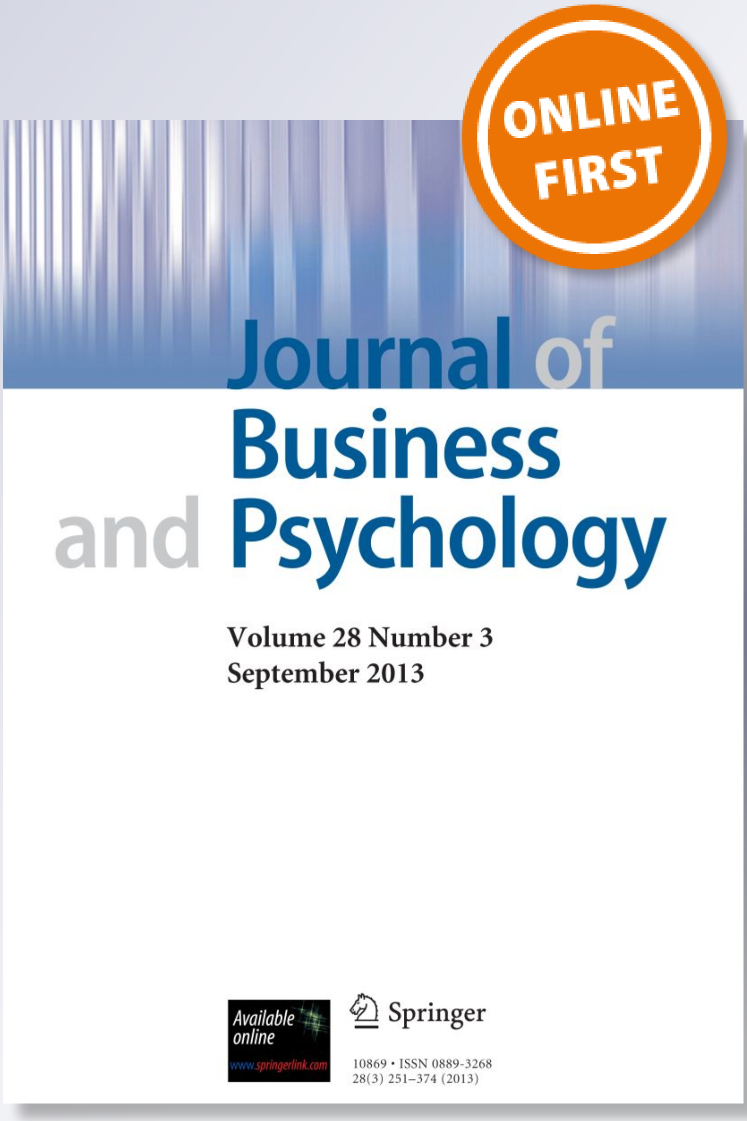
A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App)

George C. Banks, Haley M. Woznyj, Ryan S. Wesslen & Roxanne L. Ross

Journal of Business and Psychology

ISSN 0889-3268

J Bus Psychol
DOI 10.1007/s10869-017-9528-3



ONLINE FIRST

Journal of Business and Psychology

Volume 28 Number 3
September 2013

Available online
www.springerlink.com

Springer

10869 • ISSN 0889-3268
28(3) 251–374 (2013)

Journal of Business and Psychology
https://doi.org/10.1007/s10869-017-9528-3

ORIGINAL PAPER

A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App)

George C. Banks[1] · Haley M. Woznyj[2] · Ryan S. Wesslen[3] · Roxanne L. Ross[4]

Abstract

In recent decades, the amount of text available for organizational science research has grown tremendously. Despite the availability of text and advances in text analysis methods, many of these techniques remain largely segmented by discipline. Moreover, there is an increasing number of open-source tools (R, Python) for text analysis, yet these tools are not easily taken advantage of by social science researchers who likely have limited programming knowledge and exposure to computational methods. In this article, we compare quantitative and qualitative text analysis methods used across social sciences. We describe basic terminology and the overlooked, but critically important, steps in pre-processing raw text (e.g., selection of stop words; stemming). Next, we provide an exploratory analysis of open-ended responses from a prototypical survey dataset using topic modeling with R. We provide a list of best practice recommendations for text analysis focused on (1) hypothesis and question formation, (2) design and data collection, (3) data pre-processing, and (4) topic modeling. We also discuss the creation of scale scores for more traditional correlation and regression analyses. All the data are available in an online repository for the interested reader to practice with, along with a reference list for additional reading, an R markdown file, and an open source interactive topic model tool (topicApp; see https://github.com/wesslen/topicApp, https://github.com/wesslen/text-analysis-org-science, https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/R4W7ZS).

Keywords Text analysis · Topic modeling · Structural topic modeling · Thematic analysis · Content-analysis · Dictionary analysis · Natural language processing

A substantial amount of work in organizations is captured through text (McKenny, Aguinis, Short, & Anglin, 2016). Organizational stakeholders write emails, develop company websites, post on social media, upload resumes, publish press releases, and post new job descriptions (Roberts, Stewart, & Tingley, 2014a). With advancements in technology, the volume and the types of digitized text is proliferating (Blei, 2012; Grimmer, 2015). Internetlivestats.com, a website which counts Internet activity in real time, illustrates this trend. For example, on April 27, 2017, there were billions of internet users (n = 3,621,841,480), websites (n = 1,184,444,633), and active Facebook users (n = 1,897,478,984), not to mention the number of emails, tweets, and blog posts amassing daily. There is no shortage of text data to be analyzed. In response, computer-aided text analysis methods are being used to manage impressive volumes of text (Blei, 2012). As an example of the practical application of such analyses, some companies are applying these methods by analyzing employee posts on internal social media platforms to understand how their employees are feeling so that they can improve working conditions (e.g., Waddell, 2016).

We dedicate this article to Jared Borns for his insight, patience, and guidance in the data collection process. We thank the three reviewers at *Journal of Business and Psychology* as well as John Batchelor, Wenwen Dou, Katherine Frear, Tiffany Gallicano, Andy Loignon, Aaron McKenny, Bob Muenchen, Ernest O'Boyle, Jeremy Short, Anne Smith, Allison Toth, and Christopher Whelpley for their feedback on previous versions of the manuscript and our analysis. The article was pre-registered via the Open Science Framework (https://osf.io/g9wjy/?view_only=045606c4e42843f7b3d131de6d0908d0).

✉ George C. Banks
gbanks3@uncc.edu

[1] Department of Management, Belk College of Business, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA

[2] Department of Management, Longwood University, Farmville, VA, USA

[3] Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

[4] Department of Organizational Science, University of North Carolina at Charlotte, Charlotte, NC, USA

Published online: 11 January 2018

The primary goal of this article is to describe available tools and apply an example of topic modeling in an organizational science context. The article is a part of the *Journal of Business and Psychology*'s *Method Corner series*, which have covered topics ranging from relative weights analyses (Tonidandel & LeBreton, 2015) and polynomial regression (Shanock, Baran, Gentry, Pattison, & Heggestad, 2010) to other more recent topics, such as common method variance (Williams & McGonagle, 2016), mixed-effects models (Bliese, Maltarich, & Hendricks, 2017), and modeling temporal interaction dynamics (Lehmann-Willenbrock & Allen, 2017).

The current academic literature on the analysis of text is somewhat disconnected among fields, and few articles have integrated these disparate techniques. Inductive, exploratory analysis of qualitative data, in general, is used in areas of academia such as communication studies, sociology, and, to some extent, in management and political science. Each of these fields, along with computer science, has idiosyncrasies when it comes to unpacking meaning from text. The degree of automation versus human coding in the analysis process can vary significantly between fields, and some fields tend to favor certain techniques over others. The result is that some areas, such as computer science, focus on text using big data and minimum human input (Roberts, Stewart, et al., 2014a) while others, such as communication studies, focus on smaller bodies of text with a greater emphasis on human coding (Strauss & Corbin, 1998).

Methods for analyzing text thus vary by discipline and feature different strengths and limitations. The techniques used are largely contained within the individual disciplines, with limited crossover despite their widespread applicability. One contribution of the present article is to bridge the gap between literature areas (e.g., computer science, communication studies, management) by introducing topic models to organizational scientists as well as comparing it to more conventional text analysis tools (e.g., content and thematic analysis). A simplified taxonomy of extant techniques for text analysis is provided; it illustrates how techniques vary with respect to the degree of automation and descriptions of each method to give the reader a roadmap of available tools as well as an understanding of how new methods compare with other techniques.

Finally, the current article also outlines specific guidance for computer-aided text analysis. Text analysis is an iterative process that involves multiple judgment calls. Decisions such as how much text to collect, what words should be excluded from analysis, and how many topics to interpret all impact how insightful the results are (Roberts et al., 2014). Clarity on how to make these decisions and the impact of these decisions on the outcomes of the analyses have not yet been clearly articulated for an organizational science audience. Consequently, researchers may be largely unaware of assumptions that are made when cleaning and pre-processing a dataset (Denny & Spirling, 2017). We outline important steps and considerations to aid those with little (or advanced) knowledge on how to perform text analysis.

To complement the advice on best practices and the step-by-step user's guide provided, we created online resources for additional support to perform these analyses. We offer links and resources through GitHub and dataverse (see https://github.com/wesslen/topicApp, https://github.com/wesslen/text-analysis-org-science, https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/R4W7ZS) including an R markdown file and topicApp (and a Frequently Asked Questions section). We also make available a detailed reading list for both novice and advanced researchers of text analysis also provided in the aforementioned links. By leveraging readily available and free resources, readers not only have an outline how to perform text analysis but also supplementary materials to do so conveniently.

We proceed by first reviewing the literature with the goal of covering key features regarding different types of text analysis, highlighting the major strengths and weaknesses of each. Next, an illustrative example is included to provide a step-by-step guide regarding how to use one method in particular, topic modeling. This illustrative example includes a description of basic terminology and major assumptions. Also included are assumptions and decisions that researchers need to consider for text analysis, as well as recommendations to help guide researchers with these decisions. In sum, the goal of the present article is to provide researchers, who have varying levels of familiarity with text analysis, an exhaustive step-by-step guide to using the technique, an example of how it can be used, and supplementary resources to assist in the process.

Literature Review

The following section organizes and discusses various text analysis techniques used across disciplines (for a detailed reading list see the GitHub link previously mentioned). Text analysis takes many forms, each with its own assumptions, advantages, and limitations (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010). Specific techniques used within a particular discipline are largely unknown to or unused by the other disciplines, which may lead to insufficient awareness about the idiosyncrasies of each technique. Figure 1 provides a categorization of the most popular text analysis techniques used across the disciplines, organized based on the degree of computer automation. Such a categorization provides an organizing framework that allows researchers to see the tools available to analyze text data. Similarly, Table 1 summarizes the major points regarding each technique. Researchers can make informed decisions about which text analysis tool is best suited for the type of data and the research questions they have. We briefly review the techniques presented in Fig. 1 and Table 1 to situate our main focus, topic modeling, among other popular

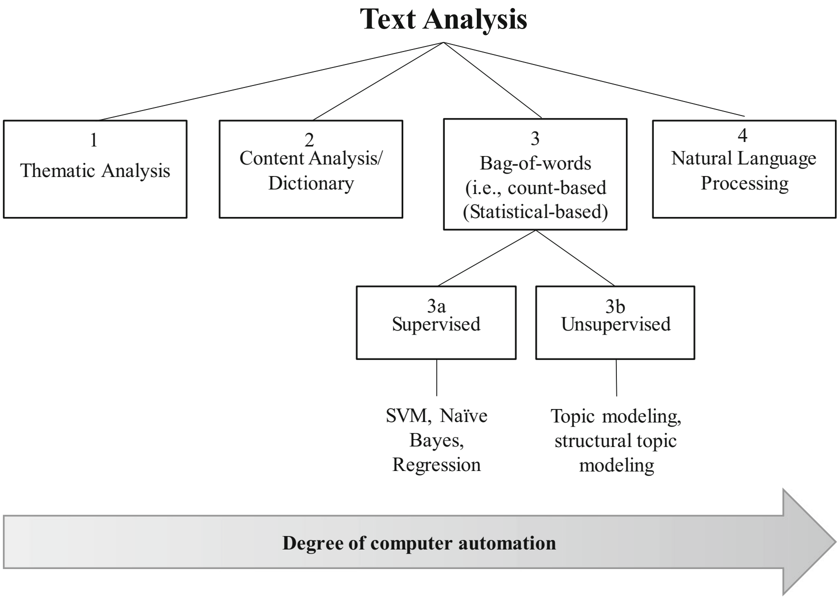


Fig. 1 Categorization of common text analysis techniques

text analysis techniques. We note that some statements in the descriptions below as well as in the figure and table are generalizations presented for simplicity purposes.

Thematic Analysis Looking first at box 1 in Fig. 1, thematic analysis is an interpretivist approach to text analysis commonly associated with grounded theory methodologies (Baumer, Mimno, Guha, Quan, & Gay, 2017). The purpose of grounded theory and thematic analysis is to create and refine theory based on the sensemaking and meaning that people assign to their own worlds. Consequently, data are most commonly in the form of transcribed interviews, notes from participant observations, and archived text, including documents, websites, and emails (Glaser & Strauss, 1967; Suddaby, 2006). Thematic analysis is frequently used in the organizational sciences and communication studies (Gioia, Corley, & Hamilton, 2013; Strauss & Corbin, 1998).

Thematic analysis involves an iterative process where the researcher develops a series of codes and categories that emerge from the text. In general, the categories are not known before analysis begins, except when seeking to refine theory; in such instances, data analysis involves a constant comparison between the literature and the data. The researcher starts with the participants' own language (called "first-order codes" or "open coding"; Gioia et al., 2013; Strauss & Corbin, 1998), and later groups similar codes together into categories (called "second-order codes" or "axial coding"; Strauss & Corbin, 1998). Computer software such as NVivo and ATLAS.ti can help to facilitate the organization of such codes and categories, though the categorization of the text is typically reliant on operational definitions of categories that are derived from human coding; thus, the amount of computer automation is often low or sometimes nonexistent.

While thematic analysis allows researchers to understand phenomena and concepts that are not particularly well known, there are a number of costs associated with the technique (Quinn et al., 2010). Because the coding is neither semi- nor fully automated and categories are unknown, it is time intensive to read through, categorize, and interpret the text, which makes it nearly impossible to analyze large amounts of data (e.g., 100,000 transcript pages). In addition, thematic analysis can be vulnerable to the errors and the biases of the researchers (Antonakis, 2017). Further, thematic analysis tends to require more substantive knowledge about the domain of interest than other approaches (Quinn et al., 2010).

Content Analysis/Dictionary-Based Methods Content analysis and other dictionary-based methods (Fig. 1, box 2) are often conducted by taking the frequency counts of words and/or phrases in a particular text (Reinard, 2008; Short, Broberg, Cogliser, & Brigham, 2010). Consequently, with dictionary-based methods, qualitative data can be used to answer more quantitatively oriented research questions because the text is often reduced to frequency word counts (McKenny et al., 2016; Reinard, 2008). The amount of human knowledge needed beforehand, as well as computer automation, ranges depending on whether the word lists emerge from the data or whether the word lists were determined a priori using theory. Similar to thematic analysis, computer software can assist in the content analysis process. Programs, like DICTION, automatically score the text using a categorization dictionary (i.e., determining themes based on words or n-grams rather than operational definitions). Other programs, like NVivo or ATLAS.ti, can be used similarly to thematic analysis, where the coding and categorizing is done by hand with help from the software to organize the data.

Table 1 A comparison of text analysis approaches across scientific disciplines

Text analysis technique	Description	Strengths	Weaknesses	Commonly used disciplines	Most common software
Thematic analysis	An iterative process where researchers develop a series of codes and categories based on operational definitions	Uses participants' own words or constructs from the extant literature; provides insight into phenomena that are not well understood	Time and labor intensive, particularly with large amounts of data; subject to researcher biases and errors; does not allow for theory testing	Communication studies; sociology; management	None; NVivo; ATLAS.ti
Content analysis/dictionary analyses	Creates word/phrase frequency counts of text	Allows for testing of quantitative research questions	Requires subject matter expert validation when using a priori word lists	Management	LIWC; DICTION; NVivo; ATLAS.ti
Bag-of-words (count-based)	A group of techniques that are used to reduce and simplify text; ignores word order, which allows for statistical properties	Simple, statistical properties, scalable	Omits word order leading to poor semantic meaning		
Supervised models (e.g., SVM, naive Bayes, logistic regression)	Researcher knows input and output; system creates an algorithm to map the connection between the two variables	Allows for a priori specifications; can produce marginal effects (e.g., rank words by most likely to predict outcome)	Requires substantial pre-existing knowledge; Potential overfitting due to many factors (words)	Computer science; political science	R, Scikit-Learn in Python
Unsupervised models (e.g., k-means clustering, topic models)	Clusters/factors/topics are automatically and empirically created	Facilitates the analysis of large amounts of data without manual annotations (labels or dictionaries)	Require interpretation; tuning parameters;	Computer science; political science	Java (Mallet), SAS; Python (gensim), R (topicmodels, stm)
Natural language processing	Models how humans understand and process language (e.g., sentiment analysis, part-of-speech tagging, translation); word order is important	Computer automated; attempt to analyze semantic meaning	Many tasks have high error rates; low level programming (Java, Python); some tasks are "black boxes"	Computer science; marketing; psychology	Java; Python (nltk)

Bag-of-Words and Natural Language Processing The final two techniques are the bag-of-words (BOW; i.e., count-based; Fig. 1, box 3) and natural language processing (NLP; Fig. 1, box 4) approaches to text analysis. These text analysis tools can involve machine learning whereby computer programs learn various text classification and categorization tasks and may complete them in a semi- or fully automated fashion. The artificial intelligence may achieve human-level expertise (or perhaps, better) without the time demands of human processing (Sebastiani, 2002). That being said, combining human expertise with artificial intelligence can have a synergistic effect.

The BOW approaches, in particular, are a group of techniques primarily used in the computer science discipline to simplify and reduce text (Blei, 2012). These methods assume that word order in a document is irrelevant; that is, documents are treated as a "bag-of-words." Ignoring word order gives the data statistical properties by counting the number of times words appear in a document (Roberts, Stewart, and Tingley, 2014a, b). The resulting data structure is called the document-term matrix in which each row in a data file/matrix represents a document and columns represent individual terms used within the document. Because BOW techniques ignore word order, they yield statistical properties (i.e., exchangability). As a result, these techniques are good at macro tasks, like classifying documents into categories, and can handle large amounts of data. For example, spam filters use a BOW approach founded on Bayesian probability models to classify certain emails as spam based on their content. However, BOW techniques are not good at micro-level tasks, like determining the semantic meaning of text.

There are two types of BOW approaches: *supervised* (Fig. 1, box 3a) and *unsupervised* (Fig. 1, box 3b). These techniques and their application vary depending upon the research context. In supervised methods, the researcher knows in advance what s/he is looking for (Roberts et al., 2014). More specifically, the researcher gives the program both the input, in this case the text, and the output (e.g., the identification of the author of the text), and the system creates an algorithm to map the connection between the two (Janasik, Honkela, & Bruun, 2009). Mosteller and Wallace (1963) provided one of the earliest examples of this approach by using simple Bayesian word probabilities to predict the authorship of 12 disputed Federalist Papers (James Madison or Alexander Hamilton). Today, techniques like naive Bayes and support vector machines (SVMs) are popular supervised algorithms used for text analysis (Manning, Prabhakar, & Hinrich, 2008).

Alternatively, unsupervised algorithms identify word clusters and topics that emerge from the data, similar to thematic analysis (Janasik et al., 2009). However, unlike thematic analysis, topic modeling uses a highly (though not completely) automated approach to determine important topics. It requires less time to analyze and less substantive knowledge about the text. Consequently, topic modeling is suitable for analyzing large amounts of data (Kobayashi1, Mol, Berkers, Kismihok, & Den Hartog, 2017), though human insight is still important to help interpret the topics that emerge. Topic modeling draws on the advantages of thematic analysis (i.e., human insight) and those of machine learning (i.e., quick analysis of large amounts of text).

In the current article, we focus our attention on how to use topic models by providing an illustrative example. Topic models, of which latent Dirichlet allocation (LDA; Blei, 2012) is one of the most popular, assumes that a document is a mixture of topics where each word in the document belongs to a single topic (Blei, 2012; Roberts et al., 2014). Topic modeling is similar to factor analysis in that it reduces the text (a conglomerate of words) to various dimensions, called topics. Structural topic modeling (STM) is an advancement on topic modeling because it allows for the inclusion of covariates or attributes about the document (e.g., gender of the author, work context, etc.). STM uses a regression framework to understand whether those covariates influence the content (i.e., how the topics are talked about) or the prevalence (i.e., which topics and how frequently topics are discussed) of certain topics in the document (Roberts et al., 2014). In this sense, STM adds more depth to the meaning derived from text by accounting for how text changes based on covariates (for an in-depth review of STM and the use of covariates see Roberts, Stewart, & Tingley, 2014b). Topic modeling and STM, of course, have limitations, such as a failure to capture low prevalence topics relative to traditional qualitative text analysis (Baumer et al., 2017).

Finally, NLP is typically the most highly automated form of text analysis (for a review, see Manning et al., 2008). This method models how humans understand and process language (Chowdhury, 2003; Collobert et al., 2011; Joshi, 1991). For example, NLP techniques can tag the parts-of-speech of words in a sentence (e.g., nouns, adjectives, etc.), translate documents from one language to another, and even use the context of a sentence to clarify the meaning of a word (Buntine & Jakulin, 2004). Consequently, unlike the BOW approach, NLP assumes that word order is important. Sentiment analysis, using cutting edge techniques like deep learning and multi-modalities (i.e., combining text and images), is one popular form of NLP when training sets are employed (Kouloumpis, Wilson, & Moore, 2011). This particular analysis classifies the overall attitude, emotion, or opinion of a text as positive, negative, or neutral. In direct contrast to thematic analysis, NLP is a fully computer-automated process and therefore requires little-to-no human insight and/or interpretation (Quinn et al., 2010). In addition, relative to techniques that require human coding (e.g., thematic analysis), NLP is fairly quick to conduct and is more systematic than other approaches. Researchers in computer science, information sciences, linguistics, and psychology, for example, utilize NLP as a text analysis tool (Chowdhury, 2003).

An Illustrative Example of Topic Modeling

We organize our discussion of best practices and the illustrative example of topic modeling into four overarching categories that researchers are likely to encounter when designing and analyzing a study using text data. These categories are outlined in Table 2. We start with considerations shared by a multitude of techniques (not just text analysis)—mainly (1) hypothesis and question formation and (2) design and data collection. We then turn to the major steps in topic modeling, including (3) pre-processing and (4) topic modeling itself. We discuss the creation of scale scores for more traditional correlation and regression analyses in Appendix.

Table 2 Steps in topic modeling needed for validation, evaluation, and interpretability

Step	Description
Step 1: Hypothesis and question formation	
(a) Hypotheses/research questions	When conducting topic modeling, hypotheses and/or research questions are developed in a fashion similar to any other research study.
Step 2: Design and data collection	
(a) Data type	Example types of data include open-ended responses in surveys and archival data such as website text, letters from CEOs, emails between co-workers, and social media (e.g., Twitter, Facebook).
(b) Sample size	Sample size depends on theoretical and methodological considerations of the hypotheses and research questions. For example, theory may specify a certain population of interest or the desire for computing correlation analyses may warrant a power analysis to determine sample size. Researchers should also consider the document-level of analysis.
(c) Quality of writing	Higher quality writing tends to help with analysis. However, even text that is of lesser quality (e.g., Twitter data) can be analyzed.
(d) Length of responses	Typically, more text is better than less text. While there is no minimum number of words needed, we recommend at least 200 characters per document.
(e) Covariates	Theoretical and methodological moderators should be considered which could be tested as "covariates" to understand if different topics emerged or are discussed differently depending on the sub-category.
Step 3: Pre-processing	
(a) Invalid records	Consider the removal of invalid records, such as responses that did not meet a minimum number of words or did not provide relevant text.
(b) Tokenize	The tokenization step involves reducing sentences to individual words or "tokens."
(c) Cleaning	The cleaning step involves creating lower case tokens, removing white space, as well as punctuation.
(d) Stop words	We recommend the removal of stop words. Stop word can be considered to be words that occur so frequently in a research context that they do not add value in the identification of topics. We suggest the removal of highly occurring words that might be random noise. A basic package of common English language words is available in the Quanteda package in R. We also recommend that researchers begin by removing additional stop words.
(e) Sparse terms	A minimum number may need to be set for a word to occur in a document for it not to be a sparse term. When running topic modeling on a few hundred documents, it would be beneficial to start with a two-document minimum and adjust accordingly if results are noisy or the analysis take too long. In instances where the sample consists of thousands of documents, it would be better to start with a higher minimum, such as five documents. In most cases, at least a five-document minimum is used to reduce noise and increase computational speed.
(f) Stemming/lemmatization	Stemming or lemmatization should be conducted in order to facilitate analysis. For instance, "Opportunity" and "opportunities" would both be changed to "opportunit." We recommend preliminary analyses be conducted before this step. For later analyses, we recommend stemming, but researchers should examine if stemming influences conclusions drawn.
(g) Uni-/bi-/tri-grams	The use of bi-grams typically aids in analysis by helping to overcome the "bag of words" problem. For instance, "North Carolina" could be used instead of simply "North" and "Carolina." However, this assumption can be examined in each analysis to determine if conclusions change with the use of uni-, bi-, or tri-grams.
Step 4: Topic modeling	
(a) Commonly used words	"Important" and frequently occurring words from the text should be identified using an iterative process involving adding and removing stop words and other pre-processing steps in order to evaluate if such steps change the types of words that appear.
(b) Number of topics	We recommend considering between 1 and 100 topics; the number of topics that emerge should ultimately be influenced by the interpretability of the topics and the need for parsimony. The exploration of the number of topics should be conducted in an iterative fashion.
(c) Examine network structure	A topic network allows one to see how correlated certain topics are. This is helpful for evaluating the dimensionality of the emerging constructs. We suggest starting with a minimum correlation of 0.10 when displaying the topic network. However, the threshold will need to be adjusted, especially depending on the number of topics (more topics require a lower value, vice versa).
(d) Working definition	Identify construct definitions from the extant literature or develop a working definition in order to facilitate the selection of words for a word list

Many of the recommendations from the authors of the current article involve judgment calls. We encourage transparency by future researchers in their judgment calls. Further, future research should evaluate if any of these judgment calls change the outcomes in their particular study. Researchers could make their data available in order to provide even greater transparency

We would like to briefly note three caveats to all the following sections. First, while we describe each step as distinct and in a particular order, in actuality, there is overlap between the steps, and they are sometimes conducted in an iterative fashion. Second, we focus on the steps that we think are the most important; however, there are, of course, other steps that may need to be considered given one's research context. Third, we cannot provide exact recommendations or rules of thumb in certain steps. Instead, we highlight potential assumptions that researchers make while using topic modeling. Many of these assumptions will need to be explored in future research (Denny & Spirling, 2017). In each individual research study, authors should report if certain analytic decisions change any of the conclusions being drawn.

Step 1: Hypothesis and Question Formation The first overarching category is hypothesis and question formation. Researchers should begin by considering a priori hypotheses and/or research questions that they are interested in testing or answering. Also, it is important to make sure that there is alignment between the hypotheses and/or research questions and the data collected. Hypotheses and research questions that are most applicable for topic modeling are those in which understanding the latent variables underlying a set of text is of interest. To help illustrate topic modeling, the current authors conducted a survey study on leader-member exchange (LMX). LMX refers to the quality of the relationship between subordinates and their supervisors (Schriesheim, Castro, & Cogliser, 1999). In particular, we developed two hypotheses and three research questions to illustrate how they could be answered with the analysis of text. For the sake of brevity, we elect to not go into greater depth, but simply present them in Table 4 in the current article as well as in Appendix.

Step 2: Design and Data Collection The next major category is design and data collection. This step involves identifying the appropriate text needed to test the hypotheses or answer the research questions and developing a method for gathering it. There are many types of design and data collection methods available to compile a text database. First, when considering the type of data (step 2a), researchers can take an archival approach where they collect existing text data such as website text, letters from CEOs, emails between co-workers, or social media data (e.g., Twitter, Facebook). Web scraping tools (e.g., Import.io, Dexio.io) can facilitate the collection of Internet-based text. Second, researchers can design a study based on interviews and focus groups; the audio is then transcribed into text that can be analyzed. As a third example, text data can be collected via a survey where there is or is not an experimental manipulation. For instance, survey participants may be presented with vignettes where information is presented in different conditions and then asked to write responses. A survey design can also be implemented to collect text data in which organizational employees or participants are asked open-ended questions. Any design that produces text can be used so long as the researcher is able to collect a sufficient amount of data, discussed further later.

For the current study, we collected text from open-ended survey responses using Amazon's Mechanical Turk (MTurk). We elected to use this method, as text from open-ended survey questions is very common in organizational science research. A link to the pre-registered study protocol (https://osf.io/g9wjy/) is available via the open-science framework.[1] We provided participants with two open-ended survey questions focused on LMX drawing upon past research (Dulebohn, Bommer, Liden, Brouer, & Ferris, 2012):

> Open-ended question #1: How would you describe your working relationship with your leader? For instance, (1) how much coaching and development does your leader provide to you, (2) how does your leader behave to show that he or she respects you, (3) how does your relationship with your leader compare to your co-workers' relationships with the same individual? Please use a minimum of 100 words in your response.
>
> Open-ended question #2: To what extent does your leader understand your problems and needs? For instance, (1) what types of support does your supervisor provide you to accomplish your work objectives, (2) how does your supervisor listen to your concerns and provide advice, (3) does your supervisor provide you with coaching and mentoring? When answering, please provide examples of how your leader responds to your problems and needs (and again use a minimum of 100 words in your response).

We also collected closed-ended response data using measures of leader vision (Pearce & Sims, 2002), perceived organizational support (Eisenberger, Hungtinton, Hutchsion, & Sowa, 1986), LMX (Bernerth, Armenakis, Feild, Giles, & Walker, 2007), employee self-determination (Spreitzer, 1995), supervisor satisfaction (Cammann, Fichman, Jenkins, & Klesh, 1983), and turnover intentions (Mitchel, 1981). As such, we were able to test our hypotheses by correlating open- and close-ended measures of LMX to leadership outcomes (see Table 4).

[1] Changes from pre-registered protocol: The final sample size (n = 585) was lower than expected (n = 1000), but was dictated by our prespecified budgetary limit. Also, we originally planned to ask participants about their time working with the leader, but dropped the question due to space concerns. We had planned to examine how occupation related to LMX. However, there were not enough respondents for the majority of the occupations ($n < 20$); given the small n there is not adequate power to detect even a small magnitude effect (e.g., d = .30). When we aggregated the occupations, the information became redundant with our industry question. Hence, our question about how LMX varied by occupation was dropped.

One important consideration in text data collection pertains to sample size (step 2b). Sample size in topic modeling is a function of both the number of documents included in the analysis (e.g., 5 versus 30 transcribed interviews) and the length of the document (e.g., a tweet versus a transcribed interview). As such, several considerations should be made in determining sample size. First, one should contemplate the theoretical context for the research study and the subsequent hypotheses and research questions. For example, if one's population of interest is text from *Fortune* 500 firm websites, one could examine text from all these websites. However, websites often have many webpages nested within websites. Drawing upon signaling theory (Connelly, Certo, Ireland, & Reutzel, 2011), one might decide that only the first and second-level webpages are of interest because firm signals on third level webpages are typically less salient. A second consideration is methodological issues. For instance, if a research team desires to conduct null hypothesis significance testing using variables that emerge from the text, a traditional power analysis could be used.

As a third point of attention, researchers should also consider how to define a *document* (Tang, Meng, Nguyen, Mei, & Zhang, 2014). For instance, paragraphs of text are nested within webpages, webpages are nested within websites, and websites are nested within organizations. As another example, a sentence is nested within a paragraph, a paragraph is nested within a social media post, and a social media post is nested within the user. Depending on the hypotheses or research questions of interest, researchers may change the level at which a document is defined (e.g., a website as a document), which could influence sample size. This is a similar concept to multi-level analyses in traditional organizational science research (e.g., employees are nested within departments, departments are nested within firms). Depending on the covariates of interest, the covariates may not be available at one particular document level.

There are two additional factors that could play a role in determining sample size when analyzing text. First, researchers must consider the quality of the writing (step 2c). Higher quality writing tends to help with analysis for two reasons. Words that are consistently spelled correctly or used in grammatically correct sentences are easier to match. It is more difficult to find patterns in the text if the same word is misspelled or used improperly in one instance and not another. Higher quality writing also aids in interpretation of the topics. If slang terms are used consistently across the sample, they will emerge in the topics. However, if the researcher is not familiar with the terms, the results may seem nonsensical and therefore uninterpretable. It is important to note that it is possible to analyze text that is perhaps of lesser quality due to slang and poor grammar (e.g., Twitter data).

Second, the length of responses and/or documents can influence the target sample size (step 2d). While there is not typically a rule of thumb in terms of the minimum number of words needed for text analysis, generally speaking, longer responses are better. For example, tweets are documents that may be too short because they contain a limited number of words. Aggregation of individual tweets to the user level may solve the document "shortness" challenge. However, aggregating to the user level reduces the sample size (i.e., the number of documents). The reverse can be done as well in which researchers break-up tweets at the user level into individual tweets to facilitate analysis. A minimum number of words cannot be recommended because this issue is context dependent.

When archival text is being analyzed (e.g., text already available in a public domain), specifying a goal amount of text is often not possible. However, in survey research, one can specify a minimum number of words in responses. In the current study, we collected data with a minimum number of 100 words. We also specified to a sub-group of participants a minimum number of 50 words. We did not see any noticeable differences in the quality and nature of the writing. In our illustrative example, we combined the responses to the two open-ended questions. Combining the two responses is substantiated because the two questions were designed to capture different aspects of LMX, and we were interested in the general LMX construct rather than a particular dimension. Thus, the document level for our sample moved from the question to the person. Our final dataset consisted of 585 rows of data, equal to the number of participants ($n = 585$).

As the next step in the design and collection of data, we recommend the inclusion of variables that can act as covariates (step 2e). "Covariates" are essentially theoretical and methodological moderators; they are used to understand if different topics emerged or are discussed differently depending on the sub-category of the documents. For example, in our analysis, we provided participants with a series of demographic questions that asked, "in what country do you currently reside?," "what is the gender of your leader?," "what is your gender?," "what is the current job title of your leader?," "what is your current job title?," "in what industry do you currently work?," and "which of the following most closely matches your job level?" (e.g., intern, entry level, associate). These demographic questions can be used as covariates to determine if the answer to any of these questions changes the content (i.e., how they talk about topics) or the prevalence (i.e., which topics) are discussed.

Pre-processing Text The next major category of a text analysis study is the pre-processing step, which is similar to data cleaning in quantitative analyses. In text analysis research, studies often gloss over pre-processing steps (see Denny & Spirling, 2017 for a useful R package (preText), and we encourage readers to view their tutorial). However, similar to any primary study, the way in which data are cleaned for analysis could potentially influence one's results. All of the steps in the pre-processing phase are meant to prepare the data for later analyses. However, pre-processing in this context is an exploratory, iterative phase and researchers may need to return to certain pre-processing steps even after conducting more advanced analyses.

To begin the pre-processing steps, we recommend that researchers start by considering the removal of invalid records (step 3a), such as eliminating responses that did not meet the minimum number of words. For example, we dropped two participants in the current study; one reported substantially fewer words than the specified minimum and the other completed the survey multiple times.

The next steps in the text analysis process involve tokenization (step 3b) and "cleaning" of the text (step 3c). Both steps are standardized and automated in the quanteda package in R. Tokenization involves reducing sentences to individual words or "tokens." Cleaning the text creates lowercase tokens and removes white space as well as punctuation.

One of the most important steps of the pre-processing phase includes the evaluation and removal of *stop words* (step 3d). Stop words[2] are words that occur so frequently in a research context that they do not add value in the identification of topics (also referred to as "exception" words in Wordstat). A basic package of common English language words is available in the quanteda package in R (it includes words such as "and," "the," and "then"). More recently, Schofield, Magnusson, & Mimno (2017) evaluated the effects of stop words within topic modeling and found that, like stemming, its effects are negligible and simply implementing after inference (model training) can produce similar results with more transparency around its role. Similarly, we follow their recommendations that researchers begin by removing standard stop words. However, additional words can be removed as one begins to interpret the findings. Note that it is important that this step follow the study of n-grams because stop words can be the key to identifying very useful phrases. One must also be careful to consider that in some instances, normally irrelevant words might be important. For example, when studying the field Information Technology, commonly referred to as "IT," the abbreviation would be considered the simple word "it" and removed.

Subject matter expertise can be very useful in identifying stop words. For instance, the word "family" may seem out of place in text discussing businesses. However, those more familiar with business literature and entrepreneurship might immediately recognize that "family businesses" may be an important point of discussion. Thus, having some knowledge of the literature can be quite valuable (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). When in doubt, we recommend that researchers exercise caution and keep words in the analyses.

One can apply Zipf's law (Newman, 2005) when identifying other stop words. This law suggests removing frequently occurring words to reduce random noise in the results. Zipf's law also applies to the removal of infrequent words at the tail of the distribution (this helps to reduce processing time and power requirements; step 3e). These infrequent words are known as *sparse terms*. A minimum number could be set for a word to occur in a document above which a word is not considered a sparse term. For instance, in a dataset with 1000 documents, one might say that to be included in an analysis a word has to appear in a minimum of 15 documents. As an example, the word "Walmart" might appear 100 times in one document (e.g., text from Walmart's website) but not in any other document (e.g., text from competitors' websites). Hence, specifying a minimum number of documents would lead to the removal of Walmart from the analysis. We recommend a two-document minimum, particularly for small sample sizes. If a word is used in only one document, it cannot be found to co-occur more than once; such terms will add unnecessary time and computational complexity to the analysis. When running topic modeling on a few hundred documents, it would be beneficial to start with a two-document minimum and adjust accordingly if results are noisy or the analysis takes too long. In instances where the sample consists of thousands of documents, it would be better to start with a higher minimum, such as five documents. In most cases, at least a five-document minimum reduces noise and increases computational speed. Removing sparse words prevents words that are idiosyncratic to a particular author (document) and otherwise meaningless words from dominating a topic.

Next, researchers should consider the role of stemming, where words are reduced to their roots (step 3f). For instance, "opportunity" and "opportunities" would both be changed to "opportunit." Stemming should be considered as an option to facilitate analysis. When determining whether or not to stem, researchers should consider the interpretability of the results, as well as the predictive validity. For instance, stemming might change the words "run," "running," and "runs" to simply "run." A more advanced approach to this problem is lemmatization. Rather than simply chopping off the end of words, lemmatization considers the root of the word as its replacement. For example, "am", "is", and "are" would be converted to their common base word "be." Unfortunately, fewer software packages include this more advanced approach (for guidance, see https://nlp.stanford.edu/software/tagger.shtml).

Although both stemming and lemmatization remain popular options in text pre-processing, research on the usefulness of each approach is mixed and largely depends on the corpus and method. For example, Manning et al. (2008) found that these approaches have the most value when applied to small

[2] Start words also exist where a researcher specifies that only certain words be included in an analysis.

samples in text classification problems to compensate for data sparseness. Ultimately, they argue that as a corpus grows, the gains of stemming become diminished. More recently, Schofield and Mimno (2016) considered the impact of stemming and lemmatization on topic modeling performance for four different sources of text (research articles, IMDb reviews, New York Times articles, and Yelp reviews). Controlling for vocabulary size, they find no empirical benefit in fit or topic coherence when stemming or lemmatization in topic modeling. In fact, they find that stemming and lemmatization can potentially diminish performance as topic modeling was already grouping root words together.

They do find specific cases, such as misspellings in Yelp reviews, in which stemming can help overcome spelling errors. However, they argue stemming should be used to help identify problems like misspellings but do not recommend stemming to resolve these problems. Moreover, they recommend if stemming is necessary for interpretation, a better (and computationally cheaper) approach is to stem after running model inference. Therefore, we recommend that during early exploratory analyses researchers do not stem or lemmatize. However, as researchers develop an understanding of the topics emerging from their data, researchers can consider stemming but should exercise caution.

Another important pre-processing step includes the consideration of n-grams, e.g., uni-, bi-, and tri-grams (step 3g). The use of n-grams typically aids in analysis and may be considered a “default” starting point. For instance, when analyzing with bi-grams, “North Carolina” would be included in analyses as one word/token instead of “North” and “Carolina” considered as separate words. However, this assumption can be examined in each analysis to determine if conclusions change with the use of uni- or bi-grams. In the current study, we implemented the use of bi-grams. In general, we recommend that researchers begin analyses with uni-grams, but that in most contexts, bi-grams should also be considered. Like stemming, ultimately the best evaluation of the value of including additional n-grams should be the out-of-sample evaluation of fit or coherence depending on the nuisances of the corpus. The current R code provided in this article could ultimately be adapted so that the researchers can specify the bi-grams a priori; however, currently the algorithm determines what bi-grams emerge.

Topic Modeling As we mentioned in the literature review, topic modeling is a framework of unsupervised machine learning algorithms that identifies clusters of words that cooccur together. We chose topic models for three reasons. First, topic modeling does not require pre-modeling annotations. Instead, topics are automatically discovered as lists of words that co-occur, allowing researchers to analyze empirically driven word lists. Second, without the need for human coding or manual annotations, topic models can be scaled to hundreds of thousands of documents and different types of corpora (e.g., emails, open-ended surveys, social media posts). Last, topic models can also serve the purpose of information retrieval, as documents are scored based on their topic likeness (probability) and thus can be ranked to identify the most representative documents. This approach enhances the interpretability of topics and the identification of outliers. For our analysis, we primarily used LDA which is the most widely used topic model (Blei, Ng, & Jordan, 2003). We now briefly review key properties about LDA, and we refer the reader to other sources for a more detailed discussion (e.g., Blei, 2012).

Topic models are based on the assumption that the input text is produced by hidden (latent) probabilistic variables which can be understood as topics. A Bayesian hierarchical mixture model is applied, which draws upon co-occurrence among words in order to determine emerging topics. Terms can be characterized by three key properties from the model. First, text can be organized into a document-term matrix that quantifies the occurrence of each word (columns) by each document (rows). The input to the LDA algorithm is thus a document-term matrix that requires specifying the number of topics. Second, the model is a Bayesian mixture model. Documents are made of a mixture (probability distribution) of topics instead of just a single topic. The LDA algorithm serves as a dimensionality reduction procedure that reduces the amount of information about each document from a large number of columns (words) to a significantly smaller number of columns (i.e., Crain, Zhou, Yang, & Zha, 2012). In other words, it “summarizes” the information in the word counts down to a reduced number of columns. This leads to the first output of the algorithm which is the document-topic matrix. In this matrix, each document is scored as a probability across all the identified topics. As a third key property, the model used can be considered as a hierarchical mixture model as it includes a hierarchy of two probability mixtures. At the top of the model, the documents compose a mixture of topics, and at the bottom, the topics are a mixture of words.

Each topic is defined as a probability distribution over words. The LDA algorithm’s second output is a word-topic matrix. This matrix delivers a conditional probability for each word (row) based on the corresponding hidden topic (column). The probability distribution can rank-order words by topic to understand the most common words within each topic. Higher probability words facilitate the interpretation of the meaning of each topic. The word-topic mixture properties of the LDA algorithm allows for words to be utilized for different topics (polysemy) and to also allow similar words to be clustered (synonym). The R code and topicApp provided in the current article employs variational inference which runs analyses using random sampling with replacement until a model converges (the algorithm is set to stop running if convergence is not achieved after 200 iterations). Variational inference is distinct from simulation or other sampling-based methods (e.g., Gibbs sampling, Markov chain Monte Carlo) but yields similar results, as both seek to identify latent variables in text (for additional reading on variational inference, see Bliese et al., 2017).

When conducting topic modeling, we recommend that researchers first conduct the pre-processing steps using the topicApp provided in this article to prepare the dataset. Next, researchers should examine commonly used words that emerge from the topic modeling (step 4a). These words are tokens that are important and frequently occurring. They can be identified by examining the "top 5" words that appear per topic, the topic network (how topics relate to other topics), individual word clouds per topic (the range of words in the word cloud can range from 1 to 200), as well as the most representative documents. We recommend the consideration of all of these sources of information when examining the emerging topics.

Step 4b involves the evaluation of the number of topics. In the topicApp provided, users can toggle between 0 topics (in which the algorithm picks the number of topics) and 100 topics. We recommend that researchers begin with a smaller number of topics and explore their data by increasing the number of topics. Examining different numbers of topics helps researchers become more familiar with their data. When choosing the number of topics, we recommend that researchers apply the principal of parsimony. That is, in order to justify increasing the number of topics or latent constructs in the text, and therefore, the complexity of the results, there has to be justification that one is better able to interpret the findings with more topics. This is consistent with the literature on construct redundancy in the organizational science literature (see Banks; Gooty, Ross, Williams, and Harrington, 2017; Banks, McCauley, Gardner, & Guler, 2016; Schmidt, 2010; Shaffer, DeGeest, & Li, 2016). Thus, with a larger number of topics, one might find redundancy among the topics. In the current study, we selected one topic (LMX) but provide three sub-dimensions that emerged from the data as illustrative examples in Table 3 (for another example, see Short et al., 2010). Given that we asked two very specific open-ended questions about LMX, this is not a surprise. If, for instance, we used website data or social media data, one might expect many more topics to emerge.

For the LMX topics, we created labels and definitions of the topics based on the words that represented them after discussing among co-authors as to what they meant. To inductively characterize and define emerging topics, a constant comparative method can be used (Glaser & Strauss, 1967). This approach involves comparing the text data to emergent topics and also associating the topics with the extant literature (Strauss & Corbin, 1990). Research teams should discuss what the emerging topics represent and develop topic labels (Cowan & Fox, 2015). Key points to discuss include "how are these topics similar to one another?," "how are these topics different?," and "if they are different, in what way?" (Cowan & Fox, 2015). The outcome of this process is to verify that the topics identified are relatively robust and characteristic of the data. The analysis is complete when researchers are satisfied with the topic labels and definitions and supportive examples are identified from the original text.

As a final step in the topic modeling phase, researchers should examine the network structure of topics (step 4c). A

Table 3 Topics that emerged from analysis of the open-ended survey text

Hypothetical dimension	Description	Example stemmed word list
Individual relationship	The extent to which a leader develops relationships with individual employees by showing concern for their feelings and needs as well as respect through coaching and development	Promot, care, relationship, time, kind, talk, need, respect, guidance, friend, answer, nice, offer, appreci, give, feel, listen, accomplish, advic, assist, attent, avail, coach, concern, develop, fair, flexibl, honest, listen, mentor, prais, respond, succeed, support, train, treat, trust
Task-oriented helping	The extent to which a leader provides assistance with a task through problem solving and advice	achieve, advis, clear, discuss, encourag, exampl, explain, goal, guid, help, idea, knowledg, learn, mistak, motiv, patient, problem, project, provid, respons, share, solve, suggest, teach, target, ask, communic, complet, ensure, fix, improv, issu, question, schedule, solu
Team performance	The extent to which a leader directs team tasks, develops relationships among members, and provides feedback in order to enhance team performance	assign, direct, feedback, group, handl, involv, meet, member, opportun, other, people, perform, procedur, recog, resourc, role, task, team, trust, understand, inform

Subject matter expert validation was completed by obtaining two additional evaluations from non-authors. Finally, analyses were computed by combining all three-word lists. Stop words used included the following: leader, much, just, lot, way, though, sometimes, amount, without, goes, pretty, however, many like, fellow, level, job, work, overall, thinks, get, thing, try, makes, make, come, comes, said, tries, someone, towards, first, really, every, put, can, done, one, manner, due, place, current, type, office, specific, often, say, even, still, actually, last, since, left, years, see, manager, supervisor, general, company, will, set, basis, certain, part, else, whatever, staff, month, field. Negative words included the following: aggressive, angry, annoy, bad, barely, bother, cannot, complaint, confused, confusing, couldn't, degrading, difficult, discriminate, doesn't, dominate, don't, excuses, failed, fault, friction, frustrated, frustrating, headache, horribly, ignore, isn't, little, mean, meaner, mess, negative, never, no, nonexistent, not, overworked, own, quit, rarely, self, selfish, slack, strict, stupid, tense, tension, threat, threatened, trouble, uninterested, wasn't, wrong

Table 4 Means, standard deviations, and correlations of all variables

Variable	Mean	SD	1	2	3	4	5	6	7
1. Vision	3.66	0.95	.91						
2. POS	4.88	1.27	.62**	.90					
3. LMX-C	4.80	1.36	.54**	.69**	.96				
4. LMX-O	.02	0.14	.15**	.21**	.25**	–			
5. EM	4.05	0.96	.22**	.41**	.42**	.07	.91		
6. SS	4.00	1.18	.51**	.62**	.74**	.27**	.46**	.92	
7. TI	2.41	1.24	− .50**	− .65**	− .59**	− .20**	− .38**	− .66**	.87

N = 584~585. Coefficients alpha are listed in the diagonal where appropriate. Hypothesis 1: Leader vision will positively relate to LMX scores (both closed- and open-ended responses on LMX): *Supported*. Hypothesis 2: LMX (both closed- and open-ended responses) will positively relate to employee (a) empowerment, (b) perceived organizational support, (c) satisfaction with supervisor, and (d) turnover intentions: *Supported*; For additional discussion regarding the creation of scale scores as well as the current hypotheses and research questions, please see Appendix

POS perceived organizational support, *LMX-C* leader-member exchange-closed-ended, *LMX-O* leader-member exchange-open-ended, *EM* empowerment (self-determination), *SS* satisfaction with supervisor, *TI* turnover intentions

* $p < 0.05$; ** $p < 0.01$ (two-tailed tests)

topic network allows one to see how correlated certain topics are (a minimum threshold can be set for the topicApp to illustrate that the topics are correlated). This is helpful for evaluating the dimensionality of the emerging constructs. We suggest looking at how correlated the topics are in the network. For instance, if one considers a topic network with six topics, and one topic is not correlated at all with the other five topics, that speaks to the uniqueness of this topic. This can aid in the interpretation of the topics.

At this point, researchers should have an understanding of the topics that emerged in the text. It is possible that researchers choose to stop in the analysis because their research questions have been answered. Stopping at this point in the process would most resemble a thematic analysis/grounded theory approach, with the added benefit of computer assistance (Baumer et al., 2017). However, researchers may create scale scores from text and conduct null hypothesis significance testing (see Table 4). We demonstrate this in Appendix.

Conclusion

The purpose of the current article is to provide researchers with the basic knowledge and tools necessary to conduct text analysis in the social sciences. In doing so, we contribute to the literature in a number of ways. First, although text analysis is common across a wide range of disciplines (e.g., computer science, political science, psychology, organizational science), the way in which text is analyzed in each discipline has been relatively contained. For example, researchers in psychology have most likely not been exposed to the text analysis techniques in computer science, and those in computer science are likely not familiar with techniques used in psychology. We bridge the knowledge gaps among disciplines by providing readers with a taxonomy of available text analysis techniques (Fig. 1 and Table 1).

Second, although text analysis and topic modeling are not necessarily new, a step-by-step guide on how to conduct the analyses does not exist. The current study addresses this gap by outlining specific steps and best practice recommendations for how to conduct a particular type of computer-aided text analysis: topic modeling. Third, we offer a handful of resources that readers can use to help become familiar with the analyses and to help conduct them. These resources include (1) an R markdown file that provides R code and accompanying explanations (available in Dataverse), (2) a topicApp that offers a user-friendly option to conduct topic model for those not familiar or comfortable with R and a FAQ section to troubleshoot potential issues, and (3) a detailed reading list (Appendix) that provides both seminal topic modeling articles (for more novice readers) as well as more detailed information on the steps and decisions outlined above (for more experienced readers). These resources can be found at https://github.com/wesslen/topicApp, https://github.com/wesslen/text-analysis-org-science, https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/R4W7ZS. Our hope is that the resources provided will lessen the intimidation associated with text analysis and R and encourage research that utilizes text analysis.

Future Directions for Text Analysis

The current article was purposely designed to illustrate text analysis and topic modeling in a simple manner. Future research in text analysis methods should continue to consider ways to integrate techniques from across disciplines. For example, research is needed that explores how topic modeling can be combined with grounded theory analysis (Baumer et al., 2017). However, there are many more advanced techniques for text analysis that we did not cover and that may provide opportunities for organizational scientists in the future. We now briefly outline three approaches for future research.

First, recent work in topic models has improved methods in interpretation, evaluation, and validation. For example, computer scientists have devised measures to aid interpretability like FREX and Exclusivity that identify distinctive words. Other new techniques like semantic coherence attempt to measure word consistency within topics as a proxy for measuring interpretability (Grimmer & Stewart, 2013). Computer scientists have also focused on methods to evaluate and compare model performance by prediction (maximize hold out likelihood) or interpretability (maximize semantic coherence). Further, recent work by political scientists improved methods to validate and ensure model stability (Roberts, Stewart, et al., 2014a). Several of the same researchers advanced methods in measuring causal and treatment effects within text (Fong & Grimmer, 2016; Roberts, Stewart, & Airoldi, 2016).

For advanced researchers, methods have been developed to evaluate the emerging model based on prediction capabilities. These techniques include the use of the held-out likelihood (Wallach, Murray, Salakhutdinov, & Mimno, 2009), semantic coherence (Mimno, Wallach, Talley, Leenders, & McCallum, 2011), residuals (Taddy, 2012), and convex hull (Lee & Mimno, 2014). All four methods can be employed in the STM package (see stm vignette; pages 12–14). Yet, see a discussion by Chang et al. (2009) about the limitations of some of these methods in terms of trading off prediction versus interpretation.

Second, given topic models' large data output and need of human-level interpretation, visualization interfaces are critically important for applied researchers to analyze topic model results. Recently, the field of visual analytics has pioneered a variety of techniques to analyze topic models including ways to interpret topics, analyze temporal and spatial trends, and graph (network)-based features (Dou & Liu, 2016). While most of these interfaces were written using JavaScript or D3 libraries, programming languages that many organizational scientists lack experience in, recently many of these libraries have become available in R through tools like Shiny and htmlwidgets, requiring only R knowledge rather than HTML, CSS, or JavaScript. In the future, we expect a rise of visualization apps as tools like topic models spread to applied researchers who are interested in analyzing results rather than building custom visualizations themselves.

Third, computer scientists have more recently shifted their focus to applying neural probabilistic models through models like word embedding and deep learning. Unlike traditional bag-of-word models that ignore word order, word embedding models identify co-occurrence on a micro-window (e.g., rolling ten-word window). In doing so, this approach directly models word context which tends to capture deeper semantic meaning than traditional bag-of-words models like topic models. Two primary examples of word embedding models are the word2vec (Mikolov, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014). Both of these models show marked improvements in identifying word similarities and analogies.

Further, these models are part of a larger framework of neural language models, which also include deep learning. Deep learning is an approach that uses large-scale neural networks that have demonstrated breakthroughs in many text analysis tasks like machine translation, text classification, and even text generation (LeCun, Bengio, & Hinton, 2015). Work on deep learning is also being expanded to convert images to text for analysis. However, one major limitation of these methods (especially for deep learning) is their "black box" nature that prevents measurements of marginal effects to explain why the model made certain predictions. Nevertheless, continued work on the interpretability and visualization of deep learning may yield potential tools for social scientists in the near future. In conclusion, we present these three avenues as areas for future research.

References

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly, 28*, 5–21.

Banks, G. C., Gooty, J., Ross, R., Williams, C., & Harrison, N. (2017). Construct redundancy in leader behaviors: A review and agenda for the future. *The Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2017.12.005.

Banks, G. C., McCauley, K. D., Gardner, W. L., & Guler, C. E. (2016). A meta-analytic review of authentic and transformational leadership: A test for redundancy. *The Leadership Quarterly, 27*, 634–652.

Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology, 68*, 1397–1410.

Bernerth, J. B., Armenakis, A. A., Feild, H. S., Giles, W. F., & Walker, H. J. (2007). Leader–member social exchange (LMSX): Development and validation of a scale. *Journal of Organizational Behavior, 28*, 979–1003.

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

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

Bliese, P. D., Maltarich, M. A., & Hendricks, J. L. (2017). Back to basics with mixed-effects models: Nine take-away points. *Journal of Business and Psychology*, 1–23.

Buntine, W., & Jakulin, A. (2004). *Applying discrete PCA in data analysis.* Paper presented at the Proceedings of the 20th conference on Uncertainty in artificial intelligence.

Cammann, C., Fichman, M., Jenkins, G. D., & Klesh, J. R. (1983). Assessing the attitudes and perceptions of organizational members. In S. E. Seashore, E. E. Lawler, P. H. Mirvis, & C. Cammann (Eds.), *Assessing organizational change: A guide to methods, measures, and practices* (pp. 71–138). New York: Wiley.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). *Reading tea leaves: How humans interpret topic models.* Paper presented at the Advances in neural information processing systems.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*, 51–89.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*, 2493–2537.

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management, 37*, 39–67.

Cowan, R. L., & Fox, S. (2015). Being pushed and pulled: A model of US HR professionals' roles in bullying situations. *Personnel Review, 44*, 119–139.

Crain, S. P., Zhou, K., Yang, S.-H., & Zha, H. (2012). Dimensionality reduction and topic modeling: From latent semantic indexing to latent Dirichlet allocation and beyond *Mining text data* (pp. 129-161): Springer.

Denny, M. J., & Spirling, A. (2017). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Available at SSRN:* https://ssrn.com/abstract=2849145.

Dou, W., & Liu, S. (2016). Topic-and time-oriented visual text analysis. *IEEE Computer Graphics and Applications, 36*, 8–13.

Dulebohn, J. H., Bommer, W. H., Liden, R. C., Brouer, R. L., Gerald, R., & Ferris, G. R. (2012). A meta-analysis of antecedents and consequences of leader-member exchange: Integrating the past with an eye toward the future. *Journal of Management, 38*(6), 1715–1759.

Eisenberger, R., Hungtinton, R., Hutchsion, S., & Sowa, D. (1986). Perceived organizational support. *Journal of Applied Psychology, 71*, 500–507.

Fong, C., & Grimmer, J. (2016). Discovery of treatments from text corpora. In In Proceedings of the Annual Meeting of the Association for Computational Linguistics.

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*, 15–31.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine.

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics, 48*, 80–83.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*: mps028.

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

Joshi, A. K. (1991). Natural language processing. *Science, 253*, 1242.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., & Den Hartog, D. N. (2017). Text classification for organizational researchers: A tutorial. *Organizational Research Methods*. https://doi.org/10.1177/1094428117719322.

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! *Icwsm, 11*, 164.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.

Lee, M., & Mimno, D. (2014). *Low-dimensional embeddings for interpretable anchor-based topic inference.* Paper presented at the Proceedings of Empirical Methods in Natural Language Processing.

Lehmann-Willenbrock, N., & Allen, J. A. (2017). Modeling temporal interaction dynamics in organizational settings. *Journal of Business and Psychology*, 1–20.

Manning, C. D., Prabhakar, R., & Hinrich, S. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2016). What doesn't get measured does exist improving the accuracy of computer-aided text analysis. *Journal of Management*: 0149206316657594.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv: 1301.3781*.

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

Mitchel, J. O. (1981). The effect of intentions, tenure, personal, and organizational variables on managerial turnover. *Academy of Management Journal, 24*, 742–751.

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association, 58*(302), 275–309.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*, 323–351.

Pearce, C. L., & Sims, H. P. (2002). Vertical versus shared leadership as predictors of the effectiveness of change management teams: An examination of aversive, directive, transactional, transformational, and empowering leader behaviors. *Group Dynamics: Theory, Research, and Practice, 6*, 172–197.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *In EMNLP, 14*, 1532–1543.

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*, 209–228.

Reinard, J. C. (2008). *Introduction to communication research* (4th ed.). Boston: McGraw-Hill.

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

Roberts, M. E., Stewart, B. M., & Tingley, D. (2014a). *Navigating the local modes of big data: The case of topic models*. New York: Cambridge University Press.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2014b). stm: R package for structural topic models. *R package version 0.6, 1*.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*, 1064–1082.

Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5*, 233–242.

Schofield, A., Magnusson, M. and Mimno, D. (2017). Pulling Out the stops: Rethinking stopword removal for topic models. *EACL, 432*.

Schofield, A., & Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics, 4*, 287–300.

Schriesheim, C. A., Castro, S. L., & Cogliser, C. C. (1999). Leader-member exchange (LMX) research: A comprehensive review of theory, measurement, and data-analytic practices. *The Leadership Quarterly, 10*, 63–113.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR), 34*, 1–47.

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods, 19*, 80–110.

Shanock, L. R., Baran, B. E., Gentry, W. A., Pattison, S. C., & Heggestad, E. D. (2010). Polynomial regression with response surface analysis: A powerful approach for examining moderation and overcoming limitations of difference scores. *Journal of Business and Psychology, 25*, 543–554.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods, 13*, 320–347.

Spreitzer, G. M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal, 38*, 1442–1465.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research*. Newbury Park, CA: Sage.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks: Sage.

Suddaby, R. (2006). From the editors: What grounded theory is not. *Academy of Management Journal, 49*, 633–642.

Taddy, M. (2012). *On estimation and selection for topic models.* Paper presented at the International Conference on Artificial Intelligence and Statistics.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). *Understanding the limiting factors of topic modeling via posterior contraction analysis.* Paper presented at the ICML.

Tonidandel, S., & LeBreton, J. M. (2015). RWA web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. *Journal of Business and Psychology, 30*, 207–216.

Waddell, K. (2016). The algorithms that tell bosses how employees are feeling. *The Atlantic*.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). *Evaluation methods for topic models.* Paper presented at the Proceedings of the 26th annual international conference on machine learning.

Williams, L. J., & McGonagle, A. K. (2016). Four research designs and a comprehensive analysis strategy for investigating common method variance with self-report measures using latent variables. *Journal of Business and Psychology, 31*, 339–359.