Spring 6-10-2017

TOPIC MODELLING METHODOLOGY: ITS USE IN INFORMATION SYSTEMS AND OTHER MANAGERIAL DISCIPLINES

Matthias Eickhoff
University of Göttingen, Germany, meickho@uni-goettingen.de

Nicole Neuss
University of Göttingen, Germany, n.neuss@stud.uni-goettingen.de

TOPIC MODELLING METHODOLOGY: ITS USE IN INFORMATION SYSTEMS AND OTHER MANAGERIAL DISCIPLINES

Research paper

Eickhoff, Matthias, University of Göttingen, Germany, meickho@uni-goettingen.de

Neuss, Nicole, University of Göttingen, Germany, n.neuss@stud.uni-goettingen.de

Abstract

Over the last decade, quantitative text mining approaches to content analysis have gained increasing traction within information systems research, and related fields, such as business administration. Recently, topic models, which are supposed to provide their user with an overview of themes being discussed in documents, have gained popularity. However, while convenient tools for the creation of this model class exist, the evaluation of topic models poses significant challenges to their users. In this research, we investigate how questions of model validity and trustworthiness of presented analyses are addressed across disciplines. We accomplish this by providing a structured review of methodological approaches across the Financial Times 50 journal ranking. We identify 59 methodological research papers, 24 implementations of topic models, as well as 33 research papers using topic models in Information Systems (IS) research, and 29 papers using such models in other managerial disciplines. Results indicate a need for model implementations usable by a wider audience, as well as the need for more implementations of model validation techniques, and the need for a discussion about the theoretical foundations of topic modelling based research.

Keywords: Topic Modelling, Literature Review, Model Validation

1 Introduction

The rise of social media platforms and the availability of news online have created textual "big data", which has outgrown the feasibility of in-depth qualitative analysis. Quantitative methods to the analysis of textual data, such as sentiment analysis (Liu, 2012), have consequently become an established tool in the methodological spectrum of information systems research. Recent developments, such as efforts towards a "web of data", will only increase the need for an automated analysis of textual content (W3C, 2013). Among the approaches to analyzing large document collections, topic models, such as Latent Dirichlet Allocation (Blei et al., 2003), have recently gained traction in applied (non-methodological) research. Debortoli et al. (2016) provide a tutorial for using topic modelling as a tool in information systems research and provide readers with an example analysis showcasing the use of this model class. The recent focus on topic modelling as a quantitative research method has enabled researchers to address questions that previously would have been considered out of reach. As noted by Rai (2016), evaluation strategies for topic modelling include the reference to expert opinion, as well as quantitative approaches, such as the comparison of models estimated using varied parameters. However, modelling the contents of document collections is a challenging task and remains an area of active research in natural language processing and computer science literature. The "unreasonable effectiveness" (Halevy et al., 2009) of current models representing large document collections continues to be a challenge regarding the question on how to use these models in social-sciences and information systems research. In research concerned with testing hypothesis on the basis of theory, it is of critical importance to be able to convince readers that the models actually represent large document collections accurately, in order to establish the trustworthiness of conclusions based upon the models

(Lincoln and Guba, 1985). In this paper, we investigate how researchers across different disciplines deal with this problem by conducting a structured review of literature in the top outlets of business related literature, on the basis of which we categorize different strategies to address this challenge. The paper is structured as follows: In section 2, the concept of topic modelling is introduced before a brief introduction to the relation between topic modelling methodology and (meta) theoretical considerations is given, based on which we discuss some of the results of the review. In section 3, the research design of this study is developed and presented. Section 4 presents and discusses the results of the review, while section 5 summarized this research.

2 Topic Models

The aim of topic modelling is to determine structures in underlying document collections. Initially, topic models were developed as an information retrieval tool, intended to make browsing large document collections easier (Salton et al., 1975). In example, topic models can be used to browse collections of scientific journals according to the subject of articles, without relying on metadata (Blei and Lafferty, 2009a). The first widely used model in this class was Latent Semantic Indexing (LSA), which as this review shows is still a popular option (Croft and Harper, 1979; Dumais et al., 1988). LSA extracts the underlying topics from a term-document matrix by applying singular value decomposition (SVD), which results in mathematically orthogonal topics. While this assumption of orthogonality contradicts human intuition about topics, topic models are essentially a data compression technique and this approach leads to the maximization of topic variance on a compressed representation of the document collection like how principal component analysis (PCA) does when used to reduce the number of features in a regression problem, which many researchers may be more familiar with. This assumption of topics' mutual exclusiveness is softened by probabilistic LSA (pLSA), which models topics as word distributions (Hofmann, 1999), leading to a notion of topics more in line with human intuition. After all, we would not assume most topics to be completely distinct from one another. This model type is extended upon by Latent Dirichlet Allocation (LDA), which differs from pLSA by imposing Dirichlet distributed priors to its word to topic and topic to document distributions (Blei et al., 2003).

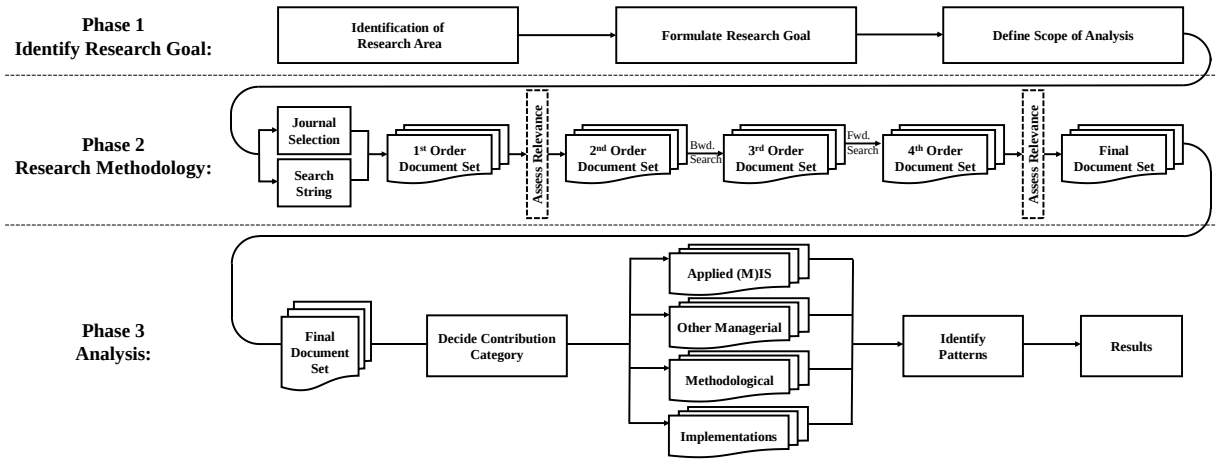

Figure 1: Research design segmented in three phases following Ngai et al. (2011). The first phase identifies the scope and goal of the presented research. Phase 2 describes the methodology of the conducted literature review. Phase 3 gives an overview of the analysis conducted based on this literature review.

Again, this is more in line with the human notion of topics, as it leads to sparse topic assignments to documents due to the sparse nature of the Dirichlet distribution. What this means is that not each document is a mixture of all topics in a model, but few topics are disproportionally more important for a document than others. A more detailed description of the intuition underlying this model type can be found in Blei (2012). This increased resemblance of human intuition does not necessarily mean that the more modern approaches outperform their precursors. In example, Bergamaschi and Po (2014) implement a plot-based movie recommender system and find that LSA outperforms LDA in their example. Since these methods became available, they have also been applied in (M)IS research. In recent years, this use has also arrived in top-tier outlets within the discipline (Kulkarni et al., 2014; Larsen and Bing, 2016; Sidorova et al., 2008). Superficially, it seems that M(IS) is not making use of the newer model types and their advantages regarding their closeness to human intuition, but determining whether this is the case requires a closer look and is one of the objectives of the following review.

2.1 Meta theoretical Foundations of Topic Modelling Research

When analyzing approaches to automated text analysis, such as topic mining, taking a step back to look at the (meta) theoretical foundation of such approaches can help in the analysis of their use. Ignatow (2015) provides an overview of the theoretical foundations of digital text analysis and argues that the meta theoretical foundations of such methods have not been sufficiently established and that applied social research using them often lacks adequate theories of language supporting the use of the method. According to Ignatow (2015), this lack of theoretical foundation stems from the unique positioning of automated text analysis between the natural and social sciences and weakens its relative positioning in comparison to exegetical methods and other inductive qualitative approaches. In principle, there are three possible meta theoretical foundations of text mining research resulting in three types of research designs. First, realist designs use models of text in a positivist framework to develop testable theories. See Elder-Vass (2014) for an extensive discussion of different variants of this approach. Second, constructivist designs use models of text to augment exegetical methods for qualitative text analysis, such as Grounded Theory (Lai and To, 2015). Third, mixed methods research designs (Venkatesh et al., 2016). Such studies often have comparably rigorous meta theoretical underpinnings because they are not conducted within the "safety" of either positivist or constructivist reference frames. While the IS, like many disciplines, traditionally focused on research designs build upon positivist mindsets, recently both qualitative (Bagozzi, 2011; Gregory, 1993; Mingers, 1995) and mixed methods research designs have become a common sight in the discipline (Ågerfalk, 2013; Venkatesh et al., 2013).

3 Research Design

As described, the goal of this analysis is to provide insight into the available methods for topic mining, and how these methods are applied both in (M)IS and other managerial disciplines. Ngai et al. (2011) present a similar analysis for the applications of data mining techniques within the domain of financial fraud detection and structure their review into three distinct research design phases, which represent a suitable research design for the case at hand. Thus, a comparable three-stage design is chosen for this study. In the first stage, the research goal is defined and the analysis is scoped. In the second phase, the research-methodology is outlined. In the third phase, we describe how the study is conducted on this basis. Figure 1 shows this process.

3.1 Phase 1: Identify a Research Goal

By determining the area of research, formulating the goal of the study, and defining the scope of the research, the studies relation to the wider research landscape is determined. In this research, the area of research is given by the search for available methodology for the training and evaluation of topic models, as well as their application in (M)IS and other managerial sciences. The goal of this study is to identify methodological opportunities for future studies and to examine how prior research has used

Category Name	Description
Methodological Foundations	A methodological contribution towards topic modelling, either a new topic modelling approach, a task specific document pre-processing logic, or an evaluation method for topic models, which is sufficiently different from other approaches included in the review.
Implementation (DSR)	An implementation of any of the above, which is made available to the public in a usable state, which means the software should be available and working.
Applied (M)IS (Empirical)	Applied research papers using topic modelling, or methodological considerations regarding topic modelling, within the IS community.
Applied Non-IS (Empirical)	Applied research papers using topic modelling, or methodological considerations regarding topic modelling, within studies from management related fields.

Table 1: Relevance criteria for literature discovered during the literature search, structured into four relevance categories.

the available methodology. To strengthen the focus of the analysis, this goal is formalized to the following three research questions:

RQ1 (Methodological pervasiveness): How widespread is the usage of topic models in the management literature and for what purposes are these models used therein?

RQ2 (Validation methodology): How do researchers address the problem of establishing trust into the results of their analysis when using topic models to analyze large document collections?

RQ3 (Interdisciplinary differences): How does the usage of topic models differ between M(IS) and other managerial disciplines?

The scope of the review is presented by an initial search within all journals included in the Financial Times 50 (FT50) ranking, which represents major outlets across numerous management-related fields. Further outlets are accepted into the study if they are deemed relevant regarding the aims of the study and are discovered by the structured literature review described in the next section. In order to formalize this relevance criterion, the following relevance definition is used throughout this research: Research is considered relevant for the scope of this study, if it falls into one of the four categories outlined in Table 1.

3.2 Phase 2: Research Methodology

Our goal in this phase is to arrive at a formalized abstraction of the conducted research process. This serves two purposes. First, the resulting design helps when conducting the study by splitting the research process into individual work units. Second, it helps readers to assess the quality and rigor of a study by providing a clear indication on how the study was conducted. In the case at hand, the first task during this phase is given by the identification of a suitable approach to the identification of literature using the relevance criterion stated in phase 1, resulting in the question what ways of literature exploration have been identified by methodological literature regarding literature reviews. Due to the continuous growth of the IS discipline, and the need for junior researchers to gain an overview of extant research, as well as the increasing difficulty to remain knowledgeable for senior researchers (Templier and Paré, 2015), a growing body of work regarding the methodology of literature reviews has evolved. Webster and Watson (2002) may be considered the starting point of this methodological discussion within IS. Since, Greenhalgh et al. (2005), Sylvester et al. (2013), Rowe (2014), and Boell and Cecez-Kecmanovic (2015) are only a small sample of this diverse toolset of methodological approaches towards literature based research.

Webster and Watson (2002) are perhaps the most notable example of guidelines to performing a structured literature search in the IS literature. They propose to divide the search for literature into three

steps. Figure 1 (phase 2) provides an overview of this approach. First, a set of outlets is identified in which to search for relevant articles. Second, the references of these articles are examined to identify prior work (backward search). Third, the results of the two prior search phases are used to perform a search for articles citing them (forward search). As noted above, the relevant outlets for the first phase have already been identified as the journals included in the FT50. To search for relevant articles in the online databases listing the journals, a search string needs to be determined, which covers a broad spectrum of work related to topic modelling. The following string is used and was determined by iterating between including more search terms and removing those, which produce non-relevant results:

"Topic Mining" OR "Topic Model*" OR "Topic Distribution" OR "Hierarchical Dirichlet Process" OR "Multinomial Asymmetric Hierarchical Analysis" OR "Latent Dirichlet Allocation" OR "Latent Semantic Indexing" OR "Latent Semantic Analysis" OR Mallet OR Gensim.

As shown, the string contains several relevant variants of "Topic*", where the star denotes the appropriate *any* search wildcard for each database. Furthermore, different topic modelling techniques are included, as well as MALLET (McCallum, 2002) and Gensim (Rehurek and Sojka, 2010), which represent two popular implementations of topic models. These two are included because, as opposed to most other topic modelling software, they do not include topic modelling in their name. As the result of the literature search showed, most papers can be identified using either "Topic Mining" or "Topic Model*". The search string is used to search for titles, abstracts, keywords, as well as the full text of papers. Initially, a longer search string was used, which also included abbreviations where applicable, such as LDA in addition to Latent Dirichlet Allocation, however the results of searches including the abbreviated terms do not provide more relevant results and instead clutter the search results with other meanings for the abbreviations, which are not related to topic modelling. Regarding the reviews' scope in time, no assumptions were made during the initial search, but results indicate that no relevant content exists before 1978. Of course, arguably, text pre-processing literature precedes this year but this literature is not specific to topic modelling as a research method. The database search using this search string resulted in 108 results (1st order document set, Figure 1 phase 2). These documents were consequently assessed using the criteria outlined in Table 1, resulting in 23 2nd order documents. On this basis, the backward search resulted in 86 additional papers, increasing the 3rd order document set to 109 candidates. The forward search added another 44 papers, resulting in 153 documents. At this state, due to the large number of documents in the analysis, we conducted another relevance check and 8 documents were removed. The remaining 145 documents were assigned to the four relevance categories outlined in Table 1, resulting in 33 "Applied IS" papers, 29 "Applied Non-IS" papers, along with 24 implementations and 59 methodological contributions. Figure 1 (phase 3) provides an overview of this categorization into methodological research, implementations, and applied research papers stemming from M(IS) or other managerial disciplines. The analytic part of this research is based on this final set of documents.

3.3 Phase 3: Analysis

Methodological work: First, the methodological works are reviewed, to arrive at an overview of the available methodology, which can be used by applied studies. To this end, the main methodological contribution of each paper is identified by examining each paper in the sample and summarizing its main contribution. Based on the sum of these contributions, the typology shown in Figure 2 is developed, which considers six archetypes of contribution. It should be noted that this is not a formal typology or taxonomy, in which the characteristics of each paper would be mutually exclusive from one another (Nickerson et al., 2013). Of course, a paper can contribute in more than one way regarding these categories. For methodological papers, a *model type* contribution is given by the introduction of a new topic model, which may be done by using an entirely new approach (Blei et al., 2003), augmenting existing approaches (Blei and Lafferty, 2006), or changing what is being modelled (Chang and Chien, 2009). *Computational or mathematical* works included in the sample concern algorithms or data structures of special importance to topic modelling.

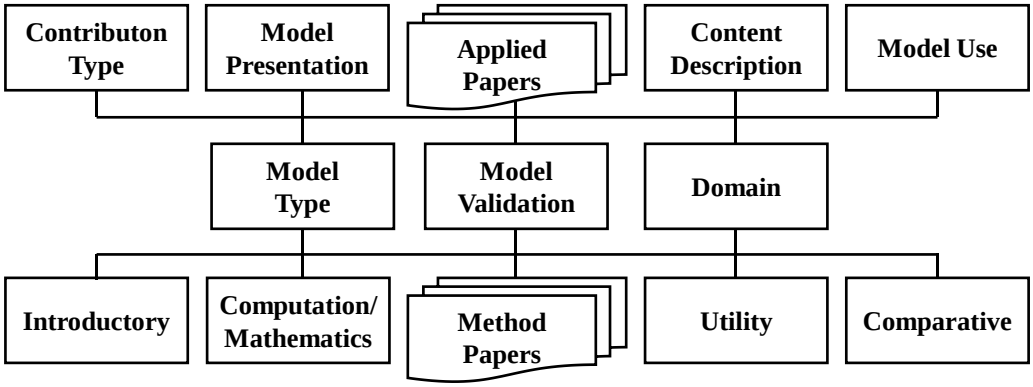

Figure 2: Assessment categories for methodological and applied contributions. The set of applied categories is identical for IS and other managerial literature. Methodological contributions are assessed using a separate but partially overlapping set of categories.

Purely statistical work, such as numeric optimization techniques, are not included in this review because, typically, they are not inherently interesting for applied work (although, of course, they are of prime importance to enable it). *Introductory* papers provide a methodological overview, or give recommendations for the use of a specific model type. Often, these papers are domain specific. *Comparative* papers assess multiple model types, to determine which model is most suitable for a specific domain or document type. *Utility* papers help researchers by providing guidance apart from modelling itself. In example, Mei et al. (2007) develop a method for automated topic-label generation. *Model validation* is concerned with assessing topic model quality and often develops or compares metrics for this task. Finally, a research *domain* is identified for each paper. The list of discovered methodological papers is presented in Appendix A.

Implementations: If methodological papers make their method publicly available or applied papers mention the used implementation, this is added as a separate reference category. Each implementation is checked regarding its public availability. An implementation is considered public if it is usable (license not considered) and a download is available. Also, since many implementations are software libraries and not stand-alone programs, the programming language used for each implementation is noted. The list of discovered implementations is presented in Table 2.

Applied work: Likewise, applied papers identified as relevant to the review are assessed. This is done by a two-pass procedure. First, like the treatment of methodological papers, each applied paper is assessed regarding its *main contribution* (not in figure)**.** These main contributions are consequently divided into *contribution types*. Second, the papers are assessed regarding their use of topic models. Again, the *model type* used in a paper is determined. *Model use* describes to what end the topic model is used in the paper. For example, if it is used to inspect topics or if the topics are used in a regression model. Consequently, *model validation* (how the model is validated) and *model presentation* (how the model is presented to readers) are derived by examining each paper in more detail. Finally, a *domain* is coded for each applied paper. As shown, the model type, model validation, and domain criteria are shared between the two paper categories, while the other criteria are distinct for each paper type. The list of discovered applied papers is presented in Appendices B and C.

4 Results and Discussion

The answer to RQ1 is presented by Figure 3, which shows the annual paper counts across the different review categories and the distribution of all papers in appendices A-C over their respective disciplines. As shown, while in earlier years most discovered contributions are methodological, more recently this relation has inversed and topic models are being used in applied studies more frequently. It is

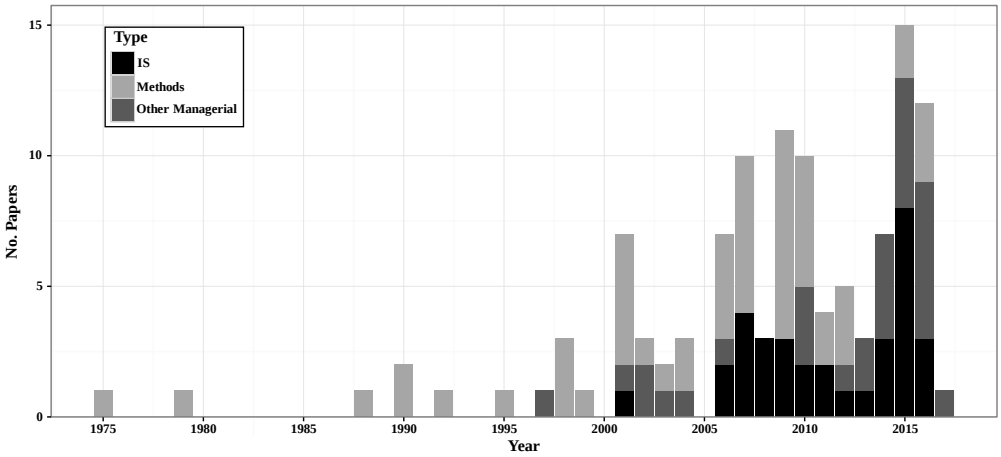

Figure 3: Annual distribution of contributions grouped by applied (M)IS and other applied managerial contributions, as well as methodological contributions. As shown, in recent years applied contributions have begun to outpace methodological works.

important to keep in mind that recent methodological work is less likely to be discovered by the backward- and forward search mechanism applied in this review. Nonetheless, this still shows the growing relevance of topic modelling methodology for both (M)IS and other managerial disciplines. **Methods (Appendix A):** As shown in the result tables and Figure 3, topic modelling methodology has become a vibrant research subject. Starting with LSA (Dumais et al., 1988) and LDA (Blei et al., 2003), which represent the two most common model types used in applied papers, 25 model types are identified in this review, many of which focus on extensions of the two archetypes. Notably, despite beginning the literature review with the FT50 journal selection as a starting point, many methodological contributions stemming from the IS domain were identified. Thus, IS seems to have been established as a reference discipline for researchers looking for methodological guidance in the use of topic models. As can be expected, computer science (CS) and statistics present the two other most important methodological disciplines for topic modelling. **Implementations (Table 2):** As noted, implementations of the methodological contributions included in the review are reported separately if they are available, to highlight methods readily available for use. Unfortunately, while most methodological contributions explain the statistical approach to their work, publicly available implementations remain the exception. BleiLab (2016) provide an example of enabling others to benefit from methodological work and release implementations and working examples when possible. When comparing, the citation counts of papers with released implementations to those without, it is obvious that this approach pays off. Accessibility remains a major problem (Ramage et al., 2009). All publicly available implementations identified in this review are either programming libraries or command line applications. In the interest of making topic models as usable as other statistical techniques, implementations with graphical user interfaces are needed. **Applied research papers:** Table 3 provides a condensed overview of the results regarding applied research contributions, and contrasts M(IS) with other managerial disciplines. Appendices B and C report the full results for these categories in more detail, such as brief descriptions for each contribution. As shown, papers applying topic modelling are dominated by the two 'basic' model types LDA and LSA. When comparing IS to other domains, LSA is favored over LDA, while this relation is reversed elsewhere. Also, IS research included twice as many discussion and review articles when compared to other managerial disciplines, with 30% of IS articles being reviews and 27% being discussion pieces. Also, while 69% of non-IS articles use topic models for content analysis, only 12% of IS articles do so. In IS research, 36% percent of papers actively validate a model, while 59% of non-IS articles do so. Thus, regarding RQ2, we find that while many applied research papers make use of the topic model validation techniques proposed by methodological contributions, many researchers who use the topic model as a part of regression models abstain from a dedicated validation of the topic model.

Citation	Title	Comment	URL
(Miller and Fellbaum, 1998)	Wordnet	Digital dictionary for word relations	https://wordnet.princeton.edu
(McCallum, 2002)	MALLET	Java, several topic models	http://mallet.cs.umass.edu
(Blei et al., 2003)	LDA-C	C, Blei et al. (2003)	https://github.com/blei-lab/lda-c
(Blei and Lafferty, 2007)	Correlated topic model	C, correlated topic model (CTM)	https://github.com/blei-lab/ctm-c
(Blei and Lafferty, 2009b)	Turbotopics	Python, multiword phrases in topics	https://github.com/blei-lab/turbotopics
(Chong et al., 2009)	Supervised LDA for classification	C++	https://github.com/blei-lab/class-slda
(Rehurek and Sojka, 2010)	Gensim	Python, several model types, flexible	https://radimrehurek.com/gensim
(Gerrish and Blei, 2010)	Dynamic and Influence Topic Model	Command line implementation	https://github.com/blei-lab/dtm
(Hoffman et al., 2010)	Online var. Bayes for LDA	Python	https://github.com/blei-lab/onlineldavb
(Wang and Blei, 2010)	Hierarchical Dirichlet Process	C++	https://github.com/blei-lab/hdp
(Crossno et al., 2011)	TopicView	-	-
(Grün and Hornik, 2011)	Topicmodels (R-Package)	R implementation of LDA	https://cran.r-project.org/web/packages/topicmodels/index.html
(Ramage and Rosen, 2011)	Stanford topic modeling toolbox	Not maintained anymore	http://nlp.stanford.edu/software/tmt/tmt-0.4
(Wang and Blei, 2011)	Collaborative modeling	C++	https://github.com/blei-lab/ctr
(Wang, 2011)	Online Hierarchical Dirichlet Process	Python	https://github.com/blei-lab/online-hdp
(Zhai et al., 2012)	Mr. LDA	-	https://github.com/lintool/Mr.LDA
(Roberts et al., 2014)	R, stm: structural topic models	-	https://cran.r-project.org/web/packages/stm/index.html
(Sievert and Shirley, 2014)	LDAvis	R Package for Visualization	https://github.com/cpsievert/LDAvis
(Chaney, 2014)	Online Topic Model Visualization	Python, browsing topics	https://github.com/blei-lab/tmv
(Gopalan et al., 2014)	COLLABTM	Nonnegative Collaborative Modeling	https://github.com/blei-lab/collabtm
(Blei, 2014)	Hierarchical latent Dirichlet allocation	C,Hier. LDA,fixed depth tree and a stick breaking prior on the depth weights	https://github.com/blei-lab/hlda
(Günther et al., 2015)	LSAfun	R Package for LSA	https://cran.r-project.org/web/packages/LSAfun/index.html
(Charlin et al., 2015)	Dynamic Poisson factorization (dPF)	Command line implementation	https://github.com/blei-lab/DynamicPoissonFactorization
(Ranganath et al., 2015)	Deep Exponential Family	Command line implementation	https://github.com/blei-lab/deep-exponential-families
(BleiLab, 2016)	Blei Group Implementations	David Blei Github repository, many implementations (see this table).	https://github.com/blei-lab

Table 2: Implementations identified by the literature review. If available, the citation of the methodological research paper is provided. If no such paper could be identified, a web reference is provided pointing to the implementation itself.

One reason for the omission of a dedicated validation may be presented by the argument that if a topic model produces topics which are useful as variables in the context of statistical analysis, this itself validates the model for the purposes of these studies. However, the presentation of the topic model in such cases should be especially careful, to establish trustworthiness of the presented analysis. However, the lack of implementations of such model types in software intended for the use by social scientists remains a major hurdle for such work. 53% of all applied articles stem from the IS domain, followed by accounting research with 10%, while general management and marketing are tied at 8% each. 39% of all applied papers use their topic model as a tool for content analysis. Models are mostly used as a variable augmenting existing regression models, or to gain a general sense of topics included in text collections.

Model Type	Word Collocation	Naïve Bayes	Hierarchical	LDA	LSA	LSA&LDA	CTM	SOM	SVD	Clusters	None
IS	3%	3%	0%	18%	45%	0%	3%	0%	3%	3%	21%
Other	0%	3%	3%	41%	24%	3%	7%	7%	0%	0%	10%

Paper Type	DSS	Information Retrieval	Review	Statistical	Text Similarity	Tool	Tutorial	Validation	Discussion	Content Analysis
IS	3%	12%	30%	0%	3%	3%	3%	6%	27%	12%
Other	0%	0%	14%	3%	0%	0%	10%	0%	3%	69%

Validation	Yes	No
IS	36%	63%
Other	59%	41%

Table 3: Overview of applied research contributions in information systems compared with other managerial domains. Note that for the purposes of this summary table, less granular categories are reported than in the detailed tables.

This type of article is much more common in other managerial disciplines, but IS research using the methodology to this end still exists. The second most common applied paper type is presented by review articles, which review research domains (Moqri et al., 2015; Sidorova and Isik, 2010) or journals (Cohen Priva and Austerweil, 2015; Wang et al., 2015). Overall, the review indicates a diverse research landscape using topic modelling as a content analysis tool, as well as many research papers using it to review entire disciplines or outlets. Regarding their use of topic models some of the discovered contributions distinguish themselves and can serve as examples providing interesting ways to describe the usage of a model in an applied paper or integrating topic models in an analysis in another interesting way, which sets them apart from papers that 'end' after a topic model has been estimated.

First, Bao and Datta (2014), who investigate risk types in corporate risk disclosures, highlight topic models' capability to simultaneously discover and quantify categories in a document collection, coupled with an extensive model evaluation, which enables readers to assess the reliability of the presented approach. Their evaluation includes both quantitative measures for model fit, comparisons to alternative topic models, as well as presentations of the chosen approach using graphs and word-clouds. Second, Paul and Girju (2009), who compare research domains using topic models, show how topic similarity between different models can be used to compare different document collections, and support their arguments using a mix of reporting the words included in their estimated topics and graphs showing the evolution of topic similarity over time. As these examples show, how a model is displayed in a contribution is crucial to establishing trust in presented results. Going one step beyond the idea of presentation, Ramage et al. (2009) argue that readers should be able to explore models for themselves. Mützel (2015), a sociologist, discusses the lack of student method training in topic modelling, and data processing in general, as another challenge hindering the integration of the method in non-technical domains but notes that non-technical fields can draw on a vast experience regarding the study of meaning, which can support the automated analysis of large data sets, raising the question of the **theoretical foundations of topic modelling**: Using topic models for the purposes of advancing theory has been one of the uses of this model type early on (Landauer and Dumais, 1997) but remains the exception when surveying the applied literature. As discussed, accessibility may present one major cause for this. Another is given by the lack of theoretical foundations of topic modelling, which makes it more challenging to establish trust in results based upon its use. While few studies explicitly state their (meta) theoretical foundations, for studies classified as *content analysis* a positivist underpinning aiming at the empirical validation of established theory is often implicitly clear. On the other hand, constructivist foundations or mixed methods approaches to the analysis of topic models remain largely unexplored. However, similarities and differences between topic modelling and human coding have been discussed (Quinn et al., 2010). Also, many studies use topic labels coded from the top words of topics as a tool to present their results. Yet, this is usually done for presentation only and not using qualitative methodology, which may be suitable for this purpose. Since qualitative researchers have developed rigorous coding techniques, this methodology can support quantitative topic modelling creating opportunities for collaboration. Thus, the combination of qualitative methodology and topic modelling remains an interesting opportunity for future research. Evans and Aceves (2016) survey text mining methodology and provide recommendations on how it can be used as a tool for theory generation in the social theory. Wagner-Pacifici et al. (2015) discuss similar issues with a focus on using big data to access knowledge about social phenomena. Ignatow (2015) remains the only article discovered in the review in which the theoretical foundations of topic modelling are discussed. However, these articles do not discuss topic models in particular. As shown, (M)IS has established itself as a reference discipline for other managerial fields regarding topic modelling methodology. The exploration of the theoretical foundations of the use and interpretation of topic models, as well as their capabilities regarding the generation and testing of social-, economic- and systems theory present an opportunity to strengthen this referential role of IS.

5 Conclusion

In this review, we surveyed the topic modelling literature regarding the methodological possibilities and uses thereof in applied research papers. To this end, we formalize our research design and conduct a structured literature review, resulting in a sample of topic modelling methodology and applied research papers in IS and other managerial disciplines. Also, we provide an overview of available implementations of topic modelling approaches. Our results indicate that while, in recent years, topic modelling has become a tool used across many disciplines and is especially prevalent in IS research, most researchers use "vanilla" LSA or LDA, instead of more specialized modelling approaches. A likely reason for this focus on two approaches is given by the lack of publicly available implementations for many methods. However, some researchers have had great success with making their implementations available (BleiLab, 2016). More such "open-access methodology" is needed to advance the use of topic modelling methodology in IS and other domains, especially regarding model validation, which many toolkits for topic modelling do not yet address as a priority and the resulting lack of methodological accessibility remains a problem (Ramage et al., 2009).

Looking at the non-IS research landscape, there is a need for modelling tools which are suited to the needs of researchers who do not use command line interfaces or software libraries, as there are no graphical user interfaces for most available implementations. A key factor in the quality of topic modelling based research is given by the presentation of the model in a paper. As discussed, looking beyond the boundaries of individual disciplines can help to identify successful solutions to this task. Also, the (meta) theoretical foundations of topic modelling remain to be established to make it easier to integrate the methodology in studies aimed at validating or expanding theory in the social and managerial sciences. One promising avenue for the creation of this theoretical foundation is presented by mixed methods, aiming at combining the advantages of modelling large document collections with qualitative approaches to content analysis. In conclusion, topic modelling has become a useful tool for many researchers, but specialized models and the development of suitable implementations for applied researchers remain largely unsolved problems offering perspectives for future research.

Appendix A: Methodological Research Contributions

Citation	Main Contribution	Type	Domain
(Salton et al., 1975)	Document storage in vector space	Computational	CS
(Croft and Harper, 1979)	Document search without prior content information	Model	IR
(Dumais et al., 1988)	LSA Model	Model	CS
(Deerwester et al., 1990)	Document retrieval using higher order term relations	Model	IS
(Spence and Owens, 1990)	Words that statistically co-occur often have a contextual association	Model	Psychology
(Cutting et al., 1992)	Clustering as an information retrieval tool	Model	IR
(Raftery, 1995)	Bayesian model selection	Validation	Sociology
(Landauer et al., 1998)	LSI: Explanation and interpretation	Validation	Interdisciplinary
(Dumais et al., 1998)	Comparison of approaches to text categorization	Comparative	IS
(Papadimitriou et al., 1998)	LSI: Evaluation of method	Validation	CS
(Hofmann, 1999)	Probabilistic-LSI: Modelling approach	Model	CS
(Lee and Seung, 2001)	Algorithmic comparison regarding non-negative matrix factorization	Computational	IS
(Park et al., 2001)	Model including prior document knowledge	Computational	IR
(Heylighen, 2001)	Comparison of word sense disambiguation approaches	Validation	IR
(Turney, 2001)	IR using pointwise mutual information (PMI-IR)	Comparative	CS
(Hofmann, 2001)	Unsupervised Learning by Probabilistic Latent Semantic Analysis	Model	ML
(Visa et al., 2002)	Document comparison by prototype matching	Model	IS
(Blei et al., 2003)	Modelling document topics using latent topics (LDA)	Model	CS
(Griffiths and Steyvers, 2004)	MCMC approach to LDA inference	Model	Interdisciplinary
(Dumais, 2004)	Overview of LSI/LSA	Model	IS
(Wei et al., 2006)	Two hierarchical agglomerative clustering (HAC) techniques	Comparative	IS
(Blei and Lafferty, 2006)	Model similar to LDA but topics change over time	Model	CS
(Teh et al., 2006)	Mixture model similar to LDA for unknown number of topics	Model	Statistics
(Teh et al., 2006)	Hierarchical Dirichlet Processes	Model	Statistics
(Wallach, 2006)	Combining n-grams and topics for document description.	Model	CS
(Blei and Lafferty, 2007)	A correlated topic model (CTM), inter-topic relations	Model	Statistics
(Mei et al., 2007)	Automated label generation for multinomial topic models	Utility	CS
(Foltz, 2007)	Book chapter: Discourse coherence and LSA	Introductory	Interdisciplinary
(Landauer, 2007)	Book chapter: Interpretation of LSA as theory of meaning.	Introductory	Interdisciplinary
(Steyvers and Griffiths, 2007)	Book chapter: Introduction to probabilistic topic models (LDA)	Introductory	Interdisciplinary
(Graesser et al., 2007)	Book chapter: Case study: Using LSA as part of a tutoring system	Theoretical	Interdisciplinary
(AlSumait et al., 2008)	Adaptive Topic Models for Mining Text Streams	Model	IS
(Wallach et al., 2009b)	Empirical evaluation methods for topic modelling.	Validation	CS
(Lin and He, 2009)	Joint Sentiment and Topic model (JST).	Model	CS
(Chang and Chien, 2009)	Sentence based Latent Dirichlet Allocation (SLDA)	Model	CS
(Wang et al., 2009)	Using topic models for multi-document summarization	Model	Comp. Ling.
(Asuncion et al., 2009)	Algorithmic comparison regarding inference in topic models	Comparative	ML
(Wallach et al., 2009a)	Comparison of structured priors for LDA	Comparative	IS
(Blei and Lafferty, 2009a)	Book chapter: Introduction to topic models	Introductory	Interdisciplinary
(Liu et al., 2009)	Joint author community and topic modelling	Model	CS
(Blei and Lafferty, 2009b)	Visualizing topics with multi-word expressions	Validation	CS
(Du et al., 2010)	Topic modelling method incorporating document segmentation	Model	ML
(Lee et al., 2010)	Comparison of topic modelling methods	Comparative	IS
(Newman et al., 2010b)	Automated evaluation of topic coherence	Validation	Comp. Ling.
(Ramage et al., 2010)	"Labeled LDA" for tweet and user characteristics	Model	IS
(Newman et al., 2010c)	Automated evaluation of topic coherence	Validation	Comp. Ling.
(Newman et al., 2010a)	Visualizing search results and document collections using topic maps	Utility	IS
(Grimmer and King, 2011)	Unsupervised clustering and evaluation thereof	Model, Evaluation	Interdisciplinary
(Newman et al., 2011)	Improving topic coherence with regularized topic models	Validation	IS
(Lu et al., 2011)	Topic modelling and multi-aspect sentiment analysis	Model	IS
(Nguyen et al., 2012)	Hierarchical nonparametric model using speaker identity	Model	Comp. Ling.
(Evangelopoulos et al., 2012)	Methodological recommendations for LSA studies	Introductory	IS
(Blei, 2012)	Overview article regarding probabilistic topic models	Introductory	CS
(Ramirez et al., 2012)	Automated topic model validation	Validation	CS
(Ignatow, 2015)	Discussion of theoretical foundations of textual analysis	Theoretical	Sociology
(Nikolenko et al., 2015)	Interval semi-supervised topic model (ISLDA) and coherence metric	Metric	IS
(George et al., 2016)	Model use cases in management research	Theoretical	Management
(Evans and Aceves, 2016)	Discussion of theory development based on text mining	Theoretical	Sociology
(Loughran and McDonald, 2016)	Overview of textual research in finance	Theoretical	Finance

Table 4: Methodological contributions identified by the structured literature review.. Domains: Information Systems (IS), Computer Science (CS), Information Retrieval (IR), Machine Learning (ML), Computational Linguistics (Comp. Ling.).

Appendix B: Applied Research Papers (Other Managerial Disciplines)

Citation	Model Type	Content	Domain	Validation	Model Use	Presentation	Description
(Landauer and Dumais, 1997)	LSA	Content Analysis	Psychology	Human benchmark	Abstraction	Statistical	LSA for analyzing Plato's problem
(Back et al., 2001)	SOM	Content Analysis	Accounting	Benchmark	Annual reports vs. quant. Data	Plots, Labels	Use of SOM for annual reports
(Kintsch and Bowles, 2002)	LSA	Content Analysis	Language	-	Metaphor comprehension	Similarities	What makes metaphors difficult to understand?
(Landauer, 2002)	LSA	Tutorial	Psychology	-	Model meaning	Example models	Introduction to LSA as a representation of learning
(Wolfe and Goldman, 2003)	LSA	Tutorial	Behavior	Guidelines	Discuss model use	LSA similarity scores	Methodological guidance for LSA use in psychology
(Kloptchenko et al., 2004)	SOM	Content Analysis	Accounting	Qualitative clustering	Explain market variation	SOM example shown	Financial reports, information regarding fut. performance
(Boukus and Rosenberg, 2006)	LSA	Content Analysis	Accounting	-	Explain market variation	Labels	LSA of FOM minutes correlated w. economic conditions
(Li, 2010)	Naïve Bayes	Content Analysis	Accounting	Cross-validation	Classify corp. Filings	Statistical	Using bayes classification for thematic and sentiment
(Quinn et al., 2010)	LDA	Content Analysis	Pol. Science	K-choice, extensive	Generate topics from political texts	Evolution	Topic modelling with political texts
(Grimmer, 2010)	Own (Hier.)	Content Analysis	Pol. Science	Over time variation	Per-author agenda	Evolution, result clustering	Measuring expressed agendas in pol. texts, new model
(Cicon et al., 2012)	LSA	Content Analysis	Finance	-	Cluster by topics	Theme clustering	Thematic analysis of corporate governance codes
(Grimmer and Stewart, 2013)	LSA,LDA	Content Analysis	Pol.Science	Validity measures	Discussion of use cases	-	Different models, assumptions, capabilities, problems
(Mohr and Bogdanov, 2013)	LDA	Tutorial	Language	-	Example model	Labels	Nontechnical introduction to topic models (LDA)
(Bao and Datta, 2014)	LDA	Content Analysis	Management	Perplexity, pred. validat.	As variable	Labels, word clouds	Identification of risk categories
(Campbell et al., 2014)	LDA	Content Analysis	Accounting	-	As variable	-	Information content of 10-K risk factor section
(Tirunillai and Tellis, 2014)	LDA	Content Analysis	Marketing	Dimension validation	Interpretation of topics/factors.	-	Consumer satisfaction dimensions (social media)
(Huber et al., 2014)	-	Review	Marketing	-	Topic evolution	Importance over time	Topics in JMR
(Huang et al., 2015)	LDA	Content Analysis	Accounting	Topic change	As variable	Labels	Analyst report topic modelling
(Kaplan and Vakili, 2015)		Content Analysis	Management	-	Ideas in patents	-	Topic modelling of patents
(Giorgi and Weber, 2015)	LDA	Content Analysis	Management	Word intrusion	Extract topics from analysts' reports	Labels	Analysts' framing repertoires and analyst evaluation.
(Cohen Priva and Austerweil, 2015)	LDA	Review	Cognition	-	Topic evolution	Top words, importance over time	Journal topic article: "Cognition"
(Wang et al., 2015)	LDA	Review	Marketing	-	Topic evolution	Labels, importance over time	50 Years "Journal of Consumer Research"
(Trusov et al., 2016)	CTM (no TM)	Content Analysis	Marketing	Accuracy	Profile clustering	Statistics for dimensions	Profiling in customer-base analysis, behavioral Targeting
(Castelló et al., 2016)	LSA	Content Analysis	Management	-	Analysis of tweet topics	First and second order topic labels	Stakeholders' sustainable development agendas
(Bellstam et al., 2016)	LDA	Content Analysis	Finance	k-choice by experiment	Topics and sentiment	Word clouds	Text-based measure of innovation using analyst reports
(Bendle and Wang, 2016)	LDA	Discussion	Management	-	Discussion of use cases	-	Discussion of LDA use cases in business
(Guerreiro et al., 2016)	CTM	Review	Ethics	Likelihood and perplexity	Key themes of research area	Discussion of each topic of interest	Review of cause-related marketing literature
(Jacobs et al., 2016)	-	Statistical	Marketing	Success rate	As variable	Statistical	Model-Based Purchase Predictions for Large Assortments
(Guo et al., 2017)	LDA	Content Analysis	Tourism	Benchmark	Compare to review ratings	Evaluation plots	Tourist satisfaction analysis

Table 5: Applied papers in other managerial disciplines (non-IS). Results indicate a strong focus on the use of topics models as a tool for content analysis, which often involves using the topic to document assignments as variables in regression models. The temporal distribution of the discovered contributions within this category indicate a rapid increase in the use of topic modelling methods. While there are several tutorials and methodological advice papers within these fields, there is still room for future research regarding a broader spectrum of model use and topic model validation. While most studies within these fields use very extensive validation techniques for other statistical methods, topic model validation has not yet been adopted to the same degree. Likewise, most studies either use LSA or LDA, while there may still be many use cases for derivatives of these methods.

Appendix C: Applied Research Papers (Information Systems)

Citation	Model Type	Content	Domain	Validation	Model Use	Presentation	Description
(Husbands et al., 2001)	SVD	IR	IS	Precision measure	Model is main contribution	Statistical	Using SVD for document retrieval
(Wei and Croft, 2006)	LDA	IR	IS	Average precision	Find similar documents	Model not shown	Using LDA for ad-hoc information retrieval
(Mihalcea et al., 2006)	LSA	Text Similarity	IS	Precision, recall, F-Score	Find similar documents	-	Corpus- and knowledge-based measures of similarity
(Wei et al., 2007)	Clustering	IR	IS	Performance metric	Model is main contribution	Statistical	Topic based query expansion for IR
(Arazy and Woo, 2007)	Collocation	IR	IS	F-score	Model is main contribution	Word collocation	Information retrieval using collocation indexing
(Sidorova et al., 2007)	LSA	Review	IS	-	Interpretation of topics/factors	Interpretation and description	Using LSA to identify research streams
(Graesser et al., 2007)	LSA	Tutorial	IS	?	Part of virtual tutor	?	Explanatory case study
(Titov and McDonald, 2008)	MG-LDA	Content Analysis	IS	LDA benchmark, metric	Model is main contribution	Labels	Extract aspects from product reviews (Multi-Grain LDA)
(Hall et al., 2008)	LSA	Review	IS	-	Interpretation of topics/factors	Interpretation and description	Using LDA to identify historical research trends.
(Sidorova et al., 2008)	LSA	Review	IS	-	Interpretation of topics/factors	Interpretation and description	Using LSA to identify research streams
(Ramage et al., 2009)	-	Discussion	IS	-	Exploration of output	Model should be explorable	Accessibility, trust in topic models (social sciences)
(Paul and Girju, 2009)	Naïve Bayes	Review	IS	-	Interpretation of topics/factors	Labels, evolution, inter-model	Topic comparison between research domains
(Chang et al., 2009)	LDA	Validation	IS	Benchmark (other metric)	Model output evaluation	Word and topic intrusion.	Quantitative metrics for semantic topic coherence
(Turney and Pantel, 2010)	-	Discussion	IS	-	Review article	-	Different text representations using vector space models
(Sidorova and Isik, 2010)	LSA	Review	IS	-	Exploration of output	Topic labels, importance	Review using LSA: Business process literature
(Aral et al., 2011)	LDA	Content Analysis	IS	Model comparison	As variable	Labels	Impact of stock recommendations on stock returns
(O'Connor et al., 2011)	-	Discussion	IS	Review article	Review article	-	Model complexity and model assumptions
(Chen et al., 2012)	-	Discussion	IS	-	Special issue about BI research	-	Overview of big data landscape, including topic models
(Jin et al., 2013)	LDA	DSS	IS	-	As variable	-	Forex trend modelling system
(Koukal et al., 2014b)	LSA	Discussion	IS	-	Literature review	-	LSA for literature reviews and prototype tool
(Kulkarni et al., 2014)	LSA	Review	IS	-	Literature review	Importance over time	Operations management research
(Koukal et al., 2014a)	LSA	Validation	IS	Purpose of article	Literature review	Benchmarks	Validation of Koukal et al. (2014b)
(Ahmad and Laroche, 2015)	LSA	Content Analysis	IS	-	Measure emotions	Statistical	Review helpfulness and emotions shown in review
(DiMaggio, 2015)	-	Discussion	IS	-	-	-	Different research perspectives in CS and social sciences
(Mützel, 2015)	-	Discussion	IS	Discussion article	Discussion article	-	Topic modelling in sociology, challenges, opportunities
(Wagner-Pacifici et al., 2015)	-	Discussion	IS	Review article	Literature review	-	Discussion: Big data in the social and cultural sciences
(Moqri et al., 2015)	LSA	Review	IS	No Full text	Literature review	No Full text	Identifying Research Trends in IS
(Chen and Zhao, 2015)	CTM	Review	IS	-	Literature review	Plots	Correlated topic model: Information systems
(Aryal et al., 2015)	LSA	Review	IS	-	Literature review	Period-comparison of key terms	Healthcare research
(Kundu et al., 2015)	LSA	Review	IS	-	Literature review	Importance over time	Supply chain management
(Müller et al., 2016)	LSA	Content Analysis	IS	Varying topic count	Interpretation of topics/factors	Term- and document loadings	Develop a typology of BPM professionals
(Rai, 2016)	LDA	Discussion	IS	-	Discussion article	-	Call for use of LDA for theory generation
(Larsen and Bing, 2016)	LSA	Tool	IS	Recall, Precision, F-Score	Construct identity	Constructs, graphs for evaluation	Addressing construct identity in literature reviews

Table 6: Applied research papers in Information Systems (IS). As shown, IS researchers have, so far, mainly used topic models for reviewing purposes in several contexts. Like researchers in other managerial disciplines, they focus on LSA and LDA for their studies. In comparison, more IS papers discuss the use of topic modelling methodology, while using the model as part of another analysis is less common. As was observed in other domains, the usage of topic models has recently spiked within the discipline.

References

Ågerfalk, P. J. (2013). "Embracing Diversity through Mixed Methods Research," *European Journal of Information Systems* 22 (3), pp. 251-256.

Ahmad, S. N., and Laroche, M. (2015). "How Do Expressed Emotions Affect the Helpfulness of a Product Review? Evidence from Reviews Using Latent Semantic Analysis," *International Journal of Electronic Commerce* 20 (1), pp. 76-111.

AlSumait, L., Barbará, D., and Domeniconi, C. (2008). "On-Line Lda: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," in: *IEEE International Conference on Data Mining*. pp. 3-12.

Aral, S., Ipeirotis, P. G., and Taylor, S. J. (2011). "Content and Context: Identifying the Impact of Qualitative Information on Consumer Choice," In: *Proceedings of the International Conference on Information Systems*, Shanghai.

Arazy, O., and Woo, C. (2007). "Enhancing Information Retrieval through Statistical Natural Language Processing: A Study of Collocation Indexing," *MIS Quarterly* 31 (3), pp. 525-546.

Aryal, A., Gallivan, M., and Tao, Y. Y. (2015). "Using Latent Semantic Analysis to Identify Themes in Is Healthcare Research," In: *Proceedings of the Americas Conference on Information Systems*, Puerto Rico: AISeL.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). "On Smoothing and Inference for Topic Models," In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*: AUAI Press, pp. 27-34.

Back, B., Toivonen, J., Vanharanta, H., and Visa, A. (2001). "Comparing Numerical Data and Text Information from Annual Reports Using Self-Organizing Maps," *International Journal of Accounting Information Systems* 2 (4), pp. 249-269.

Bagozzi, R. P. (2011). "Measurement and Meaning in Information Systems and Organizational Research: Methodological and Philosophical Foundations," *MIS Quarterly* 35 (2), pp. 261-292.

Bao, Y., and Datta, A. (2014). "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Management Science* 60 (6), pp. 1371-1391.

Bellstam, G., Bhagat, S., and Cookson, J. A. (2016). "A Text-Based Analysis of Corporate Innovation," *SSRN* 2803232.

Bendle, N. T., and Wang, X. S. (2016). "Uncovering the Message from the Mess of Big Data," *Business Horizons* 59 (1), pp. 115-124.

Bergamaschi, S., and Po, L. (2014). "Comparing Lda and Lsa Topic Models for Content-Based Movie Recommendation Systems," In: *Proceedings of the International Conference on Web Information Systems and Technologies*: Springer, pp. 247-263.

Blei, D. M. (2012). "Probabilistic Topic Models," *Communications of the ACM* 55 (4), pp. 77-84.

Blei, D. M. (2014). "Hierarchical Lda with a Fixed Depth Tree and a Stick Breaking Prior on the Depth Weights,".

Blei, D. M., and Lafferty, J. D. (2006). "Dynamic Topic Models," In: *Proceedings of the International Conference on Machine Learning*, Pittsburgh PA: ACM, pp. 113-120.

Blei, D. M., and Lafferty, J. D. (2007). "A Correlated Topic Model of Science," *The Annals of Applied Statistics*, pp. 17-35.

Blei, D. M., and Lafferty, J. D. (2009a). "Topic Models," *Text mining: Classification, Clustering, and Applications* 10 (71), p. 34.

Blei, D. M., and Lafferty, J. D. (2009b). "Visualizing Topics with Multi-Word Expressions," *arXiv* 0907.1013.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet Allocation," *The Journal of Machine Learning Research* 3, pp. 993-1022.

BleiLab. (2016). "Blei Lab Github Repository." Retrieved 12/2/2016, from https://github.com/blei-lab

Boell, S. K., and Cecez-Kecmanovic, D. (2015). "On Being 'Systematic' in Literature Reviews in Is," *Journal of Information Technology* 30 (2), pp. 161-173.

Boukus, E., and Rosenberg, J. V. (2006). "The Information Content of Fomc Minutes," *SSRN* 922312.
Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H.-m., and Steele, L. B. (2014). "The Information Content of Mandatory Risk Factor Disclosures in Corporate Filings," *Review of Accounting Studies* 19 (1), pp. 396-455.
Castelló, I., Etter, M., and Nielsen, F. Å. (2016). "Strategies of Legitimacy through Social Media: The Networked Strategy," *Journal of Management Studies* 53 (3), pp. 402-432.
Chaney, A. J. B. (2014). "Online Topic Model Visualization,".
Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models," In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 288-296.
Chang, Y.-L., and Chien, J.-T. (2009). "Latent Dirichlet Learning for Document Summarization," In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*: IEEE, pp. 1689-1692.
Charlin, L., Ranganath, R., McInerney, J., and Blei, D. M. (2015). "Dynamic Poisson Factorization," In: *Proceedings of the ACM Conference on Recommender Systems*: ACM, pp. 155-162.
Chen, H., Chiang, R. H. L., and Storey, V. C. (2012). "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* 36 (4), pp. 1165-1188.
Chen, H., and Zhao, J. L. (2015). "Istopic: Understanding Information Systems Research through Topic Models," In: *Proceedings of the International Conference on Information Systems*, Fort Worth, USA: AISeL.
Chong, W., Blei, D., and Li, F.-F. (2009). "Simultaneous Image Classification and Annotation," In: *Proceedings of the Computer Vision and Pattern Recognition, 2009. CVPR 2009*: IEEE, pp. 1903-1910.
Cicon, J. E., Ferris, S. P., Kammel, A. J., and Noronha, G. (2012). "European Corporate Governance: A Thematic Analysis of National Codes of Governance," *European Financial Management* 18 (4), pp. 620-648.
Cohen Priva, U., and Austerweil, J. L. (2015). "Analyzing the History of Cognition Using Topic Models," *Cognition* 135, pp. 4-9.
Croft, W. B., and Harper, D. J. (1979). "Using Probabilistic Models of Document Retrieval without Relevance Information," *Journal of Documentation* 35 (4), pp. 285-295.
Crossno, P. J., Wilson, A. T., Shead, T. M., and Dunlavy, D. M. (2011). "Topicview: Visually Comparing Topic Models of Text Collections," In: *Proceedings of the International Conference on Tools with Artificial Intelligence*: IEEE, pp. 936-943.
Cutting, D. K., Karger, D. R., Pedersen, J. O., and Scatter, T. J. W. (1992). "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections," In: *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318-329.
Debortoli, S., Müller, O., Junglas, I., and Brocke, J. (2016). "Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial," *Communications of the Association for Information Systems* 39 (7).
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science* 41 (6), pp. 391-407.
DiMaggio, P. (2015). "Adapting Computational Text Analysis to Social Science (and Vice Versa)," *Big Data & Society* 2 (2), pp. 1-5.
Du, L., Buntine, W., and Jin, H. (2010). "A Segmented Topic Model Based on the Two-Parameter Poisson-Dirichlet Process," *Machine Learning* 81 (1), pp. 5-19.
Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). "Inductive Learning Algorithms and Representations for Text Categorization," In: *Proceedings of the International conference on Information and knowledge management*: ACM, pp. 148-155.
Dumais, S. T. (2004). "Latent Semantic Analysis," *Annual Review of Information Science and Technology* 38 (1), pp. 188-230.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). "Using Latent Semantic Analysis to Improve Access to Textual Information," In: *Proceedings of the ACM Conference on Human factors in computing systems*: SIGCHI pp. 281-285.

Elder-Vass, D. (2014). "Debate: Seven Ways to Be a Realist About Language," *Journal for the Theory of Social Behaviour* 44 (3), pp. 249-267.

Evangelopoulos, N., Zhang, X., and Prybutok, V. R. (2012). "Latent Semantic Analysis: Five Methodological Recommendations," *European Journal of Information Systems* 21 (1), pp. 70-86.

Evans, J. A., and Aceves, P. (2016). "Machine Translation: Mining Text for Social Theory," *Annual Review of Sociology* 42, pp. 21-50.

Foltz, P. W. (2007). "Discourse Coherence and Lsa," in *Handbook of Latent Semantic Analysis*. pp. 167-184.

George, G., Haas, M. R., and Pentland, A. (2016). "From the Editors: Big Data and Data Science Methods for Management Research," *Academy of Management Journal* 59 (5), pp. 1493-1507.

Gerrish, S., and Blei, D. M. (2010). "A Language-Based Approach to Measuring Scholarly Impact," In: *Proceedings of the International Conference on Machine Learning*, pp. 375-382.

Giorgi, S., and Weber, K. (2015). "Marks of Distinction: Framing and Audience Appreciation in the Context of Investment Advice," *Administrative Science Quarterly*, pp. 1-35.

Gopalan, P. K., Charlin, L., and Blei, D. (2014). "Content-Based Recommendations with Poisson Factorization," In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3176-3184.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., and Hu, X. (2007). "Using Lsa in Autotutor: Learning through Mixed Initiative Dialogue in Natural Language," in *Handbook of Latent Semantic Analysis*. pp. 243-262.

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., and Peacock, R. (2005). "Storylines of Research in Diffusion of Innovation: A Meta-Narrative Approach to Systematic Review," *Social Science & Medicine* 61 (2), pp. 417-430.

Gregory, F. H. (1993). "Soft Systems Methodology to Information Systems: A Wittgensteinian Approach," *Information Systems Journal* 3 (3), pp. 149-168.

Griffiths, T. L., and Steyvers, M. (2004). "Finding Scientific Topics," *Proceedings of the National Academy of Sciences* 101 (suppl 1), pp. 5228-5235.

Grimmer, J. (2010). "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases," *Political Analysis* 18 (1), pp. 1-35.

Grimmer, J., and King, G. (2011). "General Purpose Computer-Assisted Clustering and Conceptualization," *Proceedings of the National Academy of Sciences* 108 (7), pp. 2643-2650.

Grimmer, J., and Stewart, B. M. (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21 (3), pp. 267-297.

Grün, B., and Hornik, K. (2011). "Topicmodels : An R Package for Fitting Topic Models," *Journal of Statistical Software* 40 (13), pp. 1-30.

Guerreiro, J., Rita, P., and Trigueiros, D. (2016). "A Text Mining-Based Review of Cause-Related Marketing Literature," *Journal of Business Ethics*, pp. 111-128.

Günther, F., Dudschig, C., and Kaup, B. (2015). "Lsafun-an R Package for Computations Based on Latent Semantic Analysis," *Behavior Research Methods* 47 (4), pp. 930-944.

Guo, Y., Barnes, S. J., and Jia, Q. (2017). "Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation," *Tourism Management* 59, pp. 467-483.

Halevy, A., Norvig, P., and Pereira, F. (2009). "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems* 24 (2), pp. 8-12.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). "Studying the History of Ideas Using Topic Models," In: *Proceedings of the Conference on empirical methods in natural language processing*: Association for Computational Linguistics, pp. 363-371.

Heylighen, F. (2001). "Mining Associative Meanings from the Web: From Word Disambiguation to the Global Brain," In: *Proceedings of the Trends in Special Language and Language Technology*, R. Temmerman (ed.), Brussels: Standaard Publishers.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). "Online Learning for Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems*, J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel and A. Culotta (eds.). Curran Associates, Inc., pp. 856-864.

Hofmann, T. (1999). "Probabilistic Latent Semantic Indexing," In: *Proceedings of the ACM SIGIR conference on Research and development in information retrieval*: ACM, pp. 50-57.

Hofmann, T. (2001). "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning* 42 (1), pp. 177-196.

Huang, A., Lehavy, R., Zang, A., and Zheng, R. (2015). "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Ross School of Business Working Paper* 1229.

Huber, J., Kamakura, W., and Mela, C. F. (2014). "A Topical History of Jmr," *Journal of Marketing Research* 51 (1), pp. 84-91.

Husbands, P., Simon, H., and Ding, C. H. (2001). "On the Use of the Singular Value Decomposition for Text Retrieval," *Computational information retrieval* 5, pp. 145-156.

Ignatow, G. (2015). "Theoretical Foundations for Digital Text Analysis," *Journal for the Theory of Social Behaviour* 46 (1), pp. 104-120.

Jacobs, B. J. D., Donkers, B., and Fok, D. (2016). "Model-Based Purchase Predictions for Large Assortments Model-Based Purchase Predictions for Large Assortments," *Marketing Science* 35 (3), pp. 389-404.

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., and Ramakrishnan, N. (2013). "Forex-Foreteller: Currency Trend Modeling Using News Articles," In: *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 1470-1473.

Kaplan, S., and Vakili, K. (2015). "The Double-Edged Sword of Recombination in Breakthrough Innovation," *Strategic Management Journal* 36 (10), pp. 1435-1457.

Kintsch, W., and Bowles, A. R. (2002). "Metaphor Comprehension: What Makes a Metaphor Difficult to Understand?," *Metaphor & Symbol* 17 (4), pp. 249-262.

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., and Visa, A. (2004). "Combining Data and Text Mining Techniques for Analysing Financial Reports," *Intelligent systems in accounting, finance and management* 12 (1), pp. 29-41.

Koukal, A., Gleue, C., and Breitner, M. (2014a). "Enhancing Literature Review Methods - Evaluation of a Literature Search Approach Based on Latent Semantic Indexing," In: *Proceedings of the International Conference on Information Systems*, Auckland.

Koukal, A., Gleue, C., and Breitner, M. (2014b). "Enhancing Literature Review Methods - Towards More Efficient Literature Research with Latent Semantic Indexing," In: *Proceedings of the European Conference on Information Systems*, Tel Aviv, Israel: AISeL.

Kulkarni, S. S., Apte, U. M., and Evangelopoulos, N. E. (2014). "The Use of Latent Semantic Analysis in Operations Management Research," *Decision Sciences* 45 (5), pp. 971-994.

Kundu, A., Jain, V., Kumar, S., and Chandra, C. (2015). "A Journey from Normative to Behavioral Operations in Supply Chain Management: A Review Using Latent Semantic Analysis," *Expert Systems with Applications* 42 (2), pp. 796-809.

Lai, L. S., and To, W. (2015). "Content Analysis of Social Media: A Grounded Theory Approach," *Journal of Electronic Commerce Research* 16 (2), p. 138.

Landauer, T. K. (2002). "On the Computational Basis of Learning and Cognition: Arguments from Lsa," *Psychology of Learning and Motivation* 41, pp. 43-84.

Landauer, T. K. (2007). "Lsa as a Theory of Meaning," in *Handbook of Latent Semantic Analysis*. pp. 3-34.

Landauer, T. K., and Dumais, S. T. (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychological Review* 104 (2), p. 211.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). "An Introduction to Latent Semantic Analysis," *Discourse Processes* 25 (2-3), pp. 259-284.

Larsen, K. R., and Bing, C. H. (2016). "A Tool for Addressing Construct Identity in Literature Reviews and Meta Analyses," *MIS Quarterly* 40 (3), pp. 529-551.

Lee, D. D., and Seung, H. S. (2001). "Algorithms for Non-Negative Matrix Factorization," In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 556-562.

Lee, S., Baker, J., Song, J., and Wetherbe, J. C. (2010). "An Empirical Comparison of Four Text Mining Methods," In: *Proceedings of the Hawaii International Conference on System Sciences*: IEEE, pp. 1-10.

Li, F. (2010). "The Information Content of Forward-Looking Statements in Corporate Filings—a Naïve Bayesian Machine Learning Approach," *Journal of Accounting Research* 48 (5), pp. 1049-1102.

Lin, C., and He, Y. (2009). "Joint Sentiment/Topic Model for Sentiment Analysis," In: *Proceedings of the ACM Conference on Information and knowledge management*: ACM, pp. 375-384.

Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills, California et al.: SAGE Publications.

Liu, B. (2012). "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies* 5 (1), pp. 1-167.

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). "Topic-Link Lda: Joint Models of Topic and Author Community," In: *Proceedings of the Annual International Conference on Machine Learning*: ACM, pp. 665-672.

Loughran, T., and McDonald, B. (2016). "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research* 54 (4), pp. 1187-1230.

Lu, B., Ott, M., Cardie, C., and Tsou, B. K. (2011). "Multi-Aspect Sentiment Analysis with Topic Models," In: *Proceedings of the International Conference on Data Mining Workshops*: IEEE pp. 81-88.

McCallum, A. K. (2002). "Mallet: A Machine Learning for Language Toolkit,".

Mei, Q., Shen, X., and Zhai, C. (2007). "Automatic Labeling of Multinomial Topic Models," In: *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 490-499.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity," In: *Proceedings of the AAAI Conference*, pp. 775-780.

Miller, G., and Fellbaum, C. (1998). *Wordnet: An Electronic Lexical Database*. Cambridge: MIT Press.

Mingers, J. C. (1995). "Information and Meaning - Foundations for an Intersubjective Account," *Information Systems Journal* 5 (4), pp. 285-306.

Mohr, J. W., and Bogdanov, P. (2013). "Introduction - Topic Models: What They Are and Why They Matter," *Poetics* 41 (6), pp. 545-569.

Moqri, M., Bandyopadhyay, S., and Cheng, H. K. (2015). "Identifying Research Trends in Is," In: *Proceedings of the Americas Conference on Information Systems*, Puerto Rico.

Müller, O., Schmiedel, T., Gorbacheva, E., and vom Brocke, J. (2016). "Towards a Typology of Business Process Management Professionals: Identifying Patterns of Competences through Latent Semantic Analysis," *Enterprise Information Systems* 10 (1), pp. 50-80.

Mützel, S. (2015). "Facing Big Data: Making Sociology Relevant," *Big Data & Society* 2 (2).

Newman, D., Baldwin, T., Cavedon, L., Huang, E., Karimi, S., Martinez, D., Scholer, F., and Zobel, J. (2010a). "Visualizing Search Results and Document Collections Using Topic Maps," *Web Semantics: Science, Services and Agents on the World Wide Web* 8 (2), pp. 169-175.

Newman, D., Bonilla, E. V., and Buntine, W. (2011). "Improving Topic Coherence with Regularized Topic Models," In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 496-504.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010b). "Automatic Evaluation of Topic Coherence," In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*: Association for Computational Linguistics, pp. 100-108.

Newman, D., Noh, Y., Talley, E., Karimi, S., and Baldwin, T. (2010c). "Evaluating Topic Models for Digital Libraries," In: *Proceedings of the Annual joint Conference on Digital libraries*: ACM, pp. 215-224.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature," *Decision Support Systems* 50 (3), pp. 559-569.

Nguyen, V.-A., Boyd-Graber, J., and Resnik, P. (2012). "Sits: A Hierarchical Nonparametric Model Using Speaker Identity for Topic Segmentation in Multiparty Conversations," In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*: Association for Computational Linguistics, pp. 78-87.

Nickerson, R. C., Varshney, U., and Muntermann, J. (2013). "A Method for Taxonomy Development and Its Application in Information Systems," *European Journal of Information Systems* 22 (3), pp. 336-359.

Nikolenko, S. I., Koltcov, S., and Koltsova, O. (2015). "Topic Modelling for Qualitative Studies," *Journal of Information Science* 0165551515617393.

O'Connor, B., Bamman, D., and Smith, N. A. (2011). "Computational Text Analysis for Social Science: Model Assumptions and Complexity," in: *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala, S. (1998). "Latent Semantic Indexing: A Probabilistic Analysis," In: *Proceedings of the ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*: ACM, pp. 159-168.

Park, H., Jeon, M., and Rosen, J. B. (2001). "Lower Dimensional Representation of Text Data in Vector Space Based Information Retrieval," *Computational information retrieval*, pp. 3-23.

Paul, M. J., and Girju, R. (2009). "Topic Modeling of Research Fields: An Interdisciplinary Perspective," In: *Proceedings of the International Conference on recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 337-342.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). "How to Analyze Political Attention with Minimal Assumptions and Costs," *American Journal of Political Science* 54 (1), pp. 209-228.

Raftery, A. E. (1995). "Bayesian Model Selection in Social Research," *Sociological methodology*, pp. 111-163.

Rai, A. (2016). "Editor's Comments: Synergies between Big Data and Theory," *MIS Quarterly* 40 (2), pp. iii-ix.

Ramage, D., Dumais, S. T., and Liebling, D. J. (2010). "Characterizing Microblogs with Topic Models," *International Conference on Web and Social Media*, pp. 130-137.

Ramage, D., and Rosen, E. (2011). "Stanford Topic Modeling Toolbox,".

Ramage, D., Rosen, E., Chuang, J., D. Manning, C., and McFarland, D. A. (2009). "Topic Modeling for the Social Sciences," in: *NIPS Workshop on Applications for Topic Models: Text and Beyond*. Whistler, Canada.

Ramirez, E. H., Brena, R., Magatti, D., and Stella, F. (2012). "Topic Model Validation," *Neurocomputing* 76 (1), pp. 125-133.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. (2015). "Deep Exponential Families," In: *Proceedings of the AISTATS*.

Rehurek, R., and Sojka, P. (2010). "Software Framework for Topic Modelling with Large Corpora," In: *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*: Citeseer.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2014). "Stm: R Package for Structural Topic Models," *R package* 1, p. 12.

Rowe, F. (2014). "What Literature Review Is Not: Diversity, Boundaries and Recommendations," *European Journal of Information Systems* 23 (3), pp. 241-255.

Salton, G., Wong, A., and Yang, C.-S. (1975). "A Vector Space Model for Automatic Indexing," *Communications of the ACM* 18 (11), pp. 613-620.

Sidorova, A., Evangelopoulos, N., and Ramakrishnan, T. (2007). "Diversity in Is Research: An Exploratory Study Using Latent Semantics," In: *Proceedings of the International Conference on Information Systems*, p. 10.

Sidorova, A., Evangelopoulos, N., Valacich, J. S., and Ramakrishnan, T. (2008). "Uncovering the Intellectual Core of the Information Systems Discipline," *MIS Quarterly* 32 (3 September), pp. 467-482.

Sidorova, A., and Isik, O. (2010). "Business Process Research: A Cross-Disciplinary Review," *Business Process Management Journal* 16 (4), pp. 566-597.

Sievert, C., and Shirley, K. E. (2014). "Ldavis: A Method for Visualizing and Interpreting Topics," In: *Proceedings of the Workshop on interactive language learning, visualization, and interfaces*, pp. 63-70.

Spence, D. P., and Owens, K. C. (1990). "Lexical Co-Occurrence and Association Strength," *Journal of Psycholinguistic Research* 19 (5), pp. 317-330.

Steyvers, M., and Griffiths, T. (2007). "Probabilistic Topic Models," in *Handbook of Latent Semantic Analysis*. pp. 424-440.

Sylvester, A., Tate, M., and Johnstone, D. (2013). "Beyond Synthesis: Re-Presenting Heterogeneous Research Literature," *Behaviour & Information Technology* 32 (12), pp. 1199-1215.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association* 101 (476), pp. 1566-1581.

Templier, M., and Paré, G. (2015). "A Framework for Guiding and Evaluating Literature Reviews," *Communications of the Association for Information Systems* 37 (1), p. 6.

Tirunillai, S., and Tellis, G. (2014). "Mining Marketing Meaning from Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," *Journal of Marketing Research* 51 (August), pp. 463-479.

Titov, I., and McDonald, R. (2008). "Modeling Online Reviews with Multi-Grain Topic Models," In: *Proceedings of the International conference on World Wide Web*: ACM, pp. 111-120.

Trusov, M., Ma, L., and Jamal, Z. (2016). "Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting," *Marketing Science* 35 (3), pp. 405-426.

Turney, P. (2001). "Mining the Web for Synonyms: Pmi-Ir Versus Lsa on Toefl," In: *Proceedings of the European Conference on Machine Learning*, pp. 491–502.

Turney, P. D., and Pantel, P. (2010). "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research* 37 (1), pp. 141-188.

Venkatesh, V., Brown, S. A., and Bala, H. (2013). "Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems," *MIS quarterly* 37 (1), pp. 21-54.

Venkatesh, V., Brown, S. A., and Sullivan, Y. W. (2016). "Guidelines for Conducting Mixed-Methods Research: An Extension and Illustration," *Journal of the Association for Information Systems* 17 (7), p. 2.

Visa, A., Toivonen, J., Vanharanta, H., and Back, B. (2002). "Contents Matching Defined by Prototypes: Methodology Verification with Books of the Bible," *Journal of Management Information Systems* 18 (4), pp. 87-100.

W3C. (2013). "The W3c Data Activity: Building the Web of Data.", from https://www.w3.org/2013/data/

Wagner-Pacifici, R., Mohr, J. W., and Breiger, R. L. (2015). "Ontologies, Methodologies, and New Uses of Big Data in the Social and Cultural Sciences," *Big Data & Society* 2 (2), p. 2053951715613810.

Wallach, H. M. (2006). "Topic Modeling: Beyond Bag-of-Words," in: *International conference on Machine Learning*. Pittsburgh, Pennsylvania, USA: ACM, pp. 977-984.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009a). "Rethinking Lda: Why Priors Matter," In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1973-1981.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b). "Evaluation Methods for Topic Models," In: *Proceedings of the Annual International Conference on Machine Learning*: ACM, pp. 1105-1112.

Wang, C., and Blei, D. (2010). "Hierarchical Dirichlet Process (with Split-Merge Operations),". Github.

Wang, C., and Blei, D. M. (2011). "Collaborative Topic Modeling for Recommending Scientific Articles," In: *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 448-456.

Wang, C. J. W. B., David M. (2011). "Online Variational Inference for the Hierarchical Dirichlet Process," In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 752-760.

Wang, D., Zhu, S., Li, T., and Gong, Y. (2009). "Multi-Document Summarization Using Sentence-Based Topic Models," In: *Proceedings of the ACL-IJCNLP Conference Short Papers*: Association for Computational Linguistics, pp. 297-300.

Wang, X. S., Bendle, N. T., Mai, F., and Cotte, J. (2015). "The Journal of Consumer Research at 40: A Historical Analysis," *Journal of Consumer Research* 42 (1), pp. 5-18.

Webster, J., and Watson, R. T. (2002). "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* 26 (2), pp. xiii-xxiii.

Wei, C.-P., Chiang, R., and Wu, C.-C. (2006). "Accommodating Individual Preferences in the Categorization of Documents: A Personalized Clustering Approach," *Journal of Management Information Systems* 23 (2), pp. 173-201.

Wei, C.-P., Hu, P. J.-H., Tai, C.-H., Huang, C.-N., and Yang, C.-S. (2007). "Managing Word Mismatch Problems in Information Retrieval: A Topic-Based Query Expansion Approach," *Journal of Management Information Systems* 24 (3), pp. 269-295.

Wei, X., and Croft, W. B. (2006). "Lda-Based Document Models for Ad-Hoc Retrieval," In: *Proceedings of the ACM SIGIR conference on Research and development in information retrieval*: ACM, pp. 178-185.

Wolfe, M. B. W., and Goldman, S. R. (2003). "Use of Latent Semantic Analysis for Predicting Psychological Phenomena: Two Issues and Proposed Solutions," *Behavior Research Methods, Instruments, & Computers* 35 (1), pp. 22-31.

Zhai, K., Boyd-Graber, J., Asadi, N., and Alkhouja, M. L. (2012). "Mr. Lda: A Flexible Large Scale Topic Modeling Package Using Variational Inference in Mapreduce," In: *Proceedings of the International conference on World Wide Web*: ACM, pp. 879-888.