Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture

Article *in* Organizational Research Methods · April 2018

DOI: 10.1177/1094428118773858

Theresa Schmiedel

University of Liechtenstein

Oliver Müller

IT University of Copenhagen

Jan vom Brocke

University of Liechtenstein

Paper accepted for publication in Organizational Research Methods

Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial with an Application Example on Organizational Culture

Theresa Schmiedel, Oliver Müller, and Jan vom Brocke

Abstract

Research has emphasized the limitations of qualitative and quantitative approaches to studying organizational phenomena. For example, in-depth interviews are resource-intensive, while questionnaires with closed-ended questions can only measure predefined constructs. With the recent availability of large textual data sets and increased computational power, text mining has become an attractive method that has the potential to mitigate some of these limitations. Thus, we suggest applying topic modeling, a specific text mining technique, as a new and complementary strategy of inquiry to study organizational phenomena. In particular, we outline the potentials of structural topic modeling for organizational research and provide a step-by-step tutorial on how to apply it. Our application example builds on 428,492 reviews of Fortune 500 companies from the online platform Glassdoor on which employees can evaluate organizations. We demonstrate how structural topic models allow to inductively identify topics that matter to employees and to quantify their relationship with employees' perception of organizational culture. We discuss the advantages and limitations of topic modeling as a research method and outline how future research can apply the technique to study organizational phenomena.

Keywords: topic modeling, structural topic model, tutorial, organizational culture, online reviews

Introduction

Organizational research follows both quantitative and qualitative research paradigms and applies various empirical methods for data collection and analysis (Currall, Hammer, Baggett, & Doniger, 1999; Yauch & Steudel, 2003). Researchers may, for example, collect qualitative data through interviews or observation and subsequently use coding techniques to build a new theory explaining a particular organizational phenomenon. Or they may collect quantitative data via surveys or experiments in order to statistically test theory-derived hypotheses about cause-and-effect relationships in organizations. Both methodological approaches come with certain limitations (Gioia, Corley, & Hamilton, 2013). For example, the generalizability of case study data and the appropriateness of using questionnaires to gain valid insights are often discussed issues (Hardy & Ford, 2014; Schein, 1990). Researchers have, thus, called for new methods to studying organizational phenomena (Taras, Rowney, & Steel, 2009).

In recent years, several disciplines have started applying methods novel to their fields in order to gain access to new data sources. Specifically, computational methods for text mining hold great potential for many disciplines, considering the vast amount of textual data available today. First applications of text mining appeared in the biomedical field (Spasic, Ananiadou, McNaught, & Kumar, 2005). But recently researchers from various other disciplines have started to apply text mining as a strategy of inquiry (Bao & Datta, 2014; Debortoli, Müller, Junglas, & vom Brocke, 2016; Janasik, Honkela, & Bruun, 2009; Michel et al., 2011; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010). Yet, while more and more researchers have started using these data-analytic techniques, much needs to be done to leverage their full potential in the organizational sciences (Tonidandel, King, & Cortina, 2016).

Accordingly, we suggest that organizational research embrace the methodological advances from other fields and propose topic modeling as a specific text mining technique to study organizational phenomena. Thus, the purpose of our research is to demonstrate the utility of a new and complementary methodological approach to study organizations. We use an application example that outlines which topics matter to employees' perceptions of corporate cultures. We show the potential of topic modeling as a methodology that can advance organizational research and that can provide a complementary solution to the prevailing issues regarding the use of extant empirical methods in the field.

Next, we provide some background on topic models in general and structural topic models in particular. We then provide details on structural topic modeling in the form of a step-by-step tutorial that contains an application example, in which we apply the method to organizational culture research and examine online reviews of Fortune 500 companies. Our approach includes identifying topics that matter to employees, quantifying the relationship of these topics with employees' perception of organizational culture, and engaging with existing literature in interpreting the findings. Finally, we outline advantages and limitations of topic modeling for organizational research and discuss application fields in organizational research to, then, conclude with a summary and outlook.

Topic Modeling as a Method for Organizational Research

Traditional Organizational Research Methods

Organizational research applies both qualitative and quantitative research methods. Qualitatively exploring organizational phenomena through techniques such as observations and interviews allows themes and structures to be inductively identified through examining patterns of individual behavior (Gioia et al., 2013; Morey & Morey, 1994; Van Maanen, 1979). A key advantage of qualitative studies is that the emerging insights on a particular organization provide a deeply grounded picture of reality that accounts for the dynamics and complexity of organizations (Sackmann, 2001). However, qualitative studies also come along with weaknesses that have engendered criticism. For example, important aspects for research may be overlooked, since social desirability can strongly influence the data collection process (Yauch & Steudel, 2003). Furthermore, qualitative organizational studies require a considerable amount of resources and, thus, typically focus on small samples only (Jung et al., 2009).

Following a quantitative research paradigm allows organizational phenomena to be measured and groups to be compared based on numerical data (Yauch & Steudel, 2003). Typically, such research assesses organizational groups through the operationalization of a set of relevant constructs (Fields, 2002). A key advantage of quantitative studies is their scalability, that is, their efficient and effective examination of large samples at comparably low costs and in comparably little time (Jung et al., 2009). Yet, scholars have criticized that the use of predefined scales to measure constructs restricts exploration, because the predetermined dimensions in survey instruments do not allow unanticipated insights to be gained (Fields, 2002; Jung et al., 2009). Deeper levels of organizations cannot be explored with surveys that assess given organizational categories and important issues may be overlooked in such deductive approaches (Jung et al., 2009; Yauch & Steudel, 2003).

As qualitative and quantitative research methods differ in their strengths and weaknesses, combining them can help overcome some of the trade-off between gaining in-depth qualitative insights and gaining large amounts of quantitative data (Jung et al., 2009). However, research on the complementary use of the two approaches is rare (Yauch & Steudel, 2003). One of the reasons for researchers' reluctance to follow mixed-methods recommendations may lie in the comprehensive effort that the combination of both approaches requires.

Summing up, the criticism regarding extant organizational research methods may originate in the complexities involved in studying organizations on a broad empirical basis via traditional methods such as case studies or surveys. Thus, researchers have called for new ways to examine organizational phenomena (Taras et al., 2009) with alternative methodological approaches that allow to study organizations inductively based on large empirical samples (Berente & Seidel, 2014; Tonidandel et al., 2016). Our study takes up on this call and provides guidance on the use of topic modeling, a specific text mining technique, to address this research gap.

Topic Modeling

With the deluge of user-generated content available on the Internet, more and more social science researchers started to make use of text mining techniques (Janasik et al., 2009). The term text mining refers to computational methods for extracting potentially useful knowledge from large amounts of text data (Fan, Wallace, Rich, & Zhang, 2006; Frawley, Piatetsky-Shapiro, & Matheus, 1992). As a specific form of text mining, topic modeling is a methodological approach to derive recurring themes from text corpora. For researchers, topic modeling represents a novel tool for analyzing large collections of qualitative data in a scalable and reproducible way.

Topic modeling can be understood as an automated method for content analysis and, thus, complements traditional content analysis approaches, characterized by four basic phases, in several ways (Duriau, Reger, & Pfarrer, 2007; Holsti, 1969; Weber, 1990). In the *data collection* phase, topic modelling enables researcher to work with a much larger corpus of documents than would be possible with manual methods; yet, the mechanics behind topic modeling algorithms require a text corpus sufficiently large to produce valid and reliable results. In the *coding* phase, standard topic modeling uses unsupervised machine learning methods that can be compared to exploratory, inductive approaches, in which codes are suggested by the data instead of predefined coding schema (Quinn et al., 2010; Urquhart, 2012); yet, extensions of standard topic modeling allow the algorithm to also be weakly supervised to form topics that contain certain researcher-defined “seed words” (e.g., Jagarlamudi, Daumé III, & Udupa, 2012). In the *content analysis* phase, manual approaches typically use frequency counts and cross-tabulations in combination with a qualitative description of themes emerging from the investigation (Duriau et al., 2007); similarly, topic modeling also combines quantitative analyses (e.g., summary statistics based on document metadata) and qualitative interpretation (based on highly-associated documents and highly-associated words) to analyze content (Quinn et al., 2010). In the *interpretation of results* phase, a strength of topic modelling is to feed identified topics into subsequent statistical analysis methods (e.g., clustering, principal components analysis, regression) (Debortoli et al., 2016). Thus, in that it analyzes text corpora on a large scale to explore potentially new concepts or new concept relations, topic modeling complements existing research methods.

Over the last ten years, probabilistic topic modeling, an unsupervised machine learning method, has received growing attention as a tool for mining large collections of texts in social science research (e.g., Bao & Datta, 2014; Müller, Junglas, vom Brocke, & Debortoli, 2016; Quinn et al., 2010). Probabilistic topic models, like Latent Dirichlet Allocation (LDA), are algorithms that are able to inductively identify topics running through a large collection of documents and to assign individual documents to these topics (Blei, 2012; Blei, Ng, & Jordan, 2003). The underlying idea of such algorithms is rooted in the distributional hypothesis of linguistics (Firth, 1957; Harris, 1954), which posits that “words that occur in the same contexts tend to have similar meanings” (Turney & Pantel, 2010, p. 142). For example, the co-occurrence of words like “sunshine”, “temperature”, “wind”, and “rain” in a set of newspaper articles can be interpreted as a marker for a common topic of these articles, namely “weather”. Hence, topic modelling algorithms like LDA take a relational approach to meaning in the sense that co-occurrences of words are important in defining their meaning and the meaning of topics (DiMaggio, Nag, & Blei, 2013). Due to their emphasis on relationality, topic models are able to capture polysemy and different uses of a word based on the contexts in which it occurs. For example, the term “bank” can co-occur with words like “money” and “credit” in one topic, and “river” and “water” in another topic – indicating two very different meanings for the same word. The focus of topic modelling on analyzing word usage patterns in a corpus to uncover its content is in strong contrast to automated content analysis approaches that try to formalize the semantics of words by means of dictionaries. Rather, the idea behind topic models is in line with the belief of many linguists and philosophers that meanings emerge out of relations between words rather than reside within single words (DiMaggio et al., 2013). For example, Wittgenstein warned against the view that the meaning of a word is defined by the object that it refers to; instead, he famously stated (Wittgenstein, 2010, Section 43): "For a large class of cases – though not for all – in which we employ the word ‘meaning’ it can be defined thus: the meaning of a word is its use in the language."

In contrast to traditional classification or clustering methods, which assign a data point to exactly one category, probabilistic topic models allow documents to belong to multiple categories with a varying degree of membership. Statistically, probabilistic topic models represent documents by a probability distribution over a fixed set of topics, and each topic, in turn, by a probability distribution over a fixed vocabulary of words. The per-document topic distribution is a matrix with one row per document and one column per topic; the cells of the matrix contain probabilities indicating the prevalence of a topic in a document (the probabilities for one document add up to 100%). Similarly, the per-topic word distribution is a matrix with one row per word and one column per topic; the cells of the matrix contain probabilities indicating the relative occurrence of a word in a topic (the probabilities for one topic also add up to 100%). Taken together, the two matrices represent a statistical summary of the contents of the complete document collection (Figure 1, black text).

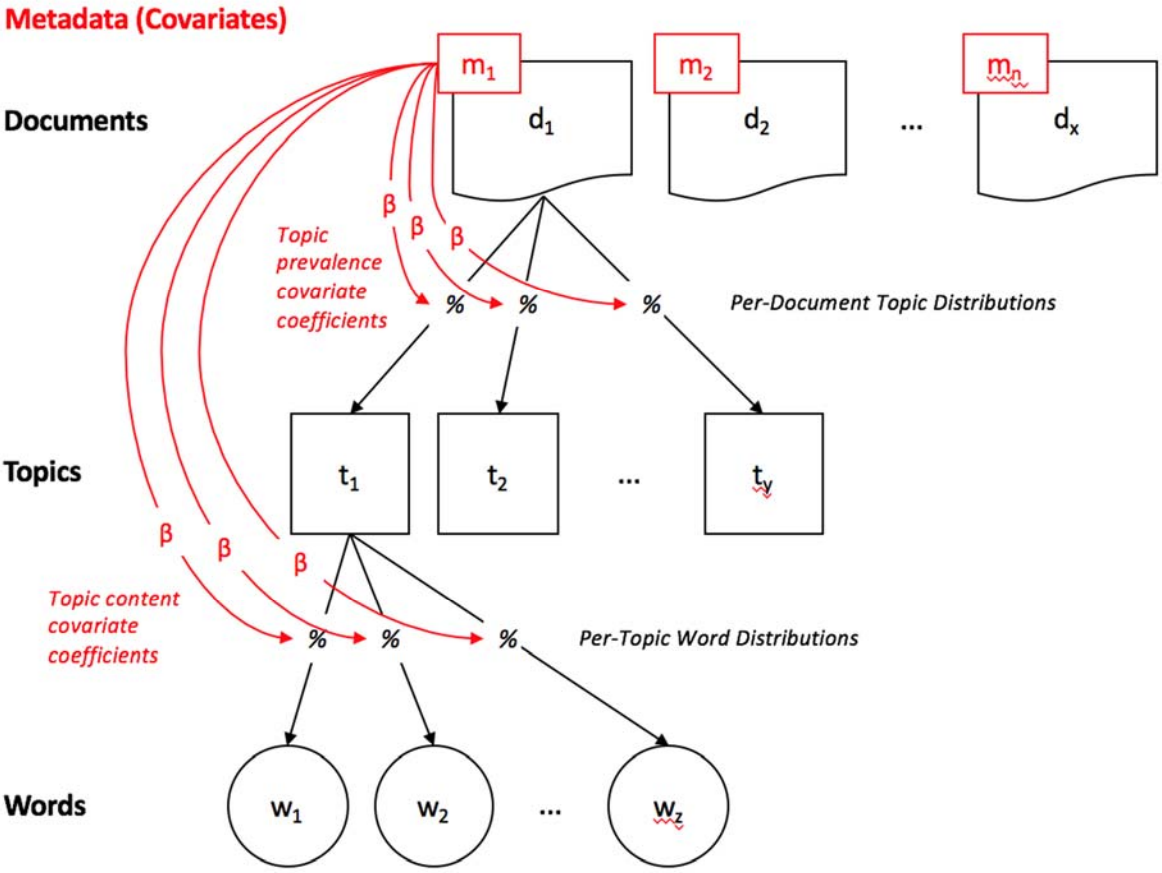

Metadata (Covariates)
Documents
m_1
m_2
m_n
d_1
d_2
...
d_x
β
β
β
Topic prevalence covariate coefficients
%
%
%
Per-Document Topic Distributions
Topics
t_1
t_2
...
t_y
β
β
β
Topic content covariate coefficients
%
%
%
Per-Topic Word Distributions
Words
w_1
w_2
...
w_z

Figure 1 *Schematic overview of structural topic modeling*

While standard probabilistic topic models can provide insights into the topical structure of a whole document collection and individual documents, they cannot easily show how document metadata (e.g., author, date) is related to the content of a document (Roberts, Stewart, & Airoldi, 2016). However, social scientists are often specifically interested in the relationship between document metadata and content; for example, online reviews posted on websites like Amazon or Yelp include a review text and additional metadata, such as a numerical rating. Building on the idea of probabilistic topic models, Roberts et al. (2016) have developed the structural topic model (STM), which allows document metadata to be incorporated into the estimation of the per-document topic distributions (Topic prevalence) and per-topic word distributions (Topic content). Figure 1 illustrates the key idea behind the standard probabilistic topic model (black text), and how the structural topic model (red text) extends this model by allowing the per-document topic distributions and per-topic word distribution to vary as a function of document-level covariates.

STM provides two main advantages compared to other topic modeling approaches. First, from a statistical point of view, considering document metadata as covariates in the topic estimation procedure is likely to improve the fit of the resulting model to the input data. Second, and more important from a social science perspective, it is often the relationship between known covariates and latent topics that is in the main focus of a study, and STM is able to provide this information in the form of coefficient estimates known from generalized linear models.

Next, we outline step-by-step how organizational researchers can apply STM to study the relationships between latent topics and observable metadata of a large collection of documents. For a more formal description of STM and its estimation the interested reader is referred to Roberts et al. (2016).

Step-by-Step Tutorial and Application Example

As in any study, defining a research question is at the outset of a topic modeling study and guides all further data collection and analysis steps. In our application example, we are interested in the following question: What organizational factors influence employee’s perception of corporate culture?

Step 1: Data Collection

Considerations

The data collection phase requires careful reflection on the availability and suitability of data for answering a particular research question using topic modeling.

Regarding *volume*, data collection needs to ensure a sufficiently large size of the text corpus, since the statistics behind topic modeling algorithms require a certain volume of text to produce accurate and meaningful results. The size of the corpus can vary depending on the amount of text files and the length of the single documents. Unfortunately, existing literature to date lacks theoretically-justified guidelines regarding minimal corpus size, but insights from empirical studies can provide some guidelines. Experimental studies suggest that the results of LDA for corpora with few documents (i.e., <100) are very difficult to interpret, even if the documents are long; the interpretability of topic models improves with increased corpus size and stabilizes at around 1,000 documents (Nguyen, 2015). Analyzing metadata of existing topic modeling studies provides additional useful insights. The cloud-based topic modeling service MineMyText.com hosts more than 400 active topic modeling projects conducted by more than 230 researchers. A descriptive analysis of these projects (Table 1) reveals that the average study comprises approx. 38,000 documents and that documents have an average length of 84 words. These statistics suggest that researchers typically use the service to analyze relatively large amounts of short texts, for example, social media posts (Note that the distributions of number of documents and words per document are heavily right-skewed). The length of the document influences both the number of topics included in the texts (see step 3) and the shape of the per-document topic distributions; for corpora with short documents, the distributions are typically dominated by one or two topics, while long documents, that bear more topics, are often characterized by a more uniform topic distribution.

Table 1 *Descriptive statistics of number of documents and words per document of 416 topic modeling projects hosted on MineMyText.com*

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Number of documents	4	231	3,279	38,582	20,000	866,115
Words per document	1	6	14	84	35	5,038

Regarding *representativeness*, data collection needs to consider appropriate sampling to ensure generalizability of the study findings. In particular, researchers need to pay attention to potential systematic biases in their sample. Such biases can occur, for example, when available data only represents a specific part of the population under study (e.g., social media selection bias) or when the data predominantly covers a specific time period. In such cases, researchers need to consider including additional data from other sources, adjusting their study focus, or proceeding and accepting respective limitations. For example, researchers could explore the distribution of documents over time and decide whether they want to exclude or down-sample documents from certain time periods to avoid that one epoch dominates the topic modeling results or crowds other documents out.

Key Decisions

In practical terms, researchers need to take the following key decisions in the data collection phase:

- *What is the best way to gain access to text documents?* While some data is publically available and can be accessed via Application Programming Interfaces (APIs), web crawlers, or file downloads (Debortoli et al., 2016), other data can only be accessed via collaborations (e.g., data from social networks or content sharing platforms). The technical data collection strategy may also influence the time frame that can be captured. Many APIs and web crawlers only provide data snapshots, and not historical data; so, if researchers are interested in a longitudinal dataset, they may need to develop scripts that periodically extract data.
- *What document metadata should be extracted?* Apart from discovering the topical structure of a text corpus (which could also be a study purpose on its own), researchers are often interested in examining the relation of latent topics with other variables. While standard document metadata (e.g., time stamp, author) can

provide insightful descriptive statistics, STM also allows numerical covariates to be included into the estimation procedure (e.g., answers to Likert-scale questions in surveys, or numerical "star" ratings in online reviews) to explore construct relations with further variables. The type of metadata to be included depends on the study purpose and the related research question. Regarding the maximum number of covariates that can be included, issues of overfitting or non-adequate statistical power are far less likely to be a limitation for topic modeling studies as compared to traditional regression analysis of numerical data, since topic modeling typically builds on much larger data samples. So, while the number of covariates that can be included depends on the sample size, a minimum of 1000 documents for conducting topic modeling hardly leads to a limitation in including covariates in practice.

Application Example

We apply structural topic models in the field of organizational culture, because research in this organizational field has explicitly called for new ways to study culture beyond the classic qualitative and quantitative approaches (Taras et al., 2009). Our application builds on data from the online review platform Glassdoor. The platform allows employees to anonymously review organizations in which they have been or are currently employed. These reviews include both text and numerical evaluations. The textual reviews provide insights into organizational aspects that are particularly relevant to employees, while the numerical parts of the review include an evaluation of the organizational culture. In our study, we examine a longitudinal dataset of 428,492 reviews of Fortune 500 companies spanning the period from 2008 to 2015. Access to the dataset was made possible through a collaboration with Glassdoor. Since the text corpus covers only data from employees who are active on this online review platform, the generalizability of our findings is limited to the perceptions of this group. Further, the reviews are not uniformly distributed over time as half of the reviews were written after 2013, which biases our findings towards the recent past. However, the goal of our application example is to illustrate the general use of structural topic modeling for organizational research and provide first insights into what factors influence employees' perception of corporate culture. Thus, we decided to proceed albeit these limitations.

Step 2: Data Preparation

Considerations

The data preparation phase focuses on getting an in-depth understanding of the data and auditing data quality, including potential steps for cleaning data.

With regards to *data understanding*, researchers should spend sufficient time on exploratory data analysis. In particular, descriptive statistics and visualizations help to get a deeper understanding of the data. Such analyses typically allow understanding the distributional properties of available metadata (e.g., age, gender, industry), which also covers understanding the percentage of missing values. However, exploratory analysis can go beyond statistics and include exploring the potential structure of the textual data (e.g., how far text is split into passages with different functions, such as abstract, main text, references).

With regards to *data quality*, textual data, by comparison with numerical data, is characterized by a lack of well-defined data structures and a higher proportion of noise. Hence, text data typically needs to undergo extensive preparatory steps before it can be passed on to the actual topic-modeling algorithm. Data cleaning comprises various steps, for example, removing duplicates or "spam". Table 2 provides an overview of standard data cleaning and preparation steps that researchers can use to prepare their data, to remove noise, and to gradually turn the unstructured textual data into a numerical representation that is amenable to subsequent statistical analysis (Miner, 2012).

Table 2 *Data cleaning and preparation steps to consider*

Step	Specification	Necessity	Considerations
Transforming document formats	Converting raw text data into the required data format for further processing	On demand	Raw text data requires transformation if the original data format cannot be processed by the chosen tool. For example, if documents are stored in individual .txt files, it is often required to consolidate them in a single .csv or .json file that represents the whole corpus before further analysis in R is possible.
Constructing metadata attributes	Deriving metadata variables from given data	On demand	If a specific research question explicitly calls for certain metadata variables that can be derived from the given data set, researchers need to take this step before examining variable relations. An example is calculating the age of a document from its date of creation.
Removing duplicates	Eliminating redundant documents	Mandatory	Large text collections often contain duplicate documents, such as, repeated posts of the same message by the same user on Twitter or repeated copies of the same email newsletter. To avoid biases, these duplicates need to be eliminated (see application example).
Tokenization	Splitting documents into sentences and sentences into words	Mandatory	Tokenization represents a key prerequisite for extracting topics from documents, since topics are derived based on word distributions. At this, researchers have to decide whether to treat strings separated by whitespaces as separate words (uni-grams, e.g., new, york, city), or to allow sequences of two (bi-grams, e.g., New York) or three (tri-grams, e.g., New York City) strings to be treated as composite words.
Stop word removal	Removing common or uninformative tokens	Recommended	Removing standard stop words (e.g., "the", "and"), lists of which are available in all major text mining tools, reduces noise in the topics and is, thus, highly recommended. Beyond, researchers need to decide whether to optionally exclude further customized stop words. We recommend to do so, if words are highly frequent in the text corpus without adding meaning to the single topics (e.g., most topics in our application example included "company", which is to be expected in company reviews).
Normalization	Turning capital letters into lower case	Recommended	Normalization helps to reduce noise in the topics, so that capitalized words and words with lower case do not appear separately in the per-topic word distributions.
Part-of-speech filtering	Identifying and filtering words by their part of speech	Optional	Filtering the text corpus to only retain parts of speech that are important to convey the content of a text (e.g, nouns, verbs, adjectives), can add clarity to the topic models. Yet, when removing function words (e.g., auxiliary verbs, pronouns), important stylistic information of the texts can get lost.
Lemmatizing/ stemming	Reducing a word to its dictionary form or to its stem	Optional	Lemmatizing or stemming also reduces noise in the data and can lead to cleaner topics. However, aggressively reducing different word forms to their common stem or dictionary form may also cause a loss of information, as one cannot distinguish between subtle differences in meaning anymore (e.g., when turning a verb from past to present tense).

Key Decisions

Translating the above considerations into practical questions, the data preparation phase includes the following key decision:

- *Which parts of the text corpus are relevant?* Based on an in-depth understanding of the data and its quality, researchers can take informed decisions regarding which data parts to include in their topic modelling approach. These decisions can comprise selecting relevant text passages, selecting a subset of the data based on metadata, and selecting appropriate data cleaning steps to prepare the data for further analysis.
- *Is it appropriate to differentiate data subsets?* If researchers would like to compare topic-modeling results of subgroups in the data, they need to ensure that the subgroups are sufficiently large for a meaningful comparison. Like in traditional regression analysis, too small subgroups can increase the risk of type II errors (Kelley & Maxwell, 2003). Thus, we refer researchers to well-established guidelines on sample sizes to ensure sufficient statistical power (Scherbaum & Ferreter, 2009; VanVoorhis & Morgan, 2007).

Application Example

After exploring the overall dataset using descriptive statistics and visualizations, we first cleaned the data by eliminating duplicate reviews (using the *duplicated* function in R), reviews that were not written in the English language[a], and reviews with missing values for the numerical corporate culture rating. In addition, we decided to focus on reviews from the 10 industries with the largest number of reviews in the sample to allow for valid comparisons of employees' cultural preferences between industry sectors.

[a] To remove non-English documents, we applied a simple but effective heuristic. We looked up all the words of a document in an English stop word list. If we did not find any match, we considered the document to be non-English. An alternative approach would be to use a language detection service, such as, Google Translation API (https://cloud.google.com/translate/docs/detecting-language).

To execute the document-level natural language processing steps, we used the statistical computing programming language R, or more specifically the *tm* (text mining) and *stm* (structural topic models) packages. Using the *textProcessor* function, we tokenized the documents into single words (uni-grams) and removed standard English language stop words (we used the standard English stop word list of the *tm* package), a small number of custom stop words (e.g., company), words with fewer than three letters, numbers, and punctuation. We decided to work with uni-grams instead of bi- or tri-grams, as the topic modeling algorithm works on the basis of co-occurrences and anyhow clusters the individual elements of composite words (e.g., "New York City") together in the same topics. Hence, using uni-gram tokenization allows for more flexibility (e.g., "New York" vs. "New York City") and results in smaller overall vocabulary sizes[b], without losing essential meaning. Finally, we stemmed all the words and converted all characters to lower case to reduce the dimensionality of the data set. After preprocessing we were left with 295,532 reviews that consist of around 14.8 million words overall (i.e., 35 words per review) and 2,592 unique words.

Step 3: Identification of Topics

Considerations

In this phase, researchers need to carefully reflect on validity and reliability criteria when extracting and interpreting topics from their data.

Regarding *construct validity*, researchers need to ensure that the topics they identify indeed represent what they claim to represent. To date, no commonly accepted methods for measuring convergent and discriminant validity of topics have emerged. Yet, researchers have developed statistical metrics that relate to these quality criteria, namely coherence and exclusivity. Semantic coherence is a measure of the internal coherence of topics and highly correlates with human judgments of topic quality (Mimno, Wallach, Talley, Leenders, &

[b] With tri-gram tokenization the phrase "New York City" would be deconstructed into at least five elements (i.e., "new", "York", "city", "New York", "NewYork City"), instead of three.

McCallum, 2011; Roberts, Stewart, & Tingley). It, thus, serves as an indicator for the validity of the identified topics. Technically, it measures how often the most probable words of a given topic actually co-occur close to each other in the original texts. Exclusivity measures the distinctness of topics by comparing the similarity of word distributions of different topics. A topic is exclusive if its top words are unlikely to appear within the top words of other topics. While semantic coherence focuses on the internal qualities of single topics, exclusivity takes the similarity between different topics of the same model into account. For example, two topics with very similar word distributions might both have high semantic coherence scores if their top words tend to co-occur within documents, but low exclusivity scores indicating that the overall topic model contains redundant information and performs bad in differentiating between concepts appearing in the text corpus.

Semantic coherence and exclusivity are both a function of the number of topics that a topic model contains. Hence, these metrics can be used to guide the selection of an “optimal” number of topics. Apart from performing a search over different topic numbers and comparing coherence and exclusivity of the resulting models, no commonly accepted rules for analytically determining this number for a given corpus have emerged so far. Yet, to guide this search we can again turn to the results of our empirical analysis of the topic modeling projects hosted on MineMyText.com (Table 3). Half of the studies hosted on the platform contain between 10 and 50 topics and the average study works with 35 topics; less than 5 percent of the studies have extracted more than 100 topics. Furthermore, the scatterplot in Figure 2 shows that there is a positive correlation between the number of documents in a corpus and the number of topics that are extracted from that corpus.

Table 3 *Descriptive statistics of number of topics of 416 topic modeling projects hosted on MineMyText.com.*

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Number of topics	2	10	20	35	50	250

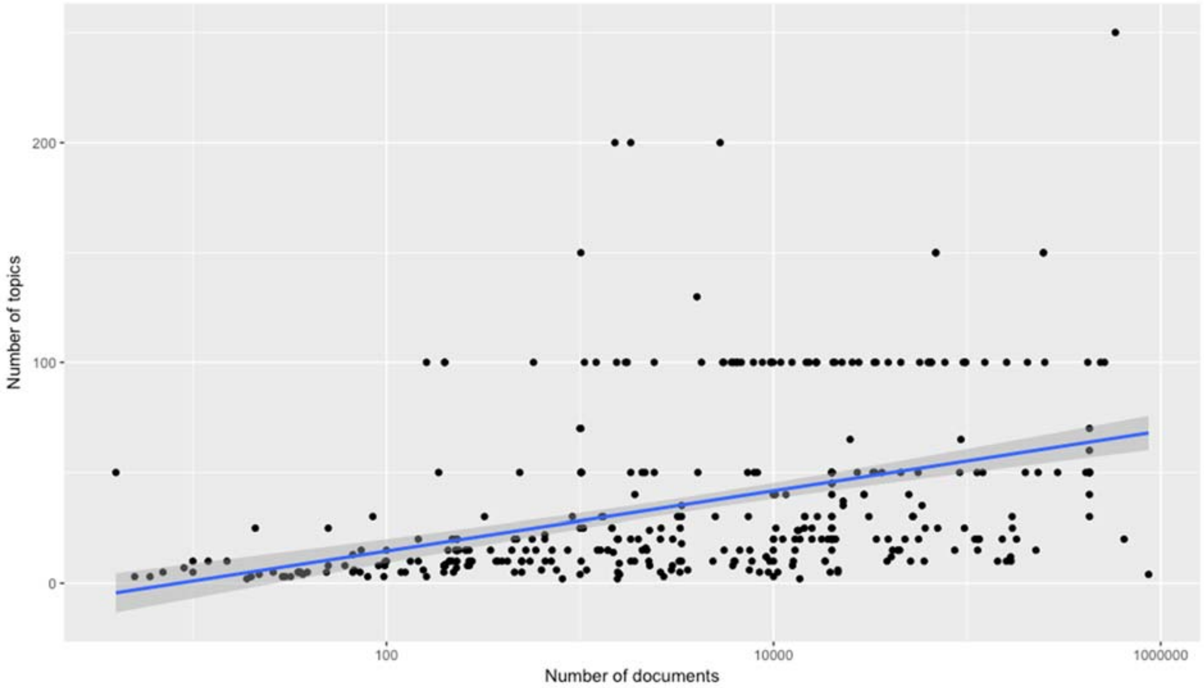


Figure 2 *Relationship between number of documents and number of topics of 416 topic modeling projects hosted on MineMyText.com*

In addition to these quantitative analyses, researchers are advised to qualitatively examine the words and documents that are strongly associated with each topic in order to interpret the meaning of the identified topics. To ensure the *reliability* of this process, researchers should use multiple coders and spend enough resources to reach consensus about the meaning of topics.

Regarding *face validity*, researchers need to carefully examine the relevance of the identified topics for their specific study context. While researchers selected a specific text corpus because they expected it to contain answers to their research question, it is very likely that the topic-modeling algorithm also extracts topics that are not within the scope of the study (e.g., topics describing the structure of text documents with terms such as “abstract”, “introduction”, “conclusion”). Therefore, researchers need to examine how far the extracted topics relate to the phenomenon of interest before proceeding.

Key Decisions

Based on the introduced considerations, the following practical decisions represent important steps in the topic identification phase:

- *How many topics are covered in the data?* For the identification of topics, the user has to specify the number of topics to be discovered from the document collection. Finding the right number of topics means to iteratively analyze the data with various amounts of topics to avoid both overloaded and overlapping topics. Metrics such as semantic coherence and exclusivity can help decide on an appropriate number of topics in the corpus.
- *What are appropriate labels for the identified topics?* Based on words and documents that are highly associated with each topic, researchers need to identify suitable labels that describe the essence of a topic. For topic labeling, researchers typically use multiple coders; cases of disagreement need close attention and require sufficient in-depth discussion until agreement can be reached.

Application Example

In our application example, we first identified the appropriate number of topics to be derived from the text corpus. Although there is no single “correct” number of topics, we tried to follow a reproducible script to arrive at a final decision. First, we examined the average semantic coherence and exclusivity of different topic models ranging from 10 to 100 topics (The upper and lower bounds of this range were motivated by the above shown statistics derived from MineMyText.com). Figure 3 shows the scores for exclusivity at the top and scores for semantic coherence at the bottom.

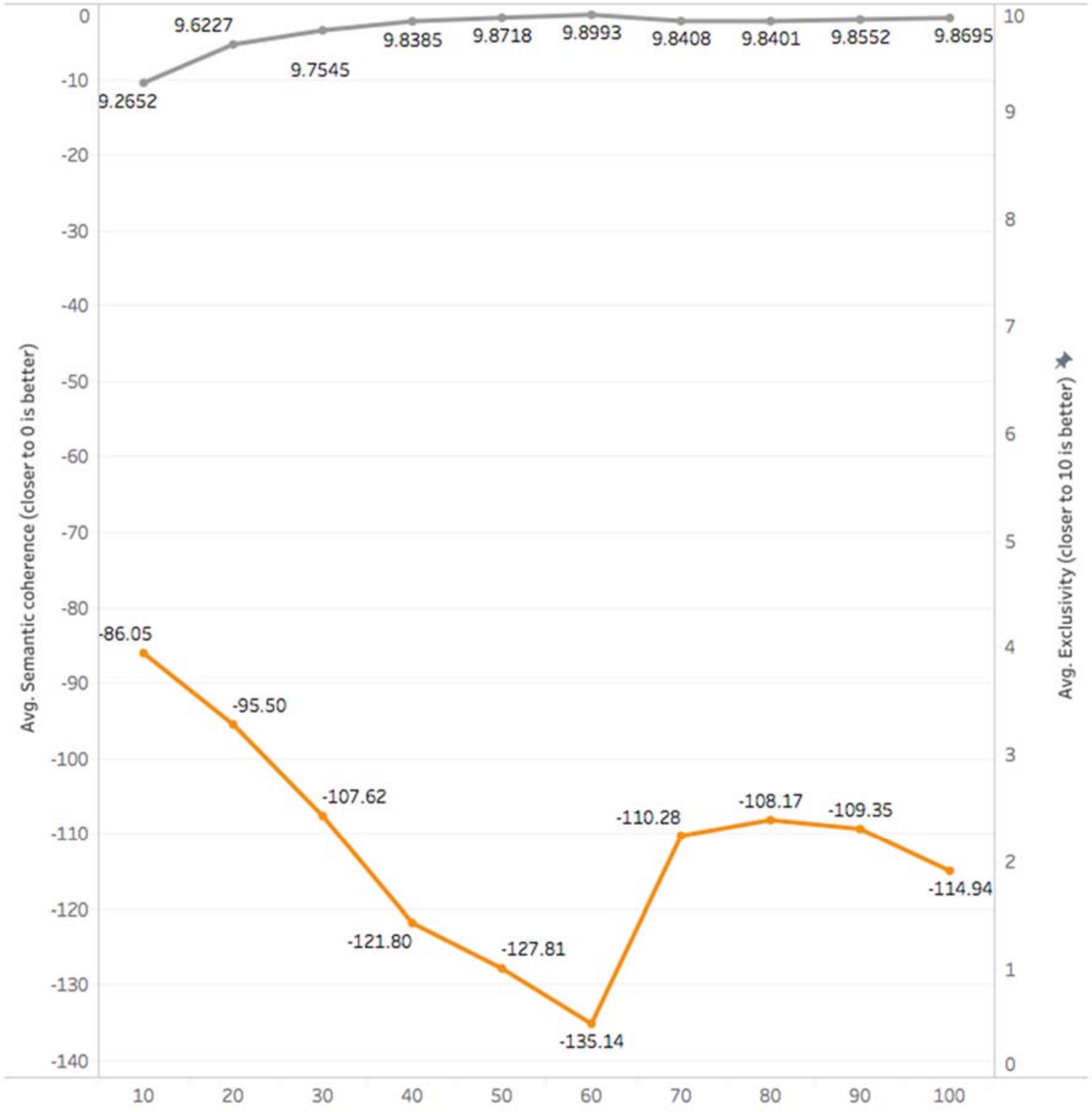


Figure 3 *Semantic coherence and exclusivity of various topic model solutions*

As one can see from the plot, no model dominates the others. While the scores for exclusivity generally improved with an increase in the number of topics, the scores for semantic coherence first declined before improving again for models with more than 60 topics. The reading of the graphs suggested that one opt either for a small (20 or 30 topics) or large (70-100 topics) model, since these yield good scores in both statistical measures.

To complement the quantitative analysis and decide on the appropriate number of topics, we qualitatively examined the interpretability of the different models. We discarded the small topic models (20 or 30 topics), as they merged similar topics and did not clearly differentiate between themes. Examining the large topic models, we found that models with 80 or more topics revealed duplicate topics that differed only in writing style. Thus, we settled on 70 topics as the best option, as the values for semantic coherence and exclusivity did not substantially improve for larger topic models and the qualitative analysis of the 70-topics model revealed clearly interpretable topics.

For the interpreting and labeling of topics, we used the *labelTopics* function of the *stm* package, which produces four different weightings of the most important words per topic (i.e., Highest Probability, FREX, Lift, and Score). The Highest Probability weighting uses the raw per-topic word probabilities; FREX uses a weighted mean of overall word frequency and the exclusivity of words to a topic; Lift uses the frequency of a term in other topics to emphasize words that are specific to a topic; and, similarly, Score uses the log frequency of terms in other topics to identify words that are specific to a topic (Roberts et al.). Two researchers independently coded each topic by examining the four word rankings for each topic and examining reviews highly associated with each topic. While we assessed all of the four word weightings, we paid most attention to the Highest Probability and Score metrics, because the other metrics ranked very rare terms (e.g. typos) high. For example, the Score words that represent Topic 67 include “employe”, “appreci”, “respect”, and “care” (note that terms are stemmed) and reviews that are closely associated with the topic refer to how far management appreciates employees. Hence, we labelled this topic “employee appreciation”.

Table 4 shows additional examples of topics, highly associated words, extracts from the most probable reviews, and the topic labels generated by the researchers. Consolidating the individual coding results, the topic labeling exercise revealed high inter-coder reliability with an inter-coder agreement of 86 percent and an average Kappa value of 0.86 (Light, 1971; Moore & Benbasat, 1991). In cases where the labels differed between researchers, the researchers discussed their findings until they reached a consensus about a label. Table A.1 in the appendix provides an overview of all 70 topics.

Table 4 *Exemplary topic labeling*

Topic ID	Highly associated terms	Exemplary review text	Topic label
4	Highest Prob: help, will, work, alway, need, peopl, everyon FREX: help, succeed, eager, pharmacist, answer, question, pharmaci Lift: havochir, unitychalleng, preparedlisten, endedif, againcustom, althiugh, instructionsinform Score: help, succeed, question, alway, will, everyon, answer	"People are very friendly and are always willing to help you with anything you need whether they work on your team or not."	help
54	Highest Prob: employe, manag, door, polici, open, concern, listen FREX: revolv, door, suggest, digniti, concern, retali, treatment Lift: consequens, epilepsi, excelretali, accomplis, packagek, worker-suggest, wouild Score: employe, door, polici, concern, open, listen, treat	"When management says they are listening they really are not. Take heed to the open door policy."	listening
46	Highest Prob: job, stress, easi, work, high, good, secur FREX: repetit, stress, bore, secur, easi, volum, monoton Lift: family-friendi, simpleit, goodpushi, timeask, freedomgreat, networkgood, constantlyfear Score: job, stress, easi, secur, bore, high, repetit	"High stress work environment with nonstop fire drills and very little reward."	stress

Next, we had to determine which topics are relevant to our research question. Thus, we first individually examined the content of all 70 topics and marked those topics that are not meaningful to answer our research question; we, then, consolidated our individual topic examinations. We found that most of the topics we identified represent organizational factors that employees value (or dislike) in companies. They include, for example, the following topics: "work-life balance", "flexibility", "employee treatment", "lying", and "home office". However, we also found topics important to employees that did not refer to organizational factors. For example, some topics refer to a specific company (e.g., "Apple", "Barnes and Noble", "Medco", "Starbucks"), a certain employee function or type (e.g., "sales", "software development team", "store managers", "store employees", "internship"), or particular industry-specific vocabulary (e.g., "consulting", "corporate clients", "store management"). Further topics described organizations on an overall level, and thus did not refer to specific organizational aspects (e.g., "great place to work", "best company"). Additionally, some topics referred to the general vocabulary of company reviews (e.g. "general review vocabulary" included terms like "pros", "worst", "none"; "work vocabulary" included terms like "work", "people", "employe"; "time vocabulary" referred to terms like "long", "term", "hour", "short"). Also, some factors were generated that did not reveal meaningful topics. We excluded all topics irrelevant to our research purpose from the further analysis. Following this examination, we focused on 45 topics that allowed us to address our research question and studied their relation to the employees' perception of organizational culture.

Step 4: Relationships between Topics and Other Variables

Considerations

The core feature of STM is that it allows to examine the relationship between the identified topics and document-level covariates.

Regarding *topic-covariate relationships*, researchers should have comprehensive background knowledge on established concept relations in their field of study to derive meaningful hypotheses about potential relationships in their data. STM allows for complex relationships between latent topics and covariates to be specified, including, for example, interaction effects or non-linear relationships using regression splines.

Regarding *coefficient estimates*, STM follows the logic of generalized linear models and offers a number of functionalities to calculate the uncertainty of the coefficient estimates. In addition, the *stm* package offers functionalities to visualize complex interaction effects or non-linear relationships using partial dependence plots, which hold all variables, except the ones under consideration, at their sample medians.

Key Decisions

In order to specify and interpret topic-covariate relationships, researchers need to take the following key decisions:

- *Which specific topics should be included in the analysis?* Ex ante analysis, researchers need to decide which topics should be included in the model specification. While researchers may have extracted a broad number of topics from their corpus, they might want to only model a subset of all possible covariate-topic relationships. For example, examining reasons for customer satisfaction based on product reviews may yield a very broad number of generally relevant topics; yet, for further analysis, researchers may only want to focus on topics relating to customer service and select these topics accordingly.
- *Which topics can be included in the interpretation?* Ex post analysis, researchers need to consider significances when selecting insightful topic-covariate relations for interpretation. The *stm* package automatically reports standard errors, t-statistics, and p-values for all coefficient estimates.

Application Example

We apply the *estimateEffect* function of the *stm* package to examine the importance of the identified organizational factors for employees' perception of organizational culture, that is, the relation of the identified topics to numerical company ratings.

We used two document-level metadata variables as covariates for our analysis: the numerical star rating of the organizational culture dimension and the industry sector the organization belongs to. Regarding culture star ratings, company reviewers can numerically assess their satisfaction with the "culture and values" of the company they review. For this purpose, reviewers have the option to rate the organization and express their satisfaction regarding "culture and values" with one (low satisfaction) to five (high satisfaction) stars.

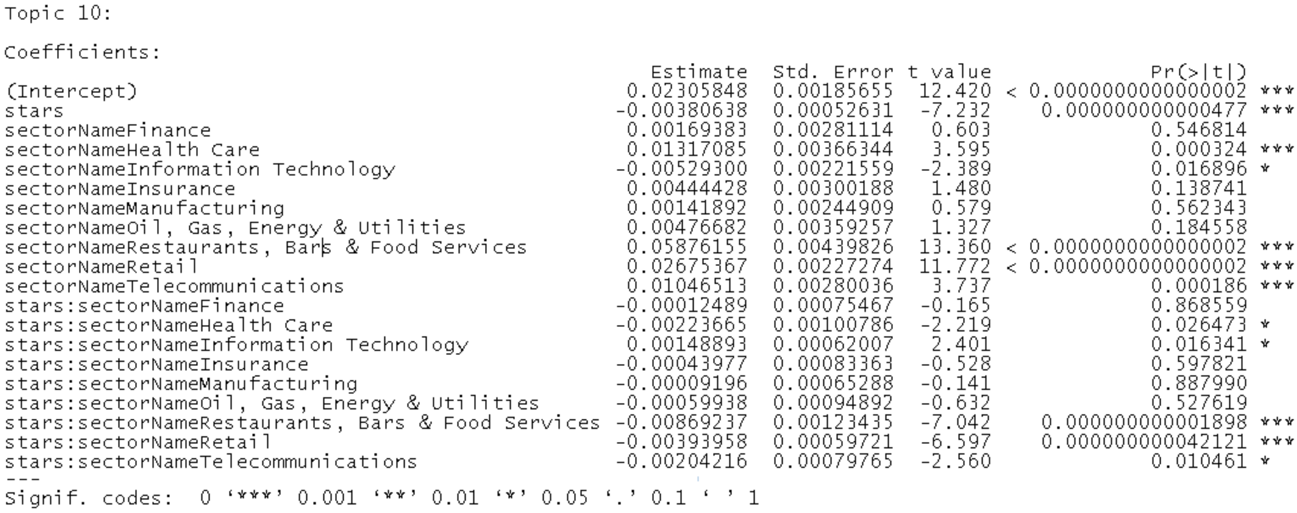
Regarding industry sector, the datasets included a specification of the industry that the reviewed companies belong to. We included this variable in our analysis to account for potential differences between sectors.

We fit an estimation model to the data based on the following generic model:

$$Prevalence_{ij} \sim \beta_0 + \beta_1 * Rating_i + \beta_2 * Industry_i + \beta_3 * Industry_i * Rating_i + \varepsilon_i,$$

where *i* indexes the *i*th review and *j* indexes the *j*th topic, $Prevalence_{ij}$ is the matrix of topic prevalence values derived from the STM analysis, β_0 is the intercept, $Rating_i$ is the numerical company rating of a review, β_1 its respective coefficient, $Industry_i$ is a categorical variable for industry sector, β_2 its respective coefficient, and ε_i is the standard error term. Besides modeling the main effects of culture star rating and industry sector, we also considered interactions between these two covariates. This is captured by the coefficient estimate β_3 and allows for an industry-moderated effect of culture star ratings on topic prevalence.

For each topic, we received an output specifying all estimates, standard errors, and p-values relating to the topic. Figure 4 shows the output for topic 10, “employee treatment”. We can see a negative association between employee treatment and the number of stars given to rate organizational culture. The negative estimate for stars shows that reviews with high star ratings are most likely not covering employee treatment as a topic, while reviews with low star ratings most likely cover employee treatment and report negatively on interactions between managers and employees.



```
Topic 10:

Coefficients:
                                                     Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)                                        0.02305848  0.00185655  12.420 < 0.0000000000000002 ***
stars                                             -0.00380638  0.00052631  -7.232    0.000000000000477 ***
sectorNameFinance                                  0.00169383  0.00281114   0.603             0.546814    
sectorNameHealth Care                              0.01317085  0.00366344   3.595             0.000324 ***
sectorNameInformation Technology                  -0.00529300  0.00221559  -2.389             0.016896 *  
sectorNameInsurance                                0.00444428  0.00300188   1.480             0.138741    
sectorNameManufacturing                            0.00141892  0.00244909   0.579             0.562343    
sectorNameOil, Gas, Energy & Utilities             0.00476682  0.00359257   1.327             0.184558    
sectorNameRestaurants, Bars & Food Services        0.05876155  0.00439826  13.360 < 0.0000000000000002 ***
sectorNameRetail                                   0.02675367  0.00227274  11.772 < 0.0000000000000002 ***
sectorNameTelecommunications                       0.01046513  0.00280036   3.737             0.000186 ***
stars:sectorNameFinance                           -0.00012489  0.00075467  -0.165             0.868559    
stars:sectorNameHealth Care                       -0.00223665  0.00100786  -2.219             0.026473 *  
stars:sectorNameInformation Technology             0.00148893  0.00062007   2.401             0.016341 *  
stars:sectorNameInsurance                         -0.00043977  0.00083363  -0.528             0.597821    
stars:sectorNameManufacturing                     -0.00009196  0.00065288  -0.141             0.887990    
stars:sectorNameOil, Gas, Energy & Utilities      -0.00059938  0.00094892  -0.632             0.527619    
stars:sectorNameRestaurants, Bars & Food Services -0.00869237  0.00123435  -7.042     0.00000000001898 ***
stars:sectorNameRetail                            -0.00393958  0.00059721  -6.597     0.00000000042121 ***
stars:sectorNameTelecommunications                -0.00204216  0.00079765  -2.560             0.010461 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4 *Exemplary output for each topic*

The industry estimates show that employees from some sectors are more likely to discuss topic 10 in their reviews than employees from other sectors. For example, employees from the sectors “Restaurants, Bars & Food Services”, “Retail”, and “Health Care” are more likely to write about employee treatment than employees from other sectors. The interaction estimates show that reviews with high star ratings differ in their coverage of topic 10 depending on the sector the employee works in. For example, employees from the IT sector are more likely than employees from other sectors to write about employee treatment when they rate the organizational culture highly, while employees from the retail sector are less likely than employees from other sectors to address employee treatment in their reviews when they rate the organizational culture highly.

Table A.2 in the Appendix provides an overview of the estimates of all topics in our analysis. To illustrate the relation between various topics and culture perceptions across industries, Table 5 visualizes exemplary estimates.

Table 5 *Exemplary estimates on the relation between topics, culture, and industries*

ID	Topics	Culture estimates* (indicating positive or negative association with topics)	Industry estimates* (indicating more or less presence of topics in reviews of an industry)		Interaction estimates* (indicating positive or negative association of culture estimates with topics depending on the industry)	
			IT	Retail	Stars:IT	Stars:Retail
26	Career opportunities	**0.0089**	0.0066	-0.0052	-0.0035	**-0.0067**
21	Work-life balance	**0.0040**	0.0058			**-0.0030**
6	Flexibility	**0.0040**				-0.0013
49	Brain drain	**-0.0027**	**0.0125**	**-0.0152**		0.0023
52	Laid back atmosphere	**-0.0034**	-0.0065	**-0.0149**	*0.0011*	**0.0030**
57	Poor management	**-0.0050**		**-0.0181**		**0.0028**

* **bold font** = significant at 0.001 level, normal font = significant at 0.01 or 0.05 level, *italic font* = significant at 0.1 level, missing value = not significant

The comparison shows that employees emphasize different topics depending on their perception of organizational culture. For example, employees who value the existing culture, in general, report positively on career opportunities, work-life balance, and flexibility; while employees who dislike the corporate culture, in general, report negatively on the management, a laid-back atmosphere, and brain drain. In the IT sector, employees are more likely to point out topics like career opportunities or brain drain than employees from the retail sector. A laid-back atmosphere is less likely to be a topic in the IT and retail sectors compared to other sectors. However, when employees from the IT and retail sectors highly value their corporate culture, they are more likely to refer to a laid-back atmosphere (topic 52) than employees from other sectors; however, they are less likely to refer to career opportunities (topic 26).

Figure 5 graphically illustrates the overall effects that relate to the topics “laid-back atmosphere” (topic 52) and “career opportunities” (topic 26). The graphic on the left side shows the relation between the topic prevalence and the culture star rating of the two topics in the IT sector, while the graphic on the right side refers to the retail sector.

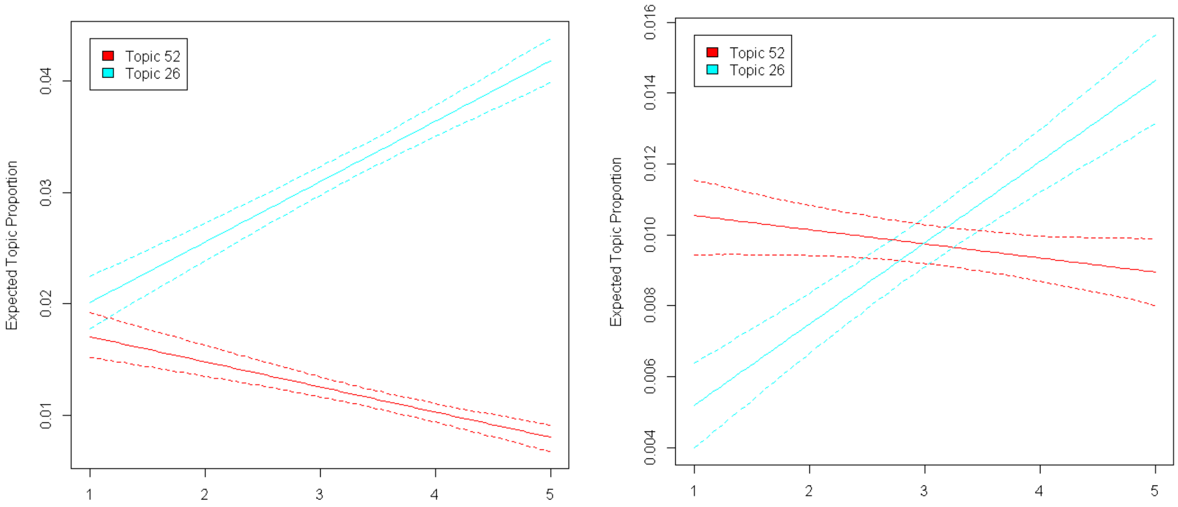


Figure 5 *Effects of culture star rating on topic prevalence (IT (left), retail (right) sector)*

In both sectors, the general shape of the curves is similar (increase of the reference to career opportunities the more employees are satisfied with the corporate culture; decrease of mentioning a laid-back atmosphere the more highly the organizational culture is rated). However, career opportunities are mentioned more often in the IT sector, with an increase in valuing the corporate culture, compared with the retail sector; furthermore, the decrease in discussing a laid-back atmosphere when people highly value their organizational culture is stronger in the IT sector than in the retail sector.

Step 5: Interpretation of Findings and Engagement with Existing Literature

Considerations

Interpreting the findings and comparing them to extant work requires researchers to reflect on potentially new insights through the exploratory research approach.

Regarding the *exploratory nature* of topic modelling, researchers need to pay particularly close attention to the most dominant findings, especially to those aspects that are surprising in a given context. Typically, topic modelling reveals a broad number of topics on an overarching theme and, while many may be relevant, only a few manage to attract attention, most likely because they are extreme or unexpected.

Regarding *new concepts or concept relations*, researchers should reflect their findings against the background of existing research. A comparison with extant state-of-the-art research may confirm existing findings or reveal that additional concepts or concept relations are relevant in a certain context.

Key Decisions

Closely related to the above considerations, researchers need to take the following key decisions in this phase:

- *Which of my findings are "interesting" and which insights do they bring?* Researchers should assess their results to identify potentially new topics and topics with a very strong positive or negative relation to other variables. A comparison between data subsets may help to identify interesting findings that yield insights for research and practice.
- *Which prior research results do my findings fit to, extend, or contradict the most?* Researchers should build on potentially interesting findings and examine how far these go beyond existing work. For example, assigning the identified topics to conceptual categories of established frameworks provides an opportunity to reveal novel insights in a certain research domain.

Application Example

Those topics that relate most positively or negatively to the perception of organizational culture provide valuable insights for organizational research and practice. Regarding factors that are strongly positively associated with culture perceptions, we can identify topics referring to career options ("career opportunity", "career development", "advancement opportunities"), topics referring to formal rewards ("salary raise", "benefits"), topics referring to the work environment ("work-life balance", "flexibility"), and topics referring to social aspects ("great people", "help"). Regarding factors that are highly negatively associated with culture perceptions, we can observe topics referring to management (“poor management”, “management layers”), topics referring to formal rewards (“paycheck”, “bonus”, “low wage”), topics referring to corporate aspects (“company strategy”, “brand name”, “brain drain”), and topics referring to social aspects (“lying”, “listening”, “employee appreciation”, “employee treatment”, “laid back atmosphere”).

Comparing topics that are positively and negatively associated with culture perceptions shows: Employees who highly value their corporate culture emphasize career options provided by the organization, while employees who dislike the organizational culture point out deficits regarding social aspects of their work. These social aspects primarily address the relevance of a respectful and open working atmosphere among employees (e.g., “employee appreciation”, “listening”). More surprisingly, the factor “laid back atmosphere” is also among those social aspects with the largest relevance regarding how employees perceive the corporate culture. At first sight, it seems rather unexpected that a “laid back atmosphere” is negatively associated with employees’ culture perception, but the results might suggest that employees do not appreciate a too easy-going, casual, unconstrained work environment.

Generally, these findings confirm the relevance of existing frameworks that apply to the organizational culture level, such as the GLOBE dimensions (House, Hanges, Javidan, Dorman, & Gupta, 2004; Jung et al., 2009), which include, for example, performance orientation (to which topics like “bonus” and “salary raise” relate), power distance (to which topics like “management layers” or “organizational hierarchy” relate), and humane orientation (to which topics like “employee treatment” or “caring” relate). Other organizational culture dimensions that our results confirm are the ones by Chatman and Jehn (1994), which include people orientation (to which topics like “great people” or “employee appreciation” relate), outcome orientation (to which topics like “paycheck” or “benefits” relate), and easygoingness (to which the topic “laid back atmosphere” relates).

Yet, our findings also yield additional insights that complement prior research. For example, previous research studied easygoingness as a neutral dimension that distinguishes cultures across industries (Chatman & Jehn, 1994). Our findings suggest that this dimension is negatively associated with culture. In other words, our findings provide hints regarding *desired* organizational cultures in various industries and, thus, extend previous findings that focused on a purely *descriptive* analysis of culture.

Examining industry-specific results in greater detail, we find that the emphasis of topics differs across industries. To illustrate this, we differentiate again between topics that are generally negatively or positively associated with culture perceptions. Regarding the former, we can observe that employees from the insurance, telecommunications, and finance sectors emphasize “lying” much more than employees from other industries. This observation may indicate that lying negatively dominates organizational cultures in these industries. Another example, (bad) “employee treatment”, seems to be much more severe in the sectors “Restaurants, Bars, and Food Services”, “Retail”, and “Health Care” than in other industries, which may indicate that companies in these sectors consider social factors less relevant for their organizational cultures than companies in other industries. Regarding factors that are generally positively related to how employees perceive their corporate culture, we can see, for example, that IT sector employees emphasize “salary raise”, “career development”, and “work-life balance” at the same time, indicating that employees from this sector value not only career options, but also a work environment that provides flexibility. In contrast, employees from the health care sector emphasize the factor “help”, which means they highly appreciate a work environment where they can rely on their colleagues.

The comparison of industry differences provides insights that can support future research in further specifying differences between organizational cultures in various industries (Chatman & Jehn, 1994; Gordon, 1991; Phillips, 1994). Particularly, future research should distinguish descriptive from desired culture dimensions to develop a more detailed understanding of organizational culture profiles in different industries and derive meaningful recommendations for culture development in practice. While aggregate analyses of organizational culture, such as ours, can provide insights on as-is and to-be culture profiles in specific industry sectors, they naturally only describe general tendencies of current and desired cultures. These insights give guidance regarding what to consider in more fine-granular examinations of organizational culture, for example, in studies that focus in-depth on a particular organization.

The exemplary overall and industry-specific results show the potential of topic modeling for gaining insights on organizational factors that matter to employees and their relation to corporate culture. Overall, our topic modeling example and related analyses provide insights on how to use this methodological approach to quantify the relation of topics important to employees and employees' perceptions of organizational culture. It illustrates how to determine dimensions that require management attention and, potentially, interventions. In broader terms, our approach shows the applicability of topic modeling to gain insights on organizational phenomena.

Discussion

Advantages of Topic Modeling in Organizational Research

Our analysis of company reviews to identify topics that matter to employees' perception of organizational culture serves as a first illustration of how to use topic modeling for organizational research. We outline several advantages of this technique for organizational research.

Compared to established organizational research methods, unsupervised topic modeling provides specific advantages for organizational research. While the use of widely applied methods like surveys or case studies represents a trade-off between examining large amounts of data in a quantitative study and gaining in-depth insights through a qualitative study (Jung et al., 2009), topic modeling covers the advantages of both, that is, it allows to examine large amounts of qualitative, textual data. In particular, the advantage over questionnaires, which have to date practically been the only means for large-scale organizational assessments, is essential, because topic modeling does not require the predefining of dimensions for an analysis. While survey research examines organizations deductively based on predetermined scales and operationalized constructs, topic modeling makes it possible to study inductively what employees feel to be most relevant to mention about an organization. Therefore, topic modeling combines the benefits of quantitatively examining culture on a large scale with the benefits of an inductive qualitative method approach (Berente & Seidel, 2014; Tonidandel et al., 2016).

As a new methodological approach, topic modeling complements existing organizational research methods through novel ways of gaining insights into large text corpora. Since the main focus of topic modeling lies in inductively examining the content of large amounts of texts, it generally supports theory building rather than theory testing, that is, it generally supports exploratory rather than explanatory research. However, STM, as a specific topic modeling technique, allows not only potentially new concepts to be examined, but also the relation of emerged topics with other variables of interest (Roberts et al.). Therefore, studies applying STM can not only explore large data sets, but also explain relations between new and established concepts in organizational research.

Another key advantage of topic modeling over established organizational research methods lies in the nature of the data that can be analyzed. Data such as the online reviews in our example is generated with no research purpose; it is not created through interview questions or questionnaires but it represents so-called “naturally occurring data” (Müller et al., 2016) and may, thus, be less biased through social expectations that can influence data which is created in research situations. Considering the already existing and rapidly increasing amount of textual data that is available from various kinds of organizations, our application example provides insights on the huge potential that topic modeling bears for organizational research.

Apart from the methodological contribution of our approach to general organizational research, the application example also contributes to organizational culture research in particular. Since topic modeling has not been applied to organizational culture research before, our application addresses the existing call for research on new ways to study culture (Taras et al., 2009) and gives guidance to researchers on how to approach the suggested new way to culture research. In addition, our findings provide insights with regard to factors that influence differences in organizational culture perceptions between industries. The identified topics that are positively and negatively related to employees' perceptions of organizational culture not only provide insights regarding prevailing cultures, but also regarding desired cultures. Future research can build on these first insights by developing a more detailed understanding of organizational culture profiles in different industries.

Limitations of Topic Modeling

While topic modeling provides many advantages for organizational research over established organizational research methods, the approach obviously also has certain limitations.

Topic modeling does not automatically yield new valid constructs or extracts significant relationships at the push of a button. The algorithms used to extract topics from textual data rather have a supporting role; in fact, researchers need to take many decisions throughout all the steps of their study, which range, for example, from selecting appropriate algorithms for their study purpose to interpreting and labeling topics. Thus, topic modeling does not fully automate the identification and measurement of constructs, but requires subjective interpretations through the researcher. For example, metrics are available that help researchers examine the validity of the identified constructs; yet, topic identification still requires manual coding and interpretation. As a research method, topic modeling is, therefore, “in the middle” between rather measurement-centric quantitative and rather interpretation-centric qualitative methods. Since the identification of constructs and their relation to other variables is at the core of topic modeling, the method clearly contributes to addressing research questions in exploratory research. It provides new opportunities to theorize on established, but also on new constructs that may be identified from large text corpora. Since large text corpora were previously not accessible for exploratory research on a large scale, topic modeling represents a new and complementary approach to existing research methods.

Another limitation refers to the nature of the data used for topic modeling. The textual data that serves as a basis for topic modeling most probably contains not only topics relevant to the field of study. Since the data is typically not generated for a specific research purpose, it very likely includes topics that are not related to the research focus. Therefore, researchers need to explore how far the identified topics relate to their field of study. While this task may be prone to subjective biases, such manual tasks are also typical in qualitative research and several techniques exist to mitigate subjectivity, for example, by involving several researchers.

Frequently, the data also comes along with potential biases, such as in online reviews, which may be biased through potentially spurious reviews (Nan Hu, Bose, Koh, & Liu, 2012), and which may be biased towards extreme opinions and contain only few moderate ratings (N. Hu, Zhang, & Pavlou, 2009). Research has suggested techniques for detecting and removing fake reviews that bias datasets, such as the elimination of duplicate reviews (Jindal & Liu, 2008; Liu & Zhang, 2012). Utilizing such techniques, as we did in our application example, helps researchers to mitigate potential biases in their data.

Furthermore, the generalizability of the findings is limited to the data set. In our application example, we examine data of Fortune 500 companies only, and approximately 80% of the Glassdoor visitors are from the United States[c]. Thus, our findings are limited to these companies and geographic regions. While such limitations are typical for all types of data analysis methods, researchers may mitigate them by triangulating the findings with additional data.

Application Fields in Organizational Research

Topic modeling offers a broad spectrum of application possibilities in organizational research. Considering the vast amount of textual data that is generated on a daily basis, organizational research should leverage the potential of various types of texts for gaining insights on organizational phenomena. While our application example used user-generated text from an online company review platform, future research may also analyze text from other sources, such as company-internal employee platforms or company-internal documents. Also, research may focus on textual data beyond the corporate sphere. Social media data, for example, but also data from daily news or research publications may represent insightful sources for future organizational research.

Table 6 provides an overview of exemplary application fields for topic modeling that can inspire organizational research. For example, text mining social network data may allow one to complement insights on job attitudes and the prediction of withdrawal behavior (Lebreton, Binning, Adorno, & Melcher, 2004). Further, organizational research may apply topic modeling to examine both job characteristics and competence profiles, similarly to how researchers in the information systems discipline have applied topic modeling (Gorbacheva, Stein, Schmiedel, & Müller, 2016; Müller, Schmiedel, Gorbacheva, & vom Brocke, 2014). In addition, organizational research may be inspired by applications in finance, where

[c] https://www.alexa.com/siteinfo/glassdoor.com

researchers used topic modeling to extract textual risk disclosures from annual reports to quantify their effect on the investors' risk perceptions (Bao & Datta, 2014); and in the area of marketing and public relations, where researchers used topic modeling for mining consumer perceptions about brands from social media data (e.g., Pournarakis, Sotiropoulos, & Giaglis, 2017). Finally, organizational research may apply topic modeling for examining existing literature and analyzing the development of topics over time (e.g., Blei, 2012).

Table 6 *Topic modeling application examples*

Type of data	Data source	Application fields	Domain
Internal company data	Social networks (e.g., Yammer)	Employee concerns, job stress, organizational culture	Human resources
External company data	Job/employee platforms (e.g., Monster.com; LinkedIn.com)	Job characteristics, competence profiles	Human resources
External company data	Annual reports (e.g., Form 10-K)	Risk disclosures	Finance
Public data	Social media (e.g., Twitter, Blogs)	Brand image	Marketing, Public Relations
Research articles	Literature databases (e.g., EbscoHost, Google Scholar)	Literature review	All

Conclusion

The purpose of our research was to demonstrate the utility of topic modeling as a new approach to studying organizational phenomena. We suggest embracing the methodological advances from other fields in organizational research. While we show the advantages of topic modeling over traditional qualitative and quantitative organizational research methods, we argue that future research should not apply topic modeling to organizational research as an ultimate remedy to the limitations of currently used research methods. However, future research should consider text-mining approaches, such as topic modeling, as complementary to well-established organizational research methods. We believe that the combination of various techniques allows organizations to be studied from new perspectives that help to gain novel insights into the field.

References

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science, 60*(6), 1371-1391.

Berente, N., & Seidel, S. (2014). *Big data & inductive theory development: Towards computational grounded theory?* Paper presented at the 20th Americas Conference on Information Systems (AMCIS), Savannah, USA.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(1), 993-1022.

Chatman, J. A., & Jehn, K. A. (1994). Assessing the relationship between industry characteristics and organizational culture: How different can you be? *Academy of Management Journal, 37*(3), 522-553.

Currall, S. C., Hammer, T. H., Baggett, L. S., & Doniger, G. M. (1999). Combining qualitative and quantitative methodologies to study group processes: An illustrative study of a corporate board of directors. *Organizational Research Methods, 2*(1), 5-36.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text mining for Information Systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems, 39*(7), 110-135.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics, 41*(6), 570-606.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational Research Methods, 10*(1), 5-34.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM, 49*(9), 76-82.

Fields, D. L. (2002). *Taking the measure of work: A guide to validated scales for organizational research and diagnosis*. Thousand Oaks: Sage Publications.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955 *Studies in Linguistic Analysis* (pp. 1-32). Oxford: Philological Society.

Frawley, W., Piatetsky-Shapiro, G., & Matheus, C. (1992). Knowledge discovery in databases: An overview. *Al Magazine, 13*(3), 57-70.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods, 16*(1), 15-31.

Gorbacheva, E., Stein, A., Schmiedel, T., & Müller, O. (2016). The role of gender in business process management competence supply. *Business & Information Systems Engineering, 58*(3), 213-231.

Gordon, G. G. (1991). Industry determinants of organizational culture. *Academy of Management Review, 16*(2), 396-415.

Hardy, B., & Ford, L. R. (2014). It's not me, it's you: Miscomprehensions in surveys. *Organizational Research Methods, 17*(2), 138-162.

Harris, Z. (1954). Distributional structure. *Word, 10*(23), 146-162.

Holsti, O. R. (1969). Content analysis. In L. Gardner & E. Aronson (Eds.), *Handbook of social psychology* (pp. 596-692). Reading, MA: Addison-Wesley.

House, R. J., Hanges, P. J., Javidan, M., Dorman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Thousand Oaks, London, New Delhi: Sage.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems, 52*(3), 674-684.

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM, 52*(10), 144-147.

Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). *Incorporating lexical priors into topic models.* Paper presented at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France.

Janasik, N., Honkela, T., & Bruun, H. (2009). Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods, 12*(3), 436-460.

Jindal, N., & Liu, B. (2008). *Opinion spam and analysis*. Paper presented at the Conference on Web Search and Web Data Mining, Palo Alto, USA.

Jung, T., Scott, T., Davies, H. T. O., Bower, P., Whalley, D., McNally, R., & Mannion, R. (2009). Instruments for exploring organizational culture: A review of the literature. *Public Administration Review, November|December*, 1087-1096.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8*(3), 305-321.

Lebreton, J. M., Binning, J. F., Adorno, A. J., & Melcher, K. M. (2004). Importance of personality and job-specific affect for predicting job attitudes and withdrawl behavior. *Organizational Research Methods, 7*(3), 300-325.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76*(5), 365-377.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. C. Aggarwal & C. X. Zhai (Eds.), *Mining Text Data*: Springer.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., . . . Aiden, E. L. (2011). Quantiative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing semantic coherence in topic models.* Paper presented at the Conference on empirical methods in natural language processing, Edinburgh.

Miner, G. (2012). *Practical text mining and statistical analysis for non-structured text data applications.*: Academic Press.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

Morey, N. C., & Morey, R. V. (1994). Organizational culture: The management approach. *National Association for the Practice of Anthropology Bulleting, 14*(1), 17-26.

Müller, O., Junglas, I., vom Brocke, J., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems, 25*(4), 289-302.

Müller, O., Schmiedel, T., Gorbacheva, E., & vom Brocke, J. (2014). Toward a typology of business process management professionals: Identifying patterns of competences through latent semantic analysis. *Enterprise Information Systems*.

Nguyen, L. (2015). *Topic modeling with more confidence: A theory and some algorithms*. Paper presented at the Pacific-Asia Knowledge Discovery and Data Mining, Ho Chi Minh City.

Phillips, M. (1994). Industry mindsets: Exploring the culture of two macro-organizational settings. *Organization Science, 5*(3), 384-402.

Pournarakis, D. E., Sotiropoulos, D. N., & Giaglis, G. M. (2017). A computational model for mining consumer perceptions in social media. *Decision Support Systems, 93*, 98-110.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science, 54*(1), 209-228.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the Amercian Statistical Association, 111*(515), 988-1003.

Roberts, M. E., Stewart, B. M., & Tingley, D. stm: R package for structural topic models. *Journal of Statistical Software*, 1-42.

Sackmann, S. (2001). Cultural complexity in organizations: The value and limitations of qualitative methodology and approaches. In C. I. Cooper, S. Cartwright & P. C. Earley (Eds.), *The International Handbook of Organizational Culture and Climate* (pp. 143-163). Chichester, UK: Wiley.

Schein, E. H. (1990). Organizational culture. *American Psychologist, 45*(2), 109-119.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*(2), 347-367.

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics, 6*(3), 239-251.

Taras, V., Rowney, J., & Steel, R. (2009). Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management, 15*, 357-373.

Tonidandel, S., King, E. B., & Cortina, J. M. (2016). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*.

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141-188.

Urquhart, C. (2012). *Grounded theory for qualitative research: A practical guide*. Thousand Oaks: Sage.

Van Maanen, J. (1979). Reclaiming qualitative methods for organizational research: A preface. *Administrative Science Quarterly, 24*(4), 520-526.

VanVoorhis, C. R. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology, 3*(2), 43-50.

Weber, R. P. (1990). *Basic content analysis* (2nd ed.). Newbury Park, CA: Sage Publications.

Wittgenstein, L. (2010). *Philosophical investigations*: John Wiley & Sons.

Yauch, C. A., & Steudel, H. J. (2003). Complementary use of qualitative and quantitative cultural assessment methods. *Organizational Research Methods, 6*(4), 465-481.

Appendices

Table A.1 *Topics resulting from structural topic modeling*

Topic ID	Topic label	Highly associated terms
1	Medco	medco, esi, stock, employe, purchas, bla, disabl
2	Performance measurement	metric, measur, perform, score, technician, scorecard, survey
3	Training	train, program, agent, comput, proper, trainer, provid
4	Help	help, succeed, question, alway, will, everyon, answer
5	Perks	free, food, cabl, coffe, gym, drink, phone
6	Flexibility	life, balanc, flexibl, great, environ, con, good
7	Organizational hierarchy	corpor, ladder, larg, climb, cultur, american, headquart
8	Starbucks	partner, starbuck, coffe, barista, tip, drink, store
9	Salary hike	good, work, salari, hike, life, onsit, balanc
10	Employee treatment	like, treat, feel, didnt, break, slave, crap
11	Benefits	great, benefit, sale, commiss, leadership, chang, bank
12	Hiring	hire, good, contractor, peopl, big, contract, money
13	Sales	sale, product, sell, rep, commiss, custom, servic
14	General review vocabulary	none, list, absolut, worst, pros, walmart, second
15	Leadership	leader, leadership, industri, senior, visionari, market, substanc
16	Talent attraction/retention	talent, innov, retain, attract, market, conserv, engin
17	Lying	dont, know, tell, say, anyth, just, fire
18	Career development	growth, career, develop, advanc, opportun, path, limit
19	Work hours	hour, schedul, holiday, week, shift, day, weekend
20	Discount card	card, discount, credit, maci, merchandis, store, cloth
21	Work-life balance	worklif, work-lif, life, compens, balanc, work, maintain
22	Management-dependent atmosphere	depend, upon, vari, heavili, frown, locat, may
23	Bonus	rais, year, bonus, increas, annual, salari, perform
24	Store management	team, member, manag, store, schedul, etl, target
25	Great people	can, think, sometim, great, realli, work, cant
26	Career opportunities	opportun, lot, differ, intern, career, learn, larg
27	*topic indeterminate*	keep, happi, promis, work, chang, toe, pile
28	Software development team	project, amazon, softwar, engin, team, develop, design
29	Best company	best, ive, ever, one, world, buy, compani
30	Management layers	layer, mani, chief, indian, tier, much, overhead
31	*topic indeterminate*	know, job, need, someon, secur, what, peopl
32	Home office	home, day, work, depot, night, hrs, week
33	Working together	togeth, act, great, page, peopl, work, stack
34	Low wage	pay, wage, low, decent, minimum, rais, hour
35	*topic indeterminate*	grow, dead, weight, sale, beat, within, cdw
36	Glass ceiling	blah, ceil, glass, nonsens, search, whose, pit
37	*topic indeterminate*	filler, endur, frito, paccar, deterior, aquisit, understat
38	*topic indeterminate*	littl, databas, harsh, extraordinarili, unmatch, jobveri, dishonesti
39	Company strategy	group, direct, clear, strateg, strategi, execut, chang

40	Health benefits	health, match, insur, tuition, medic, reimburs, pension
41	Full-/Part-time	full, time, part, posit, spent, convert, intern
42	Apple	appl, older, younger, retail, age, generat, women
43	Consulting	consult, citi, live, eastman, area, san, town
44	Internship	learn, lot, internship, busi, stuff, summer, gain
45	Employee supervision	supervis, annoy, shock, protocol, emot, moodi, paperwork
46	Stress	job, stress, easi, secur, bore, high, repetit
47	Great place to work	work, great, place, fun, nice, good, con
48	Safety	safeti, elev, injuri, e-mail, day, mainten, four
49	Brain drain	smart, peopl, microsoft, layoff, great, cultur, polit
50	Caring	care, take, employe, patient, number, good, work
51	Strategic focus	focus, strategi, global, custom, cost, strong, result
52	Laid back atmosphere	back, laid, stab, big, pictur, aecom, bread
53	Brand name	brand, name, ibm, recognit, resum, morgan, usa
54	Listening	employe, door, polici, concern, open, listen, treat
55	*topic indeterminate*	cant, truth, will, donât, letter, dish, one
56	Barnes and Noble	book, booksel, nobl, ventur, barn, joint, nook
57	Poor management	manag, upper, middl, poor, senior, level, micro
58	Advancement opportunities	great, advanc, room, benefit, opportun, move, environ
59	Store employees	associ, store, cashier, hour, payrol, floor, guest
60	Corporate clients	firm, client, booz, allen, bank, govern, advisor
61	Time vocabulary	long, term, time, hour, short, period, take
62	Store managers	store, manag, district, upper, payrol, assist, hour
63	Dress code	dress, casual, code, space, relax, cubicl, jean
64	*topic indeterminate*	job, told, month, week, call, day, anoth
65	Process orientation	core, balanc, process, rank, level, perform, system
66	Customer service	custom, servic, rude, sale, store, regist, cashier
67	Employee appreciation	employe, treat, appreci, job, loyal, respect, care
68	*topic indeterminate*	gain, experi, career, leadership, engin, role, skill
69	Paycheck	money, call, day, dish, will, save, make
70	Work vocabulary	work, peopl, manag, good, great, promot, employe

Table A.2 *Culture, industry, and interaction estimates*

Topics		Culture estimates*	Industry estimates* (left figure) and culture-industry interaction estimates (right figure)																	
			Finance		Healthcare		Information Technology		Insurance		Manufacturing		Oil, Gas, Energy & Utilities		Restaurants, Bars & Food Services		Retail		Telecom-munications	
17	Lying	**-0.0074**	0.0065				*-0.0058*	*0.0015*	0.0100						**-0.0216**	**0.0048**			0.0090	
57	Poor management	**-0.0050**			-0.0102	0.0024							0.0158	-0.0031	**-0.0206**	0.0033	**-0.0181**	**0.0028**		
67	Employee appreciation	**-0.0047**	-0.0116	0.0021			-0.0086		-0.0107	0.0022	-0.0089	0.0018	-0.0102	*0.0022*						
10	Employee treatment	**-0.0038**			**0.0132**	-0.0022	-0.0053	0.0015							**0.0588**	**-0.0087**	**0.0268**	**-0.0039**	**0.0105**	-0.0020
52	Laid back atmosphere	**-0.0034**	**-0.0092**	0.0019	**-0.0134**	**0.0028**	-0.0065	*0.0011*	**-0.0116**	0.0024	**-0.0082**	0.0015	-0.0090	0.0021	**-0.0176**	**0.0031**	**-0.0149**	**0.0030**	**-0.0116**	**0.0025**
69	Paycheck	**-0.0031**			**0.0218**	**-0.0047**			**0.0193**	**-0.0034**	0.0063	*-0.0013*	0.0097	*-0.0020*	-0.0095	0.0021			**0.0347**	**-0.0073**
23	Bonus	**-0.0028**	-0.0078	*0.0014*				-0.0016			-0.0072		**-0.0131**	*0.0017*	**-0.0219**	0.0031	**-0.0154**	0.0021	**-0.0167**	0.0026
49	Brain drain	**-0.0027**	*-0.0057*		**-0.0118**	0.0025	**0.0125**		-0.0091						**-0.0178**	0.0025	**-0.0152**	0.0023	-0.0098	
54	Listening	**-0.0025**			0.0085		-0.0061	0.0015	0.0091						*-0.0065*		0.0052			
39	Company strategy	**-0.0020**				*0.0016*	0.0066								**-0.0135**	0.0021	**-0.0103**	0.0017	*-0.0048*	0.0018
53	Brand name	**-0.0020**	**-0.0138**	*0.0011*	**-0.0216**	0.0020	**-0.0104**		**-0.0187**	0.0017	**-0.0116**		**-0.0177**		**-0.0223**	0.0020	**-0.0211**	0.0018	**-0.0205**	0.0018
34	Low wage	-0.0018			0.0093		*-0.0044*								**0.0466**	**-0.0067**	**0.0233**		0.0067	
30	Management layers	-0.0011	*-0.0034*		-0.0064		0.0047		-0.0047						-0.0083		**-0.0081**	*0.0010*	*-0.0047*	
40	Health benefits		0.0067			0.0027			**0.0106**				0.0092		**-0.0128**	0.0021	**-0.0095**	*0.0010*	0.0074	
51	Strategic focus		**-0.0127**		**-0.0152**				**-0.0162**		-0.0059		**-0.0136**		**-0.0206**		**-0.0184**		**-0.0151**	*0.0014*
19	Work hours				0.0130		*-0.0047*				0.0063				**0.0367**		**0.0188**	**0.0029**		0.0021
50	Caring				**0.0184**	-0.0026														
2	Performance measurement																		**0.0103**	-0.0016
5	Perks							0.0013							**0.0382**	**-0.0034**			**0.0133**	
66	Customer service		**0.0151**	**-0.0025**	**0.0151**										**0.0265**		**0.0292**	-0.0013	**0.0222**	**-0.0028**
7	Organizational hierarchy										*0.0040*									
48	Safety																			
16	Talent attraction/retention		-0.0047		**-0.0123**		0.0056		-0.0060						**-0.0134**		**-0.0128**		**-0.0098**	

15	Leadership						*0.0036*								-0.0048					
45	Employee supervision																			
36	Glass ceiling																			
20	Discount card																**0.0225**			
33	Working together																			
32	Home office										*-0.0041*						-0.0051			
3	Training				*0.0049*				**0.0105**											
22	Management-dependent atmosphere																			
41	Full-/Part-time																			
63	Dress code																			
65	Process orientation																			
46	Stress		0.0079												**0.0196**				*0.0059*	
12	Hiring			-0.0017			0.0051	-0.0015		-0.0019		-0.0017				-0.0026	*-0.0038*	-0.0013		
58	Advancement opportunities	**0.0025**		*0.0016*		**0.0044**				0.0028						**0.0044**		**0.0055**		**0.0033**
4	Help	**0.0025**			**0.0116**												*0.0039*			
25	Great people	**0.0038**						0.0016							0.0063					
11	Benefits	**0.0038**	0.0081	0.0022						**0.0050**									0.0095	**0.0045**
6	Flexibility	**0.0040**													*0.0048*	-0.0018		-0.0013		*-0.0011*
21	Work-life balance	**0.0040**				-0.0023	0.0058									-0.0029		**-0.0030**		-0.0018
18	Career development	**0.0050**					0.0059	**-0.0030**		-0.0021		0.0023				**-0.0041**		**-0.0027**		
9	Salary raise	**0.0058**		**-0.0034**	*-0.0059*	-0.0033	**0.0195**			**-0.0044**	*0.0055*	**-0.0030**	0.0075	-0.0033		**-0.0061**	-0.0057	**-0.0050**		**-0.0030**
26	Career opportunities	**0.0089**		-0.0030	-0.0084	-0.0028	0.0066	**-0.0035**		-0.0027					*-0.0064*	**-0.0074**	-0.0052	**-0.0067**	*-0.0060*	-0.0028

* **bold font** = significant at 0.001 level, normal font = significant at 0.01 or 0.05 level, *italic font* = significant at 0.1 level, missing value = not significant