International Journal of Research in Marketing 43 (2026) 317–343

Contents lists available at ScienceDirect

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full length article

A topic change point model for phrase-based topic inference

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

[a] School of Management, Catholic University of Eichstätt-Ingolstadt & Ohio State University, USA
[b] Fisher College of Business, Ohio State University, USA

ARTICLE INFO

Article history:
First received on 6 June 2024 and was under review for 7½ months.
Area Editor: Liye Ma
Accepting Editor: Koen Pauwels
Available online 5 June 2025

Keywords:
Text data
Topic models
LDA
Transformer
BERT
Change point model
Big data
Time-series data
Bayesian analysis

ABSTRACT

Obtaining customer insights from large and unstructured text corpora (e.g. customer reviews, blogs, tweets, written exchange in user forums or customer service emails) has attracted significant interest in marketing research. Topic models are among the most popular descriptive devices to analyze such data. We propose a new type of topic model built on observed word sequences. The proposed topic change point model assumes that text consists of sequences of words belonging to the same topic, possibly interspersed by ubiquitous terms such as stop words, and that topics change from run to run so that the end of each topic run marks a topic change point. Our model yields strings of words assigned to the same topic for phrase-based topic inference. In an extension of our model, we use parts-of-speech tags as prior information to topic change points which allows for observed syntactical structure of text to enter topic inference. We apply our model to two data sets and compare it to alternative modeling approaches, including state-of-the-art, BERT based topic models. We investigate the capability of models to predict consumer ratings which addresses their power in summarizing words so that the origin of ratings can be assessed by marketers. Implications and directions for future research are discussed.

1. Introduction

Text data arises from a process where there is a communication intent, which simultaneously influences an approximated length of a text passage and its associated topic. Sometimes it is easy to convey a thought and the passage length is small, and sometimes the topic is highly nuanced and more words are needed to be precise in conveying an idea or sentiment. Moreover, longer passages may be driven by greater passion as a person's desire to be understood increases. Text data arising in product reviews are most likely a mixture of intentions, some of which describe aspects of life that are mundane (e.g., where a product was purchased) to aspects that are important (e.g., how the product failed at the worst possible time).

Current statistical models used to understand text data in marketing are not designed to fully understand the communication intent of the writer. The latent Dirichlet allocation (LDA) model (Blei et al., 2003) is built on a bag-of-words assumption that does not account for the local allocation of words from the same topic, or topic chunking. Sentence-based analysis forces words within a sentence to come from the same latent topic (Büschken and Allenby, 2016), and the presence of sentence punctuation and conjunctions (Büschken and Allenby, 2020) have been shown to be beneficial in relaxing the requirement that topic changes occur only at observable syntactical points. However, these models still rely on the possibility of a topic change at each word in the text.

* Corresponding author.
E-mail addresses: joachim.bueschken@ku.de (J. Büschken), allenby.1@osu.edu (G.M. Allenby).

https://doi.org/10.1016/j.ijresmar.2025.05.004
0167-8116/

In this paper we develop and apply a new topic model for text analysis that more fully embraces the idea of a communication intent driving the presence of words in a document. Communication intent gives rise to an allocation of words associated with a common topic that does not change. Topic changes occur at the next allocation of words, and the text document is assumed to be constructed by repeating this generative process of the data.

The generation of a sequence of words from a common topic allows for an analysis of phrases instead of words. That is, our model allows for the analysis of specific phrases used in reviews to improve the interpretation of the latent topics. Phrase lengths can also be used to improve text analysis as it indicates the effort respondents are willing to invest in the communication. We demonstrate the advantage of these statistics that are afforded by our proposed model.

A common problem encountered in the analysis of text data is how to deal with ubiquitous terms that are typically removed from an analysis. Ubiquitous terms have a high probability of occurring and are thought to not provide information useful for interpreting topics. Examples include conjunctions (e.g., and, but, for) which are typically assumed content or context neutral but are important to the construction of coherent phrases of speech. We address the presence of these terms by incorporating a mechanism that probabilistically filters out these commonly occurring words when they are of no value to topic interpretation. We call this mechanism a "non-topic". Sometimes a term such as the word "to" is simply for grammatical correctness of a phrase or sentence. A typical example is the phrase "*we went to see a Broadway show*" from our hotel review data in which this term is necessary for correct use of the (infinitive) verb. Compare this to the phrase "*we went to Yosemite for the weekend*" from our camping tent reviews where this term indicates the use of "Yosemite" as a destination. We find that our model flexibly assigns ubiquitous terms to topics and the non-topic, depending on local contexts. Based on this, we also find that ubiquitous terms are sometimes integral to understanding a phrase and should therefore not be filtered out for the purpose of topic inference.

To support topic inference, we use parts-of-speech (POS) syntactical meta-data as prior information to topic change points. POS labels indicate the syntactical role of words (e.g. adverb, noun, adjective) within phrases of speech. We link these labels to a word-level topic change point indicator and find that the functional role of words is highly predictive of topic changes. More importantly, we find improved change point inference to directly improve topic inference as such. We demonstrate the performance of our proposed model using two product review data sets and find that it outperforms alternative approaches.

A typical application of model-based text analysis in marketing is the analysis of the relationship between customer evaluations or ratings and (latent) topics in text data originating from users or buyers. The purpose of this analysis is to explore the origin of ratings through textual features of reviews. Topic models have been shown to provide novel insights into various aspects of consumer behavior (Dellarocas et al., 2007; Lee and Bradlow, 2011; Archak et al., 2011; Büschken and Allenby, 2016). We extend this analysis by relating consumers' ratings to features of documents derived from topic runs which condition on the observed order of words.

Large language models (LLM) have attracted significant attention and demonstrated the usefulness of deep-learning approaches to understanding and generating text (Vaswani, 2017; Kenton et al., 2019). A potential disadvantage of such models is that their results are not readily interpretable and need (very) large data sets for training. Our application of text analysis presents a typical application in marketing in which a targeted data set (combination of product category, competing brands, location and price point) is analyzed with the goal of generating novel insights into consumer behavior. In particular, we are interested in exploring the origin of customer product ratings under these conditions, using interpretable model estimates for this purpose. Topic models have proven to be very capable of delivering such insights even with relatively small data sets (Zhang et al., 2024).

The remainder of our paper is structured as follows: in Section 2, we develop the proposed topic change point model (TCPM) with ubiquitous terms. Section 3 describes the data sets used in our empirical analysis. Section 4 presents results from applying our model and comparing results to alternative modeling approaches as well we topics identified by human readers. Our set of benchmark models includes a BERT-based topic model to assess an LDA-based approach to using contextual document embeddings. Section 5 summarizes our findings and offers concluding remarks.

2. A topic change point model with ubiquitous terms

In the following, we develop the proposed topic change point model formally and explain the ideas behind it. We start by providing a motivating example to outline the key features of our model. Essentially, we view text as a sequence of topic discussions where each topic appears in the form of phrases where each phrase is a particular sequence of words. Within phrases, topic words may be interspersed by ubiquitous terms whose role is primarily functional. A central feature of our model is conditioning on the observed order of words and the assumption that topic dependency is local. In very simple terms, authors of product reviews chose a topic to talk about and stick with it until they move to the next topic. For simplicity, we assume the order in which topics are discussed (and topic phrases appear) to be irrelevant. In that, our model bears resemblance to the SC-LDA (Büschken and Allenby, 2016) which is a "bag-of-sentence" model. Our model can be viewed as a "bag-of-phrases" model where change points between phrases are not observed. The key advantage of our model is that typical topic phases can be used to understand what "topics" actually talk about.

2.1. Overview and motivation

Stylized example. Fig. (1) presents a stylized example for data generated by the proposed topic change point model (henceforth: TCPM). A document is assumed to consist of topic runs (color-coded) of certain lengths, interspersed by purely functional ("ubiquitous") non-topic words ("the", "and", "but" etc.). A topic run is a sequence of topic words originating from the same topic, defined by topic change points. A phrase is the sequence of all words between change points, containing topic words and ubiquitous words. In Fig. (1), ubiquitous words are indicated by $\tau = 0$. These words originate from a common (unigram) distribution over terms. The prior probability of a non-topic word occurring is assumed to be constant across documents. This constant reflects the grammatical role of non-topic words in speech which extends specific topics or contexts that an author talks about. In a sense, non-topic words are "syntactical glue" that combines topic words into meaningful expressions or phrases and this function is independent of context. The change points in our model simply result from a run ending and a new run beginning. By definition of change points, a new run must carry a different topic than the previous run.

Our TCPM does not assume certain terms (e.g. stop words) to be non-topic words. We use a unigram for ubiquitous terms, defined over the same vocabulary as topic terms. The difference is that the prior to ubiquitous terms is assumed to be homogeneous whereas the prior to topic terms is heterogeneous and varies with local context (documents and change points). This allows for non-topic words to be any words that appear at a constant frequency, irrespective of local context.[1] In our empirical application, we show that words such as "tent" or "room" appear frequently and independent of local context in reviews of camping tents and hotels, although they are not stop words. We show that their role in our data is functional, not topical in a sense of co-occurrence dependent on local context. Our model allows to probabilistically separate functional and topical terms. The key advantage of our model is its ability to identify phrases of speech from the data indicative of a specific topic which can then be used for topic inference. We explore this feature extensively in our empirical analysis.

It is interesting to compare our model to the standard LDA model in which topics are assigned to single words IID. An important deviation of our model is that topics are assigned to strings of words so that all words in a string carry the same topic. Within each string, non-topic words may appear but these words do not carry information with respect to topics. From string to string, topics must change. A priori, topic run lengths and topic change points are unknown. Ignoring non-topic words, setting topic run length to one and removing the constraint to the topic assignment of subsequent words gives the LDA model. The assignment of topics to word strings and the constraint that any topic run cannot originate from the topic as the previous run facilitates the identification of topic runs and presents a significant departure of the TCPM from the assumptions typically made by topic models. A model in which all words in a sentence carry the same topic (Büschken and Allenby, 2016) is a special case of our topic change point model. In the sentence-based LDA, topic runs are given by incidents of punctuation and, thus, observed. However, this model does not impose topic changes at the end of a run and it also does not consider constant frequency of certain terms which may dilute topic inference and must therefore be discarded.

Literature. The last word in each topic run of our TCPM is, by definition, a topic change point. It arises from assuming that the next run in a sequence of words originates from a prior distribution over topics constrained so that the current topic is excluded. The purpose of change point models lies is detecting regime changes in time-series data. Early work in marketing on regime changes has focused on purchase data and the change between an "active" and an "inactive" state of customers (Schmittlein et al., 1987). This approach has been extended in various ways (Fader and Hardie, 2009) and has also been applied to other areas such as to new product sales (Fader et al., 2004). In these models, regime changes reside on the individual level and end in an absorbing state (customer "death").

We apply the idea of regime changes to text data which, by the nature of speech, is time-series data. We consider such changes on the level of sequences of words within documents. The proposed topic change point model conditions topic inference on the locality of each word and their order of appearance. Our model is different from the topic change point model in Zhong and Schweidel (2020) who propose a topic model with a hierarchical prior distribution to document-level topic shares. This upper-level distribution in the model is allowed to change over time. Regime changes, as result, occur on the aggregate level (prior topic distribution to sets of documents, ordered in time). A similar approach on modeling change points is taken in Gopalakrishnan et al. (2016) who consider regime changes between customer cohorts. In comparison, we specify a change point model on the level of individual documents and we assume that the process driving topic run lengths resides on the corpus level, an approach similar to Schweidel and Fader (2009).

A variant of our TCPM allows for observed parts-of-speech (POS) covariates to drive unobserved change points. In this model, we link change points augmented by our MCMC to syntactical word classifications. Our use of POS covariates to inform topic change points is similar to the use of customer satisfaction as observed prior information to customers' change from an active to an inactive state (Ho et al., 2006). An important difference to our approach, however, is that we do not assume any topic to be an absorbing state. Instead, we only assume the topic of a current run is excluded when the topic for the next run is generated, giving rise to a change point.

How does improved topic inference matter for marketers? In our empirical analysis (Section 3), we show how change points allow to associate topics with phrases of speech, not (only) frequent terms. This is simply because topic phrases are by

[1] It would actually be simple to introduce substantial prior knowledge at this point by assuming e.g. stop words to have a higher prior probability to be ubiquitous terms than non-stop words. This could be achieved by specifying the subjective prior distribution to the word probabilities for ubiquitous terms accordingly. We thank an anonymous reviewer for this observation. We considered this approach but found it unnecessary with our data.

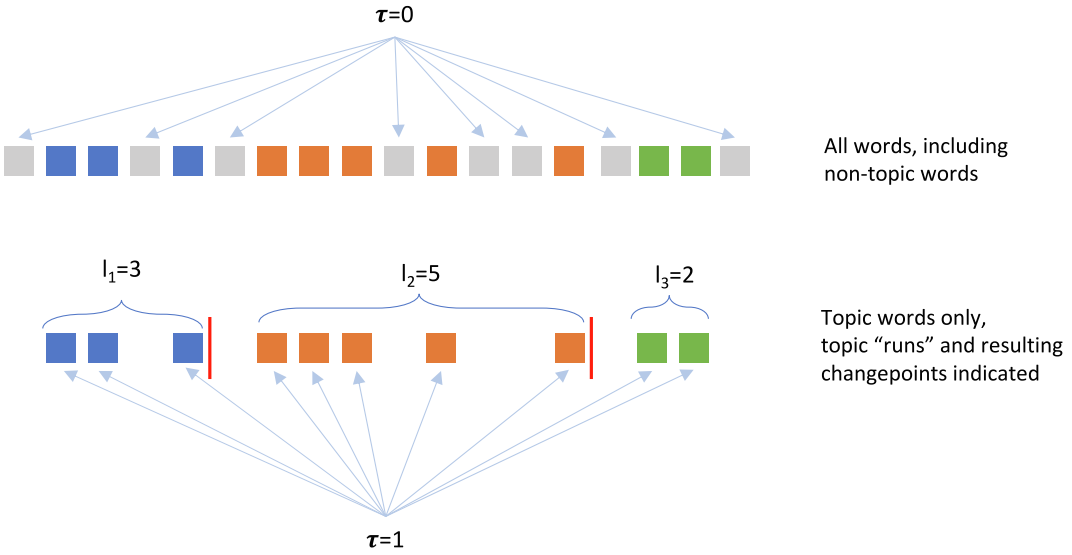


Fig. 1. Stylized example for data generated from the proposed change point model. Upper panel shows all words, including non-topic words. Lower panel shows topic words only. Topics indicated by color. Words colored grey indicate non-topic words. Vertical red bars indicate topic change points.

definition sequences of words from one topic, interspersed by non-topic functional terms. Note that word order is observed, whereas change points are not. Inference with respect to change points between runs and the topics of runs allow to find topic phrases. As we show, this lends topic terms the context necessary to understand what topics talk about. Phrases seem a more natural way to find sets of words indicative of topics compared to most frequent terms which can appear anywhere in a document.

Aside from improved topic inference, we show that topics derived from our TCPM exhibit higher predictive power of consumers' ratings. Ratings present a way to "supervise" inference of latent textual features of documents such as topic composition (Blei et al., 2003; Büschken and Allenby, 2020; Yang et al., 2022). Topics can in turn be used as predictors of ratings. We show in Section (5) that the TCPM contributes to an improved understanding of the origin of consumer ratings. When, e.g., the reason for product failure is known (from our camping tent data: breaking tent poles), marketers can move to eliminate the underlying weaknesses in product design.

2.2. Data generation process of the TCPM

For formal model development, we start by considering the data generating process of this model. Given N_d, the observed number of words in document d, the size of the vocabulary V in the corpus, fixed prior parameters $T, \alpha, \beta, \epsilon, \mu_\gamma, \sigma_\gamma, a, b$ and latent global communication intent v_d, assumed to be scalar, we assume words to be generated as follows:

1. draw $\delta : \delta \sim Beta(a, b)$: the corpus-level (i.e. constant) prior probability of a word being an ubiquitous term
2. draw $\{\phi_t\} : \phi_t \sim Dirichlet(\beta)$, IID $\forall t$: the topic-specific probabilities of words, each a vector of length V
3. draw $\Psi : \Psi \sim Dirichlet(\epsilon)$: the corpus-level probabilities of ubiquitous terms, a vector of length V
4. draw $\{\theta_d\} : \theta_d \sim Dirichlet(\alpha)$, IID for all documents d: topic shares on the document level, each a vector of length T where T is a scalar number, fixed a priori
5. draw $\{\gamma_t\} : \gamma_t \sim Gamma(\alpha_\gamma, \theta_\gamma)$, IID for each topic t
6. For $c_{d,1}$, i.e. a first string of words (and which we call a "phrase" henceforth), containing the first topic run $r_{d,1}$ in d (and omitting subscript d in the following to reduce clutter), we generate a word sequence consisting of topic words and non-topic words as follows:
   (a) draw $z_1 \sim$ Multinomial(θ)
   (b) draw topic run length $l_1 \sim Pois\left(\gamma_{z_1}\right)$, a zero-truncated Poisson distribution to ensure $l > 0$
   (c) generate topic and non-topic words so that non topic words are interspersed into the topic run at constant frequency:
      i. draw $\tau \sim Bernoulli(\delta)$
      ii. if $\tau = 1$, draw word $w \sim Multinomial(\phi_{z_1})$
      iii. if $\tau = 0$, draw word $w \sim Multinomial(\psi)$
      iv. repeat (i) to (iii) until l_1 topic words have been generated.
      v. compute length of the resulting string of words: c_1

(d) if $c_1 \geqslant v_d$: stop

7. If $c_1 < v_d$, generate c_2, the second phrase:
   (a) draw $z_2 \sim \text{Multinomial}(\theta, -z_1)$ where $-z_1$ indicates that this topic is excluded from the set of possible topics
   (b) draw topic run length $l_2 \sim Pois\left(\gamma_{z_2}\right)$
   (c) generate topic and non-topic words so that non topic words are interspersed into the topic run at constant frequency:
      i. draw $\tau \sim Bernoulli(\delta)$
      ii. if $\tau = 1$, draw word $w \sim Multinomial(\phi_{z_2})$
      iii. if $\tau = 0$, draw word $w \sim Multinomial(\psi)$
      iv. repeat (i) to (iii) until l_2 topic words have been generated
      v. compute length of the resulting string of words: c_2
   (d) if $c_1 + c_2 \geqslant v_d$ stop, otherwise repeat until $\sum c \geqslant v_d$
8. Compute the observed number of words across all phrases: $N_d = \sum c$.
9. Repeat steps $5 - 9$ $\forall d \in (2, \ldots, D)$

A number of features of this model are noteworthy. First, each phrase generated by our model must end with a topic word, not a non-topic word. That implies that the end of a phrase marks the end of a topic run and this identifies topic change points. Second, we assume runs of topic words to be interspersed by non-topic words randomly, but at a constant frequency. Note, however, that topic runs, by definition, consist of topic words only and that, therefore, non-topic words cannot impact topic inference. Third, we assume topic run lengths l to be (only) a function of $\gamma_t > 0$ which is a topic specific, corpus-level parameter to be estimated for every topic. Through this mechanism, we allow for characteristic run lengths of topics, some of which may be more verbose than others. In our empirical analysis, we find that topics vary greatly in terms of typical run lengths.

Fig. (2) presents the DAG of our model. In Appendix (A.1), we develop our model formally and in detail. For each document, the TCPM generates topic runs of length l. Topic runs are interspersed by ubiquitous terms at constant frequency, generated from ψ. The number of such non-topic words within runs is governed solely by δ which resides on the corpus level. The constraint $z_r \neq z_{r-1}$ imposes topic changes between consecutive runs so that the last word in each topic run marks a topic change point. Topics may exhibit a general tendency to generate certain run lengths, expressed by γ_t. The number of words in each document N_d is an outcome of the process, governed by the number of phrases necessary to convey global communication intent v_d. In the estimation of our model, we set $N_d = v_d$, treating global communication intent as observed, a standard practice in estimating topic models. For the LDA model, the sufficient statistics are the word counts by documents and N_d is only needed to normalize these frequencies.

The switching of topics between runs in our TCPM is induced by generating topic run lengths (implying that every run ends) and containing topics for runs so that the topic of current run cannot be identical to the topic of the previous run. Otherwise, the notion of change points would become meaningless. No other structure on topic switching is imposed. We do this for simplicity and note that topic switching could exhibit e.g. first-order dependency and leave such extensions of our model to future research.

Note that our TCPM does not employ a "supervised" approach to document-level topic shares θ (Mcauliffe and Blei, 2007; Büschken and Allenby, 2016). Supervised topic models link observed document-level labels such as consumer ratings to latent topic summaries for the purpose of predicting missing labels (Chakraborty et al., 2022). We argue that such models present a conflict between topic inference and prediction. The purpose of supervised models is to find topics which improve rating prediction, not to find topics which summarize a corpus of documents best. Our goal is to improve topic inference by way of (unobserved) topic phrases, not to predict labels, missing or otherwise. In order to not confound topic inference with predicting consumers' ratings (which in our case are not missing), we use an unsupervised approach.

2.3. Syntactical information and topic change points

In an extension of our TCPM, we assume that topic change points are related to observed syntactical classification of words (in the DAG indicated by X_d) through a regression model. For this purpose, we link the vector of (binary) word-level change points C_d, deterministically derived from z_d, l_d, τ_d, to observed parts-of-speech covariates. For a word indicating the end of a topic run: $C_{n,d} = 1$ and 0 otherwise. The purpose of this approach is to incorporate syntactical information into the model for improved inference regarding topic runs.

Given z_d, l_d, τ_d, our model implies a set of topic change points for each document. A change point is the last topic word in a topic run (Fig. 1). Thus, z_d, l_d, τ_d can be used to deterministically derive a vector of word-level change point indicators such as in the following example of a document containing 15 words:

$$z_d = (4, 4, NA, 4, 2, NA, 2, 2, 2, 1, NA, 1, 8, 8, NA)$$
$$C_d = (0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0)$$

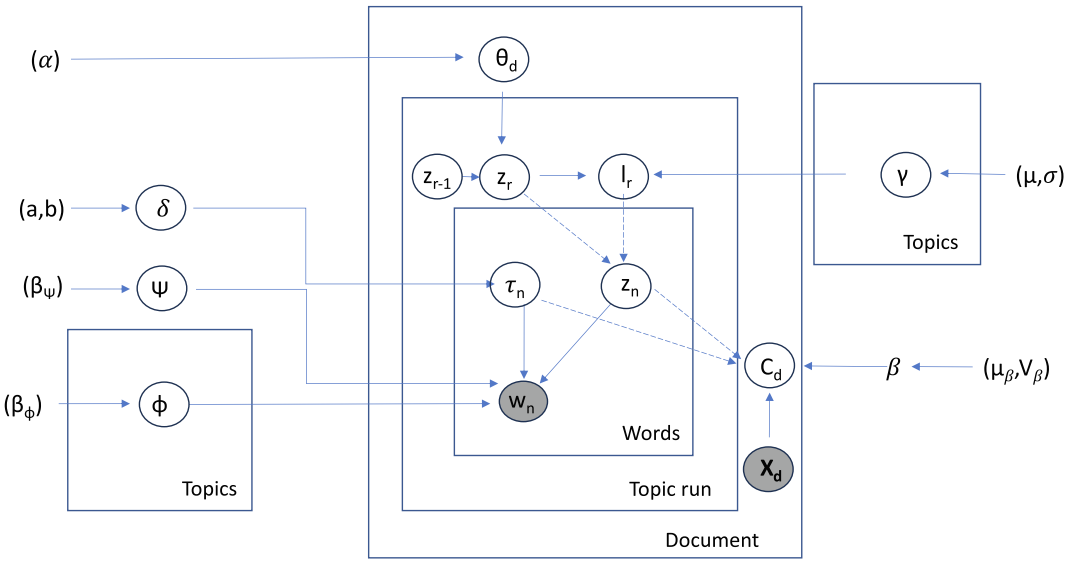


Fig. 2. Directed acyclic graph of the topic change point model with ubiquitous terms. Deterministic relationships indicated by dashed arrows. X_d contains observed POS covariates on word level. C_d indicates a vector of word-level, binary topic change point indicators. Quantities in shaded circles are observed, quantities inferred from data in clear circles.

where z_d contains topic assignments (including non-topic words, labeled "NA") and w_4, w_9, w_{12} are words with topic changes. From the DAG of our model (Fig. 2):

$$p(C_d|\tau_d, z_d, l_d, X_d, \beta) = p(C_d|X_d, \beta)$$

because of C_d being immediately derived from τ_d, z_d, l_d. We note that this distribution enters the joint distribution of knows and unknowns as a conditionally independent factor implying that, given draws of τ, z, l, inference regarding β can be conducted independently.

We link C_d across documents to word-evel parts-of-speech (POS) tags which label words with respect to their syntactical function (e.g. noun, adverb, verb) through a regression model:

$$p(c_{d,n} = 1|X, \beta, \tau_{d,n} = 1) = \frac{e^{X_{d,n}^T \beta}}{1 + e^{X_{d,n}^T \beta}} \tag{1}$$

where $X_{d,n}$ is a vector of covariates containing POS tags describing the syntactical function of word $w_{d,n}$ within a phrase. Note that:

$$p(c_{d,n} = 1|X, \beta, \tau_{d,n} = 0) = 0$$

because, in our model, non-topic words cannot be change points. We use conjunctions and incidents of punctuation (end, start of sentences) as covariates to change points, similar to Büschken and Allenby (2016) who find that these POS classifications are predictive of topic carry-over. The difference to our model lies in using POS data as indicators of topic changes across phrases of speech. The motivation for our approach is twofold. First, the (observed) syntactical structure of phrases and the functional role of words therein is crucial for understanding meaning. We exploit this by using POS as covariates in our model. (1) assumes POS labels to inform change points independent of topical context. Second, integration of POS meta-data by way of (1) implies that C_d contains likelihood information with respect to z_d, l_d, τ_d. Inference about topics in our model is therefore supported by syntactical data and not only word co-occurrence as in standard topic models. We use Apache OpenNLP to obtain POS tags on the word level for our data, using the syntactical categories listed in Appendix (A.4). Our MCMC estimation procedure for estimation of the TCPM is described in Appendix (A.2).[2]

3. Data

For our empirical analysis, we use two product review data sets previously used by Büschken and Allenby (2020). These are the luxury hotel and the camping tent data (see Table 1). This allows to compare results to prior research. For pre-

[2] An anonymous reviewer has correctly pointed out to us that this approach is computationally costly and may scale badly to larger data sets with which we fully agree. We have had success with a hybrid approach in which we ran a SC-LDA first to obtain a reasonable start solution and then switched to the MCMC for our TCPM. By our experience, this accelerates convergence greatly and therefore absorbs some of the scalability issues with our approach.

processing, we only change capital letters to small letters. An important difference to the usual pre-processing approach is that we do not prune stop words or any other frequently appearing terms, including conjunctions. This leads to larger corpora and longer documents. We only prune terms that appear less than three times in a corpus in total because such words carry no co-appearance information. This leads to larger vocabularies in all data sets because we allow for more rare terms to be included. As we will demonstrate in the empirical analysis, our approach to topic analysis lets rare terms enter topic interpretation (Section 4.3). We use punctuation to identify sentence breaks for the application of the sentence-based LDA (which is one of our benchmarks models). Both data sets exhibit the typical skew in consumer ratings towards high evaluations.

4. Model comparison

4.1. Comparison set

LDA-based topic models. To evaluate performance of the proposed TCPM, we compare the proposed TCPM to the plain-vanilla LDA, the AT-LDA with auto correlated topics (Büschken and Allenby, 2020) and the SC-LDA with a sentence constraint (Büschken and Allenby, 2016). Like our TCPM, AT-LDA and SC-LDA allow for local topic dependency and, hence, are not bag-of-words models. All these models, though, retain the capability of LDA to summarize documents via topic distributions.

We follow the approach by Büschken and Allenby (2020) and link topic carry-over in the AT-LDA to syntactical covariates. We use the non-topic for ubiquitous terms (Section 2.2) for both the LDA and SC-LDA. The AT-LDA does not easily give rise to a version with ubiquitous terms. This is because the presence of ubiquitous terms in any sequence of (topic) words would result in carry-over other than first order. Actually, the order would be a function of presence of non-topic words which are not observed. We leave development of this model to future research. Appendix (A.3) outlines our estimation procedure for both the SC-LDA and the LDA with ubiquitous terms. Note that these are also new models as result of incorporating a non-topic.

All LDA-based models in our benchmark set require fixing the number of topics T a priori. We follow the customary approach in the literature to test alternative T by way of predictive fit with respect to a hold-out data set. We use plain-vanilla LDA for both data sets and find T = 10 to be optimal for the hotel data and T = 12 for the camping tent data. Using other models in our set does not yield much different solutions. We fix the number of topics across all models equally for an even comparison.

Embedding-based topic models. To compare LDA-based models to state-of-the-art embedding models, we use c-Top2Vec proposed by Angelov and Inkpen (2024). Embeddings allow to move from sparse, one-hot representations of words in the vocabulary space to densely distributed representations in semantic space (Mikolov et al., 2013). c-Top2Vec leverages the self-attention mechanism in transformers (Vaswani, 2017). The key idea is to embed documents, words and topics in a single embedding space which allows to assess semantic similarity through distance (Angelov, 2020). A high concentration of documents in the embedding space (low distance) suggests common underlying topics. Essentially, topics are obtained through clustering document embeddings which are combinations of (contextual) word vectors. c-Top2Vec is an extension of BERTopic (Grootendorst, 2022) as it allows for document topic shares other than 0 or 1. In contrast to LDA, c-Top2Vec does not require to fix the number of topics a priori. Note, however, that c-Top2Vec as all BERT-based masked language models, is not based on a likelihood of (all) words in documents, making a likelihood-based comparison infeasible. [3]

A more direct approach to incorporate embeddings into LDA was proposed by Dieng et al. (2020). In their Embedding-LDA, the word-topic probabilities ϕ in LDA are decomposed into the product of word-embeddings and topic embeddings:

$$\phi_t = f(\Omega\alpha_t) \tag{2}$$

where Ω is a $V \times L$ matrix of word embeddings and α_t is a L-dimensional vector of topic embeddings. f is a function translating the matrix product in (2) to the simplex (e.g. softmax). This approach retains the capability of LDA to summarize documents by way of topics which are defined over the vocabulary. In comparison, c-Top2Vec results in topic vectors in the embedding space which does not lend itself to easy interpretation. The model by Dieng et al. (2020) allows to use pretrained word embeddings (i.e. a fixed matrix Ω) or to estimate Ω from the data. Using pre-trained embeddings extends the analysis to (previously) unobserved words as their topic probability can be computed by (2) independent of actual occurrence. For estimation of the Embedding-LDA model, we employ the Bayesian approach outlined in Appendix (A.6) and use pre-trained word embeddings from the English CoNNL17 corpus obtained by way of a continuous skip gram model.[4]

4.2. Evaluation of models

4.2.1. Evaluation of model fit

We evaluate model fit by computing the likelihood of the words in our reviews, integrated over topic assignments. A likelihood is not available for c-Top2Vec because BERT is a masked language model. We return to this model later when dis-

[3] Alternatively, a pre-trained large language model based on contextual embeddings such as GPT can be directly prompted to provide topics from a given data set. This is known as "zero-shot learning". This can be see as a model-free approach as the user has no control over the structure of the model behind the output from prompts. We tested this approach with one of our data sets and report some findings further down below.

[4] www.vectors.nlp.eu/repository.

Table 1
Descriptive statistics of data sets (after pre-processing).

	Camping Tents	Luxury Hotels
Number of reviews	7,983	3,216
Total number of words (corpus size)	701,623	162,181
Number of unique terms (vocabulary)	5,551	2,779
Number of words per review		
Mean	87.9	50.5
SD	115.3	44.4
Max	1,631	509
Share less than 10 words	12%	8%
Number of sentences per review		
Mean	6.2	4.3
SD	6.8	3.2
Consumer rating (5pt scale)		
Mean	4.19	4.42
SD	1.23	0.88

cussing models' power to predict the rating. In Appendix (A.7), we provide details on computing the integrated Likelihood for our models. We compute this quantity at every iteration of the MCMC chain and then compute the log marginal density (LMD). Note that this results in integrating over the (augmented) assignment of words to the non-topic for models with ubiquitous terms. LMD implicitly penalizes models with a larger parameter space because of the integration over the posterior distribution. Models with constraints on word-level topic assignments (AT-LDA, SC-LDA, TCPM) imply a posterior distribution of lower dimensionality compared to the LDA. Table (2) summarizes fit results from our models. Note that Top2Vec is not included as it does not define a likelihood for all words in the corpus.

From Table (2) we find that the TCPM outperforms all other topic models in all categories and for both data sets. We also find that allowing for ubiquitous terms improves the fit of all models. The use of POS covariates as prior information to topic change (or carry-over) further improves the fit which indicates that syntactical information contributes to the fit of a model. With respect to models allowing for local topic dependency, we find that all outperform a simple LDA implying that topics are locally dependent. However, the TCPM provides a significant improvement of fit over the AT-LDA and the SC-LDA. The implication is that both these models do not capture local topic dependency well. Our results suggest that topic dependency extends first order carry-over but does not extend to (complete) sentences. In our analysis of topic runs in the following section, we show this empirically.

4.2.2. Rating prediction

We also evaluate performance of models with respect to their capacity to predict consumer ratings. More specifically, we use R^2 with respect to the rating given topic solutions from our models. More specifically, we use the following covariates for each document:

- Share of ubiquitous terms
- Number of topic runs
- Average length of topic runs
- Topic shares

LDA-type models that allow for local topic dependency give rise to extended sequences of words assigned to the same topic. For standard LDA and Embedding-LDA, we take advantage of augmented word-topic assignments z_n to identify implied topic runs resulting from words in a sequence assigned to the same topic. We use this is to identify number and length of (implied) topic runs. Topic shares are simply estimates of θ_d which are available for each LDA-based model. We note that c-Top2Vec only provides an estimate of (the equivalent to) θ_d, but does not supply topic run-based statistics. This is because Top2Vec is build on document embeddings, which implies marginalizing over word or phrase level characteristics. Recall that c-Top2Vec extracts topics post hoc by comparing documents. We also note that a comparison of c-Top2Vec and LDA introduces a confound since Top2Vec requires prior results from a large language model trained on (very) large corpora. With the exception of Embedding-LDA, all our LDA-based models are trained from scratch. We return this issue further down below.

The capacity to explain the rating by way of topics gives rise to various downstream applications in marketing. Topics negatively associated with the rating point at reasons for service or product failure that managers can address via improved designs. Such topics may talk about specific situations in which failure occurs ("torrential downpour", "family weekend"), elucidating how consumers interact with the product. Topics positively associated with the rating, such as "great staff", provide feedback to employees as to their role in service provision. Table (3) displays predictive performance of our models with respect to the observed consumer rating.

[4] www.vectors.nlp.eu/repository.

Table 2
Fit results. Reported is LMD (per word), given model and number of topics. Number of topics in parentheses if applicable. T chosen to maximize predictive fit from Ubi-LDA. POS indicates models with hierarchical regression of topic dependency on observed POS covariates. Best-fitting model in category highlighted boldfaced.

Model	Camping Tent Reviews	Luxury Hotel Reviews
Topic models		
LDA	–5.893 (12)	–5.662 (10)
Embedding LDA	–6.159 (12)	–5.822 (10)
AT-LDA	–5.484 (12)	–5.340 (10)
SC-LDA	–5.728 (12)	–5.515 (10)
TCPM	**–5.323** (12)	**–5.128** (10)
Topic models with ubiquitous terms		
Ubi-LDA	–5.485 (12)	–5.497 (10)
Ubi-SC-LDA	–5.476 (12)	–5.161 (10)
Ubi-TCPM	**–5.271** (12)	**–5.008** (10)
Topic models with syntactical prior to topic dependency		
AT-LDA-POS	–5.483 (12)	–5.329 (10)
Ubi-TCPM-POS	**–5.260** (12)	**–4.973** (10)

Table (3) reveals a number of interesting findings. Across both data sets, we find our TCPM to be the best performing model. Whereas its advantage over c-Top2Vec is moderate (R^2 of 0.391 over 0.376) for the camping tent data, it is striking for the luxury hotel data(0.285 over 0.090). The reason for this large difference is that c-Top2Vec identifies only two topics from the hotel reviews. The two topics are closely aligned with syntactical role: one topic contains nouns, the other verbs and adverbs. Such a topic solution can have very little predictive power. From the camping tent reviews c-Top2Vec identifies 39 topics, a highly granular solution that none of our LDA-based models supports. Closer inspection of the topics reveals that top terms of most topics exhibit significant overlap (see Appendix A.9 for top terms from camping tent reviews). Results suggest that c-Top2Vec struggles to disentangle ubiquitous terms such as "tent/s" or "camping" from other terms.[5]

We find that the Embedding-LDA does not outperform a plain-vanilla LDA for both data sets (or any other LDA-type model for that matter). This suggests that there is little advantage in using pre-trained word embeddings instead of one-hot vector representations for topic inference. A reasonable explanation for this finding is that word embeddings obtained from large generic corpora do not necessarily complement actual term co-occurrence in small and (relatively) homogeneous, targeted consumer review data sets. Also, the backbone of state-of-the-art LLMs such as BERT are high-dimensional contextual embeddings which are supposed to elucidate the meaning of words. This turns out to be a very powerful approach when the goal is to make predict the next word for applications such as question–answering, translating or providing code for a particular purpose. Our goal here is different. Topic models aim at providing high-level summaries of a given corpus assuming little or no prior knowledge regarding the meaning of words. LLMs, in comparison, aim at zero-shot learning given extensive prior contextual knowledge. It should therefore come at no surprise that these models do not necessarily outperform a strictly "local approach".[6]

4.3. Model results

4.3.1. Topic run lengths

The TCPM provides estimates of topic run lengths γ_t and associated change points C_d. Note that estimates of γ are not influenced by the number of non-topic words as these are ignored for the purpose of identifying topic runs (Fig. 1). In Table (4), we show posterior means of γ from the Ubi-TCPM and the Ubi-TCPM-POS, using the camping tent data. The luxury

[5] c-Top2Vec can be estimated on the basis of two different pretrained sentence transformer LLM. We have chosen the LLM that produces a higher R^2, but actually has fewer parameters (all-MiniLM-L6-v2 with 22.7 m parameters). When fitted using all-mpnet-base-v2 with 109 m parameters, the predictive power of the topics from c-Top2Vec declines from 0.376 to 0.251 for the camping tent reviews. It appears that rating prediction does not necessarily benefit from using larger models.

[6] To investigate the performance of LLMs, we also prompted GPT to obtain topics and topic shares for our camping tent data. More specifically, we prompted GPT-4.1 to generate topics in the way of a topic description (less than 10 words) and the top 20 topic terms. We did not define a particular number of topics a priori and GPT suggested 9 topics. Recall our topic model suggests 12 topics. We then fed the 9-topic solution back to GPT to obtain document-level topic shares. Regressing the rating on these topic shares actually yields an improvement of R^2 of 13% over our TCPM for this data set (0.449 compared to 0.391). However, we find that the topics generated by GPT are exclusively topics that talk about product attributes implying that GPT ignores all other content in that data set (weather, camping trip, family, etc.). Also, results from the regression suggest that all topics are about equally important providing virtually no order over attributes with respect to consumer experiences, a highly counter intuitive result. The top 20 topic terms contain terms that are actually rare in the corpus but exhibit high semantic similarity or are synonyms (e.g. "curtain"-"divider" or "replacement"-"exchange"). Note that the term "curtain" appears only 22 times in the corpus. Such results seem inconsistent with an estimation approach based on likelihood of the data at hand. To summarize, GPT provides a topic solution that ignores content in review which provides important context to consumer experiences and which suggests that attributes cannot be meaningfully ordered with respect to their impact on consumer evaluations. We note, however, that these results should be taken with caution. Results from prompting large language models such as GPT depend heavily on the prompting strategies applied and factors such as temperature setting. Note that we did not investigate which prompting approach yields the best results. We leave this to future research.

Table 3
Share of variance of consumer rating explained by document-level topic shares. Best-performing model highlighted boldface.

Model	Luxury Hotels	Camping Tents
LDA	0.234	miss
AT-LDA	0.177	0.258
SC-LDA	0.208	0.269
TCPM	0.210	0.362
Ubi-LDA	0.198	0.326
Ubi SC-LDA	0.205	0.271
Ubi TCPM	0.270	**0.391**
AT-LDA-POS	0.090	0.201
Ubi TCPM-POS	**0.285**	0.347
Embedding-LDA	0.188	0.107
c-Top2Vec	0.090	0.376

hotel data yield similar results. All estimates of γ displayed in Table (4) are credibly larger than 1 as indicated by 95% mass of the posterior distribution right of 1. Topic run lengths larger than 1 (word) indicate a departure from independent topic assignments assumed by LDA.

Table (4) reveals that the TCPM results in topic runs longer than those implied by a first-order topic carry-over model (AT-LDA) but shorter runs compared to a model with a sentence constraint. In fact, the SC-LDA results in runs that are by a factor of 3–4 longer than those from the TCPM. The AT-LDA results in runs about half as long compared to the TCPM. These results are independent of topics. This suggests that the TCPM is a more flexible modeling approach which allows to navigate local topic dependency between two extremes represented by SC-LDA and AT-LDA. Interestingly, we find the TCPM-POS to result in longer topic runs than the TCPM. On average, runs from the TCPM-POS extend 4.2 (topic) words compared to 3.9 words from the TCPM without syntactical information to change points.

To further illustrate results from our model, Fig. (3) displays examples from two of our camping tent reviews to show how the TCPM-POS deals with word sequences. In the examples, we added instances of punctuation for clarity. The examples show that our model allows for topics to break within sentences, but that topics runs may also "jump" over a full stop.

Note that a phrase like "*bad for those cold summer nights, but good because you can stand up in the thing*" shown in Fig. (3) is considered a "hard sentence" in machine learning and sentiment classification (Chakraborty et al., 2022) because two polar different sentiments are being expressed. Our model appears to be able to deal with this quite nicely as the coordinating conjunction "but" is strongly associated with a topic change point, facilitating a topic (and possible sentiment) change at this point in the phrase. We discuss the role of syntactical classes to topic change points in more detail in Section 4.6.

4.4. Topic phrases

Through identification of topic runs, our TCPM facilitates a phrase-based approach to topic inference. Topic runs are sequences of (topic) words which all carry the same topic. These runs are typically interspersed by non-topic words, giving rise to a syntactically complete topic phrase. Note that the order of words in text is observed, whereas change points are not. The TCPM identifies where topics change and topics can then be characterized by syntactically complete phrases between change points. Figs. (4) and (5) present high-likelihood topic phrases from the topics in our data sets. A topic phrase as depicted here is a sequence of words between change points, containing words from a single topic and interspersed with non-topic words. Given phrase length, we show top topic phrases as indicated by the joint likelihood of all topic words in a phrase. For comparison, we show the top topic terms (indicated by $\phi_t(w)$) in Appendix (A.8).

Inspection of topic runs yields several noteworthy results. First, since the TCPM conditions on the observed order of words, phrases are meaningful excerpts from a text as they preserve the context in which particular words are used. In comparison, topic interpretation based most frequent terms (Appendix A.8) suffers from analysts having to reconstruct the context, giving rise to subjectivity. Second, interspersed non-topic words are critical as they complete phrases syntactically, making them readable. This is achieved without compromising topic inference. Third, relatively rare words may appear in likely topic phrases, enlarging the vocabulary of terms indicative of a particular topic. Consider runs from topic 12 of the camping tent data (Fig. 5). The word "freaking" in this phrase is a very rare term, appearing only 10 times in total in a corpus of more than 700,000 words. Its probability under any topic is close to 0. However, since top topic phrases are identified by the joint probability of topic words therein, rare terms can be element of such phrases as the example demonstrates.

4.5. Non-topic words

A novel feature of our model is to allow for ubiquitous terms in the data. This feature facilitates using all data, including terms such as stop words, without compromising topic inference. In our model, topic inference is based on topic words only. Topic and non-topic words are a priori defined over the same, complete vocabulary. Table (5) displays the top 40 non-topic words from our Ubi-TCPM, the Ubi-SC-LDA and the Ubi-LDA model.

Table 4
Posterior means of γ from TCPM from camping tent reviews. Empirical topic run lengths from Ubi-SC-LDA and AT-LDA-POS shown for comparison.

Topic	AT-LDA-POS	Ubi-TCPM	Ubi-TCPM-POS	Ubi-SC-LDA
1	1.85	4.33	4.21	15.25
2	2.03	4.14	4.72	13.66
3	1.92	3.57	3.90	13.24
4	1.67	2.98	3.53	9.54
5	1.99	3.70	4.15	13.51
6	1.65	3.30	3.69	12.90
7	2.22	5.67	5.48	4.73
8	1.78	4.80	4.83	9.41
9	1.88	3.94	4.16	9.23
10	2.85	2.27	2.82	10.80
11	1.63	6.89	6.86	12.17
12	1.97	5.18	5.18	15.16
Mean	2.03	3.87	4.21	11.16

This tent is exactly what the reviews said. Its big and heavy. That is good and bad. Bad for those cold summer nights but good because you can stand up in the thing.

My son and I have used this tent on 2 trips so far and we really love it. Its heavy and bulky so definitely not for backpacking but if you're not carrying too far it's a great tent.

Fig. 3. Examples of allocating topics to phrases from using the TCPM-POS. Topic words are color-coded, ubiquitous terms are greyed.

Table (5) reveals that terms identified as ubiquitous terms are primarily functional words such as stop words or prepositions ("the", "and", "in"). This result is largely independent of the model used. We note that high-frequency non-topic words are typically not nouns. The nouns that appear among non-topic words ("hotel", "room") are the result of our data collection. An interesting result from Table (5) is that the TCPM and the SC-LDA result in a very similar set of frequent non-topic words. Overlap among the top 40 terms is 83% indicating that these models agree on which terms should not be used for topic inference. A reasonable explanation for this finding is that topic runs from the TCPM often break at the end of sentences (Section 4.6). The LDA results in a more unique set of non-topic words with terms such as "location" or "service" appearing among most frequent ubiquitous terms. This suggests that a more flexible modeling approach results in classifying terms as ubiquitous which are not predominantly functional.

4.6. Predicting topic change points from syntax

In Fig. (6), we present results from the hierarchical regression of topic change points on syntactical covariates, using the camping tent reviews. The luxury hotel data yield similar results. Displayed are box plots of the posterior distribution of β from (1). Since POS tags are all coded as dummies, POS coefficients are directly comparable in terms of relative magnitude. Given our definition of C_d, positive coefficients indicate presence of a topic change point whereas negative coefficients indicate continuance of a topic run. Several results from Fig. (6) are noteworthy. First, the end of a sentence is strongly associated with a (subsequent) topic change ($\beta = 1.43$) and also the strongest driver of topic changes. Coordinating conjunctions ("and", "but") are positively correlated with a topic change ($\beta = 0.48$). Büschken and Allenby (2016) report similar results from their analysis. They find that coordinating conjunctions reduce the probability of a topic carryover. Fig. (6) reveals that many syntactical elements of speech are associated with a continuance of topic runs. Determinants ("the") and personal pronouns ("I", "we") exhibit strong effect on topic run continuance. These syntactical parts of speech are typically part of noun or verb phrases ("*I have never experienced such bad service*") and should therefore not be associated with topic changes. Because verb phrases often end on verb forms or nouns, we find these POS to be associated with topic change points, i.e. the end of topic runs (e.g. VPB: $\beta = 1.43$, NN: $\beta = 0.67$).

In summary, linking topic runs to POS categories reveals that observable syntactical structure of text is strongly associated with topic runs. We find this association to be independent of actual topics. In other words, the syntactical function of words within phrases of speech is predictive of the topic relationships among words. Our model exploits this predictive power for improved inference regarding topic runs.

5. Consumer rating analysis using topic phrases

The proposed TCPM identifies topic phrases which can be used to summarize documents. Given estimates of z_d, l_d, τ_d, a document can e.g. be described by the number of topic phrases, the number of words in these phrases and the share of ubiquitous terms. This is different from a summary e.g. based on θ_d (Mcauliffe and Blei, 2007; Büschken and Allenby, 2016) which is a measure of the (normalized) number of word level i.i.d. topic draws irrespective of whether the role of words is topical or

Topic 1: *Room issues*
“the ac in my room did not work and i could hear street noise"
“we had mildew in the shower tile and bathroom door would not close"
“the room did have smell to it and the carpet looked to have a few stains"

Topic 2: *Everything great*
“front desk service was very helpful with directions and the maid service was very friendly and accommodating“
“excellent location very helpful staff room size comfortable for three people clean and well maintained"
"nicely appointed the beds where extremely comfortable the hotel staff were friendly and helpful"

Topic 3: *Value*
„this hotel was great value for the money“
“is expensive but no more so than any other quality chain hotel"
“is still expensive but the hotel is great"
“the hotel is in a wonderful location although the room was a little pricey"
“for the rate the hotel was expensive"
“very clean and good location for the money"

Topic 4: *Nearby attraction*
“in every way the location is right in the middle of times square and the views"
“city views from the rooms either north or south"
“hotel is located right in the heart of the hustle and bustle of times square"

Topic 5: *Great experience*
“the hotel surpassed all our expectations the location was superb
“it was fantastic a great hotel for families“
“it was exactly what we were looking for well"
“it was an excellent experience and satisfied all our expectations"

Topic 6: *Amenities*
“plus extras of coffee in room breakfast included flat screen tv microwave and mini refrigerator"
“the breakfast offer quite a few choices and there was a lot of seating"
“the complimentary breakfast was fabulous eggs sausage toast waffles cereal fresh fruit"
“the hotel provided a white noise machine and ear plugs for those"

Topic 7: *Walking distance*
“just a few steps to broadway shows eating and shopping most everything is within walking distance including central park the"
“close to the subway lines macys madison square garden and the empire state building times square and rockefeller center were an easy walk“
“convenient location close to subways penn station and within walking distance to 5th avenue shopping rockefeller center times square"

Topic 8: Return & *recommend*
“i would stay there again without any hesitation and i already have told all my friends“
“would definitely recommend it to others will definitely use it in the future“
“definitely put this on our keep list and would go back again“

Topic 9: *Check-in*
“at both check in and check out we“
“we arrived at the hotel and check in“
“and when i arrived there i was told“
“we got to the hotel early but they were"

Topic 10: *You*
“if you want to be"
“if you are going to"
“if you want to go"
“all you have to do"

Fig. 4. Topic phrases with interspersed non-topic words from luxury hotel data. Topic phrases identified by change points and selected on basis of typical run length and (total) word likelihood, given topic. Topic labels by authors.

functional. In this section, we explore the relationship between consumer ratings and document summaries based on topic phrases and non-topic words.

We derive phrase-based statistics from our model to summarize documents and post hoc map these summaries to consumers' overall evaluations. Managerially, this analysis explores the origin of ratings by way of phrased-based document summaries. Such summaries can be directly linked to characteristic topic phrases (Figs. 4, 5). Note that this analysis departs in various ways from the standard supervised LDA approach which uses (normalized) word-level topic counts. First, our model allows to condition on (unobserved) topic runs, disentangling the effect of a particular topic appearing in a document and the number of words used in a run. Second, we can investigate the role of ubiquitous terms with respect to the rating. Third, the prior to (run-level) topic assignments is constrained so that subsequent runs cannot originate from the same topic.

Table (6) displays results from the hotel data using covariates obtained by way of the Ubi-TCPM-POS model. Our set of covariates also includes marginal topic shares θ_d. θ_d is the interaction effect of topic runs and average topic run lengths as

Topic 1: *Screens and ventilation*
- "the ceiling in the main body of the tent is screened"
- "the bottom half on the zipper has a flap over"
- "zipper window closures to keep the heat in"
- "for ventilation or to run electrical cords through"
- "without the rain coming through the windows the front awning"

Topic 2: *First set-up*
- "set it up in about 60 seconds with very little effort we"
- "took a little longer than expected to set up the first time"
- "the tent up securely in 15 30 minutes all by myself"

Topic 3: St*ow away*
- "trying to fit in back in the bag that it came in"
- „able to fit it back into the bag it came"
- "get it folded back up and into the carrying bag"

Topic 4: *You*
- "when you take it camping with you if you have"
- "so you have to make sure you have"
- "it would be nice if you did not"
- "so if you get it that will be"

Topic 5: *Coleman tent*
- "ask for a better tent for the money"
- "buy another tent that is not a coleman
- "the issues with coleman customer service and was told"
- "for the poor quality of this coleman tent in"

Topic 6: *Great tent*
- "very happy with this tent it is very compact with a nice"
- "this is an awesome tent roomy and a great buy"
- "this is a great family tent and a great price"

Topic 7: *Rain*
- "night in the rain and it did not let any water in"
- "through a rain storm and it was dry as a bone it"
- "in the rain and it stayed dry during a heavy rain storm"
- "in wind rain and is well ventilated during the day lot of"

Topic 8: *Camping trip*
- "my wife and i used this camping with our kids 7 and"
- "used this for the first time for a camping trip in vermont"
- "took this tent out for the first time the weekend"

Topic 9: *Easy set-up*
- "was easy to set up and easy to take down"
- "and was extremely easy to put up and take down"
- "easy to set up and pretty easy to tear down"

Topic 10: *Recommend*
- "and we couldnt believe it it was just amazing and we recommend"
- "and i really love it i am glad i did"
- "i would recommend buying it"
- "i have ever owned and i highly recommend"

Topic 11: *Room*
- "room for two queen size air beds and plenty of room between"
- "enough room inside for two queen air beds and plenty of gear"
- "lots of room even with two queen size air mattresses and gear"

Topic 12: *Poles*
- "but one of the tent poles for the rainfly was broken out"
- "one of the middle poles was broken at the joint"
- "freaking rainfly pole broke on the second weekend"
- "one of the poles cracked in the night and tore"
- "one of the poles broke after the second use"

Fig. 5. Topic phrases with interspersed non-topic words from camping tent data. Topic phrases identified by change points and selected on basis of typical run length and (total) word likelihood, given topic. Topic labels by authors.

their multiplication (and normalization) gives the topic shares. In Table (6), β_{r_t} indicates the expected change in rating if a topic run is added to a document. Note that adding a run is, by definition of the TCPM, associated with an additional topic change point. β_{l_t} indicates the expected change in rating if a (topic) word is added to a topic run in a document. β_{ubi} indicates the change in rating of the share of ubiquitous terms increases by 1 % (point).

From Table (6) several results are noteworthy. The associated change in rating given an additional topic run and the change induced by extending the run typically exhibit the same sign across topics. For example, if a topic run about "room issues" is added, the rating declines on average by 0.05 (on the 5-pt scale). If the length of that run increases by 1 (topic) word, the rating declines by 0.01. Similarly, an additional run from the Check-in topic decreases the rating by 0.08 points and an additional word reduces the rating by 0.004 points.

We find topics to vary greatly in terms of their effect of run lengths on the rating. Topic 10 ("You") exhibits the smallest effect ($\beta_{l_{10}} = -0.004$) whereas Topic 2 (" Great experience") exhibits the largest ($\beta_{l_5} = 0.04$) implying that an additional word from Topic 5 carries a weight 10 times larger than a word from Topic 10.

From a managerial perspective (e.g. identifying service failures), luxury hotel guests are particularly sensitive to issues with check-in which exhibits the largest (negative) influence on the rating. This is followed by the room issues and value. The influence of a "Value" topic run is only at 50% credibility level different from 0, reflecting the rather balanced view of reviewers that 5-star NY hotels are expensive but do deliver adequate value for money (see topic phrases in Fig. 4). However, at these runs get longer, their impact on the rating is credibly negative ($\beta_{l_3} = -0.015$).

An interesting observation from Table (6) is the influence of ubiquitous terms. We find that, conditional on topic composition, the rating declines as the share of ubiquitous terms increases ($\beta = -0.40$). Thus, a 20% share increase of functional terms in a review is associated with a drop of the rating by (almost) one point. In other words, the use of functional terms is more prevalent in reviews with a bad rating, indicating that reviewers discussing a bad experience feel the need to be more verbose.

Table 5
Non-topic words in luxury hotel review data. Displayed are the top 40 non-topic words from the TCPM, the SC-LDA and the LDA model. Rank given by posterior mean of $\psi^{(word)}$, given model.

Rank	Ubi-TCPM	Ubi-SC-LDA	Ubi-LDA
1	the	the	the
2	and	and	and
3	a	a	to
4	was	was	a
5	in	in	hotel
6	hotel	to	in
7	of	hotel	was
8	to	for	it
9	it	of	for
10	is	it	room
11	for	we	with
12	with	very	there
13	at	is	location
14	all	with	very
15	room	at	but
16	there	not	of
17	but	but	not
18	on	were	from
19	this	on	great
20	as	this	i
21	were	all	at
22	not	as	all
23	we	great	on
25	that	i	had
26	also	room	an
27	our	so	would
28	from	from	also
29	i	be	rooms
30	very	had	so
31	stay	stay	service
32	are	staff	really
33	or	time	could
34	just	there	by
35	an	good	time
36	had	an	when
37	which	are	good
38	my	our	every
39	really	times	back
40	nice	you	about

6. Discussion

In this paper, we propose a new topic model for textual data that accounts for local topic dependency of words by allowing for topic runs. Topic runs are sequences of words originating from the same topic, possibly interspersed by instances of ubiquitous terms. Our model constrains topics to change from run to run, giving rise to a topic change point model. Change points mark the end of one topic run and the beginning of the next. We link unobserved change points to observed syntactical parts-of-speech data and show that syntactical word classes are often strongly associated with presence or absence of topic change points. For example, we find determinants, pronouns or adjectives to be associated with continuance of topic runs whereas nouns, verbs and incidents of punctuation are associated with endings of topic runs. From our experience, the use of observable POS tags as prior information to change points helps to more reliably identify meaningful topic phrases. We integrate ubiquitous terms into the analysis which are assumed to appear at a constant frequency within a corpus into our model. This renders pruning e.g. stop words from text data prior to model-based analysis unnecessary.

Through identification of change points between topic runs, our model facilitates topic inference using characteristic topic phrases. Because topic runs condition on the observed order in which words appear and, hence, preserve the local context of words, phrases are natural representations of how consumers express certain themes in their reviews. In doing so, our model complements topic inference which is typically based on most frequent terms. Often, frequent terms are nouns or verbs and interpretation becomes difficult because the context in which these words are used is missing. Using consumer reviews of luxury hotels and camping tents, we show that our model results in run-based phrases that describe topics in a coherent and easy to interpret way. We also find that characteristic topic phrases contain rare terms which extends topic inference across the whole vocabulary. An analysis based on most frequent terms, by definition, ignores rare terms.

We show that the proposed TCPM outperforms other topic models that allow for local topic dependency with respect of fit to observed words. Our TCPM also outperforms a state-of-the-art topic model, based on LLMs, that uses contextual doc-

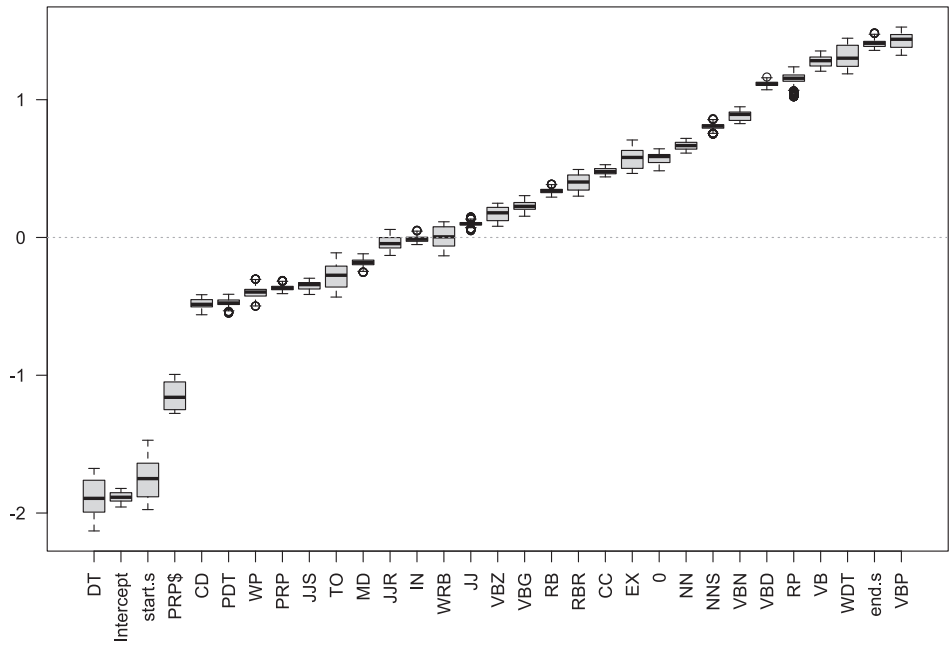


Fig. 6. Posterior distribution of coefficients from regression of change points on parts-of-speech tags. Results are from camping tent reviews. Labels on x-axis refer to POS classification (Appendix A.4. Positive coefficients indicate that syntax generates a topic change point, negative coefficients indicate that topi.c runs continue.

ument embeddings for topic inference. Given the relative simplicity of LDA-type models, this is a surprising result. A reasonable explanation for this finding is that LLMs are trained for making prediction, e.g. of sentence or document classes, not summarizing documents. Also, training occurs in very arge, highly diverse data sets. The goal of LDA is obtaining high-level summaries of documents in relatively small homogeneous data sets. It is not prediction of the next word or sentence. Apparently, LDA can achieve this goal very effectively without introducing embeddings. This and the elegance at which LDA can handle small data sets and rare terms speaks to the usefulness of the model.

Our model facilitates an analysis of the relationship between consumer ratings and topics by disentangling the role of the number of topic runs in a review and the number of words in each run. It also allows to conditions on the number of ubiquitous terms in a review and thereby controlling for words in reviews whose role is primarily functional. These features of our model allow for a more granular view of the origin of consumer ratings. In the standard approach, the rating is modeled as a function of word-level, i.i.d. topic counts (Mcauliffe and Blei, 2007; Büschken and Allenby, 2016), ignoring the actual sequence of words and their (syntactical/topical) function. We find that our approach significantly improves the power of topics to predict consumers' ratings over the standard approach.

The change point property of the proposed TCPM results from a simple constraint on run-level topic assignments where (only) the previous topic is excluded when the next topic run is generated. An alternative way of modeling topic change is to model transition probabilities, e.g. in a Markov process. This would allow to "learn" how reviewers of a particular product of service typically construct a sequence of topic discussions within their reviews, given that such similarities across reviews exist. Another shortcoming of the proposed TCPM is the assumption of communication intent that resides on the topic level. However, how topics evolve or how the "arc" of topics within a document is structured is outside the domain of our model. A more sophisticated approach would be to assume communication intent to originate from the author in the sense of (all) topics being determined simultaneously and dependently before any single topic run in a document is actually created. We leave these and other issues for future research.

Funding and competing interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

Table 6
Results from Bayesian regressing the observed rating on topic run-based covariates, share of ubiquitous terms and topic shares, using the Ubi-TCPM-POS model. Results from luxury hotel reviews (see topics from Fig. 4). $\theta_1 0$ not used as covariate because of co-linearity. Reported are posterior means and quantiles (2.5%, 97.5%) and Bayesian credibility intervals CI (mass under posterior left/right of 0, given sign of β).

Covariate	Posterior mean	2.5%	97.5 %	CI
Intercept	4.264	4.051	4.479	1.000
β_{ubi}	−0.396	-0.627	-0.165	1.000
β_{r_1}: *Room issues*	−0.051	-0.078	-0.024	1.000
β_{l_1}	−0.010	-0.023	0.004	0.915
β_{r_2}: *Everything great*	0.036	0.000	0.073	0.974
β_{l_2}	0.028	0.018	0.037	1.000
β_{r_3}: *Value*	0.000	-0.027	0.027	0.503
β_{l_3}	−0.015	-0.028	-0.001	0.983
β_{r_4}: *Nearby attractions*	0.025	-0.004	0.053	0.955
β_{l_4}	0.006	-0.007	0.020	0.828
β_{r_5}: *Great experience*	0.010	-0.020	0.040	0.740
β_{l_5}	0.042	0.023	0.060	1.000
β_{r_6}: *Amenities*	−0.014	-0.054	0.025	0.762
β_{l_6}	−0.003	-0.019	0.013	0.656
β_{r_7}: *Walking distance*	0.042	-0.006	0.092	0.957
β_{l_7}	0.009	0.000	0.019	0.971
β_{r_8}: *Return & recommend*	0.020	-0.018	0.059	0.847
β_{l_8}	0.016	0.004	0.028	0.995
β_{r_9}: *Check-in*	−0.076	-0.109	-0.044	1.000
β_{l_9}	−0.004	−0.015	0.007	0.756
$\beta_{r_{10}}$: *You*	−0.005	-0.037	0.026	0.630
$\beta_{l_{10}}$	−0.004	-0.018	0.011	0.685
β_{θ_1}	−0.653	-0.933	-0.375	1.000
β_{θ_2}	0.366	0.119	0.612	0.999
β_{θ_3}	−0.214	-0.473	0.043	0.949
β_{θ_4}	0.227	-0.044	0.501	0.950
β_{θ_5}	0.482	0.217	0.747	1.000
β_{θ_6}	−0.210	-0.526	0.102	0.902
β_{θ_7}	0.288	-0.011	0.585	0.971
β_{θ_8}	0.378	0.103	0.642	0.997
β_{θ_9}	−0.280	-0.593	0.021	0.965
σ^2	0.550	0.525	0.579	1.000
R^2	0.285			

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Appendices

A.1. Formal development of the topic change point model

A.1.1. Joint distribution

The data generation process of the TCPM implies the following joint distribution of knowns and unknowns of the topic change point model. We start by considering a single document.The joint distribution of knowns and unknowns of the model, given fixed priors and known quantities factorizes as follows (omitting subscript d to reduce clutter) given the hierarchical structure of the model suggested by the DAG (Fig. 2):

$$\begin{aligned} &p(w,z,\tau,\Phi,\psi,\theta,l,\gamma|\alpha,\beta,\delta,\epsilon,a,b,N) = \\ &p(w|\Phi,\Psi,z,\tau)\times p(C_d|\tau_d,z_d,l_d,X_d,\beta)\times p(l|\gamma,z)\times p(z|\theta^*)\times \\ &p(\tau|\delta)\times p\left(\Phi|\beta_\phi\right)\times p\left(\Psi|\beta_\psi\right)\times p(\delta|a,b)\times p\left(\gamma|\mu_\gamma,\sigma_\gamma\right)\times p\left(\beta|\mu_\beta,V_\beta\right) \end{aligned} \tag{3}$$

where θ^* indicates a constrained prior distribution over topics ($z_r \neq z_{r-1}$ for $r > 1$). C_d indicates a vector of binary word-level topic change point indicators. We explain the origin of this indicator and its conditioning on observed covariates X_d in Section 2.3. The likelihood of a word, conditional on τ, z is:

$$p(w_n|\Phi,\Psi,z_n,\tau_n)=\begin{cases}\Phi_{z_n}^{(w_n)} & \text{if} \quad \tau_n=1\\ \Psi^{(w_n)} & \text{if} \quad \tau_n=0\end{cases} \tag{4}$$

Conditional on τ, z the joint probability of all words in a document factorizes as follows:

$$p(w|\Phi,\Psi,z,\tau)=\prod_{n:\tau_n=0}p(w_n|\Psi)\prod_{n:\tau=1}p\left(w_n|\phi_{z_n}^{(w_n)}\right) \tag{5}$$

The second factor on the RHS of (5) expresses the probability of all "topic words" conditional on topic assignments. Reordering by topic runs to account for local dependency of topic assignments yields:

$$\prod_{n:\tau=1}p\left(w_n|\phi_{z_n}^{(w_n)}\right)=\prod_{r=1}^{R}p(w_{1_r},w_{2_r},\ldots,w_{l_r}|\phi_{z_r})=\prod_{r=1}^{R}\prod_{n=1}^{l_r}p\left(w_{n_r}|\phi_{z_r}^{(w_{n_r})}\right)$$

where $(w_{1_r},w_{2_r},\ldots,w_{l_r})=\{w_r\}$ is the set of words in topic run r and, within runs, z_r is a constant. Note that the $\{z_n\}_{n=1}^{N_d}$ are not independently determined but derived from $\left\{(z_r)_{\times l_r}\right\}_{r=1}^{R}$, i.e. each z_r is repeated (identically) l_r times.

The joint distribution of topic runs in a document is then:

$$p(w,l,z|\Phi,\theta,\tau=1)=\left[\prod_{r=1}^{R}\prod_{n=1}^{l_r}p\left(\{w_r\}|\phi_{z_r}^{(w_r)}\right)\times p(l_r|\gamma,z_r)\times p(z_r|\theta_d,z_{r-1})\right]p(C|X,\beta)$$

where $\{w_r\}$ is the set of words in run r. The above follows from the assumption that data generation for a run starts by drawing a topic for a new run, then generating run length given topic, and finally, given topic and run length, generating all words in a run from the same topic. The end of each run marks a topic change point implying that the current topic is excluded from the set of possible topics for the next run. The topics of runs $r=2,\ldots,R$ are constrained so that $z_r\neq z_{r-1}$. We write this down as:

$$p(z_r|\theta_d,z_{r-1})$$

and note that this is, except for $r=1$, a truncated multinomial distribution.

Identification of the TCPM with ubiquitous terms hinges upon differences in ψ,ϕ. To see this, consider the marginal likelihood of a (single) word, integrated with respect to z,τ while ignoring the likelihood contribution from topic runs (and dropping indices for clarity):

$$\begin{aligned}p(w|\psi,\phi,\theta,\delta)&=\delta\sum_{t=1}^{T}p(w|\phi,z=t)p(z=t|\theta)+(1-\delta)p(w|\psi)\\&=\delta\sum_{t=1}^{T}\phi_t^{(w)}\theta_t+(1-\delta)\psi^{(w)}\end{aligned} \tag{6}$$

It is clear that, if $\phi_t=\psi$ for any t, the two parts of the mixture likelihood in (6) are identical and the weighted sum is a constant for any δ. That is, $\delta,\psi,\phi$are not identified if ubiquitous terms overlap with topics which is an empirical issue. We return to this issue in Section 3.

A.1.2. Constrained topic assignments inducing change points

In our model, the end of each topic run is by definition a topic change point. We incorporate change points into our model by excluding the current topic z_r from the set of topics for the next run. In the LDA, the topic assignment z_n of word w_n, given number of topics T, originates from a multinomial distribution (suppressing index d):

$$p(z_n=t|\theta)=\theta_t$$

For the TCPM, we assume that a topic is simultaneously assigned to all words in a particular sequence of words but are, in principal, generated the same way. However, excluding the previous topic assignment and, thus, imposing topic change points between strings, leads to a point truncation of the multinomial:

$$p(z_r=t|\theta,z_{r-1})=\frac{1}{P(z_r|\theta_t,z_{r-1})}\theta_t=\frac{\theta_t}{1-\theta_{z_{r-1}}} \tag{7}$$

Across runs in a document, (7) gives rise to T truncated multinomial distributions each of which condition on a unique preceding topic $\{1,\ldots,T\}$. For example, for all topic runs where topic $t=1$ is excluded as it was assigned to the previous run, we arrive at the joint truncated multinomial distribution conditional on $z_{r-1}=1$:

$$p(\{z_r\}|\theta,z_{r-1}=1)\propto\prod_{t\neq1}\left(\frac{\theta_t}{1-\theta_1}\right)^{m_t} \tag{8}$$

where m_t indicates the number of times topic t was assigned to a run following a run with $t = 1$. This can also be expressed as:

$$p(\{z_r\}|\theta, z_{r-1} = 1) = \left(\frac{\theta_1}{1-\theta_1}\right)^0 \prod_{t\neq 1}\left(\frac{\theta_t}{1-\theta_1}\right)^{m_t}$$

which implies that $z_r = z_{r-1}$ does not occur (zero count m_t). Across all runs in a document, and all possible topic assignments of preceding runs, we obtain the product of T truncated multinomial distributions:

$$p(\{z_r\}|\theta, \{z_{r-1}\}) = \prod_{t\neq 1}\left(\frac{\theta_t}{1-\theta_1}\right)^{m_{t|1}} \times \prod_{t\neq 2}\left(\frac{\theta_t}{1-\theta_2}\right)^{m_{t|2}} \times \ldots \times \prod_{t\neq T}\left(\frac{\theta_t}{1-\theta_T}\right)^{m_{t|T}} \quad (9)$$

in which $m_{t|1}$ indicated the number of times topic run assignment t was "observed" to follow a run with topic assignment $z_{r-1} = 1$ etc. As an example, consider the following sequence of topic assignments across $R = 8$ topic runs in a document ($T = 4$):

$$z_1 = 1 \rightarrow z_2 = 4 \rightarrow z_3 = 3 \rightarrow z_4 = 2 \rightarrow z_5 = 1 \rightarrow z_6 = 4 \rightarrow z_7 = 1 \rightarrow z_8 = 2$$

Eq. (9) then becomes:

$$p(\{z_r\}|\theta, \{z_{r-1}\}) =$$
$$\left(\tfrac{\theta_2}{1-\theta_1}\right)^1\left(\tfrac{\theta_4}{1-\theta_1}\right)^2\left(\tfrac{\theta_1}{1-\theta_2}\right)^1\left(\tfrac{\theta_2}{1-\theta_3}\right)^1\left(\tfrac{\theta_1}{1-\theta_4}\right)^1\left(\tfrac{\theta_3}{1-\theta_4}\right)^1 =$$
$$\left(\tfrac{1}{1-\theta_1}\right)^3\left(\tfrac{1}{1-\theta_2}\right)\left(\tfrac{1}{1-\theta_3}\right)\left(\tfrac{1}{1-\theta_4}\right)^2\theta_1^3\theta_2^2\theta_3^1\theta_4^2$$

It is interesting to note that, in the last line of the above expression, the exponents of factors $\theta_1, \theta_2, \ldots$are counts of topic occurrences in the document example, marginalized with respect to the topics of previous runs, and the exponents to factors $\frac{1}{1-\theta_1}, \frac{1}{1-\theta_2}, \ldots$are counts of topics marginalized with respect to subsequent runs. Compared to the LDA, the counts of exponents to factors $\frac{1}{1-\theta_1}, \frac{1}{1-\theta_2}, \ldots$present additional information with respect to topic shares θ_d. This additional information results from the constraint that topics change between runs.

A.2. MCMC

In the following, we outline our estimation procedure for the proposed TCPM and the SC-LDA and the LDA with ubiquitous terms. For sampling steps not outlined here, we use prior results from the literature (Büschken and Allenby, 2016; Johnson and Albert, 2006; Blei and Lafferty, 2006; Blei et al., 2003). Sampling of τ is discussed as part of the MCMC for the Ubi-SC-LDA. Note that estimation of τ, δ is independent of which models in our analysis are used.

A.2.1. Sampling z, τ

We employ an independence chain MH sampling procedure to simultaneously update topic assignments $\{z_n\}_{n=1}^{N_d}$ and ubiquitous term indicators $\{\tau_n\}_{n=1}^{N_d}$ word-by-word. Note that the complete set of topic assignments $\{z_n\}_{n=1}^{N_d}$ determines topic run lengths l and change points in a document. An important feature of our approach is to limit the number of moves of the MCMC chain relative to the typical set. On the word level, $T + 1$ moves are possible. On the topic run or document level, the number of possibles moves that would have to be considered are by orders of magnitude larger. It is in this situation that MH schemes are prone to fail fully exploring the posterior distribution.

To develop our sampling scheme, we start by considering the joint posterior distribution of z_d, τ_d. Omitting subscript d to reduce clutter:

$$p(z, \tau|\Phi, \Psi, \gamma, \theta, \delta) \propto p(w|\Phi, \Psi, z, \tau) \times p(C_d|X_d, \beta) \times p(l|\gamma, z) \times p(z|\theta^*) \times p(\tau|\delta) \quad (10)$$

where θ^* indicates a set of constrained prior probabilities. For the sampling scheme we generate a joint proposal on the word level $\left(z_n^{(cand)}, \tau_n^{(cand)}\right)$ which we then evaluate across all words and topic runs in a document. We generate a candidate by way of a single draw from a Multinomial distribution with probabilities:

$$g_n \sim Multinom(1/(T+1), \ldots, 1/(T+1),)$$

where g_n can take on values $1, 2, \ldots, T + 1$ and T is the number of topics. Then:

$$\tau_n^{(cand)} = \begin{cases} 1 & \text{if} \quad g_n > 1 \\ 0 & \text{if} \quad g_n = 1 \end{cases}$$

and $z_n^{(cand)} = g_n - 1$ if $\tau_n^{(cand)} = 1$. This is equivalent to drawing $\left(z_n^{(cand)}, \tau_n^{(cand)}\right)$ step wise using the same probabilities. To see this, consider that the above implies:

$$p(\tau_n^{(cand)}=1)=1-\frac{1}{T+1}=\frac{T}{T+1}$$

so that:

$$p(z_n^{(cand)},\tau_n^{(cand)}=1)=p(z_n^{(cand)}|\tau_n^{(cand)}=1)\times p(\tau_n^{(cand)}=1)=\frac{1}{T}\times\frac{T}{T+1}=\frac{1}{T+1}$$

The key idea of our sampler is that changing the (non) topic assignment of a word simultaneously changes topic runs, their lengths and (consequently) change points. Consider our example in Fig. (1) in which the 18 words of a document are (currently) allocated to 3 topic runs with topic run assignments $z_2 \neq z_1$ and $z_3 \neq z_2$. Because of their position at the end of topic runs, w_5 and w_{14} mark topic change points. Given a joint proposal to $w_6 : \tau_{w_6}^{(new)}=1, z_{w_6}^{(new)}=orange=z_2$ (a move from $\tau_{w_6}^{(old)}=0$), the following elements (10) in change:

- $p(w_6|\phi_{z_2})$ compared to $p(w_6|\Psi)$
- $p\left(l_2=6|\gamma_{z_2}\right)$ compared to $p\left(l_2=5|\gamma_{z_2}\right)$
- $p(\tau_{w_6}=1|\delta)$compared to $p(\tau_{w_6}=0|\delta)$
- $p(z_2|\theta^*)$compared to 1where θ^* indicates the probabilities for the set of topics constrained by the topic of the previous run. The proposal has no effect on priors $p(z_1|\theta_d), p(z_3|\theta_d, z_2)$ and likelihood contributions $p(l_1|\gamma, z_1), p(l_3|\gamma, z_3), p(w_{r_1,r_2}|\Phi, \Psi, \tau)$ because runs r_1, r_3 remain unchanged. Note that the proposal does not change the world-level topic change-points. The implication is that in the MH step for the proposal we need to evaluate (everything else cancels, including the contribution from change points, given POS covariates):

$$\alpha=\underbrace{\frac{p(w_6|\phi_{z_2})p\left(l_2=6|\gamma_{z_2}\right)}{p(w_6|\psi)p\left(l_1=5|\gamma_{z_2}\right)}}_{\text{likelihoodcomparison}}\times\underbrace{\frac{p(z_2|\theta^*)p(\tau=1|\delta)}{p(\tau=0|\delta)}}_{\text{priorcomparison}}\times\underbrace{\frac{p(\tau^{(old)}|\delta)}{p\left(z_2^{(new)}|\theta\right)p(\tau^{(new)}|\delta)}}_{\text{candidategeneratingdensity}}$$

Note that the candidate generating density is in our case a symmetric distribution which implies that it cancels out. Accepting this move implies that w_6 is a topic word and topic run r_2 has length: $l_2=6$, not 5. Note that the candidate for z, given $\tau^{(new)}=1$ originates from an unconstrained prior. The prior probability of z_2, however, is constrained because $r>1$. This implies that candidate generating density and priors do not cancel in the Metropolis ratio.

Another instructive case arises from proposing $\tau_{w_{10}}=1, z_{w_{10}}=blue=z_1$. This move implies proposing two additional topic change points at w_9, w_{10} (see Fig. 10. Consequently, in (10), the following elements change:

- $p(w_{10}|\phi_{z_1})$ compared to $p(w_{10}|\psi)$
- $p\left(l_2=3|\gamma_{z_2}\right)p\left(l_3=1|\gamma_{z_1}\right)p\left(l_4=2|\gamma_{z_2}\right)$compared to $p\left(l_2=5|\gamma_{z_2}\right)$
- $p(\tau_{w_{10}}=1|\delta)$compared to $p(\tau_{w_{10}}=0|\delta)$
- $p(z_1|\theta^*)$compared to 1
- $p(C_9=1|X_9,\beta)$ compared to $p(C_9=0|X_9,\beta)$
- $p(C_{10}=1|X_{10},\beta)$ compared to $p(C_{10}=0|X_{11},\beta)$

The additional likelihood contributions $p(l|\gamma, z)$ reflect the proposed changes to topic runs as result of the word-based proposal to w_n. Given the proposed move of $\tau_{w_{10}}, z_{w_{10}}$, in the MH step we have to evaluate:

$$\alpha=\underbrace{\frac{p(w_{10}|\phi_{z_1})p\left(l_2=3|\gamma_{z_2}\right)p\left(l_3=1|\gamma_{z_1}\right)p\left(l_4=2|\gamma_{z_2}\right)p(C_{9,10}=1|X,\beta)}{p(w_{10}|\psi)p\left(l_2=5|\gamma_{z_2}\right)p(C_9=0|X,\beta)}}_{\text{likelihoodcomparison}}\times$$
$$\underbrace{\frac{p(z_1|\theta^*,z_2)p(\tau=1|\delta)}{p(\tau=0|\delta)}}_{\text{priorcomparison}}\times$$
$$\underbrace{\frac{p(\tau^{(old)}|\delta)}{p\left(z_2^{(new)}|\theta\right)p(\tau^{(new)}|\delta)}}_{\text{candidategeneratingdensity}}$$

as everything else cancels out. It is clear that splitting up "old" topic runs and introducing additional change points is penalized by the implied additional likelihood components $p(l|\gamma, z)$ if topics are characterized by longer word sequences (high γ). A penalty would also result from $w_{9,10}$ classified into a POS category that is found to be negatively associated with change points.

In our sampling of z, τ, we cycle through the $\{w_n\}_{n=1}^{N_d}$ in random order, proposing changes to τ_n, z_{w_n} independently. As their is no constraint on proposal density $p\left(z_{w_n}^{(cand)}|\theta\right), z_{w_n}$ may move in all directions possible. The proposal may be identical with the old value, leading to automatically "accepting" the candidate. Note that this scheme only consider moves of z_n, τ_n. Hence, we also implicitly condition on the number of observed words N_d (or v_d). An independence chain MH scheme relies on generating proposals that fully cover the posterior distribution. Ideally, the proposal density has fatter tails than the posterior distribution to assure to not build mass only in regions where the posterior has mass. One approach is by generating proposals from the prior distribution(s). Doing this would require to draw $z_n^{(cand)}$ from θ_d. In our experience, this can lead to always accepting $z_n^{(cand)} = t$ draws when $\theta_{d,t} \to 0$ (and the ratio of proposal densities then goes to infinity) which is not a desirable property. The likelihood of change points (1) enters the update of z, τ as an additional (and conditionally independent) likelihood contribution when jointly evaluating candidates $z_n^{(cand)}, \tau_n^{(cand)}$.

A.2.2. Sampling θ_d

In the following, we demonstrate how to generate samples from θ_d, the topic shares of document d. For this, we basically follow the procedure outlined by Johnson and Willsky (2012). We start by re-writing (9):

$$
\begin{aligned}
&p(\{z_r\}|\theta, \{z_{r-1}\}) = \\
&\left(\tfrac{1}{1-\theta_1}\right)^{\Sigma m_{t|1}} \prod_{t\neq 1} \theta_t^{m_{t|1}} \times \left(\tfrac{1}{1-\theta_2}\right)^{\Sigma m_{t|2}} \prod_{t\neq 2} \theta_t^{m_{t|2}} \times \ldots \times \left(\tfrac{1}{1-\theta_T}\right)^{\Sigma m_{t|T}} \prod_{t\neq T} \theta_T^{m_{t|T}} = \\
&\prod_{t\neq 1} \theta_t^{m_{t|1}} \times \prod_{t\neq 2} \theta_t^{m_{t|2}} \times \ldots \times \prod_{t\neq T} \theta_T^{m_{t|T}} \times \left(\tfrac{1}{1-\theta_1}\right)^{m_1} \times \left(\tfrac{1}{1-\theta_2}\right)^{m_2} \times \ldots \times \left(\tfrac{1}{1-\theta_T}\right)^{m_T}
\end{aligned}
$$

with $m_1 = \sum m_{t|1}$ etc. Note that, by the property of the geometric series and using the fact that $|\theta| < 1$:

$$\sum_{k_{1j}} \theta_1^{k_{1j}} = \frac{1}{1-\theta_1} \quad \text{for} \quad k_{1j} = \{0, 1, 2, \ldots\}$$

so that:

$$\left(\frac{1}{1-\theta_1}\right)^{m_1} = \left(\sum_{k_{1j}} \theta_1^{k_{1j}}\right)^{m_1} = \prod_{j=1}^{m_1} \sum_{k_{1j}} \theta_1^{k_{1j}}$$

For example, let $m_1 = 3$:

Johnson and Willsky (2012) define a new joint distribution of counts of topic assignments m and auxiliary (unknown) variables $k = (\{k_{1j}\}, \ldots, \{k_{Tj}\})$ with $k_{tj} = \{0, 1, 2, \ldots\}$:

$$
\begin{aligned}
&p(m, k|\theta) = \\
&\prod_{t\neq 1} \theta_t^{m_{t|1}} \prod_{t\neq 2} \theta_t^{m_{t|2}} \ldots \prod_{t\neq T} \theta_T^{m_{t|T}} \times \prod_{j=1}^{m_1} \theta_1^{k_{1j}} \ldots \prod_{j=1}^{m_T} \theta_T^{k_{Tj}}
\end{aligned}
$$

Integrating over the auxiliary variables k yields:

$$
\begin{aligned}
&\sum_k p(m, k|\theta) = p(m|\theta) = \\
&\prod_{t\neq 1} \theta_t^{m_{t|1}} \prod_{t\neq 2} \theta_t^{m_{t|2}} \ldots \prod_{t\neq T} \theta_T^{m_{t|T}} \times \sum_j \left(\prod_{j=1}^{m_1} \theta_1^{k_{1j}}\right) \ldots \sum_j \left(\prod_{j=1}^{m_T} \theta_T^{k_{Tj}}\right)
\end{aligned}
$$

A.2.3. Sampling γ

The posterior distribution of the rates generating topic run lengths:

$$p(\gamma_t|else) \propto p(\{l_t\}|\gamma, z) \times p\left(\gamma_t|\mu_\gamma, \sigma_\gamma\right) \tag{11}$$

where $\{l_t\}$ indicates the run lengths of the set of topic runs in the corpus assigned to topic $z = t$. Furthermore:

$$p(\{l_t\}|\gamma, z) = \prod_{r:z=t} p(l_{t,r}|\gamma_t)$$

We note that the RHS of this expression, given N_d, is the product of double-truncated Poisson distributions. The lower truncation point is 0 as topic runs of length 0 are not admissible. Such runs would be meaningless as any document contains an infinite number of them. The upper boundary is given by N_d. No single topic run can exceed the (observed) number of words in a document. We sample from (11) via a RW-MH step IID for each γ_t using the log-normal prior.

A.2.4. Subjective prior distributions

Our MCMC estimation approach for the TCPM employs the following subjective fixed prior distributions:

1. $V_\beta = 10I, \mu_\beta = 0$ where I indicates the identity matrix of dimensionality $T+1$
2. $V_{\beta_{POS}} = 10I, \mu_{\beta_{POS}} = 0$ with I of dimensionality number of POS categories + 1
3. $\sigma^2 \sim InverseGamma(3,3)$
4. $\phi_t \sim Dirichlet(\beta_\phi)$ with $\beta_\phi = 0.5\ \forall t$
5. $\psi \sim Dirichlet(\beta_\psi)$ with $\beta_\psi = 0.5$
6. $\theta_d \sim Dirichlet(\alpha_\theta)$ with $\alpha_\theta = 1$
7. $\delta \sim Beta(5,5)$
8. $\gamma_t \sim Gamma(2, 0.2) \forall t$

A.3. SC-LDA with ubiquitous terms

A.3.1. Joint distribution

We consider a topic model for text data in which a single topic is assigned to all "topic words" in each sentence and where ubiquitous terms appear at a constant frequency within sentences and across documents. This model is based on the SC-LDA topic model (Büschken and Allenby, 2016). For the SC-LDA with ubiquitous terms the joint distribution of knowns and unknowns on the sentence level is:

$$\begin{aligned} &p\left(\{w\}_s, z_s, \theta_d, \phi, \psi, \alpha, \beta^{(\phi)}, \beta^{(\psi)}\right) = \\ &p\left(\{w\}_s^{(topic)} | \phi_{z_s}, \tau_s\right) \times p(z_s|\theta_d) \times p(\theta_d|\alpha) \times p\left(w_s^{(non-topic)} | \psi, \tau_s\right) \times p(\tau_s|\delta) \times \\ &p\left(\phi|\beta^{(\phi)}\right) \times p\left(\psi|\beta^{(\psi)}\right) \times p(\alpha) \times p\left(\beta^{(\phi)}\right) \times p\left(\beta^{(\psi)}\right) \times p(\delta) \end{aligned} \tag{12}$$

where:

$\{w\}_s$ set of (all) words in sentence s of a document.
$\{w\}_s^{(topic)}$ set of topic words in sentence s as indicated by $\tau_s = 1$.
$\{w\}_s^{(non-topic)}$ set of non-topic words in sentence s as indicated by $\tau_s = 0$.
z_s is the (single) topic assignment of sentence s
τ_s is the vector of indicators of topic words in sentence s
θ_d is a vector of prior topic probabilities for document d
ϕ_t is a vector of word probabilities, given topic t
ψ is a vector of word probabilities for words appearing at constant frequency.
$\alpha, \beta^{(\phi)}, \beta^{(\psi)}, \delta$ are fixed priors of θ_d, ϕ_t, ψ and τ, respectively.
Eq. (12) factors into:

$$\begin{aligned} &p\left(\{w\}_s, z_s, \theta_d, \phi, \psi, \alpha, \beta^{(\phi)}, \beta^{(\psi)}\right) = \\ &\prod_{n:\tau_{n,s}=1} p(w_{n,s}|\phi_{z_s}) \times \prod_{n:\tau_{n,s}=0} p(w_{n,s}|\psi) \times \prod_n p(\tau_{n,s}|\delta) \times p(z_s|\theta_d) \times p(\theta_d|\alpha) \\ &p\left(\phi|\beta^{(\phi)}\right) \times p\left(\psi|\beta^{(\psi)}\right) \times p(\alpha) \times p\left(\beta^{(\phi)}\right) \times p\left(\beta^{(\psi)}\right) \times p(\delta) = \\ &\left[\prod_{n:\tau_{n,s}=1} p(w_{n,s}|\phi_{z_s}) \times p(\tau_{n,s}|\delta)\right] \times p(z_s|\theta_d) \times p(\theta_d|\alpha) \times \\ &\left[\prod_{n:\tau_{n,s}=0} p(w_{n,s}|\psi) \times p(\tau_{n,s}|\delta)\right] \times \\ &p\left(\phi|\beta^{(\phi)}\right) \times p\left(\psi|\beta^{(\psi)}\right) \times p(\alpha) \times p\left(\beta^{(\phi)}\right) \times p\left(\beta^{(\psi)}\right) \times p(\delta) \end{aligned} \tag{13}$$

A.3.2. Draw of τ

The above implies (dropping index s in the following to reduce clutter):

$$p(\tau_n|else) = \frac{p(w_n|\phi, \psi, z_s) \times p(\tau_n|\delta)}{\sum_\tau p(w_n|\phi, \psi, z_s) \times p(\tau_n|\delta)}$$

as everything else in (13) is a constant. Since τ_n can only take 2 values:

$$p(\tau_n = 1|else) = \frac{\phi_{z_s}^{(w_n)}\delta}{\phi_{z_s}^{(w_n)}\delta + \psi^{(w_n)}(1-\delta)}$$
$$p(\tau_n = 0|else) = \frac{\psi^{(w_n)}(1-\delta)}{\phi_{z_s}^{(w_n)}\delta + \psi^{(w_n)}(1-\delta)}$$

We assume that z_s always exists, independent of whether a sentence actually contains topic words. Thuas, we always draw a value of z_s. This facilitates the draw of *tau* conditional on z independent of whether a sentence actually contains "topic" words.

A.3.3. Draw of z_s

From the above:

$$p(z_s|else) \propto \prod_{n:\tau_{n,s}=1} p(w_{n,s}|\phi_{z_s}) \times p(z_s|\theta_d)$$

as everything is a constant. This implies:

$$p(z_s = t|else) \propto \prod_{n:\tau_{n,s}=1} \phi_t^{(w_n)} \times \theta_{d,t}$$

which is equivalent to the full conditional posterior of z from the SC-LDA which assumes all words in a sentence to be "topic" words. The above conditions on set $n : \tau_{n,s} = 1$ not being empty or the number of topic words in a sentence being larger than 0. If the set is empty, implying that there is no likelihood information from topic words, the conditional posterior of z_s is simply the prior:

$$p(z_s|else) \propto p(z_s|\theta_d)$$

we use this when all words in a sentence are assigned to the non-topic category to always facilitate a full conditional draw of τ_n.

A.3.4. LDA with ubiquitous terms

The LDA with ubiquitous terms is achieved by considering IID topic assignments on the word level. Conditional on $\tau_w = 1$, the only difference to the SC-LDA is in the draw of z_w:

$$p(z_w|\tau_w = 1, else) \propto p(w_{n,s}|\phi_{z_w}) \times p(z_w|\theta_d)$$

as everything else is a constant. This implies:

$$p(z_w = t|else) \propto \phi_t^{(w_n)} \times \theta_{d,t}$$

which is equivalent to the full conditional posterior of z from the standard LDA which assumes all words to be from the "topic" category.

A.4. POS Tags

For the hierarchical regression of change points on parts-of-speech, we use the following syntactical categories (when observed):1. CC Coordinating conjunction2. CD Cardinal number3. DT Determiner4. EX Existential there5. FW Foreign word6. IN Preposition or subordinating conjunction7. JJ Adjective8. JJR Adjective, comparative9. JJS Adjective, superlative10. LS List item marker11. MD Modal12. NN Noun, singular or mass13. NNS Noun, plural14. NNP Proper noun, singular15. NNPS Proper noun, plural16. PDT Predeterminer17. POS Possessive ending18. PRP Personal pronoun19. PRP$ Possessive pronoun20. RB Adverb21. RBR Adverb, comparative22. RBS Adverb, superlative23. RP Particle24. SYM Symbol25. TO to26. UH Interjection27. VB Verb, base form28. VBD Verb, past tense29. VBG Verb, gerund or present participle30. VBN Verb, past participle31. VBP Verb, non-3rd person singular present32. VBZ Verb, 3rd person singular present33. WDT Wh-determiner34. WP Wh-pronoun35. WP$ Possessive wh-pronoun36. WRB Wh-adverb

A.5. Simulation study: empirical identification of the Ubi-TCPM

We demonstrate statistical identification of our model by way of a simulation study. We first note that the TCPM is trivially identified if topic-specific sets of terms (i.e. terms with high probability, given a topic) and non-topic terms are unique. That is, if frequent terms under any topic as well as the ubiquitous terms exhibit not overlap, observed words identify the topic runs and, consequently, change points. For our simulation study, we set model parameters as follows:

$$
\begin{aligned}
V &= 500 \\
D &= 500 \\
T &= 4 \\
\delta &= 0.66 \\
\beta_{\phi,t,v} = 0.05 \quad \forall v \in (1,\ldots,V), \forall t \in (1,\ldots,T) \\
\beta_{\psi,v} &= 0.005 \quad \forall v \in (1,\ldots,V) \\
\alpha_{\theta_t} &= 0.66 \quad \forall t \in (1,\ldots,T) \\
\gamma_t &\sim U(0.2,1) \quad \forall t \\
N_d \sim U(120,280) \quad \forall d \in (1,\ldots,D)
\end{aligned}
$$

Given this set-up, about 5% of the terms per topic in V exhibit a probability of larger than $1/V$ implying that about 20 terms are "characteristic" for each topic and frequently co-occurring. We repeat generating Φ, Ψ until the 20 most frequent terms under topics and ubiquitous terms are (all) unique. In the set of documents created, about 5% exhibit "extreme" topic share distributions with one topic share approaching 1. The γ generated from N_d, θ_d, γ range from 1 to 200 across topics and documents with across-document averages ranging from 2.4 to 9.7. On average, a document contains 19 topic runs (implying 18 topic changes) with an average length of 14 words per topic run. In total, the simulated data set contains about 100,000 words over a total of about 10,000 topic runs.

Table (7) summarizes recovery of the topic run lengths rates and also presents hit rates of the topic assignments of runs and their lengths. Identifying both correctly implies that topic change points are correctly identified. Fig. () shows the recovery of Φ, Θ by plotting posteriors means from our MCMC against true values. Fig. () shows that, given our simulation set-up, the topic shares of documents cannot be recovered perfectly whereas the MCMC recovers the topic probabilities very well. We note that recovery of Θ can be improved greatly by increasing the number of words per document and/or shortening (average) topic runs. Both increase the number of topic runs per document and, hence, the number of "data points" informing θ_d.

Table (7) also shows that the MCMC recovers the true topic assignments and the topic runs very well. Hit rates for both variables exceed 80%. We compute "hits" as follows. Consider the following estimate of topic run lengths and its comparison to the true run lengths given $R^{(true)} = 6$ runs:

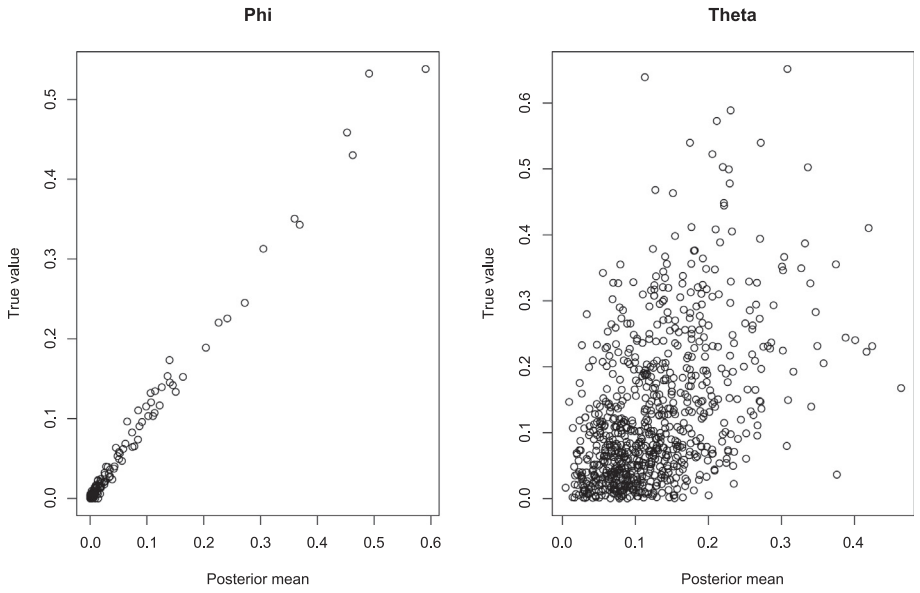
$$
\begin{aligned}
\{l_r\}^{(est)} &= 3,4,2,5,3,2,2 \\
\{l_r\}^{(true)} &= 3,4,2,5,3,4,0
\end{aligned}
$$

where the last entry in $\{l_r\}^{(true)}$ is to create vectors of equal length. In $\{l_r\}^{(est)}$, the last topic run is split into two runs each of length $l_r = 2$. A direct comparison is now possible and gives a hit rate of $5/7 = 71.4\%$ in this case. In an equal fashion, we compute hit rates for the topic assignments of the runs. In our simulations, we find that "misses" are often due to splitting up topic runs in the MCMC as in the above example. This can be minimized by creating smaller sets of topic-specific vocabularies and less overlap among the sets.

Table 7: Results from simulation study. Presented are recovery results for γ, z_r and l_r. For the topic assignments (run level) and the topic run lengths, posterior hit rates are shown.

Parameter	True Value	Posterior Mean	Posterior SD
Recovery of topic run lengths coefficients			
γ_1	7.23	7.22	0.25
γ_2	10.61	9.98	0.38
γ_3	10.44	10.33	0.31
γ_4	7.58	7.39	0.25
γ_5	7.72	7.54	0.25
γ_6	6.15	5.81	0.25
Recovery of topic assignments and topic run lengths (hit rates)			
z_r	1	0.954	0.030
l_r	1	0.894	0.027

Fig. 7: Recovery of ϕ_t and θ_d across all topics and documents. Shown are posterior means from the MCMC against true values. .



Legend:

A.6. Embedding-topic model

We follow the approach proposed by Dieng et al. (2020) but apply a Bayesian approach to estimating the embedding-topic model noting that, conditional on word- topic assignments z, the update of ϕ is independent of all other parameters of the LDA model. We use pre-trained word-embeddings and condition on Ω. This allows for a straightforward draw of α_t in our sampling scheme:

1. Propose candidate $\alpha_t^{(cand)}$ by way of a random walk
2. Compute $f(\Omega\alpha_t)$ where f is the softmax
3. Accept/reject $\alpha_t^{(cand)}$ based on the Metropolis–Hastings ratio, given "observed" word-topic assignments

We employ a weakly informative prior for $\alpha_t \sim N(0, 10I_L)$ where I_L is the $L \times L$ identity matrix. Our approach can also be applied to ubiquitous terms where:

$$\psi_t = f\left(\Omega\alpha^{(ubi)}\right)$$

with $\alpha^{(ubi)}$ as topic embedding of ubiquitous terms.

A.7. Likelihood-based model comparison

The likelihood integrated over topic assignments for the LDA and the Embedding LDA this is for word n in document d:

$$p(w_{n,d}|\phi,\theta_d) = \sum_t p(w_{n,d}, z_{n,d} = t|\phi,\theta_d)p(z_{n,d} = t|\theta_d) = \sum_t \phi_t(w_{n,d})\theta_{t,d}$$

Given LDA's bag-of-words assumption, the likelihood of a corpus is then:

$$\prod_{d=1}^{D}\prod_{n=1}^{N_d} p(w_{n,d}|\phi,\theta_d) = \prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_t \phi_t(w_{n,d})\theta_{t,d}$$

We consider various topic models that account for local topic dependency. For the SC-LDA, the probability of all words in sentence s of document d, integrated with respect to topic assignments:

$$p(w_{s,d}|\phi,\theta_d) = \sum_t \prod_{n=1}^{N_{s,d}} p(w_{n,d}, z_{s,d} = t|\phi,\theta_d)p(z_{s,d} = t|\theta_d)$$

which arises from all words in a sentence originating from a single topic draw resulting in local topic dependency. For the proposed topic change point model, conditional on "observed" topic runs, this simply becomes:

$$p(w_{r,d}|\phi,\theta_d^*) = \sum_t \prod_{n=1}^{N_{r,d}} p(w_{n,d}, z_{r,d}=t|\phi,\theta_d^*) p(z_{r,d}=t|\theta_d^*)$$

where r indicates a topic run, $N_{r,d}$ the number of words in that run and θ^* indicates the constrained prior distribution over topics. Similarly, the AT-LDA results in topic runs through (possibly multiple, consecutive) incidents of topic carry-over.

To compute fit for models with ubiquitous terms, we condition on τ which allows to split the likelihood (in case of the Ubi-LDA) into two components:

$$p(w_d|\phi,\theta_d,\tau_d) = \prod_{n=1}^{N_d^{(topic)}} \sum_t \phi_t(w_{n,d})\theta_{t,d} \prod_{n=1}^{N_d^{(non)}} \psi(w_{n,d})$$

We apply this to all models in our evaluation set noticing that, given τ, the (joint) probability of non-topic words and topic words factors out.

A.8. Top topic terms

Table 8: Most frequent topic terms from camping tent reviews and Ubi-TCPM. Labels at bottom by authors.

Term rank	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	the	up	the	to	the	tent	the	camping	up	i	of	the
2	of	it	to	you	tent	is	in	for	to	we	room	of
3	in	set	in	not	a	a	a	tent	easy	it	a	tent
4	tent	the	it	be	this	this	rain	my	set	have	and	poles
5	door	a	back	it	of	the	and	this	it	had	for	on
6	on	in	bag	if	Coleman	great	it	a	and	was	queen	and
7	a	time	tent	can	for	it	we	and	very	that	our	are
8	porch	about	up	do	tents	for	of	the	was	not	with	that
9	front	took	out	will	on	was	was	our	down	would	air	one
10	windows	first	into	have	that	very	dry	in	put	so	two	with
11	area	by	get	i	one	its	night	trip	take	but	plenty	a
12	room	with	fit	would	other	good	tent	we	as	did	fit	in
13	is	tent	and	get	is	but	water	first	is	used	2	stakes
14	rainfly	to	put	that	with	with	no	it	well	love	in	pole
15	for	two	all	does	my	as	wind	on	setup	recommend	mattress	to
16	screen	myself	a	but	in	that	with	of	held	were	size	broke
17	screened	than	on	so	amazon	nice	all	time	stand	bought	the	down
18	and	down	my	did	from	there	on	family	roomy	am	we	all
19	all	but	from	they	reviews	not	weather	i	spacious	his	family	rainfly
20	rain	this	your	could	an	big	had	year	able	got	mattresses	at
21	side	put	came	make	but	so	but	years	its	when	to	side
22	that	setting	able	need	all	well	rained	used	easily	just	had	for
23	floor	for	down	are	at	huge	stayed	last	great	like	space	use
24	back	less	its	sure	rainfly	price	were	times	pretty	also	gear	two
25	top	little	getting	use	new	little	not	old	quick	been	kids	top
26	out	takes	them	want	no	only	storm	weekend	fairly	they	all	some
27	zipper	people	away	see	than	has	at	with	how	because	my	first
28	open	we	keep	go	quality	size	during	3	nice	could	4	zipper
29	inside	me	our	how	as	really	inside	2	fast	only	enough	out
30	from	more	when	should	instant	quality	winds	at	super	are	inside	were
31	with	at	around	when	buy	awesome	some	days	together	really	comfortably	off
32	over	setup	inside	know	some	perfect	there	4	assemble	do	have	stake
33	main	of	right	much	their	large	through	husband	simple	still	us	other
34	window	my	of	going	me	happy	heavy	day	but	as	adults	plastic
35	at	i	without	keep	another	much	out	over	quickly	think	lots	seams
36	are	an	go	buy	after	roomy	any	after	can	can	bags	ground
37	sides	putting	case	say	product	an	us	summer	so	use	or	from
38	middle	15minutes	place	able	about	thing	day	went	goes	what	3	door
39	cover	20minutes	trying	as	price	like	got	use	really	purchased	sleeping	bottom
40	an	few	way	what	best	bit	leaks	wife	tent	highly	still	broken
Label	Protection	First set-up	Stow away	"You"	Coleman	Great tent	Rain	Camping trip	Easy set-up	Recommend	Room	Poles

Table 9: Most frequent topic terms from luxury hotel reviews and Ubi-TCPM. Labels at bottom by authors.

Term rank	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	i	was	a	the	we	breakfast	to	i	to	to
2	was	and	is	of	was	for	square	stay	we	you
3	not	very	hotel	in	a	the	and	new	room	be
4	that	staff	was	square	great	a	from	to	for	not
5	did	the	it	times	our	was	close	york	in	if
6	it	clean	great	on	had	free	walking	would	i	are
7	have	were	for	hotel	hotel	no	subway	again	our	get
8	only	hotel	location	at	room	room	times	this	a	can
9	they	room	but	is	were	in	distance	in	were	have
10	but	helpful	this	right	it	and	walk	hotel	us	go
11	the	friendly	very	floor	location	of	a	recommend	check	want
12	room	comfortable	room	time	stay	on	station	city	my	we
13	we	great	good	view	for	or	within	my	at	do
14	would	location	its	from	this	internet	central	stayed	and	i
15	had	nice	as	street	i	with	blocks	will	when	your
16	were	rooms	nice	middle	an	wifi	restaurants	at	they	for
17	no	service	little	all	wonderful	good	broadway	definitely	had	everything
18	could	good	with	hilton	with	service	all	we	out	place
19	so	bed	small	heart	perfect	included	of	here	the	so
20	been	excellent	price	location	view	which	away	there	on	that
21	a	beds	in	marriott	experience	hot	shopping	time	early	could
22	one	are	than	restaurant	loved	parking	easy	for	desk	stay
23	my	well	rooms	marquis	my	coffee	state	back	day	see
24	on	with	expensive	this	enjoyed	complimentary	attractions	have	up	will
25	there	extremely	well	city	excellent	is	penn	trip	me	when
26	problem	large	more	one	us	food	is	next	2	back
27	even	quiet	that	best	everything	fridge	macys	highly	night	it
28	this	so	big	located	very	at	location	be	with	did
29	is	desk	too	top	made	etc	building	go	arrived	up
30	be	spacious	money	out	time	day	for	staying	after	would
31	like	is	not	42nd	overall	there	empire	our	that	they
32	bathroom	front	perfect	lobby	stayed	extra	many	first	front	walk
33	other	breakfast	size	manhattan	trip	buffet	very	nyc	of	take
34	about	really	has	8th	amazing	water	the	visit	one	wanted
35	just	comfy	hotels	noise	fantastic	charge	block	ive	got	need
36	than	courteous	worth	across	as	morning	access	year	time	going
37	thing	bathroom	bit	nyc	also	that	grand	return	it	where
38	work	pleasant	so	outside	place	two	great	hotels	an	close
39	should	our	much	area	which	continental	are	friends	before	anywhere
40	however	wonderful	the	center	so	bar	short	marriott	rooms	any
Label	Room issues	Everything great	Location/ Value	Nearby attractions	Great experience	Amenities	Walking distance	Return & recommend	Check-in	"You"

Table A.9. Top topic terms from camping tent reviews using c-Top2Vec

Table 10: Top 10 terms of the topics in camping tent reviews obtained from c-Top2Vec.

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
Word 1	awning	going	decided	rooms	enjoyed	product	tents	tents	tents	tents
Word 2	love	complete	below	assemble	satisfied	products	tent	tent	tent	tent
Word 3	style	up	best	storage	rating	buying	stars	stake	campground	campground
Word 4	enjoyed	decided	am	assembly	complete	brand	star	campground	camp	camp
Word 5	view	top	corners	folding	below	buy	campground	stakes	campsite	ventilation
Word 6	cover	done	top	spacious	standing	quality	rainfly	staking	camper	campsite
Word 7	below	below	held	footprint	up	enjoyed	camper	windy	camping	breeze
Word 8	beautiful	along	box	built	pleased	price	fly	staked	canopy	camping
Word 9	season	doing	along	setup	liked	material	campsite	camp	wind	downpour
Word 10	place	picture	above	room	closed	items	lantern	wind	cot	camper

	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
Word 1	tents	tents	tents	tents	tents	tents	tents	coleman	tents	tents
Word 2	rained	tent	tent	tent	tent	tent	tent	tents	rained	poles
Word 3	tent	campground	campground	campground	campground	campground	campground	campground	rain	tent
Word 4	raining	cot	camp	zipper	wind	coleman	poles	tent	raining	campground

Table A9 (*continued*)

Word 5	rain	campsite	camper	zippers	rainfly	camper	awning	rainfly	rains	pole
Word 6	rains	cots	campsite	tarp	winds	camp	camp	camper	rainstorm	canopy
Word 7	moisture	canopy	cot	camper	storms	campers	camper	camp	downpour	camper
Word 8	rainstorm	camper	canopy	canopy	windy	cot	rainfly	mattress	rainy	camp
Word 9	rainy	coleman	campers	camp	rainstorm	poles	canopy	lantern	tent	campsite
Word 10	campground	us	cots	cot	storm	campsite	folding	poles	campground	awning
	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
Word 1	tents	tents	tents	tents	tents	rainfly	tents	tents	tents	tents
Word 2	rainfly	tent	tent	tent	tent	tents	tent	tent	tent	tent
Word 3	tent	campground	wind	campground	campground	rained	rained	campground	campground	downpour
Word 4	campground	camp	winds	coleman	camper	rainstorm	raining	camp	rooms	leakage
Word 5	rainstorm	camper	storms	poles	camp	tent	porch	rooms	camp	moisture
Word 6	coleman	campsite	windy	camper	awning	raining	rain	campsite	beds	campground
Word 7	fly	campers	campground	camp	campsite	rain	rainstorm	camping	mattress	rains
Word 8	canopy	rainfly	storm	campers	campers	rains	rainfly	cot	cot	leaks
Word 9	rained	camping	rainstorm	campsite	ventilation	moisture	campground	camper	campsite	rained
Word 10	rain	cot	breeze	awning	canopy	rainy	rains	cots	cots	raining
	Topic 31	Topic 32	Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38	Topic 39	
Word 1	tents	tents	tents	tents	tents	tents	tents	tents	tents	
Word 2	tent	tent	tent	tent	tent	tent	tent	tent	tent	
Word 3	campground	campground	mattress	campground	campground	campground	campground	rained	campground	
Word 4	folding	rooms	beds	camp	canopy	camp	camp	moisture	cot	
Word 5	cot	campsite	rooms	rooms	cot	camper	campsite	raining	canopy	
Word 6	fold	camp	campground	camper	awning	campsite	camper	rainstorm	cots	
Word 7	camp	cot	cot	cot	camp	cot	camping	rains	tarp	
Word 8	campsite	canopy	mattresses	campers	tarp	campers	rainfly	rain	awning	
Word 9	collapsed	cots	cots	campsite	cots	camping	awning	downpour	camp	
Word 10	cots	footprint	camp	cots	campsite	rainfly	campers	rainfly	campsite	

References

Angelov, Dimo (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Angelov, Dimo, & Inkpen, Diana (2024). Topic modeling: Contextual token embeddings are all you need. *Findings of the Association for Computational Linguistics: EMNLP, 2024*, 13528–13539.

Archak, Nikolay, Ghose, Anindya, & Ipeirotis, Panagiotis G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science, 57*(8), 1485–1509.

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

Blei, D. M., Ng, Yew-Kwang, & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

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

Büschken, Joachim, & Allenby, Greg M. (2020). Improving text analysis using sentence conjunctions and punctuation. *Marketing Science, 39*(4), 727–742.

Chakraborty, Ishita, Kim, Minkyung, & Sudhir, K. (2022). Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. *Journal of Marketing Research, 59*(3), 600–622.

Dellarocas, Chrysanthos, Zhang, Xiaoquan Michael, & Awad, Neveen F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing, 21*(4), 23–45.

Dieng, Adji B., Ruiz, Francisco J. R., & Blei, David M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics, 8*, 439–453.

Fader, Peter S., & Hardie, Bruce G. S. (2009). Probability models for customer-base analysis. *Journal of Interactive Marketing, 23*(1), 61–69.

Fader, Peter S., Hardie, Bruce G. S., & Huang, Chun-Yao (2004). A dynamic changepoint model for new product sales forecasting. *Marketing Science, 23*(1), 50–65.

Gopalakrishnan, Arun, Bradlow, Eric T., & Fader, Peter S. (2016). A cross-cohort changepoint model for customer-base analysis. In *Marketing Science -(Articles in advance)*. .

Grootendorst, Maarten. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

Ho, Teck-Hua, Park, Young-Hoon, & Zhou, Yong-Pin (2006). Incorporating satisfaction into customer value analysis: Optimal investment in lifetime value. *Marketing Science, 25*(3), 260–277.

Johnson, Matthew James, Alan S Willsky. 2012. Dirichlet posterior sampling with truncated multinomial likelihoods. arXiv preprint arXiv:1208.6537.

Johnson, Valen E., & Albert, James H. (2006). *Ordinal data modeling*. Springer Science and Business Media.

Kenton, Jacob Devlin Ming-Wei Chang, Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of naacL-HLT, vol. 1. Minneapolis, Minnesota, 2.

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

Mcauliffe, Jon, David Blei (2007). Supervised topic models. Advances in neural information processing systems 20.

Mikolov, Tomáš, Yih, Wen-tau, & Zweig, Geoffrey (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

Schmittlein, David C, Morrison, Donald G, & Colombo, Richard (1987). Counting your customers: Who-are they and what will they do next? *Management science, 33*(1), 1–24.

Schweidel, David A., & Fader, Peter S. (2009). Dynamic changepoints revisited: An evolving process model of new product sales. *International Journal of Research in Marketing, 26*(2), 119–124.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Yang, Yi, Zhang, Kunpeng, & Fan, Yangyang (2022). sdtm: A supervised bayesian deep topic model for text analytics. *Information Systems Research*.

Zhang, Dongcheng, Zhang, Kunpeng, Yang, Yi, & Schweidel, David A. (2024). Tm-okc: An unsupervised topic model for text in online knowledge communities. *MIS Quarterly, 48*(3).

Zhong, Ning, & Schweidel, David A. (2020). Capturing changes in social media content: a multiple latent changepoint topic model. In *Marketing Science*. .