocorrelated topics across words (AT-LDA)

a. Without covariates and

b. Using conjunctions and punctuation as covariates to word-to-word topic change.

The LDA model is a common approach to analyze text data and a natural benchmark model for us to use. Our implementation of the LDA model (and all other models) includes the cut-point regression model in Equations (5) and (6) but retains the assumption that topics are generated i.i.d. from the prior topic distribution, indexed by θ_d for each word in a document. The SC-LDA restricts all words within a sentence to originate from a single topic (Büschken and Allenby 2016). In our case, sentences are given by observed use of full stops, exclamation points, and question marks in the reviews. The CPC-LDA uses observed conjunctions and punctuation to define similar but smaller locales in reviews in which topics are constrained to be identical. For example, the CPC-LDA model assumes that between the word "and" and the word "but" (or a question mark), all words in a review carry the same topic. Thus, the SC-LDA model is a special case of the CPC-LDA model that only considers specific forms of punctuation but not conjunctions as local boundaries for topics. For both the SC-LDA and CPC-LDA models, we consider the option that topics carry over (*sticky topics*) across local boundaries, giving rise to two additional models.

In the AT-LDA model with covariates, the amount of autocorrelation is affected by covariates in (3), reflecting syntactic content. We fit the AT-LDA model with covariates allowing for the effect of covariates to be topic specific. This allows, for example, for the effect of a comma on topic change of the subsequent word to be different across topics. In the AT-LDA model without covariates, we estimate a topic-specific baseline probability for topic carryover only and thus ignore all conjunction/punctuation (C/P) information in terms of topic change.

Our set of models allows us to analyze the role of structural information in text in a detailed way. The standard LDA model makes no use of conjunctions or punctuation in text because both are treated as noise with respect to topic discovery and, consequently, are discarded from the data prior to running the models. Thus, the LDA model can be viewed as a null model that assumes that structural information is irrelevant to topic discovery. The SC-LDA model uses punctuation to a priori define sections in which topics are identical so that, compared with the LDA model, the benefit of assigning homogeneous topics to all words in sentences can be evaluated. The CPC-LDA model, compared with the SC-LDA model, allows us to observe the benefit of breaking (typically longer) sentences into shorter sections with homogeneous topics, defined by conjunctions. The AT-LDA model with covariates uses conjunctions and punctuation as covariates to a probabilistic topic carryover across words. Compared with the LDA model with "hard" constraints (SC/CPC-LDA models), homogeneous topic assignments to strings of words are possible but not imposed. Instead, the extent to which topics carry over across words within local sections of text is learned from the data (Table 2). In Section 4.3, we show how results from these two assumptions differ.

For model comparison, we report the (log) marginal likelihood of the text data (Table 3), allowing for direct comparison of the models using Bayes' factors. For this measure, we use all reviews. Because the number of topics is not a parameter of our models, we estimate each model for alternative T and choose the best-fitting model in terms of predictive fit for all further analysis. Note that the likelihood reported is based on the corpus of words. It does not account for the observed rating, for which we report a separate fit statistic in Table 8. Across all data sets, we find that imposing hard constraints on topic assignments on locales defined by punctuation or conjunctions is not supported by our data (Table 3). Compared with results from the standard (supervised) LDA model, both the SC-LDA and CPC-LDA models do not improve in-sample fit to the text data. More flexible ways of modeling (LDA, AT-LDA) are preferred.

An interesting result is obtained from the dog food data: With respect to in-sample fit, the standard LDA model outperforms all other models, including the proposed AT-LDA model. For the other three data sets, we obtain the opposite result. It appears that the LDA model is a useful tool for topic detection when reviews are short (Table 1). In summary, applying the set of models to our data sets suggests that, in longer reviews, (1) topics exhibit serial dependency, (2) this autocorrelation of topics is not well modeled by imposing prior constraints via observed structural elements of text, and (3) using this structural information as covariates to word-to-word topic change improves the fit of the model. The latter is evidenced by a significant improvement in fit of the AT-LDA model when covariates are used (Table 3).

Table 4 reports out-of-sample fit results. For holdout data, we use a randomly selected sample of 20% of reviews from each data set. We report the log of the average likelihood obtained by first computing the posterior mean probability of observing each word in holdout reviews and then aggregating the (log) probabilities across all words in holdout reviews. We find that the proposed AT-LDA model with covariates outperforms all other models with respect to predictive fit across all four data sets, providing evidence for not eliminating structural elements of text in preprocessing and linking this information to the topic flow in customer reviews. This result also implies that models of local topic chunking are not preferred for our data.

Table 3. Fit Results

Model category	Model	Restaurant reviews	Camping tent reviews	Luxury hotel reviews	Dog food reviews
Bag of words	LDA	−963,344 (10)	−1,878,546 (15)	−421,002 (11)	**−361,274** (8)
Topic chunking	SC LDA	−1,064,867 (6)	−2,164,162 (14)	−427,964 (10)	−449,155 (8)
	Sticky SC LDA	−1,064,805 (6)	−2,151,230 (14)	−425,981 (10)	−456,477 (8)
	CPC LDA	−1,012,657 (8)	−2,006,406 (14)	−410,286 (10)	−425,301 (8)
	Sticky CPC LDA	−1,052,217 (8)	−2,092,006 (14)	−409,896 (10)	−443,896 (8)
Topic carryover	AT-LDA without covariates	−920,380 (10)	−1,743,012 (16)	**−347,457** (10)	−387,500 (8)
	AT-LDA with covariates	**−916,784** (10)	**−1,731,181** (16)	−347,824 (10)	−383,425 (8)

Notes. Reported are log marginal densities of the data, given model, and number of topics (number of topics in parentheses). The best-fitting model is highlighted. Covariates refer to use of conjunctions and punctuation as prior information to topic carryover.

Table 4. Predictive Fit Results

Model category	Model	Restaurant reviews	Camping tent reviews	Luxury hotel reviews	Dog food reviews
Bag of words	LDA	–221,900	–481,697	–93,614	–129,378
Topic chunking	SC LDA	–221,709	–479,977	–92,957	–129,550
	Sticky SC LDA	–221,640	–478,655	–92,986	–129,365
	CPC LDA	–221,713	–479,509	–92,952	–129,423
	Sticky CPC LDA	–221,594	–478,683	–92,962	–129,289
Topic carryover	AT-LDA w/o covariates	–221,600	–478,753	–92,976	–129,307
	AT-LDA w covariates	**–220,661**	**–475,974**	**–92,785**	**–128,802**

Notes. Reported is the log average likelihood of holdout data. The number of topics is the same as in Table 3. The best-fitting model is highlighted. Covariates refer to use of conjunctions and punctuation as prior information to topic carryover.

4.2. Model Estimates from the AT-LDA Model

4.2.1. Local Topic Dependency. The AT-LDA model provides direct estimates of (local) topic dependency via ψ_t, the topic-specific probability of a word-to-word topic carryover. This parameter indicates how topics extend across words independent of the prior topic distribution $p(z|\theta_d)$. In Table 5, we report results with respect to local topic dependency from our data sets.

Results from our model indicate that topics exhibit dependency across words. This is evidenced by topic carryover probabilities of approximately 0.5, indicating that, for example, a topic *run* of length three has probability $0.5^2 = 0.25$. For some topics, we find ψ to be close to 0.7 (e.g., restaurants, camping tents), suggesting that extended sequences of words generated by carryover of a single topic have a nonmarginal probability. For example, in the restaurant data, we find sequences of repeated topic carryover of up to 34 words (camping tents: 28 words). In summary, our results suggest that topics in our data sets exhibit significant serial dependency, implying that the i.i.d. assumption of the LDA model is untenable. This also suggests that inference regarding θ_d may be biased when the standard LDA model is applied, In our analysis of models with respect to predicting customer ratings (Section 5), we return to this issue.

4.2.2. Influence of Syntactic Covariates on Local Topic Carryover. An important element of our model with autocorrelated topics is the relationship of syntactic elements of text and local topic dependency. The model allows for conjunctions and punctuation to change the probability of topic carryover. Table 6 shows how ψ changes as a result of the presence of selected covariates, marginalized over topics and the presence of other covariates (note that, e.g., a sentence may begin with the word "and," which leads to the presence of multiple covariates). For our illustration, we use results from the camping tent data. Table 6 reveals that full stops, commas, and question marks have a significant influence on the probability of a topic carryover. The presence of a full stop in front of a word reduces the probability of a topic carryover from the previous word from 51% to 4%, a 92% reduction in probability. A comma reduces this probability from 48% to 21%. A question mark cuts ψ by more than half (from 46% to 21%); an exclamation point or closed parenthesis drives it to nearly zero, indicating a possible topic change point. We obtain similar results from conjunctions (Table 6): The presence of "because" or "but" significantly reduces ψ, whereas "and" has relatively little effect. These results suggest that structural elements in text are indicators of possible topic change. In Online Appendix A.1, we present more detailed results of the hierarchical regression of topic carryover on the structural covariates (Equation (3)). These results provide further evidence that structural covariates exhibit significant influence on word-to-word topic carryover. In our analysis, we find that the influence of covariates on topic carryover is different across topics and that some covariates, given topic, increase the probability of a carryover. Our results suggest that a sentence constraint that implicitly assumes (with respect to topics) homogeneous effects of full stops, etc. on topic change does not reflect the flow of topics well.

Table 5. Estimates of Local Topic Dependency from AT-LDA

Parameter	Restaurants	Camping tents	Luxury hotels	Dog food
ψ	0.42 (0.19; 0.66)	0.47 (0.29; 0.65)	0.48 (0.36; 0.59)	0.40 (0.11; 0.56)

Notes. Reported are posterior means of ψ, averaged over topics. Numbers in parentheses indicate minimum and maximum of posterior means across topics, respectively. Estimates are based on the model with covariates and marginalized over the presence of covariates.

Table 6. Marginal Topic Change Probabilities Given Incidents of Punctuation and Conjunctions in the Camping Tent Data

Structural covariate	.	,	?	!	)	"Because"	"But"	"And"	"Once"
Absent	0.509	0.475	0.460	0.463	0.462	0.461	0.465	0.464	0.460
Present	0.043	0.208	0.211	0.052	0.063	0.144	0.009	0.391	0.140

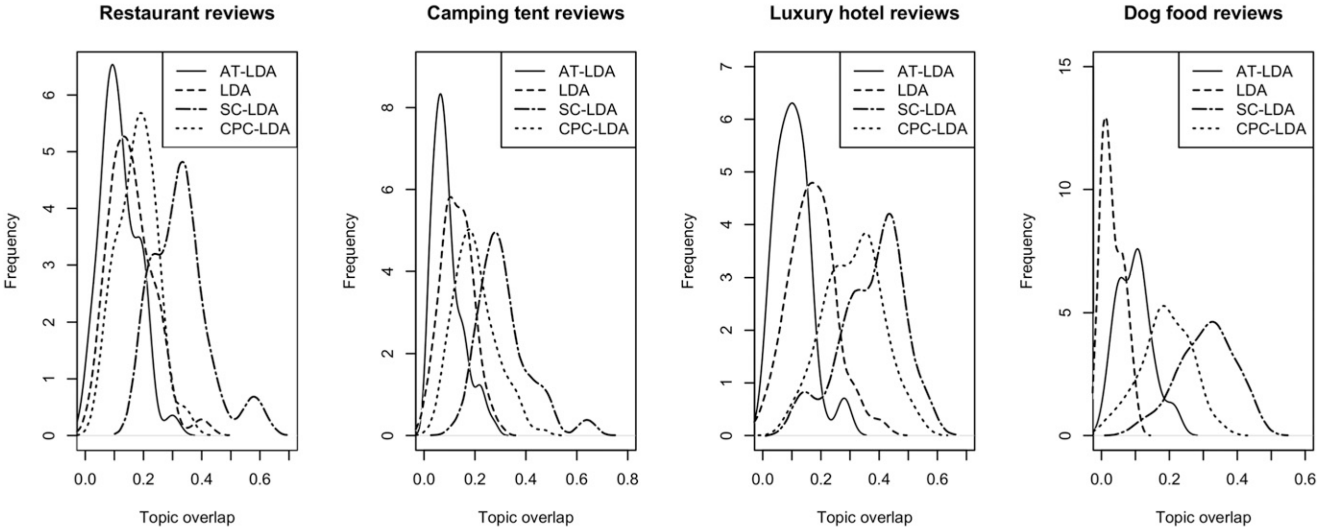
4.3. Topic Analysis

The intent of our analysis is to discover topics that are relevant to consumer ratings (Section 5). Topics, however, emerge ex post from the analysis and are not necessarily unique or coherent (Morstatter and Liu 2016). We begin our analysis of topics obtained from our data by considering the uniqueness and coherence of topics identified by the different models and then examine the effect of sentence conjunctions and punctuation on topic carryover.

4.3.1. Topic Uniqueness. The successful analysis of customer reviews should identify distinct and interpretable topics for improving customer experience. Uniqueness is important because our analysis of customer satisfaction aims to identify specific issues requiring action and further investigation. We investigate topic uniqueness by analyzing the degree to which topics are characterized by unique words. A perfectly unique topic does not share any of its most frequent words with another topic, and vice versa. For each model and data set, we compute the number of times the top 50 words from a topic appear among the top 50 words of the other topics for a given model. This gives rise to $T(T-1)/2$ overlap scores ranging, after normalization, from 0 (0 of 50 in common) to 1 (50 of 50 in common). Figure 2 plots the frequency of shared words of the topics emerging from the LDA, AT-LDA, SC-LDA, and CPC-LDA models for our three data sets, revealing that the AT-LDA model results in topics exhibiting less overlap of the top 50 words for the restaurant, luxury hotel, and camping tent data. For the dog food data, we find the LDA model to generate the most unique topics in terms of most frequent words. Again, this points to the LDA model as being a useful approach to modeling topics for less complex data.

4.3.2. Topic Coherence. *Coherence* refers to the semantic relationship among the most frequent words of a single topic. In a coherent topic, the most frequent words jointly describe a single theme. Compared with uniqueness, which is an *across-topic measure*, coherence is a *within-topic measure* of topic quality with no reference to other topics. Coherence is necessary for topics to be interpretable and for finding meaningful topic labels.

Figure 2. Uniqueness of Topics from the Three Data Sets Using Word Overlap Frequency



Our investigation of topic coherence starts by presenting topics resulting from the proposed model with autocorrelated topics and covariates. Table 7 shows the most frequent words for the camping tent data. Topic 12 talks about breaking poles ("pole," "broke"), ripped ground or plastic sheets, and resulting problems when staking out the tent. Topic 9 talks about a particular occasion or camping trip when the tent was used. Words such as "camping," "first," "trip," "weekend," "summer," "went," and "camp" all point at when or how the tent was used by a customer for the first time. In a similar fashion, one could proceed and identify the underlying theme for all topic and data sets in the way of a human expert. It is clear, however, that this is not an objective way of labeling a particular topic.

A formal linguistic analysis of coherence is based on semantic similarity of words making up a particular topic (Jiang and Conrath 1997). This approach uses structural lexical taxonomies of words (e.g., WordNet) that define cognitive synonyms of words and hierarchical relationships among words. On the basis of such strong a priori knowledge, the similarity of any pair of words can be computed. If the average pairwise similarity of all words in a topic is high, a topic is found to be coherent. Different approaches have been suggested for this purpose: Lowest common subsumer (Jiang and Conrath 1997) in the hierarchy of synsets (e.g., "breakfast" and "lunch" share the immediate superordinate "meal"), the number of nodes on the shortest path in the hierarchy from one term to the other, or the distance between two terms in the taxonomy, given the length of edges on that path. By the logic of distance of words in lexical taxonomies, topics consisting exclusively of words that are synonyms or exhibit an immediate super/subordinate relationship are highly coherent. From our analysis of text-based customer reviews, we find that this is rarely the case. Consider, for example, a topic corresponding to items purchased in restaurant reviews. Word pairs such as "chicken" and "barbecue" and "sauce" and "onion" exhibit relatively low textual similarity. The same is true for pairs such as "bought" and "tent" and "problem" and "rainfly" obtained from the camping tent data (Table 7). This suggests that methods based on semantic similarity do not apply well to an analysis of coherence of topics in customer reviews.

In the following, we employ two different approaches to evaluate topic coherence. The first approach is empirical and model based and will be explained in more detail later. A topic can be viewed as coherent if its distribution over words is different from suitable benchmark distributions of words (Fang et al. 2016). The second approach is human evaluation of topics, which is often viewed as a gold standard in measuring topic coherence (Chang et al. 2009, Newman et al. 2010, Lau et al. 2014, Morstatter and Liu 2016). We present results from human evaluation to measure topic coherence in Online Appendix A.3. As the model-based approach, we find human evaluation of topics from the AT-LDA model to be more coherent than those from the standard LDA model.

Table 7. Camping Tent Data

Rank	Topic 1	Topic 2	Topic 4	Topic 6	Topic 7	Topic 9	Topic 10	Topic 12	Topic 13	Topic 15	Topic 16
1	no	room	tent	tent	very	camping	rain	pole	tent	two	i
2	not	mattress	coleman	area	tent	time	wind	tent	great	2	love
3	tent	air	i	porch	well	first	weather	broke	good	times	used
4	water	queen	amazon	front	nice	trip	night	down	price	3	bought
5	problem	plenty	reviews	screened	easy	tent	heavy	stakes	size	family	tent
6	rainfly	two	i've	screen	roomy	weekend	winds	rainfly	love	years	husband
7	seams	fit	used	tarp	good	night	storm	not	huge	4	i'm
8	rain	size	bought	room	spacious	next	during	stake	awesome	5	wanted
9	leaks	2	new	floor	setup	last	high	out	perfect	few	really
10	inside	space	instant	over	really	use	windy	ground	family	year	thought
11	through	enough	purchased	rainfly	happy	day	day	side	big	days	wife
12	leak	gear	buy	door	pretty	used	through	top	best	old	took
13	waterproof	inside	not	great	big	second	not	plastic	quality	took	wish
14	issues	bag	another	top	large	go	cold	ripped	deal	6	can
15	without	around	product	nice	made	summer	tent	over	money	people	still
16	leaked	lots	replacement	sleeping	quality	went	bad	two	far	kids	expected
17	seam	still	many	under	enough	camp	light	door	buy	couple	far
18	issue	tent	eight-person	keep	seems	long	hard	tie	thing	ten	purchased
19	drop	side	again	part	super	week	strong	fly	excellent	tent	liked
20	floor	comfortably	made	main	sturdy	rained	hot	zipper	amazing	three	son

Notes. Most frequent words (top 20) from selected topics using the AT-LDA model with C/P covariates ($T = 16$). Topics 3, 5, 8, 11, and 14 are omitted for readability.

A statistical model-based approach to measuring topic coherence is given by the Kullback–Leibler (KL) divergence of ϕ_t with respect to a benchmark distribution Q:

$$D_{KL}(\phi_t \| Q) = \sum_{v=1}^{V} \phi_{v,t} log \frac{\phi_{v,t}}{q_v}, \tag{7}$$

where $Q = (q_1, \dots, q_V)$. Different choices for Q are available. An "uninformative" prior guess for Q is $q_v = \frac{1}{V}$, a uniform distribution. Equation (7) then measures the entropy of ϕ_t relative to all terms being equally likely. A data-driven choice for Q is based on $q_v = \frac{c_v}{C}$, the relative frequency of terms in the corpus. Essentially, this is ϕ, the word probabilities marginalized with respect to topics. Marginal probabilities present the least coherent "topic" available from the data as, in this observed "topic," all latent topics coappear. Figure 3 presents the distribution of statistical coherence scores based on KL divergence for our models and our data sets. Note that for each topic from a model applied to one data set, we obtain a KL divergence score given a particular choice of Q. So, for example, applying the AT-LDA model to the camping tent data results in 16 KL divergence scores relative to the uniform distribution. Figure 3 shows the distribution of these scores for given models and our four data sets. In general, a uniform distribution as choice of Q results in higher KL divergence scores than using the observed word frequency. Figure 3 reveals that the proposed model with autocorrelated topics generates topics, relative to marginal ϕ, that are statistically more coherent than any other model for the restaurant data, the camping tent data, and the hotel data. For the dog food data, the LDA model generates the most coherent topics. We find the same ranking of models relative to a uniform prior distribution of words.

5. Analysis of Customer Ratings

Central to our analysis is to discover how latent topics in reviews relate to customers' overall evaluations. We consider two analyses. The first is an analysis of the distribution of topics across documents for different ratings. We find that the distribution of topic probabilities changes markedly in moving from a one-star to a five-star rating, and we also show that the topic probabilities θ_d can be used to identify customer reviews of interest for further non-model-based analysis. The second analysis results from the cut-point regression model (Equation (5)), which provides a simultaneous mapping of all topic probabilities to the rating.

5.1. Distribution of Topic Probabilities Given Customers' Ratings

Figure 4 displays the distribution of topic probabilities for the restaurant data and the camping tent broken down by customer rating. For this analysis, we computed the posterior means of the topic shares at the document level and then computed across-document means of the topic shares given the different rating categories (one to five stars). The left panel of Figure 4 displays how (on average) topic shares change in documents when moving from a

Figure 3. Topic Coherence Scores Based on Kullback–Leibler Divergence

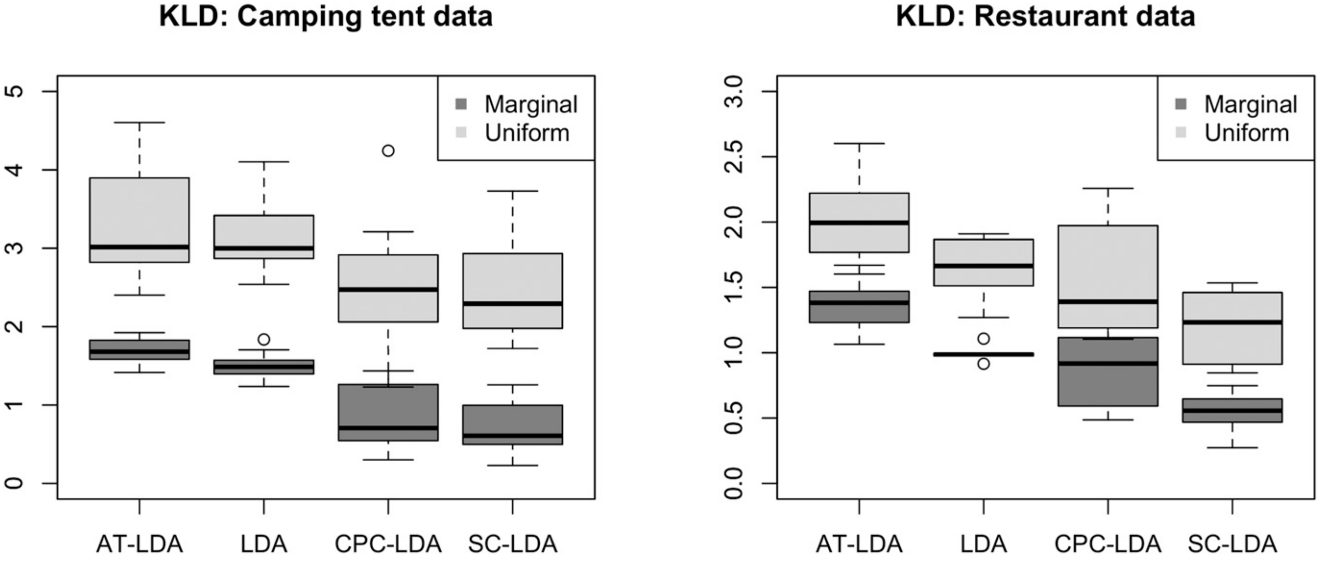
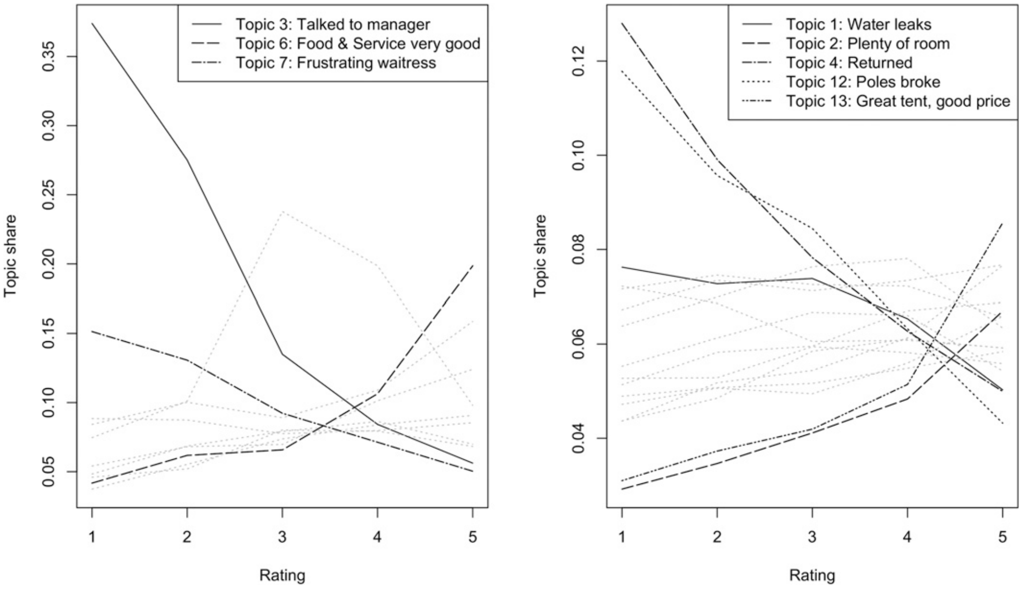


Figure 4. Distribution of Topics by Customer Rating for Restaurant Data (Left) and Camping Tent Data (Right)



Notes. Topics indicated by solid lines exhibit larger changes in shares when ratings are different. Labels of topics displayed in the legend are obtained by human evaluators (see Online Appendix A.3).

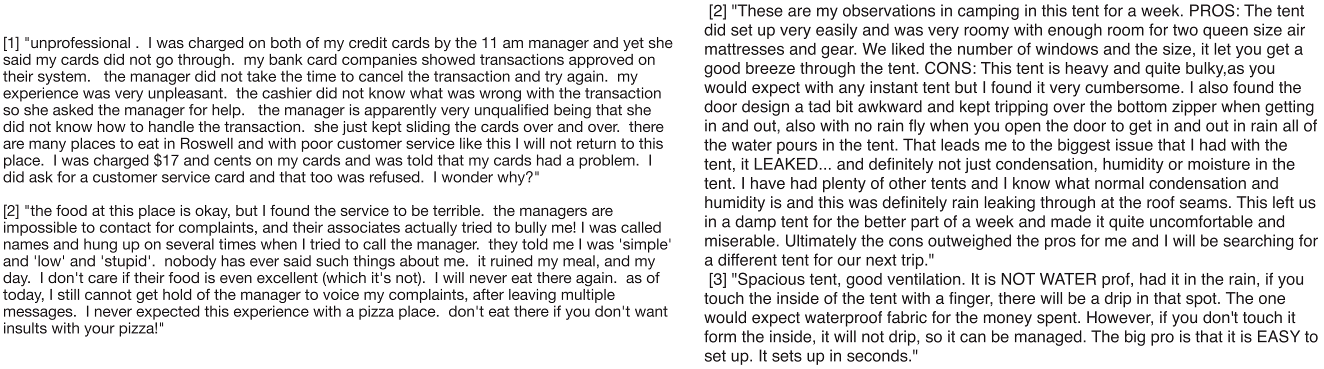
one-star rating to a five-star rating for the restaurant data, and the right panel shows the same for the camping tent data. From Figure 4, we see that the shares of many topics change given changes in ratings, indicating an empirical relationship of topic shares and ratings. We explore this relationship in more detail when discussing results from the cut-point regression (Section 5.2). From the restaurant data, we can see that low ratings are associated with a prevalence of topics 3 and 7 and the high ratings have a higher likelihood of topic 6 being present. Three- and four-star ratings of restaurants have a higher probability of being associated with topic 1. For all other topics, we observe a more uniform distribution of topic shares across ratings. For the camping tent data (Figure 4, right), we can see that higher shares of topics 1, 4, and 12 are associated with a higher likelihood of a bad rating, whereas higher shares of topics 2 and 13 are associated with good ratings. For both data sets, we find most topic shares to change monotonically with a change in rating. In part, this is caused by the assumption of a linear model (Equation (6)). In conclusion, from both data sets, we find that topic probabilities θ_d are indicative of the rating, which implies that they can be used to identify example reviews for each topic. As an example for topic-based document retrieval, we show in Figure 5 the two restaurant reviews richest in topic 3 (i.e., highest share of topic "bad service + talked to manager") and the two camping tent reviews richest in topic 1 ("water leaks"). By examining the original reviews, an analyst can obtain information about the strengths and weaknesses that are associated with the product under review.

5.2. Cut-Point Regression Results

The relative importance of topics can be assessed by using an ordinal probit model in which the latent customer rating τ_d is regressed on the topic probabilities θ_d (Equation (5)). Table 8 reports results from this regression. Results from sticky constrained models and the AT-LDA model without covariates are omitted for brevity. As a fit measure, we use an R^2-like measure of fit computed in the standard way but using the latent continuous rating τ as dependent variable. In our model, τ is augmented data, partially informed by the (also) latent θ. Thus, our fit measure for this regression is not based on exogenously given observations of the covariates. However, this holds for any model in our analysis, implying that no model is a priori at an advantage. Of course, models may differ a posteriori in how well they result in estimates of θ when the topic shares are assumed to map to the rating in a homogeneous fashion across reviews (i.e., common $\beta^{(reg)}$). In this sense, R^2 is additional information of the extent to which a model results in unbiased estimates of θ (see Section 4.2).

From Table 8, we find models of local topic dependency to outperform simple LDA models across all data sets. This suggests that models of topic dependency result in estimates of θ_d that map more closely to the customer rating. For the hotel and dog food data sets, we find the SC-LDA model to perform best among all models. The proposed AT-LDA model

Figure 5. Example of (Original) Customer Reviews of Restaurants Rich in Topic 3 (Left) and Camping Tent Reviews Rich in Topic 1 (Right)



[1] "unprofessional . I was charged on both of my credit cards by the 11 am manager and yet she said my cards did not go through. my bank card companies showed transactions approved on their system. the manager did not take the time to cancel the transaction and try again. my experience was very unpleasant. the cashier did not know what was wrong with the transaction so she asked the manager for help. the manager is apparently very unqualified being that she did not know how to handle the transaction. she just kept sliding the cards over and over. there are many places to eat in Roswell and with poor customer service like this I will not return to this place. I was charged $17 and cents on my cards and was told that my cards had a problem. I did ask for a customer service card and that too was refused. I wonder why?"

[2] "the food at this place is okay, but I found the service to be terrible. the managers are impossible to contact for complaints, and their associates actually tried to bully me! I was called names and hung up on several times when I tried to call the manager. they told me I was 'simple' and 'low' and 'stupid'. nobody has ever said such things about me. it ruined my meal, and my day. I don't care if their food is even excellent (which it's not). I will never eat there again. as of today, I still cannot get hold of the manager to voice my complaints, after leaving multiple messages. I never expected this experience with a pizza place. don't eat there if you don't want insults with your pizza!"

[2] "These are my observations in camping in this tent for a week. PROS: The tent did set up very easily and was very roomy with enough room for two queen size air mattresses and gear. We liked the number of windows and the size, it let you get a good breeze through the tent. CONS: This tent is heavy and quite bulky,as you would expect with any instant tent but I found it very cumbersome. I also found the door design a tad bit awkward and kept tripping over the bottom zipper when getting in and out, also with no rain fly when you open the door to get in and out in rain all of the water pours in the tent. That leads me to the biggest issue that I had with the tent, it LEAKED... and definitely not just condensation, humidity or moisture in the tent. I have had plenty of other tents and I know what normal condensation and humidity is and this was definitely rain leaking through at the roof seams. This left us in a damp tent for the better part of a week and made it quite uncomfortable and miserable. Ultimately the cons outweighed the pros for me and I will be searching for a different tent for our next trip."

[3] "Spacious tent, good ventilation. It is NOT WATER prof, had it in the rain, if you touch the inside of the tent with a finger, there will be a drip in that spot. The one would expect waterproof fabric for the money spent. However, if you don't touch it form the inside, it will not drip, so it can be managed. The big pro is that it is EASY to set up. It sets up in seconds."

with structural covariates is best for the camping tent and restaurant data sets, both of which contain more words per review, more and longer sentences (Table 1), and also more incidents of covariates per document (Table 2) than the dog food data set. Interestingly, the number of unique words in a corpus does not seem to play a large role. Recall that we find the SC-LDA model to perform well for the dog food data set, which contains the largest vocabulary (Table 1). This suggests that the advantage of the AT-LDA model with respect to predicting the customer rating is not related to the lexical complexity of the text but rather to the number and length of phrases in the reviews.

Tables 9 and 10 summarize results from regressing customers' overall ratings on the latent topics. Note that the regression model is not a priori identified because the topic shares from an LDA-type model, by definition, sum to one. We postprocess the coefficients by setting the coefficient of one conveniently chosen topic to zero. As labels for the top three topics, we picked to the most frequent label supplied by human evaluators (Section 4.3.2). For the remaining topics, we obtained labels manually by generating phrases from the most frequent words.

Tables 9 and 10 reveal that regression coefficients of many topics are empirically associated with the rating, as evidenced by more than 95% of their mass away from zero. It is also interesting to note that several topics map negatively on the rating. For example, in the tent data set, an increase in the topic "interaction with manager or owner" of 10% is associated with a decline of the (latent continuous) rating of –1.7. This decline is significantly higher than first differences among cut points, implying that this increase is equivalent to a decline in the rating of one or more rating scale points. Note that regression coefficients are defined as the change in y as x changes by one unit. In our case, given the use of topic shares as covariates, this would be the difference between 0% and 100% share of a topic. Hence, to compute the impact of a 1% change in the *share* of a topic, the coefficients in Tables 9 and 10 have to be divided by 100. From the camping tent data set, we find that the topic "poles and stakes broke" has a similarly high negative influence on the rating. A 10% increase in this topic changes the rating by –1.9, a two-point decrease in the ordinal rating on average. As a consequence, a review in which this topic has a 50% share has the lowest rating possible. From a managerial perspective, results from the cut-point regression suggest many possible avenues to improve customer experience. For restaurant managers, avoiding frustration with waiters and escalation of conflict ("people wanted to talk to manager or owner") are of foremost importance. In comparison, disappointment with food plays a much smaller role. Of course, a positive experience with service and food has a positive influence on the rating, but it cannot compensate for a frustrating service experience. From the camping tent

Table 8. Explained Variance of Customer Rating

Model category	Model	Restaurants	Camping tents	Luxury hotels	Dog food
Bag of words	LDA	0.631	0.603	0.441	0.626
Topic chunking	SC-LDA	0.652	0.683	0.656	**0.782**
	CPC-LDA	0.628	0.694	0.559	0.750
Carryover	AT-LDA	0.794	0.714	0.658	0.736

Notes. Reported is an R^2-like measure of the (latent, continuous) customer rating (Equation (6)). The number of topics is the same as in Table 3. The best model is highlighted.

Table 9. Restaurant Data: Results from Topic Regression, Using Best-Fitting Model (AT-LDA with C/P Covariates, T = 10)

Parameter	Topic	Posterior mean	Credibility level
Covariates			
β_0	Intercept	0.249	0.662
β_1	Really good sandwich	−1.133	0.924
β_2	This is the best pizza place	0.922	0.894
β_3	People wanted to talk to manager or owner	**−7.911**	1.000
β_4	Things ordered	1.013	0.875
β_5	Various items on menu	0.424	0.707
β_6	Food and service very good	**5.600**	1.000
β_7	Frustration with waitress	**−3.425**	1.000
β_8	Layout of restaurant	0[a]	—
β_9	Will not go back	**−1.570**	0.966
β_{10}	First dinner at this restaurant	−0.612	0.753
Cut-points			
c_1		−1.643[a]	—
c_2		**−1.163**	1.000
c_3		**−0.578**	0.998
c_4		0.128[a]	—
R^2		0.794	

Notes. Reported are posterior means of coefficients and credibility level. Credibility level is posterior mass away from zero. Regression coefficients credibly different from zero on 95% level are in boldface.
[a] Parameter fixed for identification.

data set, it seems highly critical to supply poles and stakes that do not brake when used and to make a tent rainproof in heavy weather. Both are major sources of disappointment for users. Positive attributes are room offered (topic: "tent has plenty of room for people"; topic: "number of people") and easy setup (topic: "tent can be set up easily"), both of which make a tent a much more useful tool. In conclusion, we find that topics obtained from customer reviews identify a large and diverse set of concrete attributes of products or services that, by linking them to the rating, are ranked by customer importance.

6. Concluding Remarks

In this paper, we examine the use of an autocorrelated topic model for analyzing text data. Topics are autocorrelated when they can carry over from word to word in speech. In our empirical analysis, we find that the proposed model with autocorrelated topics outperforms standard topic models and models of topic chunking across different data sets. The reason for this result is that topic carryover is a regular feature of customer review text data for which standard topic models cannot account. Although the i.i.d. assumption of topic assignment in LDA models has been criticized in the literature as unrealistic, we provide model-based evidence for violation of this assumption and a way to solve this problem.

In our application of the model to different data sets, we examine the role played by conjunctions and punctuation in signaling topic change. The difference between these two categories of covariates is that conjunctions are joiners of speech, and incidents of punctuation present natural separators of speech. Because we incorporate this information as covariates in the model, we can use it without compromising inference with respect to the topics themselves. In our empirical analysis, we find these syntactic covariates to be highly predictive of topic carryover. Typically, conjunctions and punctuation are removed prior to the model-based analysis of text data. The primary motive for this preprocessing is that such data are not diagnostic with respect to topics. Although this is true, our results suggest that syntactic covariates are highly diagnostic of topic *changes* and, through this

Table 10. Camping Tent Data: Results from Topic Regression, Using Best-Fitting Model (AT-LDA with C/P Covariates, T = 16)

Parameter	Topic	Posterior mean	Credibility level
Covariates			
β_0	Intercept	**0.865**	0.990
β_1	Problems with water leaks and rainfly	**−5.443**	1.000
β_2	Tent has plenty of room for people	**4.503**	1.000
β_3	I can't recommend this tent	−1.774	0.996
β_4	Returned tent to Amazon	**−7.335**	1.000
β_5	Needs better instructions	**−2.136**	1.000
β_6	Issues with porch and screen	−0.563	0.795
β_7	Very nice tent	0.021	0.520
β_8	Issues with door, zipper or window	**−2.928**	1.000
β_9	Occasion tent was used	**−3.000**	1.000
β_{10}	Heavy weather with winds and storm at night	**−1.474**	0.980
β_{11}	Tent can be set up easily	**3.098**	1.000
β_{12}	Poles and stakes broke	**−8.542**	1.000
β_{13}	Great tent, good price	**8.159**	1.000
β_{14}	Tent kept dry inside during rain	−0.110	0.607
β_{15}	Number of people	**1.596**	0.991
β_{16}	"I love it"	0[a]	—
Cut-points			
c_1		−1.773[a]	—
c_2		**−1.306**	1.000
c_3		**−0.922**	1.000
c_4		−0.286[a]	—
R^2		0.714	

Notes. Reported are posterior means of coefficients and credibility level. Credibility level is posterior mass away from zero. Regression coefficients credibly different from zero on 95% level are in boldface.
[a] Parameter fixed for identification.

mechanism, are useful in analyzing the latent structure of text. In short, our results present a strong case not to discard these data.

From a practical perspective, we find that a model with autocorrelated topics results in topics that map better to ratings of customer satisfaction if reviews are more complex (more words, larger vocabulary). This result can guide managers in moving away from simpler models when these do not suffice. Compared with results obtained via the standard approach to topic analysis (LDA model), our model with autocorrelated topics generally results in a larger and more diverse set of topics when applied to data sets comprised of larger sentences and more words. In our analysis of the customer review data, we find that the AT-LDA model identifies topics more focused on specific themes (e.g., rainfly issue with tent, breaking poles, frustrating interaction with waiter or waitress) and that the LDA model does not identify with similar clarity. This is because the LDA model has a tendency to allocate ubiquitous words (food, menu items) more uniformly across topics. This tendency also reduces its ability to explain customer satisfaction ratings. However, the increased power of a topic model to explain customer ratings when topics are serially dependent and its ability to predict new text data are only two of several pieces of convergent evidence that we present in favor of the proposed model. We find topics from the AT-LDA model (1) to exhibit less overlap with respect to the most frequent terms, (2) to be statistically more distinct in comparison with benchmark word probabilities, and (3) to be, according to human evaluators, easier to interpret and, hence, more coherent.

To conclude, this research suggests that the analysis of serial topic dependency presents a fruitful area of future research to advance the use of topic models for the rapidly increasing amount of text data in marketing. The models proposed here are relatively simple in that they considered first-order topic dependency across an observed sequence of words only. Yet our model outperforms a model with i.i.d. topic draws and models of topic chunking across all four data sets in terms of out-of-sample fit (Table 4). Results from our model suggest that topic dependency often extends across longer sequences of words (*topic chunking*), suggesting the need for a more complex model of serial topic dependency.

References

Blei DM, Lafferty JD (2007) A correlated topic model of science. *Ann. Appl. Statist.* 1(1):17–35.

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

Büschken J, Allenby GM (2016) Sentence-based text analysis for customer reviews. *Marketing Sci.* 35(6):953–975.

Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM (2009) Reading tea leaves: How humans interpret topic models. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Advances in Neural Information Processing Systems*, vol. 22 (Curran Associates, Red Hook, NY), 288–296.

Chib S (1998) Estimation and comparison of multiple change-point models. *J. Econometrics* 86(2):221–241.

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.

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

Fang A, Macdonald C, Ounis I, Habel P (2016) Topics in tweets: A user study of topic coherence metrics for twitter data. *Eur. Conf. Inform. Retrieval* (Springer, Cham, Switzerland), 492–504.

Gopalakrishnan A, Bradlow ET, Fader PS (2016) A cross-cohort changepoint model for customer-base analysis. *Marketing Sci.* 36(2):195–213.

Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB (2004) Integrating topics and syntax. Saul LK, Weiss Y, Bottou L, eds. *Advances in Neural Information Processing Systems*, vol. 17 (Curran Associates, Red Hook, NY), 537–544.

Humphreys A, Wang RJ-H (2017) Automated text analysis for consumer research. *J. Consumer Res.* 44(6):1274–1306.

Jiang JJ, Conrath DW (1997) Semantic similarity based on corpus statistics and lexical taxonomy. Preprint, submitted September 7, https://arxiv.org/abs/cmp-lg/9709008.

Johnson VE, Albert JH (2006) *Ordinal Data Modeling* (Springer Science and Business Media, New York).

Lau JH, Newman D, Baldwin T (2014) Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), 530–539.

Mcauliffe JD, Blei DM (2007) Supervised topic models. Platt JC, Koller D, Singer Y, Roweis ST, eds. *Advances in Neural Information Processing Systems*, vol. 20 (Curran Associates, Red Hook, NY), 121–128.

Meyer CF (1987) *A Linguistic Study of American Punctuation* (Peter Lang, Bern, Switzerland).

Morstatter F, Liu H (2016) A novel measure for coherence in statistical topic models. *54th Annual Meeting Assoc. Comput. Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), 543–548.

Nallapati R, Allan J (2002) Capturing term dependencies using a language model based on sentence trees. *Proc. 11th Internat. Conf. Inform. Knowledge Management* (ACM, New York), 383–390.

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

Newman D, Lau JH, Grieser K, Baldwin T (2010) Automatic evaluation of topic coherence. *Human Language Tech.: 2010 Annual Conf. North Amer. Chapter Assoc. Comput. Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), 100–108.

Say B, Akman V (1996) Current approaches to punctuation in computational linguistics. *Comput. Humanities* 30(6):457–469.

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.

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

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