

## ePub<sup>WU</sup> Institutional Repository

Martin Reisenbichler and Thomas Reutterer

Topic modeling in marketing: recent advances and research opportunities

Article (Published)  
(Refereed)

*Original Citation:*

Reisenbichler, Martin and Reutterer, Thomas (2019) Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, 89 (3). pp. 327-356. ISSN 1861-8928

This version is available at: <http://epub.wu.ac.at/6513/>

Available in ePub<sup>WU</sup>: September 2018

ePub<sup>WU</sup>, the institutional repository of the WU Vienna University of Economics and Business, is provided by the University Library and the IT-Services. The aim is to enable open access to the scholarly output of the WU.

This document is the publisher-created published version.



# Topic modeling in marketing: recent advances and research opportunities

Martin Reisenbichler<sup>1</sup> · Thomas Reutterer<sup>1</sup>

© The Author(s) 2018

## Abstract

Using a probabilistic approach for exploring latent patterns in high-dimensional co-occurrence data, topic models offer researchers a flexible and open framework for soft-clustering large data sets. In recent years, there has been a growing interest among marketing scholars and practitioners to adopt topic models in various marketing application domains. However, to this date, there is no comprehensive overview of this rapidly evolving field. By analyzing a set of 61 published papers along with conceptual contributions, we systematically review this highly heterogeneous area of research. In doing so, we characterize extant contributions employing topic models in marketing along the dimensions data structures and retrieval of input data, implementation and extensions of basic topic models, and model performance evaluation. Our findings confirm that there is considerable progress done in various marketing sub-areas. However, there is still scope for promising future research, in particular with respect to integrating multiple, dynamic data sources, including time-varying covariates and the combination of exploratory topic models with powerful predictive marketing models.

**Keywords** LDA · Machine learning · Marketing research · Topic modeling

**JEL Classification** M30 · C00

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11573-018-0915-7>) contains supplementary material, which is available to authorized users.

---

✉ Thomas Reutterer  
thomas.reutterer@wu.ac.at

Martin Reisenbichler  
martin.reisenbichler@wu.ac.at

<sup>1</sup> Institute for Service Marketing and Tourism, Vienna University of Economics and Business, Building D2, Entrance A, Welthandelsplatz 1, 1020 Vienna, Austria

## 1 Introduction

There is an ongoing trend among marketing scholars (e.g., Flach 2001, pp. 205; Shaw et al. 2001, pp. 127) and practitioners (e.g., Nimeroff 2017) to adopt machine learning techniques in a diverse field of application domains. This trend is intertwined with the digitalization of our economy and the increasing availability of “big” and unstructured data, such as large amounts of texts or other inherently sparse, high-dimensional data (Kahn et al. 2010, p. 4). The focus in this paper is on so-called topic models, a specific model class which recently emerged as a versatile tool to analyze such marketing data.

Stemming from the early ideas of traditional cluster analysis, which are particularly relevant for marketing research (Punj and Stewart 1983, p. 135; Reutterer 2003, pp. 52), and enriched by the idea of probabilistic modeling and mixed membership (Blei 2012, pp. 78; Galyardt 2015, pp. 40), popular topic models like LDA (Latent Dirichlet Allocation) are a flexible, unsupervised machine learning approach to soft-cluster big data (e.g., Blei et al. 2003, pp. 993; Blei and Lafferty 2009, p. 77). Applications in marketing research comprise but are not limited to consumer profiling (e.g., Blanchard et al. 2017, pp. 408; Trusov et al. 2016, pp. 413), to the assessment of buying patterns and purchase predictions (e.g., Hruschka 2016, p. 7; Jacobs et al. 2016, pp. 389), to discovering online communities and topics (e.g., Ngyen et al. 2015, pp. 9603), and more. In business settings, the insights derived by applying topic models can, for instance, help increasing the effectiveness and efficiency of online ads by fitting them thematically to web pages (e.g., Le et al. 2008, pp. 889), or assist in building recommender systems for online market platforms (e.g., Christidis and Mentzas 2013, pp. 4373). Despite its rising popularity, to the best of our knowledge, there is currently no systematic overview of the state of research of topic modeling in marketing, which is exactly what this article aims to contribute.

After a search in Google, Google Scholar, and various library databases using keywords like “topic model”, “topic modeling marketing”, “LDA”, and scanning for relevant literature in the initially found publications, we detected a total of 61 papers, ranging from 2008 until the end of 2017. Ten of these were published in core marketing journals (Amado et al. 2017; Blanchard et al. 2017; Büschken and Allenby 2016; Calheiros et al. 2017; Hruschka 2014; Jacobs et al. 2016; Schröder 2017; Song et al. 2017; Tirullinai and Tellis 2014; Trusov et al. 2016). Since the trend of using the method is currently spanning across numerous disciplines (e.g., Schmidt 2013), the remaining 51 articles were published in journals from other fields, but explicitly assess marketing relevant research topics. Due to the openness of the framework (Airoldi et al. 2015, p. 4), and the applicability to a large variety of datasets (Blei 2012, p. 83), the field seems to be highly disordered and diverse. This directly leads to our research questions, trying to find a classification system for the publications focused on methods: (RQ1) What are the applied methodological strategies? Specifically, we assess procedures of (1) data retrieval, (2) implementing and extending, and (3) evaluating topic models, utilized as a core clue to the orientation of publications and to the current state of

the field. Due to the still ongoing evolution of the methodological approach (Blei et al. 2003, pp. 1015; Galyardt 2015, pp. 42), a substantial amount of published research tend to be merely experimental than focusing on substantial results (e.g., Jacobs et al. 2016; Phuong and Phuong 2012; Wang et al. 2015). Additionally, because of the large variation of models (Airoldi et al. 2015, pp. 1–567) resulting from the ability to relax basic assumptions of the approach (Blei 2012, pp. 82), we expect a diversity in assessed objects and research interests. Therefore, we try to find fields of research by looking at the data, the examined objects, and the research interests: (RQ2) What are current fields of research? Our third and final research question aims at connecting the former two questions by deriving major gaps and providing possible future directions: (RQ3) What are major gaps and future directions of research?

The subsequent sections of the paper are organized as follows: In Sect. 2, we describe the building blocks of LDA (Blei and McAuliffe 2010, pp. 1; Blei 2012, pp. 78), which today is the basic approach in topic modeling and kind of “a springboard for many other topic models” (Blei and Lafferty 2009, p. 72). Also, we illustrate commonly used extensions of LDA (e.g., Blei and Lafferty 2007, 2009; Blei and McAuliffe 2010; Do and Gatica-Perez 2010; Hoffman et al. 2010), approaches of evaluating (e.g., Newman et al. 2010), and intertwined with that, essential critique directed at the method (e.g., Schmidt 2013). These aspects are crucial in understanding topic modeling applications in marketing research (Sects. 3 and 4). In Sect. 3, we develop a classification system, both, based on the theoretical literature on the subject, and on the examined papers, which focuses on methodological strategies. In Sect. 4, we adapt that scheme to explicitly summarize trends and patterns in current research. Additionally, we try to find sub-patterns by introducing fields of current research, derived from the examined data, objects, and research interests. Lastly, in Sect. 5, we conclude our work by summarizing gaps and future possibilities in that promising area.

## 2 Topic modeling

### 2.1 Latent Dirichlet Allocation

LDA, as introduced by Blei and Jordan (2003a) and Blei et al. (2003) is a method originally developed for soft-clustering large quantities of discrete textual data, in order to find latent structures (Blei 2012, pp. 77; p. 80). The model assumes that textual documents consist of topics, which in turn consist of words from a vocabulary (all words in the documents) (Airoldi et al. 2015, p. 4; Blei 2012, pp. 78; Galyardt 2015, p. 39). The aim is to automatically discover the hidden (not observable) topics in a collection of documents, and how each document exhibits them (Blei and Lafferty 2009, p. 73; Blei 2012, p. 79). The topics are “a recurring pattern of co-occurring words” (Brett 2012), or, more formally speaking, “a distribution over a fixed vocabulary” (Blei 2012, p. 78). To unveil the topic structure in addition to that, the hidden per-document topic distribution needs to be rendered (Blei 2012, pp. 79). For example, “a news article may address multiple topics rather than fitting neatly a

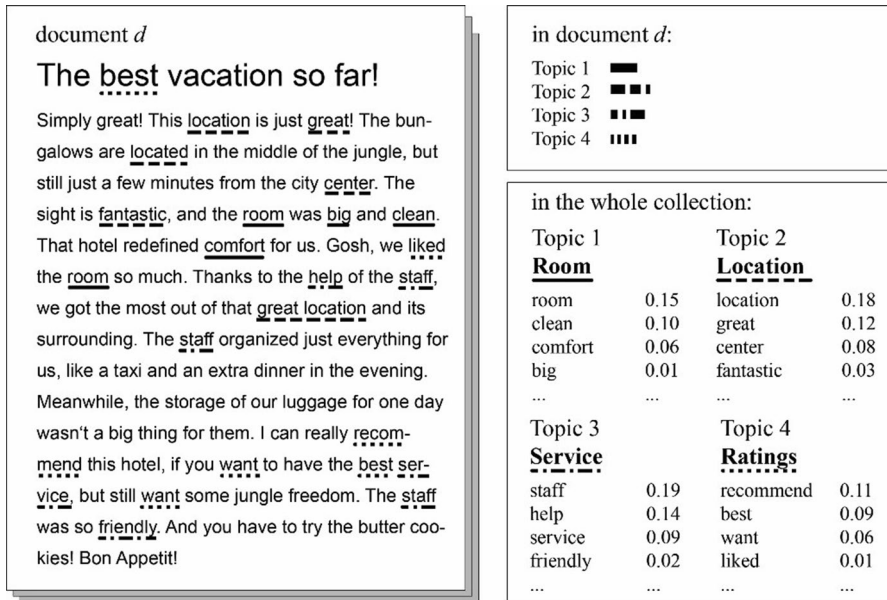


Fig. 1 Exemplary topics in an online hotel review (Own Source)

single category” (Galyardt 2015, p. 39) and “each document exhibits the topics in different proportion” (Blei 2012, p. 78). Thus, LDA belongs to the family of mixed membership models (Galyardt 2015, p. 43). Contrary to classical clustering methods, where membership is a binary variable, every unit (word) partially belongs to all clusters (topics) with different probabilities, expressed by a (membership) “vector of continuous non-negative latent variables that add up to 1” (Airoldi et al. 2015, p. 4). Similarly, every topic partially belongs to all documents with varying probabilities. Due to that, the basic modeling output in LDA are groups of words (i.e., topics) with their membership probabilities and the proportions of these topics in each document (Blei and Lafferty 2009, p. 74; e.g., Park and Ha 2016, p. 1495). Normally, researchers constrain the output to a low number of the highest probable words and topics (up to 20) and assign a topic label. Figure 1 illustrates this, using an online posting as an example with the highest probable words within four retrieved topics and the assigned topic labels.

To achieve that, LDA computes the posterior distribution (conditional distribution) of the hidden variables performing data analysis on the joint probability distribution over hidden and observed variables. Figure 2 indicates the observable variable (all words)  $W_{d,n}$  within the collection of words in documents  $N$  and the collection of documents  $D$ .  $K$  is the number of all topics, which is usually set by the researcher. There are numerous hidden variables and dependencies. For example, a specific word  $w_{d,n}$  depends on the hidden “topic assignment for the  $n$ th word in document  $d$ ” (Blei 2012, p. 80) ( $z_{d,n}$ ), and on all topics  $\beta_{1:K}$  (where each  $\beta_k$  is a multinomial distribution over all words for a topic  $k$ ).  $z_{d,n}$  depends on  $\theta_d$ , which is the (multinomial) per-document topic distribution (Blei 2012, pp. 79). Additionally, LDA uses a

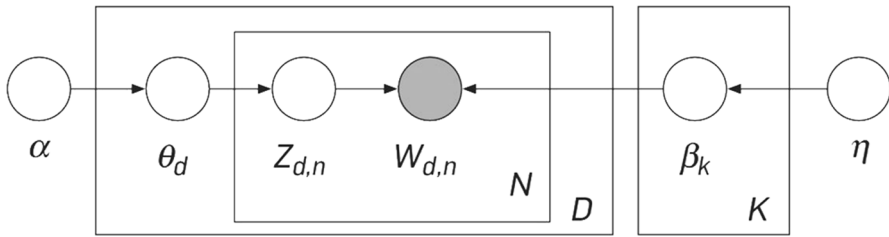


Fig. 2 Plate notation of LDA (Blei 2012, p.81)

Dirichlet prior on the per topic word distribution  $\eta$  and a Dirichlet prior on the per-document topic distribution  $\alpha$ , which in practice are mostly set up to be symmetric (Blei and Lafferty 2009, p. 74; Galyardt 2015, p. 40; Wallach et al. 2009, pp. 1).

More formally, Blei (2012, p. 80) defines these dependencies in the generative process for LDA, which depicts the joint distribution of hidden and observable variables in the model:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \left( \prod_{n=1}^N p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right).$$

The computational problem is to calculate the conditional (posterior) distribution of the topic structure (Blei 2012, p. 81), where the numerator is the joint distribution of all random variables:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}.$$

It is intractable to directly calculate the posterior due to the denominator, which is the marginal probability of observations. In theory, it would be calculated by summing “the joint distribution over every possible instantiation of the hidden topic structure” (Blei 2012, p. 81), which is exponentially large (Blei 2012, p. 81). Therefore, one needs to infer or approximate it. For inference Blei et al. (2003, pp. 1003; pp. 1019) originally used a convexity-based variational approximation with EM (expectation maximization), where the posterior is approximated by optimizing a distribution to be close to it (Blei and McAuliffe 2010, p. 3). Since then, MCMC (Markov chain Monte Carlo) algorithms (like Gibbs) have become prevalent (e.g., Wallach et al. 2009, p. 2), where samples from the posterior are collected, to approximate it with an empirical distribution (Blei 2012, p. 81). However, both procedures need initialization. A common procedure is to randomly initialize inference (Boyd-Graber et al. 2015, pp. 231), triggering an iterative process. Each iteration provides an update for the algorithm, working slowly towards a local optimum, estimating what the probability of recreating the data is, given the current values of the latent variables (Boyd-Graber et al. 2015, p. 232; Underwood 2012).

## 2.2 Comparing LDA to related methods

The just described basic LDA is clearly related to various clustering and other exploratory data compression techniques. Of particular relevance is model-based clustering, using finite mixture models (FMM) (e.g., McLachlan and Peel 2000, pp. 1–39; Titterton et al. 1985, p. 8), the Products of Experts (PoE) model as introduced by Hinton (2002, p. 1), and exploratory factor analytic (EFA) procedures (e.g., Muthen 1978, p. 407).<sup>1</sup> Below, we briefly focus on similarities and differences between these methods and LDA.

LDA is defined as a generative mixed membership model (Galyardt 2015, pp. 40), and thus represents a generalization (Sammuto and Webb 2011, p. 689), or a special application (Sun et al. 2012, p. 260) of (generative) finite mixture models. FMMs assume that  $L$  observed subjects (with observations contained in a vector  $y$ ) of an overall (empirical) distribution stem from several latent groups (sub-distributions) ( $C_1, \dots, C_S$ ) in proportions  $\pi_1, \dots, \pi_S$ , each encompassing the same distributional family, but with varying values (Frühwirth-Schnatter 2006, p. 5; Wedel and Kamakura 1999, p. 76; Wedel 2002, p. 364). Let  $v$  be a latent indicator vector for the multinomial group memberships of subject  $l$  and  $\phi$  denote the parameters of the probability density function  $f_S(y|\phi_S)$ . When  $\Phi = \{\pi, \phi\}$  denotes the parameters in the model, the unconditional (overall) probability distribution of  $y$  is:

$$f(y|\Phi) = \sum_S \pi_S f_S(y|\phi_S).$$

Once the parameters are obtained via maximizing the corresponding likelihood function (e.g., using an expectation maximization procedure; Wedel and Kamakura 1999, p. 80), the posterior probability of a subject  $l$  being member of group  $C_S$  can be derived using the Bayes rule (Wedel 2002, p. 365):

$$\hat{p}_k = f(v_S|y; \hat{\Phi}) = \frac{f(v_S|\hat{\pi})f(y|v_S; \hat{\phi})}{\sum_h f(v_h|\hat{\pi})f(y|v_h; \hat{\phi})} = \frac{\hat{\pi}_S f_S(y|\hat{\phi}_S)}{\sum_h \hat{\pi}_h f_h(y|\hat{\phi}_h)}.$$

Thus, due to unobserved heterogeneity in the data, FMMs assign a probability to the cluster assignment and thus belong to the family of (probabilistic) fuzzy clustering methods (Clement and Bořow-This 2007, pp. 169). The above described LDA is related to FMMs, such that technically speaking, in LDA, every document is modeled with an FMM (Teh and Jordan 2010, p. 170), with the mixture proportions being drawn uniquely for each document, but the mixture components (i.e., the topics) are shared across the text corpus (Blei et al. 2010, pp. 2). That is, the mixture proportions in FMMs,  $\pi$ , are equivalent to the topic proportions in a specific document,  $\theta_d$ , in LDA, and the mixture components in FMMs are equivalent to the

<sup>1</sup> We thank one of the anonymous reviewers for pointing us to this important aspect.

topics,  $\beta_{1:K}$ , which are shared across the corpus, in LDA. Specifically, the generative process of words in a specific document in LDA requires multiple draws (of topics and of words) from this finite mixture. When considering a different document, there is again a finite mixture, “with the same mixture components (the topics), but with a different set of mixing proportions (the document-specific vector  $\theta_d$ )” (Teh and Jordan 2010, p. 170). Broadly speaking, there are further similarities between both methods. For example, the determination of a suitable number of topics / mixture components to obtain reasonable results (Melnykov and Maitra 2010, p. 88; Teh and Jordan 2010, p. 260). Also, similar to LDA, FMMs are quite flexible in certain aspects, like in choosing the family of distributions under consideration (Costa Filho 2010, p. 9).

Another important related technique to LDA is the PoE model to learn sharp distributions in high dimensional spaces by multiplying and renormalizing sub-distributions (called experts). Because of its capability to detect specific subsets of dimensions in a high dimensional space and to better conjointly describe them, PoE can improve the performance of FMMs for certain applications (Hinton 2002, pp. 1; Salakhutdinov and Hinton 2009). The overall distribution,  $P(x)$ , represented by a PoE is as follows:

$$P(x) = \frac{1}{Q} \prod_{i=1}^M f_i(x|\sigma_i).$$

with  $f_i(x|\sigma_i)$  denoting un-normalized experts and  $Q$  the overall normalization constant (Welling et al. 2002, p. 2). Translated into the conceptual framework of LDA, a “product of topics” could give a higher probability to a cross-cutting specific word, where the word is not as strongly predicted by each individual topic, thus better accounting for interrelations between topics (Salakhutdinov and Hinton 2009, p. 1). On the other hand, due to multiplying every expert, there is no explicit mixture proportion (i.e.,  $\pi$  in FMM, and  $\theta_d$  in LDA), but the experts themselves are the weights, which does not make sense for the LDA framework when directly transferring this into it, due to the need of a per-document topic distribution,  $\theta_d$ . However, there are attempts to integrate PoEs into topic modeling. For example, Gormley et al. (2012, pp. 783), conceptualize each single topic as  $P(x)$ , consisting of experts, thus adding an additional layer to standard LDA, to e.g., avoid overparameterization, and to account for topic relations.

Finally, LDA is also related to a third family of methods, namely, exploratory or model-based factor analysis for binary variables used for dimensional reduction. Similar to LDA, factor analytic models aim at compressing high-dimensional data sets into a smaller set of (latent) common factors (equivalent to the topics in LDA), while conserving as much of the original information as possible (Bartholomew et al. 2011, p. 209; Fabrigar and Wegener 2012, p. 10). There are many methodological variants for finding latent groups for binary data (e.g., Hruschka 2016, p. 3; Muthen 1978, pp. 551; Muthen and Christoffersson 1981, pp. 407). While similar by idea, there are also some notable differences to LDA, regarding the conceptualization, structure, and output. For example, in LDA, the



co-occurrences of words in probability distributions that are part of the method's structure, form the topics (i.e., the factors), not correlations between variables.

### 2.3 Extensions of the basic LDA

Due to its open framework, LDA is highly extendable and can easily be applied to various kinds of data (Airoldi et al. 2015, p. 4; Blei 2012, p. 83). The prerequisite is a large set of documents, each consisting of discrete units, which are distributed unevenly. Whatever the documents and units specifically might be, plays a minor role from that perspective. For example, mixed membership models in general have been applied to texts (e.g., Wang and Blei 2011, pp. 450), surveys (Gross and Manrique-Vallier 2015, pp. 119), political voting data (Gormley and Murphy 2015, p. 441), population genetics (Shringarpure and Xing 2015, p. 397), image analyses (Cao et al. 2014, pp. 8959), image and text analyses (Blei and Jordan 2003a, p. 128), and more (e.g., Blei and Lafferty 2009, p. 71). Data in marketing relevant research exemplarily encompass purchase histories and consumer data (Hruschka 2016, p. 7; Ishingaki et al. 2015, p. 17; Jacobs et al. 2016, p. 397), the internet browser's cookies (Trusov et al. 2016, pp. 409), and mobile apps usage data (Do and Gatica-Perez 2010, p. 3).

Additionally, in LDA, even basic assumptions like the so-called “bag-of-words” property can be relaxed. The latter assumption characterizes the fact that the basic LDA ignores the order of words in documents and the order of documents in a text corpus (Blei 2012, pp. 82). Also, basic statistical assumptions like the assumed distributions were altered.<sup>2</sup> This flexibility leads to a substantial amount of extensions of LDA (e.g., Airoldi et al. 2015, p. 4; Balasubramanian and Cohen 2015, p. 256; Blei and Jordan 2003a, pp. 2). In fact, the model has served as a basis for the advent of numerous other types of topic models (Blei and Lafferty 2009, p. 72). These include the Correlated Topic Model (CTM) for discovering correlations between topics (Blei and Lafferty 2007, p. 17; Blei and Lafferty 2009, pp. 82), Dynamic LDA for modeling topics as changing over time (Blei and Lafferty 2009, p. 82), the Supervised Topic Model (sLDA), where an additional response variable (e.g., the rating of a text) is integrated for better fitting on the data (Blei and McAuliffe 2010, pp. 2), Online LDA, which reduces the computational time needed in LDA for massive document streams (Hoffman et al. 2010, p. 2), the Author Topic Model (ATM), which associates each author with a topic probability (Rosen-Zvi et al. 2004, p. 487), and the Author-Recipient Topic Model (ART), which further extends this idea by “building a topic distribution for every author recipient pair” (Balasubramanian and Cohen 2015, p. 260). For instance, other models also include hierarchies between documents like the Hierarchically Supervised Latent Dirichlet Allocation (HSLDA) (Wood and Perotte 2015, pp. 311). Various network topic models mainly differ in the view of what a network is, what a link consists of, if the links are conceptualized as within or between documents, and if additional factors (e.g., time dependency) are included (e.g., Airoldi et al. 2015, p. 7).

<sup>2</sup> For example, a logistic normal prior distribution is utilized in CTM to be able to correlate topics (Paisley et al. 2015, p.207).

These form a substantial amount of the latest attempts in the field to analyze complex, multi layered real world data. Examples are so-called Relational Topic Models (RTMs) (Chang and Blei 2009, p. 81), and Block-LDA (Balasubramanian and Cohen 2015, pp. 255). However, only a small fraction of these more recently introduced models are currently in use (see for example: Table 6).

## 2.4 Procedures and criteria for model evaluation

Stemming from the variability of the approach, in conjunction with a diverse range of data and research interests, there are numerous procedures and metrics used to evaluate the models. These are applied at various stages of the modeling process and include manual, semi-automated, and fully automated methods (Roberts et al. 2015, pp. 12). Likewise, evaluating topic models could include the computational performance (e.g., Jacobs et al. 2016, pp. 394), indicators for optimal parameter settings, model fit (in sample and predictive out of sample), and the assessment of the clustering output (topics). In general, a widely used practice is to run the same model with a different number of topics, and by varying other model parameters (e.g., the prior distributions), or to render different topic models on the same data for comparison (e.g., Hruschka 2014, p. 270; Roberts et al. 2015, p. 18; Wang et al. 2015, p. 4). This may enable the researcher to find a good model, feasible parameter settings, reasonable topics, and to assess the stability of the output and of the covariate effects (Roberts et al. 2015, pp. 14; pp. 22; p. 24). When doing that, a quite common procedure is to split the data into a training set and a hold-out validation set. This enables scholars to examine how predictive models behave on unseen data, e.g., utilizing hold-out-likelihood (e.g., Blei and McAuliffe 2010, p. 8). However, a substantial amount of research has gone into the assessment of the clustering output (topics). One way to examine topics is to evaluate semantic coherence, which is a summary measure to capture “the tendency of a topic’s high probability words to co-occur in the same document” (Mimno et al. 2011 quoted from Roberts et al. 2015, pp. 13; see also Newman et al. 2010, p. 100). Newman et al. (2009, pp. 3) use a slightly different approach. They propose a model to measure the likelihood of co-occurrence of words, by using external text data sources to provide regularization instead of the internal text data in the documents. Another important aspect of topics is exclusivity, which “captures whether those high probability words are specific to a single topic” (Roberts et al. 2015, p. 14). A manual approach to condense both procedures is to ask human raters if topics are interpretable and can be associated with a single concept (Newman et al. 2009, pp. 2). Also, one can validate that topics capture a single concept by “reading several example documents” (Roberts et al. 2015, pp. 12), or by comparing already present categories to the automatically generated clusters (Roberts et al. 2015, pp. 12).

## 2.5 Limitations and critique

Despite the above-mentioned recent advances and extensions of the basic LDA, the method is not immune against criticism and limitations. First, there are obvious problems like intervening variables (such as author and environment specific

covariates).<sup>3</sup> Despite recent advances allowing to infer the number of topics from the data, such as the Hierarchical Dirichlet Process as an extension of LDA (Teh et al. 2006, p. 1575), it is still prevalent to choose the number of topics beforehand (Blei and Lafferty 2009, p. 81) and to employ post hoc procedures to validate the suitability of choice. However, if the number of topics is chosen wrong, this can result in a poor performance (Tang et al. 2014, p. 7). Another eligible critique is the need for extensive parameter optimization before running (Asuncion et al. 2009, p. 30), possibly more arranging topic models to fit the needs of the researcher, than capturing what is really there (Schmidt 2013). Also, the underlying bag-of-words assumption (where information on word order is lost) has been criticized for oversimplifying documents (Shafiei and Milios 2006, p. 1).

Putting these deficiencies aside, Schmidt (2013) questions assumptions that are at the heart of the method. More specifically, analysts assume that topics are coherent (i.e., they share some common aspect) and stable (i.e., they apply to several documents the same way), leading them to the opinion that the co-occurrence patterns of words are “more meaningful than the words that constitute them” (Schmidt 2013) and appropriately capture a concept by being semantically coherent. However, as Schmidt (2013) nicely illustrates, in some instances, the top few words characterizing a topic are not necessarily a decent summary of the large amount of words constituting the whole probability distribution. Transferred to language processing, the most frequent words don’t necessarily create the meaning. To a certain extent, these problems can be at least partly solved by techniques like word removal, or changing the bag-of-words assumption to incorporate more information. However, these are also related to the practice of looking just on the top words in the output of the model. Newman et al. (2009, p. 2) express a similar concern on this issue by saying that some topics learned by a model “(while statistically reasonable) are not particularly useful for human use”. Intertwined with that, Crain et al. (2012, pp. 144–148) note that LDA tends to learn broad (more diffuse) topics, where adding concepts to the same topic are favored if these share the same aspects. Thus, the suggestion of Schmidt (2013) is to put extensive effort into visualizing and validating the model before interpreting the results. This, of course, is considerably easier when analyzing position data on a map than with words and their respective semantic implications. Tang et al. (2014, pp. 4–8) point to limiting factors of LDA using a posterior contraction analysis. For example, they depict that a small number of documents (no matter how much words these contain) makes it impossible to guarantee a valid topic identification. The underlying topics need to be well separated for good LDA performance in the sense of Euclidian metric, which is the case if, for example, these are concentrated at a small number of words.

Finally, there are numerous papers which compare the performance of LDA to other related methods in certain problem domains, pointing towards possible deficiencies of LDA for these. Even if LDA is optimized (e.g., in terms of the number

---

<sup>3</sup> Intervening variables like author characteristics (e.g., psychological factors, self selection), user interface specific features in online reviews (e.g., a constraint in the number of words to be written), etc., often determine what is written in the documents and add a predictive element to topic models.

of topics), other methods are capable of outperforming it for specific tasks, data and setups. The methods under consideration comprise a multitude of quantitative methods. For example, Hruschka (2016, pp. 8) compared the relative performance of LDA, CTM, a Binary Factor Analytic Model, Restricted Boltzmann Machines (RBMs) and Deep Belief Nets (DBNs) for predicting purchase incidences in a market basket. The author shows that Binary Factor Analysis vastly outperformed both topic models, and was itself outperformed by RBMs and DBNs. Schröder (2017, pp. 31) used a Multidimensional Item Response Theory Model (MIRT) to analyze market baskets for identifying latent traits of households and predicting purchase behaviour. Based on AIC (Akaike Information Criterion), and AICc (Corrected Akaike Information Criterion), MIRT outperforms LDA for both, the binary and the polytomous purchase data scenarios. Salakhutdinov and Hinton (2009, pp. 6) introduce the Replicated Softmax Model, to automatically model low dimensional latent semantic representations in academic and in newspaper articles. Compared to LDA, their model makes better predictions and has a higher retrieval accuracy.

### 3 Approaches and applications in marketing research

In this section, we derive a structured review of main methodological implications for applying topic modeling in marketing. By utilizing conceptual articles as well as analyzing empirical work, we derive the following characterizing dimensions to categorize prior applications of topic models in the field of marketing: Data structures and data retrieval (3.1), topic model implementation and extensions (3.2), procedures used for model evaluation (3.3). In the course of developing Sects. 3.1, 3.2, 3.3, we describe typical methodological strategies employed by the relevant literature along with a number of characteristic examples compiled in Tables 1, 2, 3, 4, 5 to answer our research question RQ1. Subsequently, we combine these findings in Table 6 and Table A1 in the appendix, categorizing all available publications into the scheme developed in Tables 1, 2, 3, 4, 5. By doing so, Table 6 provides an integrated view on the extent that a strategy is utilized in a specific field of research (RQ2).

#### 3.1 Data structures and data retrieval

Most of the data used in topic modeling is unstructured text in various forms like websites and online advertisements (e.g., Le et al. 2008, pp. 889), social media postings (e.g., Schieber et al. 2011, pp. 3), online product reviews (e.g., Tirullinai and Tellis 2014, pp. 465), and more (for a more complete overview see Table A2 in the appendix). Of course, there are also other forms of data like images (e.g., Cao et al. 2014, pp. 8959), purchase records (e.g., Jacobs et al. 2016, pp. 397), mobile apps usage records (Do and Gatica-Perez 2010, pp. 3) and traces of internet browsing behavior (Schröder et al. 2017, pp. 40; Trusov et al. 2016, pp. 409). Some of these comprise discrete units to be crawled and to be pre-processed beforehand (e.g., words, products, etc.) (e.g., Blanchard et al. 2017, pp. 403; pp. 407; Büschken and

Allenby 2016, p. 958; Ishingaki et al. 2015, pp. 5; Knights et al. 2009, pp. 244; Schieber et al. 2011, p. 3; Wang et al. 2015, p. 3). For others, scholars need to extract these units automatically or manually before the actual application of a topic model. An example of an automatized extraction is the application of algorithms for the recognition of shapes and patterns prior to any attempts of clustering (e.g., Cao et al. 2014, pp. 8958; Wang et al. 2015, p. 3). A rather manual approach is proposed by Do and Gatica-Perez (2010, pp. 4), who generate artificial words for mobile apps usage, which describe categorical groups, manually predefined by the authors, like the level (high, medium, low), or the time of the day (night, morning, afternoon, evening) of mobile apps usage. For the sake of completeness, however, it should be noted that a third distinct approach is to use synthetic data in a simulation study. In the papers examined by us, this kind of approach was solely used to evaluate topic models before the actual main analyses by knowing the structure of the artificially generated data. There are two ways of doing that. Firstly, one takes an algorithm (often a topic model) to create data and distributions (e.g., Blanchard et al. 2017, p. 402; Büschken and Allenby 2016, p. 971; Knights et al. 2009, p. 243). For example, Knights et al. (2009, p. 243) used LDA to generate documents from synthetic topics with each topic following a random distribution over a set of symbolic words (Knights et al. 2009, p. 243), to subsequently evaluate their Compound Topic Model. Büschken and Allenby (2016, p. 971) evaluate the efficiency of their Sentence-constrained LDA (SC-LDA) model, where a sentence is constrained to have one topic against basic LDA. Secondly, Ishingaki et al. (2015, pp. 13) create purchase records based on marketing variables from a real customer base (Ishingaki et al. 2015, pp. 13), which makes sense for their research aim, in which they try to connect these to latent purchase patterns. By doing so, they intend to assess model performance in terms of the estimators precision and computational time needed for predicting household purchase patterns. We briefly summarized these strategies in Table 1 in conjunction with a few exemplary publications.

In terms of LDA, the critical definition of documents, topics and words vastly differs between papers (Table A2 in the appendix—column: conceptualization of data). Obviously, it varies with the examined data. For example, in the paper of Sun et al. (2013, p. 2–4), documents are a user's purchase history with each purchased product being a word in this document and topics (i.e., the clusters to be retrieved) are the customers' purchasing preferences. Schieber et al. (2011, pp. 4) try to model topics and individuals' opinions about products in Twitter. Accordingly, they define a document as a single Tweet, which consists of words and (possibly) contains a few topics. However, the research interest is another important determinant of how the data is conceptualized. Sticking to the example provided, as already mentioned, a document can be set up as a single posting of a user (Schieber et al. 2011, p. 5). In contrast, Weng et al. (2010, p. 264), who try to identify influential users on Twitter by the following structure and topic similarities, classify all postings of a user as a document (Weng et al. 2010, p. 264). For Paul and Girju (2009, pp. 1412), who intend to find differences in services related to traveling, an online discussion or thread (involving postings of several users) is a document. There are further assumptions that reflect themselves in the methodological conceptualization of the data. For

**Table 1** Data structures and data retrieval for topic modeling in marketing

Units are already present	Extraction of units beforehand	Synthetically generated data
<p>The units to be pre-processed and used in topic models are already present beforehand like words in texts, products in purchase records, etc.</p> <p>E.g., Blanchard et al. (2017, p. 403; pp. 407); Büschken and Allenby (2016, p. 958); Ishingaki et al. (2015, pp. 5); Knights et al. (2009, pp. 244); Schieber et al. (2011, p. 3); Wang et al. (2015, p. 3);</p>	<p>Automatized recognition of discrete units from high dimensional data using algorithms, lexica, etc.</p> <p>E.g., Cao et al. (2014, pp. 8958); Wang et al. (2015, p. 3);</p> <p>Extraction of discrete units using manually predefined groups / categorical reduction of high dimensional data</p> <p>E.g., Do and Gatica-Perez (2010, pp. 4);</p>	<p>Algorithmic generation of entirely artificial data</p> <p>E.g., Blanchard et al. (2017, p. 402); Büschken and Allenby (2016, p. 971); Knights et al. (2009, p. 243);</p> <p>Generating artificial data based on / including real data</p> <p>E.g., Ishingaki et al. (2015, pp. 13);</p>

example, scholars often restrict a certain unit in a document (e.g., a sentence) to have a single topic (e.g., Büschken and Allenby 2016, p. 954).

### 3.2 Topic model implementation and extensions

In most publications either basic LDA (e.g., Chen et al. 2013, p. 1), common adaptations like Labeled LDA (e.g., Ramage et al. 2010, p. 132), sLDA (e.g., Blei and McAuliffe 2010, pp. 2), temporal LDA (e.g., Wang et al. 2012, p. 124), the Author Topic Model (e.g., Do and Gatica-Perez 2010, pp. 4), CTM (e.g., Trusov et al. 2016, pp. 413), or other custom adjustments of LDA (e.g., Büschken and Allenby 2016, pp. 957; Paul and Girju 2009, p. 1410; Tirullinai and Tellis 2014, p. 468) are used. However, noticeably less commonly, scholars utilize more exotic models like the User Aware Sentiment Topic Model (USTM) (Yang et al. 2015, pp. 415–417), or the Visual Sentiment Topic Model (VSTM) (Cao et al. 2014, pp. 8959). Apart from that, when taking a broader perspective, in the papers under consideration, there is a continuum of scientific strategies in how to use topic models. One easily to implement approach consists of utilizing the method for the clustering output to represent the actual research results in an exploratory manner (e.g., Cao et al. 2014, p. 8964; Christidis and Mentzas 2013, pp. 4375; Karpienko and Reutterer 2017, p. 17; Luo et al. 2015, pp. 1185; Schröder et al. 2017, pp. 42; Sun et al. 2013, p. 7; Wang et al. 2015, p. 3; Yang et al. 2015, pp. 419).

As a consequence, scholars often perform further analyses and visualizations of certain aspects of the data. For example, it is quite common to use a topic model as an in-between-step for an overall model or research aim (e.g., Cao et al. 2014, pp. 8959; Christidis and Mentzas 2013, pp. 4373; Luo et al. 2015, pp. 1180; Sun et al. 2013, pp. 4; Yang et al. 2015, pp. 420). For instance, Luo et al. (2015, pp. 1180) try to find marketing topics in social media postings and their influence, which is measured by the reaction of users. They utilize LDA to get a topic probability vector for each micro-blog-post and subsequently calculate the topic influence, the topic response, and the topic trends by implementing various custom formulas for further processing. The topic influence is a measure of the proportion of microblog-posts to contain a certain topic, topic response indicates how much customers actively engage in reposting, and the latter indicates the development of the former two over time and across topics (Luo et al. 2015, p. 1182). Different in context but similar by idea, Karpienko and Reutterer (2017, pp. 11) apply LDA to the abstracts of a large compilation of academic marketing articles to derive latent topics of scholarly interest as a preprocessing step before inferring communities of topic combinations, using a version of social network analysis. Subsequently, the authors study the evolution of marketing topics over time and academic journals. Yet another example for using LDA in conjunction with other data compression techniques is the study by Schröder et al. (2017, pp. 42), who use LDA to derive latent shopping interests based on users' website browsing behavior. Based on the derived latent shopping interests, the authors examine the existence of different online shopper segments using *k*-means clustering and study implications on shopping behavior. Sun et al.

(2013, p. 2), who try to predict customers' propensities to join group-purchasing events on an online social platform, use LDA to capture the purchasing preferences of customers beforehand. They propose two models, which use this output for further calculation. The first one, PCIM (product-centric inference model), tries to apprehend if the specific product determines that a customer would join a group purchasing event, given the user's purchasing preferences. Accordingly, the second one, GICIM (group-initiator-centric model), assumes that the group initiator in the social network plays the decisive part in that process, when accounting for the user's and initiator's topic mixture (Sun et al. 2013, p. 5). By doing so, they extend the application of topic models from an exploratory approach to one which supports hypotheses testing. Dan et al. (2017, pp. 42) also build a predictive model based on the output derived by LDA for the case of online hotel reviews. Using the latent topics derived from a sentence-constrained LDA version, they develop a latent rating regression approach for making inferences on the relative contribution of each latent topic on guests' overall hotel evaluations. Another example of that kind of utilizing the method represents a practical approach. Christidis and Mentzas (2013, pp. 4373) try to build a topic based recommender system for buyers and sellers on an e-auction platform, consisting of buyer item recommendations and seller text and item recommendations. They use basic LDA to extract the probabilities of words in topics and topics in documents (i.e., the item descriptions / items in the online marketplace). Subsequently, they calculate the cosine similarities between items (using the topic distributions of each item) and similarities between topics and terms to establish the recommendation functionality.

Obviously, a vital step in current research is to incorporate the output of a complex method or model into topic models (e.g., Cao et al. 2014, pp. 8959; Wang et al. 2015, pp. 2; Yang et al. 2015, pp. 415–418). Self-evidently, this is highly associated with more sophisticated forms of data extraction and preparation. An example for that kind of approach is the USTM framework, which aims at modeling user metadata, topical aspects, and sentiments in online consumer reviews (Yang et al. 2015, p. 414), to aggregate the opinions of various market segments. To capture the sentiments, the authors utilize seed words and two sentiment lexica and incorporate the sentiment information by using asymmetric Dirichlet priors to assign e.g., positive words with a higher probability for positive topics (Yang et al. 2015, p. 414–419). Another example is the Image-regulated Graph Topic Model (IGTM), where the authors utilize the SIFT feature algorithm and  $k$ -means clustering, to extract 500 discrete visual words from images, and build a bag of visual words model for detecting weighted relationships between images via an image relation graph, which consists of nodes (images) and edges (similarities between images). By using further variables, IGTM aims at jointly modeling text and images to enrich topic detection with sentiments (Wang et al. 2015, pp. 2). They allocate a topic assignment and an image assignment for each word in a document, where each topic is a multinomial distribution over words and an image is a multinomial distribution over topics (Wang et al. 2015, pp. 2). There are further instances like Cao et al. (2014, pp. 8958) using an Adjective Noun Pairs (ANPs) based detector for image annotations and Visual Sentiment Ontology (VSO) detectors to construct SentiBank. Subsequently, they



incorporate this information in their topic models, trying to enhance the prediction of sentiments in images.

However, at the same time, these authors mark a fourth noticeable implementation strategy, which is to combine a set of topic models in a framework, to improve prediction. More specifically, Cao et al. (2014) intend to build a topic model, which analyzes the distributions of visual sentiment features in topics. Since some non-discriminative (i.e., topic irrelevant) sentiment features have high probabilities in topics, they introduce a background topic model and additional estimators to distinguish these from discriminative ones.

However, it is noteworthy that scholars often tend to use a few strategies (Table 2) in conjunction, meaning that these are not mutually exclusive (e.g., Cao et al. 2014; Sun et al. 2013; Yang et al. 2015).

In addition to these various implementation strategies, authors extend topic models (or use common extensions) (e.g., Blei and Jordan 2003a, p. 3; Blei 2012, pp. 82) (Table 3). These are intertwined with each other and vary with the research aim, modeling purpose and data.

### 3.3 Evaluation procedures

Beyond data structure and implementation, the examined papers employing topic models in marketing can be distinguished along another dimension, namely, the way they evaluate results. All papers examined (except for 11, which do not evaluate the topic model at all<sup>4</sup>) utilize quantitative techniques. The range of these includes H-Scores, Cosine Similarities (Wang et al. 2015, pp. 3), the Kullback–Leibler Divergence (Cao et al. 2014, pp. 8962), Mean Average Precision Scores (Phuong and Phuong 2012, p. 67), the Log Marginal Density (Büschken and Allenby 2016, pp. 962), the Log-Likelihood (Park and Ha 2016, p. 1493; Paul and Girju 2009, p. 1415), Perplexity (Yang et al. 2015, pp. 417), RMSE (Root Mean Square Error) (Iqbal et al. 2015, p. 4), descriptive statistics and correlations (Do and Gatica-Perez 2010, p. 7), Fleiss’k (Ramage et al. 2010, p. 134), and more. Additionally, a considerable portion of publications explicitly or implicitly use qualitative techniques (e.g., Tirullinai and Tellis 2014, p. 470) since it is a common, but in case the model validation relies to a high degree on it, inappropriate practice to use the interpretation of generated topics for model validation (Chang and Blei 2009, p. 3). In general, scholars either provide indicators that encompass the whole model or analyze components of it (e.g., Liu et al. 2010, pp. 204; Yang et al. 2015, pp. 418) (or both). To put this into perspective on an aggregate level, we employ the methodological categories discussed in Sect. 2 (e.g., Roberts et al. 2015, pp. 12), and summarize these in Table 4.

Within the evaluation techniques depicted in Table 4, scholars perform comparisons with various goals in mind. More precisely, in about 80% of publications, the

<sup>4</sup> Authors who do not evaluate their methods: Amado et al. (2017); Chen et al. (2013); Dan et al. (2017); Ha et al. (2017); Heinrich (2015); Herzig et al. (2014); Iqbal et al. (2015); Pathak et al. (2008); Song et al. (2017); Tran et al. (2015); Xie et al. (2012).

**Table 2** Types of topic model implementation

Exploratory baseline	Topic models as part of a more complex model or research objective
Using the output of topic models as research results	Using the output of other methods / models as vital input for topic models
<p>E.g., basic LDA to cluster textual online reviews from PatientsLikeMe.com (Park and Ha 2016, pp. 1494); Further examples: Cao et al. (2014, p. 8964); Christidis and Mentzas (2013, pp. 4375); Karpienko and Reutterer (2017, p. 17); Luo et al. (2015, pp. 1185); Schröder et al. (2017, p. 42); Sun et al. (2013, p. 7); Wang et al. (2015, p. 3); Yang et al. (2015, p. 419);</p>	<p>E.g., USTM, where an additional Dirichlet Prior is constructed from sentiment databases (Yang et al. 2015, pp. 415–418); Further examples: Cao et al. (2014, p. 8959); Wang et al. (2015, p. 2);</p> <p>E.g., using the clustering output of LDA as input for further models (PCIM &amp; GICIM) (Sun et al. 2013, pp. 4); Further examples: Cao et al. (2014, pp. 8959); Christidis and Mentzas (2013, pp. 4373); Dan et al. (2017, pp. 46); Karpienko and Reutterer (2017, pp. 11); Luo et al. (2015, pp. 1180); Schröder et al. (2017, pp. 42); Yang et al. (2015, pp. 420);</p>
	Using several topic models in a framework
	E.g., VSTM, which consists of a foreground and a background topic model (Cao et al. 2014, p. 8959);

**Table 3** Topic model (LDA) extensions

Integrating additional variables	Changing the inference method	Changing basic assumptions	Introducing constraints
Incorporating extra information into the model in terms of additional variables / parameters	Changing the inference method (Variational Approximation with EM in original LDA) (Blei et al. 2003, p. 1003) to optimize the predictive performance, the rate of convergence and the computational effectiveness in respect to e.g., the data, the number of topics and hyperparameter settings (Asuncion et al. 2009, pp. 28)	Changing basic assumptions entailed in LDA to adapt to specific data and research interests	Optimizing topics to learn by putting constraints into the model in respect to certain purposes and assumptions of the specific research endeavour
E.g., in the form of components, covariates, prior distributions, etc. (e.g., Büschken and Allenby (2016, p. 958); Phuong et al. (2010, p. 132); Trusov et al. (2016, pp. 413); Yang et al. (2015, p. 415–418));	E.g., using MCMC (like Gibbs), Laplace, MAP (Maximum a Posteriori), etc. (e.g., Büschken and Allenby (2016, p. 969); Phuong and Phuong (2012, p. 66); Ramage et al. (2010, p. 133); Trusov et al. (2016, p. 416); Wang et al. (2012, pp. 125); Yang et al. (2015, p. 417));	E.g., assumed distributions (e.g., Trusov et al. 2016, pp. 415), bag-of-words (Yang et al. 2015, p. 418), that the order of documents does not matter (e.g., Wang et al. 2012, pp. 124), etc. Further examples: Büschken and Allenby (2016, p. 958); Phuong and Phuong (2012, pp. 66); Ramage et al. (2010, p. 132);	E.g., one topic per sentence (Büschken and Allenby 2016, pp. 955), correspondence between topics and labels (Ramage et al. 2010, pp. 132), etc. Further examples: Phuong and Phuong (2012, pp. 65); Trusov et al. (2016, pp. 413);

**Table 4** Evaluating topic models

Computational performance	Optimal parameter settings	Model Fit (in sample and predictive out of sample)	Analysis of clustering output	Analysis of the estimator (for inference)
Evaluating the computational performance in terms of e.g., computational time, scalability, etc	Evaluating/determining the optimal parameter settings, in terms of e.g., number of topics, prior values, etc	Evaluating the predictive performance of the model (as a topic quality indicator), e.g., in sample and out of sample, using real or synthetic data	Evaluating the clustering output (topics) in terms of e.g., semantic coherence, exclusivity, etc	Analyzing the estimator for topic inference, e.g., in terms of number of iterations for convergence
E.g., Ishingaki et al. (2015, pp. 13);	E.g., Büschken and Allenby (2016, pp. 961); Park and Ha (2016, pp. 1493); Sun et al. (2013, pp. 6); Tirullinai and Tellis (2014, p. 467; p. 471);	E.g., Büschken and Allenby (2016, p. 962; p. 974); Cao et al. (2014); Park and Ha (2016, p. 1494); Sun et al. (2013, pp. 8); Tirullinai and Tellis (2014, p. 471);	E.g., Büschken and Allenby (2016, pp. 964); Cao et al. (2014); Park and Ha (2016, pp.1493); Tirullinai and Tellis (2014, pp. 472); Wang et al. (2015, pp. 3);	E.g., Büschken and Allenby (2016, pp. 971); Ishingaki et al. (2015, pp. 14); Sun et al. (2013, p. 6);

authors conduct some form of comparison for methodological evaluation. We categorized them into 5 types and include them in Table 5 along with a brief description. These five comparison types rely on human ratings / scores (e.g., Tirullinai and Tellis 2014, pp. 470), external reports and categories (e.g., Tirullinai and Tellis 2014, pp. 470), traditional clustering techniques (e.g., Trusov et al. 2016, p. 417), specific metrics in a field (e.g., Weng et al. 2010, pp. 267), and topic models, either in the form of different types of topic models (e.g., Jacobs et al. 2016, pp. 398), or as variations of the same model (e.g., Christidis and Mentzas 2013, p. 4377; Trusov et al. 2016, p. 417).

Taking a broader perspective, one approach to assess the clustering output is to evaluate the outcome in retrospect (e.g., Tirullinai and Tellis 2014, pp. 472), another is to compare predictions of the model from limited data to hold-out data (e.g., Jacobs et al. 2016, p. 397). Another evaluation strategy is to use synthetically generated data (with known distributions beforehand) to compare expected with actually retrieved results (e.g., Blanchard et al. 2017, p. 402; Büschken and Allenby 2016, p. 971; Knights et al. 2009, p. 243).

#### 4 Topic modeling research in marketing

The 61 reviewed papers contain a diverse set of research aims—methodologically, theoretically and practically. Having broad areas of marketing research in mind (e.g., Parasuraman et al. 2007, p. 54; Proctor 2005, pp. 9), our attempt to categorize the

**Table 5** Types of comparisons for model evaluation

Types of comparisons	Sources
Human ratings/scores	E.g., Tirullinai and Tellis (2014, pp. 470);
Comparing results of an automated process to human ratings/scores and evaluations	
External reports and categories	E.g., Tirullinai and Tellis (2014, pp. 470);
Comparing the clustering output of topic models to external reports (e.g., consumer reports) or already present categories	
Traditional clustering techniques	E.g., Trusov et al. (2016, p. 417);
Comparing the output of topic models to traditional customer segmentation and clustering techniques	
Specific metrics, associated with a field	E.g., Weng et al. (2010, pp. 267);
Comparing a topic model to specific algorithms, associated with a research field (like page rank, in degree, etc.)	
Topic models <sup>a</sup>	E.g., Christidis and Mentzas (2013, p. 4377);
Comparing a topic model to other topic models	Hruschka (2014, p. 270); Jacobs et al. (2016, pp. 397); Tirullinai and Tellis (2014, p. 471);
Comparing (mathematical, componential, parameterwise (like the number of topics)) variations of the same topic model	Trusov et al. (2016, p. 417); Wang et al. (2012, pp. 127);

<sup>a</sup>Comparison on the same data, on in sample & predictive (hold-out data) and on different datasets

publications is guided by a focus on the data, the examined objects, and the research interests from a marketing managerial point of view. The former two help to detect main categories (Table 6), while the latter helps to further subdivide them (table A1 in appendix) (RQ2). By taking these features into account, we discovered a total of 7 main applied research fields in marketing: (1) online textual consumer reviews and services research, (2) sales / retailing, (3) social media, (4) images & cross media, (5) online advertisements, (6) research in marketing literature, and (7) public relations. While the former three represent about 89% of publications, the latter four comprise of 7 papers. In general, 10 of 61 papers were published in (mostly quantitative) marketing journals, of which 5 can be assigned to sales / retailing (Blanchard et al. 2017; Hruschka 2014; Jacobs et al. 2016; Schröder 2017; Trusov et al. 2016), 3 to online textual consumer reviews and services research (Büschken and Allenby 2016; Calheiros et al. 2017; Tirullinai and Tellis 2014), 1 to social media (Song et al. 2017), and 1 to research in marketing literature (Amado et al. 2017). Furthermore, we transferred the sum (numbers), and the relative importance (color) of the methodological strategies exerted by scholars into a matrix (Table 6) to better detect patterns and predict trends in research. When glancing at the 7 retrieved fields of research, we intend to detect noteworthy shifts in these patterns in comparison

**Table 6** Synopsis of topic modeling applications in marketing

Number of Papers*	Aim (Chp. 1)	Data Retrieval (Table 1)	Implementing TM (Table 2)	Extending TM (Table 3)	Evaluating TM (Table 4)	Comparison for Evaluation to ... (Table 5)	Extracted Main Fields of Research																					
								Mainly Model Validation	Mainly Exploratory Research	Units are Already Present	Automatized Recognition of Discrete Units	Extraction of Manually Predefined Groups	Algorithmic Generation of Entirely Artificial Data	Generation of Data Based on / Including Real Data	Output of TMs as Research Results	Output of TM for Further Model / Research Aim	Output of Other Methods / Models as Input for TM	Several TMs in a Common Framework	Integrating Additional Variables	Changing Inference Method	Changing Basic Assumptions	Introducing Constraints	Computational Performance	Optimal Parameter Settings	Model Fit (In Sample & Predictive Out of Sample)	Closer Analysis of Clustering Output	Analysis of the Estimator (for Inference)	Human Ratings / Scores
61	46	41	54	12	7	5	1	50	40	15	3	36	47	31	25	5	37	38	26	6	12	7	7	5	28	43	[Sum of All Papers]	
25	20	16	25	5	1	3		22	13	8		16	19	13	13		18	15	15	1	4	5	1	3	14	20	Online Textual Consumer Reviews & Services Research	
12	10	11	6	1	5	1	1	9	8	1		7	9	6	3	3	10	11	2	4	1			5	4	11	Sales / Retailing	
17	11	11	17	3		1		12	15	3	2	10	14	9	7	2	5	8	5	1	4	2	1	2	6	7	Social Media	
2	2		1	2				2	1	2	1	1	1	1	1				1	2						1	1	Images & Cross-Media Detection
2	2		2					2	2			1	2	1	1		2	1	2					2	2		2	Online Ads
2		2	2	1	1			2	1	1							1	1									1	Research in Marketing Literature
1	1	1	1					1				1	1	1			1	1								1	1	Public Relations

\*Numbers indicate the quantity of papers, the color indicates the relative weight within the category: black means 100% of papers made use of a method, white means no paper did

to the overall impression. By doing that, we integrate the methodological strategies we found in the previous sections (RQ1), by depicting their relative importance in respect to the frequency within specific fields of research, paving the way to answer RQ3.

#### 4.1 Research objectives and evaluating topic models

By simultaneously looking at Table 6 and table A1 in the appendix, we can see that topic modeling extensively depends on somewhat arbitrary factors, like setting the hyperparameters, the number of topics (etc.), and is experimental to a certain degree. Consistently, such uncertainties are visible in an emphasis on model validation. 41% of all papers aim at both (model validation and explorative research), but 34% mainly focus on evaluating their model and the remaining ones (23%), with the exception of one publication which entails no aim, implement the latter. Closely related is the extensive use of evaluation procedures with a focus on (1) assessing model fit of the proposed models, (2) optimal parameter settings, and (3) a detailed analysis of the clustering output. Scholars also introduce comparisons to examine their models (80% of papers). Concisely, they tend to compare (1) variations of the same topic model and (2) their model to other topic models as the dominant strategies. However, an outstanding fact is the frequent inclusion of human ratings and scores (20%). Despite these efforts, there are 11 publications in which there is no applied evaluation technique at all. Scholars in sales / retailing emphasize more on model fit and optimal parameter settings than on an analysis of the clustering output. Intertwined with that is the relatively prominent use of comparing variations of the same topic model. For social media, scholars evaluate topic models less often. One reason for that is that a few authors focus on utilizing topic models as part of a more complex model or research aim, evaluating the overarching model (not topic models directly) (e.g., Herzig et al. 2014, pp. 52). This also reflects itself in the “implementing TM” columns in Table 6. Apart from that, in the online textual consumer reviews section, optimal parameter settings, an analysis of the clustering output, and model fit are predominant, which seems to be a reasonable pattern when looking at the aims of research and the data.

#### 4.2 Utilized data structure

In terms of employed data, scholars heavily rely on already present units in the data (89%). However, a considerable amount (mostly in addition) performs an automatized recognition of units from high dimensional data (20%), and 11% extract manually predefined categories beforehand. It is noteworthy that seven papers don't rely on already present units but solely use another form of the depicted data retrieval strategies (e.g., Cao et al. 2014, pp. 8958; Do and Gatica-Perez 2010, pp. 4). Scholars also utilize artificially generated data for evaluation purposes in 6 publications. The importance of already present units in the data repeats itself in all the specific fields of research, except for sales / retailing, where scholars often deal with manual categorizations of products and brands (e.g., Do and Gatica-Perez 2010, p. 4).

### 4.3 Model implementation and extensions

Methodologically, scholars most often present the output of topic models (82%), but mostly also use this output for further processing in a more complex model (66%). However, in a noticeably smaller amount of publications, just one of the two strategies is exclusively utilized (41%), and 11 papers do not present the output of topic models at all, just using it for a research aim that goes beyond the scope of just deriving topics from the data. A popular strategy is to use the output of another method or model as input for topic models in addition (25%). As previously mentioned, for social media, scholars tend to use the output of topic models for further processing noticeably more often than in other fields of research, almost inverting established methodological patterns. There is an emphasis on utilizing the output of topic models as research results in the Online Textual Consumer Reviews and Services Research branch. These patterns are intertwined with field-specific research interests. In general, scholars tend to use LDA (e.g., Park and Ha 2016, pp. 1491), common extensions and / or perform customizations (e.g., Trusov et al. 2016, pp. 413). Therefore, a total of 87% of papers contain a customized topic model (in comparison to original LDA). Speaking of extending topic models, scholars mostly use MCMC (64%) (and within that Gibbs) for inference, in contrast to Blei et al. (2003). Strikingly, extended topic models seem to be prevalent with 59% of publications introducing additional variables, 51% changing basic assumptions, and 41% introducing constraints. This overall pattern is more or less consistent in the specific fields of research.

## 5 Future directions of topic modeling in marketing

When looking at the recent development of topic modeling research, starting with the first article, which appeared in 2008, we have seen an upward trend in the transition into marketing. As depicted above (Table 6), the reviewed articles cover important research areas. Still, there are interesting phenomena and unexplored fields of research in marketing, which haven't been analyzed to date. Below, we highlight some shortcomings in the previous literature, which should offer future researchers in this exciting field many promising research opportunities. By doing so, we address our third research question (RQ3).

### 5.1 Offline, high dimensional data and additional time-varying information

The already covered research areas seem to be connected to a certain extent to relatively easy to get and to process (online and digital) data. An additional inclusion of offline data and of high dimensional data (e.g., Do and Gatica-Perez 2010, pp. 4) by using sophisticated strategies of data retrieval could further extend the field. Also, as shown by Blei and Lafferty (2006, pp. 5), who propose a Dynamic Topic Model to analyze changes of topics over time, it is worth it to consider the inclusion of time-varying information. Adopting topic modeling frameworks that



allow to integrate additional information, in particular time-dependent marketing covariates, should have great potential in the field of marketing research. Furthermore, integrating some form of guidance in the process of topic generation can be helpful for interpreting the derived solutions (see e.g., Andrzejewski and Zhu 2009, pp. 43; Blei and McAuliffe 2010, pp. 2).

## 5.2 Topic models and complexity in marketing research

Modeling marketing phenomena entails complexity, reflecting itself in methodologically elaborated conceptualizations and procedures (see the “implementing TM” and “extending TM” columns in Table 6). We expect this aspect to proceed and intensify in the future—both on the implementation level of topic models, as well as on the level of extending the method itself. We highlighted numerous examples of implementing the method. For example, an intriguing approach is presented by Sun et al. (2013), shifting topic modeling from an explorative method to one that supports hypotheses testing, by integrating the output of the topic model into two subsequent models. As we noted above (Table 6), extending the method itself is quite an important strategy in dealing with marketing problems, entailing the introduction of additional variables, constraints and changing basic assumptions & the inference method respectively. This reflects itself in the literature to a certain extent. Several authors also see the need for an improvement of inference algorithms, the invention of tools to more easily develop and implement topic models, and more automatization (e.g., in choosing the number of topics) (Blei et al. 2010, p. 1; p. 10; Blei 2012, pp. 82; Blei 2014, p. 25). Some of these efforts are already in the implementation phase. For example, Bart (2009, pp. 1) technically changes Gibbs Sampling to get a faster performing inference method.

## 5.3 Presenting and visualizing research results

Despite the large variety of techniques and tools available to suitably visualize the outcome of topic models, such as topic proportions and topic distributions across corpora (e.g., Chaney and Blei 2012, pp. 420; Grün and Hornik 2011, p. 16), many scholars in the field continue to call for developing new algorithms to visualize topics and present data and corpora (e.g., Chaney and Blei 2012, p. 419). Additionally, they stress the need for simplified, yet more sophisticated and interactive frameworks for scholars & practitioners (e.g., Blei 2012, p. 84; Blei 2014, p. 25; Kjellin and Liu 2016, pp. 485–460; Zinman and Fritz n.d., p. 2).

## 5.4 Shortcomings in marketing research using topic models

Researchers utilize numerous methods and measures to validate modeling outcomes. However, there are some shortcomings as well. Firstly, some authors do not use the full spectrum of evaluation techniques, making their research results somewhat

opaque. Secondly, despite indicators that point into the right direction, most of the authors fail to meet the critical problems mentioned by Schmidt (2013), both, in assessing the clustering output, and in model validation (Sect. 2). There are numerous other problems needing to be solved. As already mentioned above, Tang et al. (2014, pp. 4–8) point to some limiting factors of LDA in the form of conditions when applying the method and Crain et al. (2012, pp. 144–148) and Newman et al. (2009, p. 2) elaborate on problems of learning topics. Blei (2012) also sees the need for the development of further methods for evaluating and selecting topic models, when confronted with a particular problem domain or dataset. As we have also mentioned in Sect. 2.5., some related methods have been demonstrated to outperform LDA-type models in specific tasks, using specific data, or setups, which calls for future research to gain a more thorough understanding on the relative advantages of competing methods.

**Acknowledgements** Open access funding provided by Vienna University of Economics and Business (WU).

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (2015) Introduction to mixed membership models and methods. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) Handbook of mixed membership models and their applications. CRC Press, Florida, p 3
- Amado A, Cortez P, Rita P, Moro S (2017) Research trends on big data in marketing: a text mining and topic modeling based literature analysis. *Eur Res Manag Bus Econ* 24:1–7
- Andrzejewski D, Zhu X (2009) Latent Dirichlet Allocation with Topic-in-Set Knowledge. *SemiSupLearn'09. Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*. pp. 43–48. ISBN: 978-1-932432-38-1
- Asuncion A, Welling M, Smyth P, The YW (2009) On smoothing and inference for topic models. pp. 27–34. <https://arxiv.org/abs/1205.2662>
- Balasubramanian R, Cohe WW (2015) Block-LDA: jointly modeling entity-annotated text and entity-entity links. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) Handbook of mixed membership models and their applications. CRC Press, Florida, p 255
- Bart E (2009) Speeding up Gibbs sampling by variable grouping. *NIPS Workshop on applications for topic models: text and beyond*. <https://www.parc.com/publication/3410/speeding-up-gibbs-sampling-by-variable-grouping.html>
- Bartholomew DJ, Steele F, Moustaki I, Galbraith JI (2011) Analysis of multivariate social science data, 2nd edn. CRC Press, Taylor & Francis Group, Boca Raton, London
- Blanchard SJ, Aloise D, Desarbo WS (2017) Extracting summary piles from sorting task data. *J Mark Res PrePrint*, Unedited
- Blei DM (2012) Probabilistic topic models. *Surveying a suite of algorithms that offer a solution to managing large document archives*. *Commun ACM* 55(4):77–84
- Blei DM (2014) Build, compute, critique, repeat: data analysis with latent variable models. *Annu Rev Stat Appl* 1:203–232. <https://doi.org/10.1146/annurev-statistics-022513-115657>
- Blei DM, Jordan MI (2003a) Modeling annotated data. *SIGIR'03*. ACM 1581136463/03/0007
- Blei DM, Lafferty JD (2006) Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, PA
- Blei DM, Lafferty JD (2007) A correlated topic model of science. *Annu Appl Stat* 1(1):17–35

- Blei DM, Lafferty JD (2009) Topic models. In: Srivastava A, Sahami M (eds) Chapman and Hall/CRC. data mining and knowledge discovery series. Taylor and Francis Group, LLC, New York, p 71
- Blei DM, McAuliffe JD (2010) Supervised topic models. NIPS Proceedings
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. *J Mach Learn Res* 3(2003):993–1022
- Blei DM, Carin L, Dunson D (2010) Probabilistic topic models: a focus on graphical model design and applications to document and image analysis. *IEEE Signal Process Mag*. <https://doi.org/10.1109/msp.2010.938079>
- Boyd-Graber J, Mimno D, Newman D (2015) Care and feeding of topic models: problems, diagnostics, and improvements. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) *Handbook of mixed membership models and their applications*. CRC Press, Florida, p 225
- Brett MR (2012) Topic modeling: a basic introduction. <http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/>
- Büschken J, Allenby GM (2016) Sentence-based text analysis for customer reviews. *Market Sci* 35(6):953–975
- Büschken J, Allenby GM (2017) Improving text analysis using sentence conjunctions and punctuation. SSRN. <https://ssrn.com/abstract=2908915>. Accessed 31 Jan 2017
- Calheiros AC, Moro S, Rita P (2017) Sentiment classification of consumer-generated online reviews using topic modeling. *J Hosp Market Manag* 26(7):675–693. <https://doi.org/10.1080/19368623.2017.1310075>
- Cao D, Ji R, Lin D, Li S (2014) Visual sentiment topic model based microblog image sentiment analysis. *Multimed Tools Appl* 2016(75):8955–8968. <https://doi.org/10.1007/s11042-014-2337-z>
- Chaney AJB, Blei DM (2012) Visualizing topic models. *Assoc Adv Artif Intell*. pp 419–422
- Chang J, Blei DM (2009) Relational topic models for document networks. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 5 of JMLR
- Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM (2009) Reading tea leaves: how humans interpret topic models. <http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf>. Accessed 28 July 2017
- Chen AT, Sheble L, Eichler G (2013) Topic modeling and network visualization to explore patient experiences, pp 1–4. [http://faculty.washington.edu/atchen/pubs/Chen\\_Sheble\\_Eichler\\_VAHC2013.pdf](http://faculty.washington.edu/atchen/pubs/Chen_Sheble_Eichler_VAHC2013.pdf). Accessed 16 Jan 2017
- Cho YS, Steeg GV, Galstyan A (2015) Mixed membership blockmodels for dynamic networks with feedback. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) *Handbook of mixed membership models and their applications*. CRC Press, Florida, p 527
- Christidis K, Mentzas G (2013) A topic-based recommender system for electronic marketplace platforms. *Expert Syst Appl* 40:4370–4379
- Clement M, Boßow-This S (2007) Fuzzy clustering mit Hilfe von mixture models. In: Albers S, Klapper D, Konradt U, Walter A, Wolf J. *Methodik der empirischen Forschung*. 2., überarbeitete und erweiterte Auflage. Gabler. pp 167–182
- Costa Filho IG (2010) Mixture models for the analysis of gene expression: integration of multiple experiments and cluster validation. Dissertation. University of Berlin
- Crain SP, Zhou K, Yang SH, Zha H (2012) Dimensionality reduction and topic modeling. In: Aggarwal CC, Zhai CX (eds) *Mining text data*. Springer, Heidelberg
- Dan N, Bellio R, Reutterer T (2017) A note on latent rating regression for aspect analysis of user-generated content. Working paper. Department of Marketing WU Vienna
- Do T-M-T, Gatica-Perez D (2010) By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In: *Proceedings of the 9th international conference on mobile and ubiquitous multimedia (MUM'10)*, 1–3 Dec. ISBN: 978-1-4503-0424-5
- Fabrigar LR, Wegener DT (2012) *Exploratory factor analysis. Understanding statistics*. Oxford University Press, New York
- Flach PA (2001) On the state of the art in machine learning: a personal review. *Artif Intell* 131:199–222
- Fox EB, Jordan MI (2015) Mixed membership models for time series. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) *Handbook of mixed membership models and their applications*. CRC Press, Florida, p 417
- Frühwirth-Schnatter S (2006) *Finite mixture and Markov switching models*. Springer, New York
- Galyardt A (2015) Interpreting mixed membership. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) *Handbook of mixed membership models and their applications*. CRC Press, Florida, p 39

- Gormley IC, Murphy TB (2015) Mixed membership models for rank data. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) Handbook of mixed membership models and their applications. CRC Press, Florida, p 461
- Gormley MR, Dredze M, Van Durme B, Eisner J (2012) Shared components topic models. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp 783–792
- Gross JH, Manrique-Vallier D (2015) A mixed membership approach to political ideology. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) Handbook of mixed membership models and their applications. CRC Press, Florida, p 117
- Grün B, Hornik K (2011) Topicmodels: an R Package for fitting topic models. *J Stat Softw* 40:1–20
- Ha T, Beijnon B, Kim S, Lee S, Kim JH (2017) Examining user perceptions of smartwatch through dynamic topic modeling. *Telematics Inform* 34:1262–1273
- Han HJ, Mankad S, Gavirneni N, Verma R (2016) What guests really think of your hotel: text analytics of online customer reviews. *Cornell Hosp Rep* 16(2):3–17
- Heinrich K (2015) Integration von Topic Models und Netzwerkanalyse bei der Bestimmung des Kundenwertes. In: Wissensgemeinschaften 2015. Technische Universität Dresden. Verlag der Wissenschaften. Dresden
- Herzig J, Mass Y, Roitman H (2014) An author-reader influence model for detecting topic-based influencers in social media. *ACM 978-1-4503-2954-5/14/09. HT '14*, September 1–4
- Hinton GE (2002) Training products of experts by minimizing contrastive divergence. *GCNU TR* 2000-004
- Ho T, Do P (2015) Analyzing the changes in online community based on topic model and self-organizing map. *IJACSA* 6(7):100–108
- Ho Q, Xing EP, Airoldi EM (2015) Analyzing time-evolving networks. In: Blei DM, Airoldi EM, Erosheva EA, Fienberg SE (eds) Handbook of mixed membership models and their applications. CRC Press, Florida, p 489
- Hoffman MD, Blei DM, Bach F (2010) Online learning for latent Dirichlet allocation. *NIPS Proceedings*
- Hruschka H (2014) Linking multi-category purchases to latent activities of shoppers: analysing market baskets by topic models. *Market ZFP* 36(4):267–274
- Hruschka H (2016) Hidden variable models for market basket data. University of Regensburg, Regensburg
- Hu B, Ester M (2013) Spatial topic modeling in online social media for location recommendation. *RecSys '13*. *ACM 978-1-4503-2409-0/13/10*
- Hu Z, Wang C, Yao J, Xing E, Yin H, Cui B (2013) Community specific temporal discovery from social media. [arXiv:1312.0860v1](https://arxiv.org/abs/1312.0860v1). Accessed 3 Dec 2013
- Iqbal HR, Ashraf MA, Nawab RMA (2015) Predicting an author's demographics from text using topic modeling approach. Notebook for PAN at CLEF 2015
- Ishingaki T, Ternui N, Sato T, Allenby GM (2015) Topic modeling of market responses for large-scale transaction data. Data science and service research discussion paper. Discussion paper No. 35. Center for Data Science and Service Research Graduate School of Economic and Management. Tohoku University
- Iwata T, Sawada H (2013) Topic model for analyzing purchase data with price Information. *Data Min Knowl Disc* 2013(26):559–573. <https://doi.org/10.1007/s10618-012-0281-y>
- Jacobs BJD, Donkers B, Fok D (2016) Model-based purchase predictions for large assortments. *Market Sci* 35(3):389–404
- Jeong B, Yoon J, Lee JM (2017) Social media mining for product planning: a product opportunity mining approach based on topic modeling and sentiment analysis. *Int J Inform Manag*. <https://doi.org/10.1016/j.ijinfomgt.2017.09.009>
- Jo Y, Oh A (2011) Aspect and sentiment unification model for online review analysis. *WSDM'11*. *ACM 978-1-4503-0493-1/11/02*
- Kahn A, Baharudin B, Hong Lee L, Kahn K (2010) A review of machine learning algorithms for text-documents classification. *J Adv Inform Technol* 1(1):4–20
- Karpienko R, Reutterer T (2017) An empirical study of journal positioning and the evolution of marketing subareas. Working paper. Department of Marketing WU Vienna
- Kjellin PE, Liu Y (2016) A survey on interactivity in topic models. *IJACSA* 7(4):456–461
- Knights D, Mozer MC, Nocolov N (2009) Detecting topic drift with compound topic models. association for the advancement of artificial intelligence. [www.aaai.org](http://www.aaai.org). Accessed 18 July 2018

- Lakkaraju H, Bhattacharyya C, Merugu S, Bhattacharya I (2009) Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In: Proceedings of the 18th International Conference on World Wide Web (WWW 2009), pp 131–140
- Le D-T, Nguyen C-T, Coltech Q-TH, Phan X-H, Horiguchi S (2008) Matching and ranking with hidden topics towards online contextual advertising. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
- Lehmann DR (2005) Journal evolution and the development of marketing. *J Public Policy Market* 24(1):137–142
- Liu L, Tang J, Han J, Jiang M, Yang S (2010) Mining topic-level influence in heterogeneous networks. *CIKM'10*, October 25–29, 2010, Toronto
- Lu B, Ott M, Cardì C, Tsou BK (2011) Multi-aspect sentiment analysis with topic models. 11th IEEE International Conference on Data Mining Workshops, pp 1–8
- Luo J, Pan X, Zhu X (2015) Identifying digital traces for business marketing through topic probabilistic model. *Technology analysis and strategic management*. ISSN: 0953-7325 (Print) 1465-3990 (Online) Journal homepage. <http://www.tandfonline.com/loi/ctas20>. Accessed 24 Apr 2017
- McLachlan GJ, Peel D (2000) *Finite mixture models*. Wiley, New York, Chichester, Weinheim, Brisbane, Singapore, Toronto. ISBN 0-471-00626-2
- Melnykov V, Maitra R (2010) Finite mixture models and model-based clustering. *Stat Surv* 4:80–116
- Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 262–272
- Moghaddam S, Ester M (2012) On the design of LDA models for aspect-based opinion mining. In: Proceedings of the 21st ACM international conference on information and knowledge management (CIKM'12). ACM 978-1-4503-1156-4/12/10, pp 803–812
- Mukherjee A, Liu B (2012) Aspect extraction through semi-supervised modeling. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp 339–348
- Muthen B (1978) Contributions to factor analysis of dichotomous variables. *Psychometrika* 43(4):551–560
- Muthen B, Christofferson A (1981) Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika* 46(4):407–419
- Newman D, Karimi S, Cavedon L (2009) External evaluation of topic models. In: Proceedings of the 14th Australasian Document Computing Symposium. 4 Dec 2009
- Newman D, Lau JH, Grieser K, Baldwin T (2010) Automatic evaluation of topic coherence. *Human language technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. 100–108
- Ngyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. *Expert Syst Appl* 42(2015):9603–9611
- Nimeroff J (2017) How machine learning will be used for marketing in 2017. *Forbes technology council*. *Forbes*. <https://www.forbes.com/sites/forbestechcouncil/2017/03/10/how-machine-learning-will-be-used-for-marketing-in-2017/#74029c4e6d3d>. Accessed 8 June 2017
- Paisley J, Blei DM, Jordan MI (2015) Bayesian nonnegative matrix factorization with stochastic variational inference. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) *Handbook of mixed membership models and their applications*. CRC Press, Florida, p 205
- Parasuraman A, Grewal D, Krishnan R (2007) *Marketing research*, 2nd edn. Houghton Mifflin Company, Boston
- Park KB, Ha SH (2016) Mining user-generated contents to detect service failures with topic model. *Int J Comput Electr Autom Control Inform Eng* 10(8):1491–1496
- Pathak N, DeLong C, Banerjee A (2008) Social topic models for community extraction. The 2nd SNA-KDD Workshop'08 (SNA-KDD'08). ACM 978-1-59593-848-0
- Paul M, Girju R (2009) Cross-cultural analysis of blogs and forums with mixed-collection topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp 1408–1417, Singapore, 6–7 August 2009. c 2009 ACL and AFNLP
- Phuong DV, Phuong TM (2012) A keyword-topic model for contextual advertising. *SoICT*, pp 63–70. ACM 978-1-4503-1232-5/12/08. <https://doi.org/10.1145/2350716.2350728>
- Proctor T (2005) *Essentials of marketing research*, 4th edn. Pearson Education Limited, Harlow
- Punj G, Stewart DW (1983) Cluster analysis in marketing research: review and suggestions for application. *J Mark Res* 20(2):134–148

- Putri IR, Kusumaningrum R (2017) Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia. *J Phys*. <https://doi.org/10.1088/1742-6596/801/1/012073>
- Rabinovich M, Blei DM (2014) The inverse regression topic model. In: Proceedings of the 31st International Conference on Machine Learning. *JMLR: W&CP*. Vol. 32
- Rahman MdM, Wang H (2016) Hidden topic sentiment model. *WWW 2016*, pp 155–165. *ACM 978-1-4503-4143-1/16/04*
- Ramage D, Dumais S, Liebling D (2010) Characterizing microblogs with topic models. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media
- Reutterer T (2003) Bestandsaufnahme und aktuelle Entwicklungen bei der Segmentierungsanalyse von Produktmärkten. *J für Betriebswirtschaft* 53(2):52–74
- Roberts ME, Stewart BM, Tingley D (2015) Navigating the local modes of big data: the case of topic models. Draft. June 2015. Prepared for “Computational social science: discovery and prediction”, pp 1–55. <https://scholar.harvard.edu/files/dtingley/files/multimod.pdf>. Accessed 3 Jan 2018
- Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: *UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence*. p 487–494
- Salakhutdinov R, Hinton G (2009) Replicated softmax: an undirected topic model. *Advances in neural information processing systems* 22. *NIPS Proceedings 2009*
- Sammut C, Webb GI (2011) *Encyclopedia of machine learning*. Springer, New York
- Schieber A, Hilbert A, Sommer S, Heinrich K (2011) Analyzing customer sentiments in microblogs—a topicmodel-based approach for Twitter datasets. In: Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, Michigan August 4th–7th
- Schmidt BM (2013) Words alone: dismantling topic models in the humanities. *J Digit Humanit* 2(1). <http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/>. Accessed 31 May 2017
- Schröder N (2017) Using multidimensional item response theory models to explain multi-category purchases. *Market ZFP* 39(2):27–37. <https://doi.org/10.15358/0344-1369-2017-2-27>
- Schröder N, Falke A, Hruschka H, Reutterer T (2017) Analyzing browsing and purchasing across multiple websites based on latent Dirichlet allocation. *ALLDATA 2017*. ISBN: 978-1-61208-552-4
- Shafiei MM, Milios EE (2006) Latent Dirichlet co-clustering. In: Proceedings of the Sixth International Conference on Data Mining (ICDM'06). 0-7695-2701-9/06. pp 1–10
- Shaw MJ, Subramaniam C, Tan GW, Welge ME (2001) Knowledge management and data mining for marketing. *Decis Support Syst* 31:127–137
- Shringarpure S, Xing EP (2015) Population stratification with mixed membership models. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) *Handbook of mixed membership models and their applications*. CRC Press, Florida, p 397
- Si J, Li Q, Qian T, Deng X (2014) Users' interest grouping from online reviews based on topic frequency and order. *WorldWideWeb 2014(17)*:1321–1342. <https://doi.org/10.1007/s11280-013-0239-z>
- Song L, Lau RYK, Kwok RCW, Mirkovski K, Dou W (2017) Who are the spoilers in social media marketing? Incremental learning of latent semantics for social spam detection. *Electron Commer Res* 17:51–81
- Sun Y, Deng H, Han J (2012) Probabilistic models for text mining. In: Aggarwal CC, Zhai CX (eds) *Mining text data*. Springer, Heidelberg, pp 260–296
- Sun F-T, Griss M, Mengshoel O, Yeh Y-T (2013) Latent topic analysis for predicting group purchasing behavior on the social web. [http://repository.cmu.edu/cgi/viewcontent.cgi?article=1157&context=silicon\\_valley](http://repository.cmu.edu/cgi/viewcontent.cgi?article=1157&context=silicon_valley). Accessed 18 Apr 2017
- Sweet TM, Thomas AC, Junker BW (2015) Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) *Handbook of mixed membership models and their applications*. CRC Press, Florida, p 463
- Tang J, Meng Z, Nguyen XL, Mei Q, Zhang M (2014) Understanding the limiting factors of topic modeling via posterior contraction analysis. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. *JMLR: W&CP* vol. 32
- Teh YW, Jordan MI (2010) Hierarchical Bayesian nonparametric models with applications. In: Hjort NL, Holmes C, Müller P, Walker SG (eds) *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, pp 158–207
- Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. *J Am Stat Assoc* 101(476):1566–1581. <https://doi.org/10.1198/016214506000000302>

- Tirullinai S, Tellis GJ (2014) Mining marketing meaning from online chatter: strategic brand analysis of big data using latent Dirichlet allocation. *J Market Res* 51(4):463–479
- Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. WWW 2008. IW3C2. ACM 978-1-60558-085-2/08/04
- Titterton D, Smith A, Makov U (1985) Statistical analysis of finite mixture distributions. Wiley. ISBN 0-471-90763-4
- Tran T, Ho T, Do P (2015) Detecting communities and surveying the most influence of online users. *ACSIJ* 4(6):172–178 (ISSN: **2322-5157**)
- Trusov M, MA L, Jamal Z (2016) Crumbs of the cookie: user profiling in customer-base analysis and behavioral targeting. *Market Sci* 35(3):405–426
- Underwood T (2012) Topic modeling made just simple enough. <https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/>
- Wallach HM, Mimno D, McCallum A (2009) Rethinking LDA: Why priors matter. <http://dirichlet.net/pdf/wallach09rethinking.pdf>. Accessed 20 July 2017
- Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'11). ACM 978-1-4503-0813-7/11/08, pp 448–456
- Wang H, Lu Y, Zhai CX (2011) Latent aspect rating analysis without aspect keyword supervision. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'11). ACM 978-1-4503-0813-7/11/08, pp 618–626
- Wang Y, Agichtein E, Benzi M (2012) TM-LDA: efficient online modeling of latent topic transitions in social media. KDD'12, August 12–16, 2012, Beijing, China. Copyright 2012 ACM 978-1-4503-1462-6/12/08
- Wang Z, Li L, Zahng C, Huang Q (2015) Image-regulated graph topic model for cross-media topic detection. *ICIMCS'15*, August 19–21, 2015, Zhangjiajie, Hunan, China
- Wedel M (2002) Concomitant variables in finite mixture models. *Stat Neerl* 56(3):362–375
- Wedel M, Kamakura WA (1999) Market segmentation, vol 2. Springer Science + Business Media, New York
- Welling M, Hinton G, Osindero S (2002) Learning sparse topographic representations with products of student  $t$  distributions. *Advances in neural information processing systems*. Vol. 15. Vancouver, Canada
- Weng J, Lim E-P, Jiang J, He Q (2010) TwitterRank: finding topic-sensitive influential twitterers. *WSDM'10*, February 4–6, 2010, New York City, New York, USA. Copyright 2010 ACM 978-1-60558-889-6/10/02, pp 261–270
- Wood F, Perotte A (2015) Mixed membership classification for documents with hierarchically structured labels. In: Airoldi EM, Blei DM, Erosheva EA, Fienberg SE (eds) *Handbook of mixed membership models and their applications*. CRC Press, Florida, p 305
- Xie Y, Gao Y, Gou J, Cheng Y, Honbo D, Zhang K, Agrawal A, Choudhary A (2012) Probabilistic macro behavioral targeting. *DUBMMMS'12*, October 29, 2012, Maui, Hawaii, USA. ACM 978-1-4503-1707-8/12/10, pp 7–10
- Yang Z, Kotov A, Mohan A, Lu S (2015) Parametric and non-parametric user-aware sentiment topic models. *SIGIR'15*, August 09–13, 2015, Santiago, Chile. ACM 978-1-4503-3621-5/15/08, pp 413–422
- Yin Z, Cao L, Han J, Zhai C, Huang T (2011) Geographical topic discovery and comparison. WWW 2011 — session: spatio-temporal analysis. March 28–April 1, 2011, Hyderabad, India. ACM 978-1-4503-0632-4/11/03, pp 247–256
- Zinman A, Fritz D (n.d.). Data portraiture and topic models. pp 1–5. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.4544>. Accessed 31 Aug 2017