Article

Organizational Research Methods
2018, Vol. 21(3) 733-765

DOI: 10.1177/1094428117722619
journals.sagepub.com/home/orm

Text Mining in Organizational Research

Vladimer B. Kobayashi[1], Stefan T. Mol[1], Hannah A. Berkers[1], Gábor Kismihók[1] and Deanne N. Den Hartog[1]

Abstract

Despite the ubiquity of textual data, so far few researchers have applied text mining to answer organizational research questions. Text mining, which essentially entails a quantitative approach to the analysis of (usually) voluminous textual data, helps accelerate knowledge discovery by radically increasing the amount data that can be analyzed. This article aims to acquaint organizational researchers with the fundamental logic underpinning text mining, the analytical stages involved, and contemporary techniques that may be used to achieve different types of objectives. The specific analytical techniques reviewed are (a) dimensionality reduction, (b) distance and similarity computing, (c) clustering, (d) topic modeling, and (e) classification. We describe how text mining may extend contemporary organizational research by allowing the testing of existing or new research questions with data that are likely to be rich, contextualized, and ecologically valid. After an exploration of how evidence for the validity of text mining output may be generated, we conclude the article by illustrating the text mining process in a job analysis setting using a dataset composed of job vacancies.

Keywords

text mining, dimensionality reduction, clustering, topic modeling, classification, validation, job analysis

Organizations are increasingly turning to big data and analytics to help them stay competitive in a highly data-driven world (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2013). Although difficult to assess let alone verify (Grimes, 2008), around 80% of data in organizations are commonly estimated to consist of unstructured text. The abundance of text data opens new avenues for research but also presents research challenges. One challenge is how to manage and extract meaning from a massive of amount of text since reading and manually coding text is a laborious exercise.

[1]Leadership and Management Section, Amsterdam Business School, University of Amsterdam, Netherlands

Corresponding Author:
Stefan T. Mol, Amsterdam Business School, University of Amsterdam, Valckenierstraat 59, 1018 XE Amsterdam, Netherlands.
Email: s.t.mol@uva.nl

To take full take advantage of the benefits of doing research with "big" text data, organizational researchers need to be familiarized with techniques that enable efficient and reliable text analysis.

Text mining (TM) is "the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text" (Kao & Poteet, 2007, p. 1). Knowledge is derived from patterns and relationships and can be used to reveal facts, trends, or constructs (Gupta & Lehal, 2009; Harlow & Oswald, 2016). A related technique which organizational researchers may be more familiar with is computer-aided text analysis (CATA). To date, most studies employing text analysis in organizational research are CATA-based (Kabanoff, 1997; McKenny, Short, & Payne, 2013; Short, Broberg, Cogliser, & Brigham, 2010). CATA (McKenny et al., 2013) is a special class of TM. Whereas most CATA procedures extract patterns by counting word/term frequencies, TM generally also capitalizes on other textual properties, such as grammar and structure, and employs techniques from natural language processing, computational linguistics, corpus linguistics, machine learning, and statistics. Hence, TM is more powerful and can be used for a wider range of purposes than CATA.

We have three aims for this article. First, to illustrate how TM might enhance the field, we provide examples of questions in organizational research where TM may help generate insight. Second, we illustrate TM and its methods and provide practical recommendations at each step in the TM process. Third, we demonstrate the application of TM to job analysis by showing how TM can automatically extract job information to derive job skill constructs from job vacancies.

Extant TM reviews and tutorials (Ghosh, Roy, & Bandyopadhyay, 2012; Gupta & Lehal, 2009; Solka, 2008) have mainly targeted readers with a strong affinity with programming and machine learning, and hence have focused on technical aspects of the TM process. Here, our intended audience comprises organizational researchers. We describe the key steps in TM research and aim to enhance organizational researchers' understanding of the concepts behind TM, the different steps involved, and strategies for evaluating the validity of TM-based outcomes. Some technical details have been left out, though references for further reading are provided. With this article we hope to inspire investigations that apply TM to the analysis and understanding of organizational phenomena.

Examples of the Uses of TM

Though to date TM has been largely used for exploratory purposes (i.e., focusing on describing or mapping new phenomena) it can also be applied to explanatory/theory-driven research (i.e., hypothesis testing on the interrelationships between constructs). An example of exploratory TM can be found in Singh, Hu, and Roehl (2007), who identified emerging research streams in human research management by analyzing published literature in this area. However, the latter use is likely more appealing to organizational researchers, as they will likely aspire to validate domain knowledge, models, frameworks, or theories in their research.

Existing knowledge and theory can be empirically tested using TM. An example is a study by Yarkoni (2010) in which he investigated the relationship between personality and language use. Specifically, he counted the frequency of words from 66 psychologically relevant categories (such as "positive emotions," "hearing," "sexuality," and "swear words"; see Pennebaker, Francis, & Booth, 2001) in 694 blogs and correlated them to the five factor model dimension scores obtained by surveying 576 bloggers. He showed that "personality plays a pervasive role in shaping the language people use" (Yarkoni, 2010, p. 371). Another example of such theory-driven work is the study of Guo, Li, and Shao (2015), who developed features derived from the theory of cognitive situational models to cluster documents. In applying the four dimensions of the cognitive situational model (i.e., protagonist, temporality, spatiality, and activity) they managed to reduce feature size and were able to analyze complex semantics. These two studies would have been difficult to conduct with mainstream qualitative research methods, due to the laboriousness of manually coding more than 115,000 words per blog or grouping 825,992 articles.

TM could also be applied to build on work that examines the motive content of leaders or company visions (Kirkpatrick, Wofford, & Baum, 2002). Company reports or CEO statements and speeches could form the textual data to start from (see also Table 1). TM can be used to examine patterns found in data at a single point in time (i.e., cross-sectional) or to investigate changes in patterns over time (i.e., longitudinal; J. Hu, Sun, Lo, & Li, 2015; J. Lee & Hong, 2013). An example of the latter is analyzing business model evolution from annual reports (J. Lee & Hong, 2013). Another application is to analyze open-ended survey responses as was done by Roberts et al. (2014), who presented two illustrations in their introduction of structural topic models. In one illustration they examined how political affiliation influences views on immigration. In another, they analyzed free text containing players' description of their strategies and related them to their game contributions. Theeboom, Van Vianen, Beersma, Zwitser, and Kobayashi (in press) applied TM to explore how coaching criteria differ according to length of coaching experience and whether a coach has a psychology background or not by analyzing coaches' responses to the question of what indicates whether coaching has been successful.

Ultimately, the research question will dictate whether the use of TM is appropriate, and if so the type of text data needed, and the choice of TM technique. Researchers can draw inspiration from existing studies to decide which technique is most suited to reach their specific objective. For instance, in choosing which technique can help identify leadership themes in a corpus of company's mission and vision statements, it could be useful to examine the technique applied to organize news into news themes (Radev, Otterbacher, Winkel, & Blair-Goldensohn, 2005). Table 1 provides a summary of the wide range of questions to which TM can be applied. It contains brief descriptions of TM techniques along with the specific questions they are designed to answer, and includes existing and potential applications of each TM technique.

Key Steps in Text Mining Research

TM generally entails three steps, namely, (a) text preprocessing, (b) application of TM operations, and (c) postprocessing (Y. Zhang, Chen, & Liu, 2015). Figure 1 provides a diagram of the different steps in the TM process and the steps, along which our discussion is organized. Text preprocessing may be further subdivided into text data cleaning and text data transformation (e.g., converting unstructured text into mathematical structures which can serve as input to various TM operations). TM operations refer to the application of pattern mining algorithms with the overriding goal to model characteristics of text. Finally, postprocessing involves interpreting and validating knowledge derived from TM operations. Below we explain each step in detail. Hereafter, all mentions of the word *data* refer to text data. Moreover, we use the words *document* and *text* interchangeably. The appendix provides a glossary of key terms.

Text Preprocessing

Text data collection. Before initiating the TM process one should have text data and the first step in data collection is to decide on the most suitable data source(s). Potential sources include the web, enterprise documents (e.g., memos, reports, and hiring offers), personal text (e.g., diaries, emails, SMS messages, and tweets; Inmon & Nesavich, 2007), and open-ended survey responses. TM requires that text must be in digital form or that it can be transcribed to this form. In case it is not, nondigital text (e.g., handwritten or printed documents) may be digitalized using optical character recognition techniques (Borovikov, 2014). Web text data are collected from websites either through web application programming interfaces (APIs) or web scraping (i.e., automatic extraction of web page content; Olston & Najork, 2010).

Table 1. Summary of Questions That Text Mining Can Address.

Question	Name	Definition	Specific Techniques	Example	Text Representation	Examples of Potential Applications in Organizational Research
How do I assign text to predefined categories?	Text classification	Using an initial set of labeled text, train a classifier that can automatically sort text into existing categories.	Classification algorithms from data mining such as naive Bayes, support vector machines, neural networks, nearest neighbors, random forest, and boosting	• Distinguishing between positive and negative product reviews (Dave, Lawrence, & Pennock, 2003; Popescu & Etzioni, 2007) • Subjective genre classification of product reviews (M. Hu & Liu, 2004; Pang & Lee, 2008) • Assigning semantic attributes to product descriptions (Ghani, Probst, Liu, Krema, & Fano, 2006) • Annotating clinical documents with semantic tags (Jang, Song, & Myaeng, 2006)	Vector space model (i.e., individual terms are used as features); kernel-based methods such as support vector machines deal with text treated as strings; can use other types of features but text is still represented as vectors	Predicting performance and charisma using leaders' collected speeches and biographies (House, Spangler, & Woycke, 1991)
How do I extract topics from a corpus of documents?	Topic modeling	Identify patterns in word frequencies and use the patterns as a basis to define "topics." For each document possible topics are determined.	Latent Dirichlet allocation model and probabilistic latent semantic analysis	• Topic modeling to extract latent evidence during the analysis phase of digital forensic investigations (Waal, Venter, & Barnard, 2008) • Topic models to enhance the feature set for scientific titles classification (Vo & Ock, 2015)	Vector space model where the words are weighted by their frequencies	Analyzing underlying motives or leadership themes from coded interview data, formal vision statements, and company mission or vision statements (Kirkpatrick, Wofford, & Baum, 2002)
How can I form groups of text?	Text clustering	Define a concept of text similarity. Use the concept to group documents together. Each group is called a cluster. Documents in the same cluster are more similar than documents in different clusters.	K-means, hierarchical clustering, biclustering, and nonnegative matrix factorization	• Clustering clinical trial records to narrow down search results about existing protocols (Korkontzelos, Mu, Restificar, & Ananiadou, 2011) • Organizing collections of legal documents and assisting automatic generation of legal taxonomies (Conrad, Al-Kofahi, Zhao, & Karypis, 2005)	Vector space model; can use other types of features but text is still represented as vectors	Investigating patterns of communication between different parties through the analysis of emails or other virtual communication between employees within firms (Holton, 2009; Nenkova & Bagga, 2003)

(continued)

Table 1. (continued)

Question	Name	Definition	Specific Techniques	Example	Text Representation	Examples of Potential Applications in Organizational Research
How can I summarize text and extract keywords and key sentences?	Text summarization	Measuring the importance of each sentence (word) in a document and using a threshold to determine which sentences (words) to retain and which to delete	Content selection using pattern matching, hidden Markov models, and keyword or key phrase extraction	• Biographical summarization (Saggion & Gaizauskas, 2005) • Summarizing web page content for display on small screens of handheld devices (Buyukkokten, Garcia-Molina, & Paepcke, 2001) • Keyword extraction in publications to narrow down and organize query results to support systematic reviews (Ananiadou, Rea, Okazaki, Procter, & Thomas, 2009)	Vector space model; text is treated in terms of strings	Automatic summarization of companies' code of conduct to gain greater understanding of what is or is not currently included in organizational policy on ethical behavior in the workplace
How can I analyze trends in text?	Keyword extraction over time, dynamic topic modeling, and clustering with temporal information	Find interesting terms or topics and analyze changes of usage or prevalence in documents indexed by time.	Most frequent term extraction, and dynamic topic modeling	• Analyze trends in SMS messages by tracking the use of specific keywords (Jonsson, Nugues, Bach, & Gunnarsson, 2010) • Analyzing information in software repositories to model software project progress (J. Hu, Sun, Lo, & Li, 2015) • Tracing significant historical trends in the field of cognition (Cohen Priva & Austerweil, 2015)	Text is treated in terms of strings; vector space model	Analyzing changes in emphasis in companies' code of conduct in reaction to specific events Identifying emergent skills in a corpus of job vacancies (Smith & Ali, 2014)
How can I find other documents that are similar to the one I have?	Distance and similarity	Given a document, find other similar documents	Distance metrics and similarity measures	• Information retrieval (Frakes & Baeza-Yates, 1992) • Input to clustering (Jain, Murty, & Flynn, 1999)	Vector space model	Analyzing conversation between people to facilitate exchange of rewarding information or detection of dangerous activities (Wang & Chen, 2008)

(continued)

Table 1. (continued)

Question	Name	Definition	Specific Techniques	Example	Text Representation	Examples of Potential Applications in Organizational Research
After transforming documents using the vector space model, how can I cope with many variables?	Dimensionality reduction techniques	Reduce the number of variables while preserving relative similarity among documents	Feature selection techniques based on thresholding (e.g., information gain) and feature transformation techniques such as principal component analysis, latent semantic analysis, and random projection	• Most dimensionality reduction techniques promote computational efficiency and a more compact representation of text data (Bingham & Mannila, 2001; Chen, Huang, Tian, & Qu, 2009; Forman, 2003). They have been applied to improve both classification performance and classification interpretability.	Vector space model	The output from this can be used in other techniques such as in classification or clustering

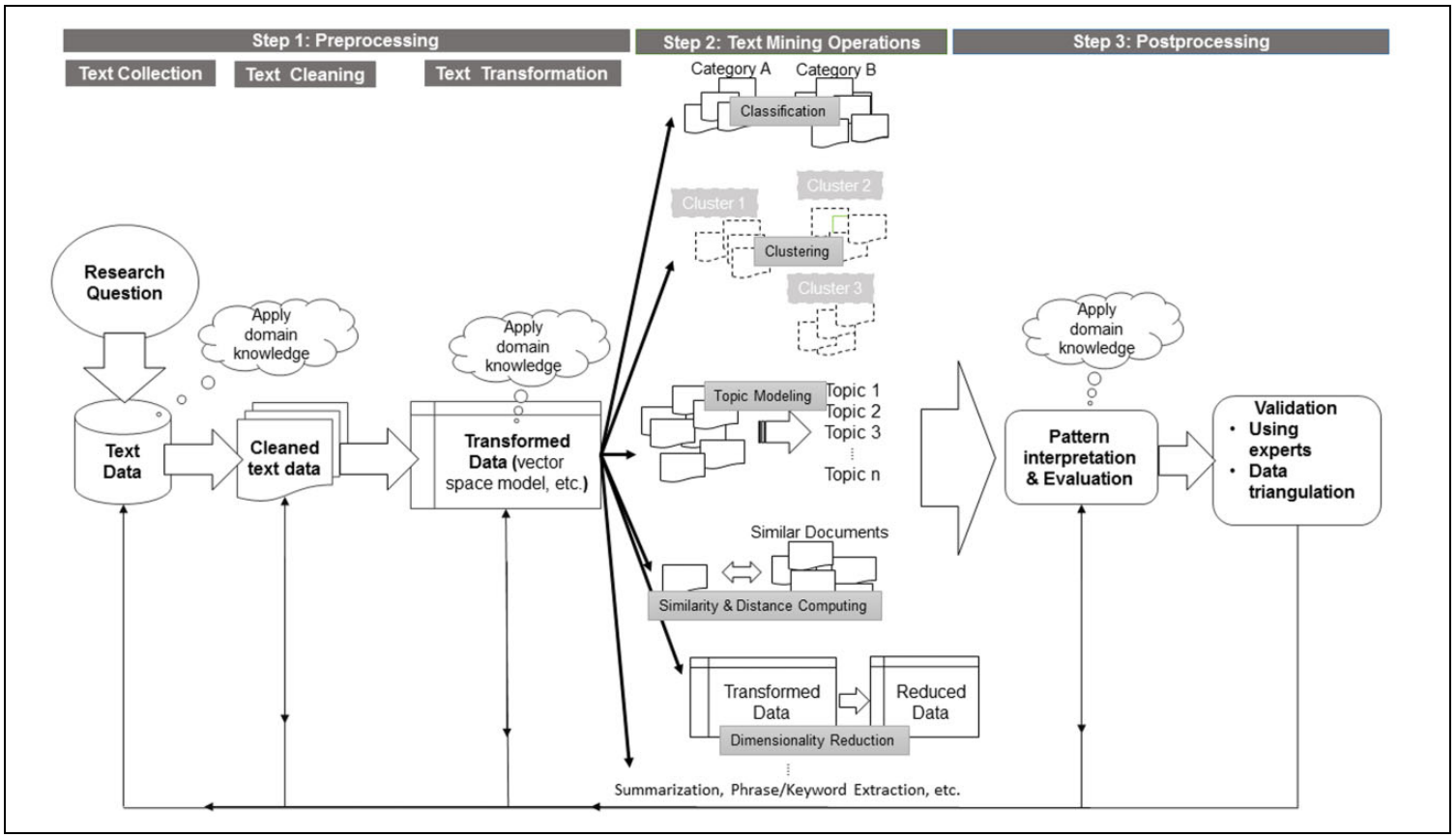

Figure 1. Flowchart of the text mining process.

It is important to be aware of legal and ethical issues associated with data access, particularly web scraping. Website contents are often protected by copyright law and lawsuits may ensue if agreement under fair use is violated (see, for instance, Associated Press, 2013). Also, privacy issues may preclude the use of certain types of personal text data without permission, such as web forms, surveys, emails, and performance appraisals (Van Wel & Royakkers, 2004). Another potential issue to be aware of is that of data storage. In small projects, collected text data can be temporarily stored in a local file system (e.g., on a computer). However, in large-scale text analytics, especially when the data come from different sources, merging, storing, and managing data may require integrated database systems or data warehouses (Inmon, 1996).

Practical Recommendations. As APIs provide an efficient and legal means of obtaining data from the web, researchers who use text from the web first need to find out whether the target website offers an API. One way to find APIs is to use a search platform for APIs such as the ProgrammableWeb (ProgrammableWeb, n.d.). When an API is not available, the next option is web scraping. R has libraries that can automate the web scraping process such as "rvest." Other useful packages include "RCurl" (for http requests), "XML" (for parsing HTML and XML documents), and "stringi" (for text manipulation). If web scraping is not allowed, researchers should ask data owners if they are willing to share their data through remote connections to their databases. Text documents in databases may be fetched using standard query language (SQL).

Text data cleaning. Data cleaning enhances data quality, which in turn enhances the validity of extracted patterns and relationships. Cleaning is done by retaining only the relevant text elements (Palmer, 2010). Standard cleaning procedures for text include deletion of unimportant characters (e.g., extra whitespaces, formatting tags, etc.), "text segmentation," "lowercase conversion," "stop word removal," and "word stemming." For open-ended survey responses (and other informally produced texts such as SMS texts or personal emails), in our experience it may be useful to run a spelling check to correct misspelled words. For web documents, HTML or XML tags must be removed since these do not add meaningful content. Thus the end result is text data stripped of all low content words and characters.

Text segmentation (Huang & Zhang, 2009) is the process of dividing text into sentences and words. Stop words such as conjunctions and prepositions (e.g., and, the, of, for) are words that have low information content and do not contribute much to the meaning in the text. "Stemming" homogenizes the representation of semantically similar words (e.g., representing the words "ensures," "ensuring," and "ensured" by "ensure"). Since these techniques delete words, they also serve to reduce the size of the vocabulary.

Practical Recommendations. A popular stemming algorithm was developed by Porter (Porter, 1980; Willett, 2006). Most of the other aforementioned procedures can be performed by applying string processing. For example in R, the Text Processing part (R Programming/Text Processing, 2014) of the R Programming wiki provides information on how to implement text processing procedures. The "tm" library, the core framework for TM in R, has functions for stop word removal and stemming. The website RANKS NL (n.d.) provides lists of stop words for many human languages. There are cases where it is not appropriate to apply stop word removal and stemming, for example, in short text classification (Faguo, Fan, Bingru, & Xingang, 2010).

An example output after applying lowercase transformation, stop word removal, stemming, and punctuation removal can be found in Table 2. Instead of just deleting "/," it is replaced by a whitespace, otherwise "send/receive" would be merged into the single word string "sendreceive." Extra whitespaces resulting from deleting characters or words are removed.

Table 2a. An Illustration of Text Preprocessing Applied to Original Text.

	Original text	Processed text
D1	Ability or experience in reviewing and authoring aircraft flight manuals, apps spec, and pilot's guides	abil experi review author aircraft flight manual app spec pilot guid
D2	Work with KEMP Management to gain approval for new product concepts/ideas	work kemp manag gain approv product concept idea
D3	Handle client queries and/or requests	handl client queri request
D4	3-5 years of supervisory or product management experience required	3-5 year supervisori product manag experi requir
D5	Understanding of XML, parsing, send/receive, and experience with web services	understand xml pars send receiv experi web servic
D6	Responsible for developing and maintaining quality management procedures and systems	respons develop maintain quality manag procedur system
	Additional text	
D7	Experience with J2EE technology components (e.g., JSP, Servlets, XML, and web services) is a requirement	
D8	Minimum 5 years of experience in marketing or product management roles	
D9	Handling consultant and client queries	
D10	Define customer applications for the product and design product positioning to support these applications	
D11	3-5 years of experience in the engineering and/or maintenance field strongly preferred	

Note: The 6 preprocessed texts were obtained after applying stop word removal, stemming, and punctuation removal except for intraword dashes and stripping extra whitespaces.

Table 2b. Document-by-Term Matrix Constructed From the First Six Texts of Table 2.

	3-5	Abil	aircraft	app	approv	author	client	experi	manag	product
D1	0	1	1	1	0	1	0	1	0	0
D2	0	0	0	0	1	0	0	0	1	1
D3	0	0	0	0	0	0	1	0	0	0
D4	1	0	0	0	0	0	0	1	1	1
D5	0	0	0	0	0	0	0	1	0	0
D6	0	0	0	0	0	0	0	0	1	0

Note: This table is truncated due to space limitations.

Text transformation. Text transformation is a quantification strategy in which text is transformed into mathematical structures. Most analytical techniques require text to be transformed into a matrix structure, where the columns are the variables (also referred to as features) and the rows are the documents. One way to construct this matrix is to use the words or terms in the vocabulary as variables. The resulting matrix is called a "document-by-term matrix" in which the values of the variables are the "weights" of the words in that document. In many applications, this is a straightforward choice since words are the basic linguistic units that express meaning. The raw frequency of a word is the count of that word in a document. Thus in this transformation, each document is transformed into a "vector," the size of which is equal to the size of the vocabulary, with each element representing the weight of a particular term in that document (Scott & Matwin, 1999).

Word frequency, in itself, may not be useful if the task is to make groupings or categories of documents (Kobayashi, Mol, Berkers, Kismihók, & Den Hartog, 2018). Consider the word *study* in a

corpus of abstracts of scientific articles. If the objective is to categorize the articles into topics or research themes then this word is not informative as in this particular context almost all documents contain this word. A way to prevent the inclusion of terms that possess little discriminatory power is to assign weights to each word with respect to their specificity to some documents in a corpus (Lan, Tan, Su, & Lu, 2009). The most commonly used weighting procedure for this is the inverse document frequency (IDF; Salton & Buckley, 1988). A term is not important in the discrimination process if its *IDF* is 0, implying that the word is present in every document. In fact, *IDF* can also be the basis to select stop words for the categorization task at hand. Words that have a low *IDF* have little discriminatory power and can be discarded. When multiplied, the word raw frequency (tf) and *IDF* yield the popular *TF-IDF*, which simultaneously takes into account the importance of a word and its specificity (Frakes & Baeza-Yates, 1992).

Representing text as a document-by-term matrix presupposes that word order information is not crucial in the analysis. Although unsophisticated, it is noteworthy that transformations that ignore word order information perform better in many applications than transformations that account for it (Song, Liu, & Yang, 2005; W. Zhang, Yoshida, & Tang, 2008). The main computational challenge for document-by-term matrix representation is how to deal with the resulting dimensionality, which is directly proportional to the size of the vocabulary. One can use different data dimensionality reduction methods to reduce the number of variables (e.g., variable selection and variable projection techniques) or employ specific techniques suited for data with high dimensionality. These techniques will be highlighted in the Text Mining Operations section.

Once text is transformed, techniques such as regression analysis and cluster analysis can be applied. Combining variables helps tackle substantive questions about the text (see also the Text Mining Operations section). For instance, if resumes of job applicants are used as a data source then the presence of the words *experience* and *year* together with a number can be used to deduce an applicant's work experience.

Practical Recommendations. The output from word segmentation provides the vocabulary. One may start by creating a document-by-term matrix. In R, the "tm" library has a function that can generate a document-by-term matrix, with an additional option for specifying weights. For example, consider the six preprocessed texts in Table 2. Part of the document-by-term matrix constructed from the texts using raw frequency weighting is shown in Table 2. The complete matrix has 40 columns, equal to the number of unique words found in the 6 texts.

Text Mining Operations

Though text transformation precedes the application of analytical methods, these two steps are closely intertwined. The document-by-term matrix from the text transformation step serves as the input data for most of the procedures in this section. Sometimes, when results are unsatisfactory, the researcher may consider changing or enlarging the set of variables derived from the transformation step (Lewis, 1992a; Scott & Matwin, 1999) or choosing another analytical method. Usually different combination of data transformation and analytical techniques are tried and tested and the one that yields the highest performance is selected.

Most TM operations fall into one of five types, namely, (a) dimensionality reduction, (b) distance and similarity computing, (c) clustering, (d) topic modeling, and (e) classification (Solka, 2008). Below we discuss each technique prior to discussing how to assess the credibility and validity of TM outcomes.

Dimensionality reduction. Document-by-term matrices tend to have many variables. It is usually desirable to reduce the size of these matrices by applying dimensionality reduction techniques. Some of the benefits of reducing dimensionality are more tractable analysis, greater interpretability

of results (e.g., it is easier to interpret variable relationship when there are few of them), and more efficient representation. Compared to working with the initial document-by-term matrices, dimensionality reduction may also reveal latent dimensions and yield improved performance (Bingham & Mannila, 2001). Two general approaches are commonly used to reduce dimensionality. One is to construct new latent variables and the second is to eliminate irrelevant variables. In the former case, new variables are modeled as a (non)linear combination of the original variables and may be interpreted as latent constructs (e.g., the words *years*, *experience*, and *required* may be merged to express the concept of work experience in job vacancies).

Singular value decomposition (SVD) is a classic tool that underlies techniques such as latent semantic analysis (Landauer, Foltz, & Laham, 1998) and principal component analysis (PCA; Jolliffe, 2005). The SVD method decomposes a matrix X of size $p \times n$ (where p is the number of variables and n is the number of documents) into a product of three matrices, that is, $X = U \sum V^T$. One of these is a diagonal square matrix ($\sum$) which contains the singular values (Klema & Laub, 1980). Reducing the number of dimensions involves retaining the first few largest singular values. Usually, this implies choosing latent dimensions and recovering the underlying dimensionality of the data because at times, true dimensionality is obscured by random noise.

LSA is commonly used to detect synonymy (i.e., different words that have the same meaning) and polysemy (i.e., one word used in different yet related senses) among words. PCA is effective for data reduction as it preserves the variance of the data. Parallel analysis (Ford, MacCallum, & Tait, 1986; Hayton, Allen, & Scarpello, 2004; Montanelli & Humphreys, 1976) is the recommended strategy to choose how many dimensions to retain in PCA. A disadvantage of both LSA and PCA is that it may be difficult to attach meaning to the constructed dimensions. Another technique is random projection, where data points are projected to a lower dimension while maintaining the distances among points (Bingham & Mannila, 2001).

An alternative approach to reduce dimensionality is to eliminate variables by using variable selection methods (Guyon & Elisseeff, 2003). In contrast to projection methods, variable selection methods do not create new variables but rather select from the existing variables by eliminating those that are uninformative or redundant (e.g., words that occur in too many documents might not be useful for categorizing documents). Three types of methods are available: filters, wrappers, and embedded methods. Filters assign scores to variables and apply a threshold to scores to delete irrelevant variables. Popular filters are TF-IDF thresholding, information gain, and the chi-square statistic (Forman, 2003; Yang & Pedersen, 1997). Wrappers select the best subset of variables in conjunction with an analytical method. In embedded methods, searching the best subset of variables is accomplished by minimizing an objective function that simultaneously takes into account model performance and complexity. Model performance can be measured for example by prediction error (in the case of classification) and complexity is quantified by the number of variables in the model. The smallest subset of variables yielding the lowest prediction error is the preferred subset.

Practical Recommendations. The dimensionality reduction stage is usually initiated by applying LSA. The LSA results (i.e., the reduced data set) can be used as input to clustering and classification. Alternatively, one can apply one of the filter methods to trim out unimportant variables. One advantage of filters as compared to LSA is interpretability since no new variables are constructed. Moreover, filter methods are faster to run. The R package "lsa" provides functionality for running LSA.

Consider the 11 texts in Table 2, which, after data cleaning, are transformed into a document-by-term matrix. Running LSA on the transpose of the document-by-term-matrix, we retained 2 dimensions. The resulting LSA space represented as a matrix is presented in Table 3. Observe that the value for "product" in Document 11 is 0.32 although Document 11 does not contain the word "product." This is due to the presence of the word "experience" in the other two documents (Document 4 and Document 8) that also contain "product." Since "experience" is present in

Table 3a. Lower Rank Approximation of the Term-by-Document Matrix Obtained From Table 5 Using LSA by Retaining 2 Dimensions.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
3-5	0.23	0.16	0.00	0.35	0.26	0.07	0.34	0.31	0.00	0.22	0.25
abil	0.15	–0.01	0.00	0.11	0.18	0.01	0.23	0.08	0.00	–0.10	0.12
aircraft	0.15	–0.01	0.00	0.11	0.18	0.01	0.23	0.08	0.00	–0.10	0.12
experi	0.88	0.22	0.00	0.94	1.02	0.15	1.30	0.78	0.00	0.05	0.81
product	–0.03	0.95	0.00	0.94	–0.08	0.28	–0.07	0.95	0.00	1.97	0.32
web	0.41	–0.05	0.00	0.28	0.49	0.02	0.62	0.21	0.00	–0.30	0.32
work	–0.01	0.13	0.00	0.12	–0.03	0.04	–0.03	0.13	0.00	0.28	0.04
xml	0.41	–0.05	0.00	0.28	0.49	0.02	0.62	0.21	0.00	–0.30	0.32
year	0.31	0.29	0.00	0.55	0.35	0.11	0.46	0.49	0.00	0.45	0.37

Note: This table is truncated.

Table 3b. Sample Topics Extracted From the 11 Texts for LDA and CTM.

	Latent Dirichlet Allocation (LDA)			Correlated Topic Model (CTM)		
	Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
Terms	product manag applic experi year	abil aircraft app approv author	experi client handl queri servic	experi manag year product 3-5	abil aircraft app author develop	product applic handl queri client
Documents	4, 6, 8, 10	1, 2	3, 5, 7, 9, 11	2, 4, 7, 8, 11	1, 6	3, 5, 9, 10

Document 11, LSA expects to find "product" in this document. This is how LSA deduces meaning from words (which is also useful for the identification of synonyms).

Distance and similarity computing. Assessing the similarity of two or more documents is a key activity in many applications such as in document retrieval (e.g., document matching), and recommendation systems (e.g., for finding similar products based on product descriptions or reviews). Numerous measures that operate on vector representations may be employed to assess distance or similarity. An example of the latter is the cosine measure that is used extensively in information retrieval (Frakes & Baeza-Yates, 1992). The values for this measure range from –1 (two vectors point in opposite direction) to 1 (two vectors point in the same direction); 0 means that the two vectors are orthogonal or perpendicular (or uncorrelated). This measure assesses the similarity of two documents based on the frequencies of terms they share, which are taken to indicate similarity of content. This measure has been applied to document matching (Frakes & Baeza-Yates, 1992) and detecting semantic similarity (Mihalcea, Corley, & Strapparava, 2006).

Of the various distance measures, Euclidean and Hamming distance measures are most commonly employed. Unlike similarity measures, higher values for distance measures implies dissimilarity. Also distance measures have to satisfy certain properties such as nonnegativity and triangle inequality. In most cases similarity measures can be converted to distance measures and vice versa.

Clustering. Many tasks in TM involve organizing text in groups such that documents belonging to the same group are similar and documents from different groups are not (Jain, Murty, & Flynn, 1999;

Steinbach, Karypis, & Kumar, 2000). The process of grouping is called *clustering*. The main use of text clustering is either to organize documents to facilitate search and retrieval or to impose an automatic categorization of documents. For example, text clustering has been used to detect crime patterns (e.g., location, type of crime, weapons) in crime reports (Bsoul, Salim, & Zakaria, 2013), to organize and deepen the taxonomy of legal practice areas (Conrad, Al-Kofahi, Zhao, & Karypis, 2005), and to improve the performance of a document retrieval system or web-based search engine by creating a taxonomy of documents and grouping the search query results (Osinski & Weiss, 2005). To perform text clustering the researcher needs to define distance between texts (e.g., Euclidean distance). The distance measure can be computed from the original set of variables or from the reduced set of variables (e.g., after application of dimensionality reduction techniques such as LSA).

Most clustering algorithms are categorized as either hierarchical or partitional (Steinbach et al., 2000). Hierarchical clustering algorithms either treat each object as its own cluster and then gradually merge clusters until all objects belong to a single cluster (i.e., agglomerative) or by first putting all objects under one cluster and recursively splitting clusters until each object is in its own cluster (i.e., divisive). The merging (or splitting) of clusters is depicted by a tree or dendrogram. For partitional clustering the user has to specify the number of clusters a priori and clusters are formed by optimizing an objective function that is usually based on the distances of the objects to the centers of the clusters to which they have been assigned. The popular k-means algorithm is an example of partitional clustering (Derpanis, 2006). One key challenge in clustering is the determination of how many clusters to form. Since clustering is an exploratory technique, a common strategy is to experiment with different number of clusters and use cluster evaluation measures to decide. Examples of quality measures are the Dunn index and the silhouette coefficient (Rendón, Abundez, Arizmendi, & Quiroz, 2011).

Practical Recommendations. One can start with k-means or a hierarchical approach such as the complete linkage or Ward's method (El-Hamdouchi & Willett, 1989). If a researcher has a clear idea of how many clusters to create, then k-means is a good start. If a researcher has no idea as to how many clusters to construct, then she may use hierarchical clustering to see whether interpretable groupings emerge. The "cluster" and "mclust" packages in R run most of the clustering techniques described here and the "proxy" package offer various distance and similarity measures.

Using the reduced dimension set from LSA and a distance metric derived from cosine similarity we applied hierarchical clustering based on Ward's method on the 11 texts (see Table 2). The resulting dendrogram is shown in Figure 2. The dendrogram basically shows two clusters: one cluster is about customer and product management and the other pertains to technical requirements on technology use.

Topic models. Topic models automatically extract topics from documents. These topics can indicate underlying constructs or themes. In machine learning and natural language processing, topic models are probabilistic models that are used to discover topics by examining the pattern of term frequencies (Blei, Ng, & Jordan, 2003). Its mathematical formulation has two premises: A topic is characterized by a distribution of terms and each document contains a mixture of different topics. The most likely topic of a document is therefore determined by its terms. For example, when an open-ended survey response, contains words such as *pay*, *compensation*, *salary*, and *incentive*, one might label its topic as "rewards or pay systems."

Perhaps the most popular topic models are the latent Dirichlet allocation (LDA; Blei et al., 2003) model and the correlated topic model (CTM; Blei & Lafferty, 2007). LDA and CTM both operate on the document-by-term matrix (Porteous et al., 2008). CTM will yield almost the same topics as LDA. The main difference between the two is that in LDA topics are assumed to be uncorrelated, whereas in CTM topics can be correlated. In comparing LSA with LDA, the latter has been found to be particularly suitable for documents containing multiple topics (S. Lee, Baker, Song, & Wetherbe, 2010).

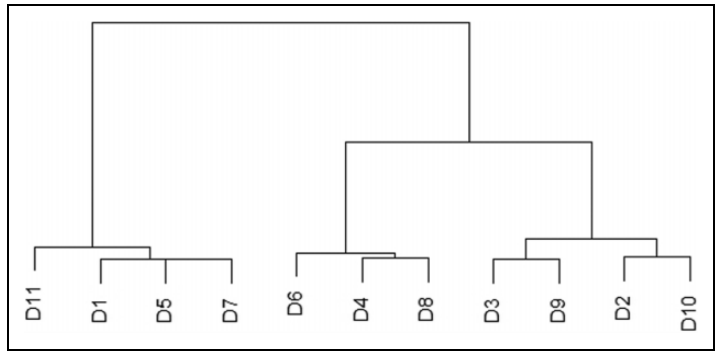

Figure 2. Cluster dendrogram of 11 texts.

Practical Recommendations. For topic extraction, the recommended initial approach is to try LDA. It may also be useful to investigate the assignment of documents to topics. Code to run topic models is available in the "topicmodels" package in R. For example, we ran LDA and CTM on the example text above in Table 2 (see Table 3 for the results). The top terms listed in each topic form the basis for topic interpretation. For example, the top terms of Topic 1 indicate that this topic is about product management, whereas Topic 2 is more about ability on aircraft apps, and Topic 3 about the handling of clients. Examining the most likely topic for each document we observe that Documents 3, 9, and 11 have Topic 3 as the most likely topic since these documents are focused on dealing with customers.

Classification. Classification is the assignment of objects to predefined classes or categories. Logistic regression is perhaps the best known classification method. The goal is to construct a model that can predict the category of a given document. Example applications of text classification are spam or ham classification of emails (Youn & McLeod, 2007), authorship identification (Houvardas & Stamatatos, 2006), thematic categorization (Phan, Nguyen, & Horiguchi, 2008), and identification of sentiments in product reviews (Dave, Lawrence, & Pennock, 2003; M. Hu & Liu, 2004; Pang & Lee, 2008; Popescu & Etzioni, 2007). For a fuller discussion and tutorial on text classification, we refer the reader to Kobayashi et al. (2018).

Evaluation. Model evaluation helps us choose which among competing models best explains the data (Alpaydin, 2014). Model evaluation needs to address issues related to underfitting and overfitting. Underfitting happens when the model does not adequately represent the relationships present in the data (i.e., high variance). Overfitting occurs when a model performs well on data used to build it but poorly on new data (i.e., high bias). Hence, a model generalizes well if it also demonstrates good performance on new data (Mitchell, 1997). A common way to assess the quality of the model's generalizability is to use hold-out data (Alpaydin, 2014). The procedure involves repeatedly splitting the corpus of documents to create a training and a test set either by randomly sampling documents from the corpus or by partitioning the corpus. Documents in the training set are used to fit the model and the generalizability of this model is assessed using the documents in the test set. Procedures that evaluate a model by partitioning the corpus are K-fold cross validation and a resampling procedure called bootstrapping (Kohavi, 1995). Measures to assess generalizability are commonly referred to as evaluation metrics. Since different values of the metric for each unique split will be obtained, values are usually averaged across splits. Using cross-validation and bootstrapping, one can build confidence intervals and assess the true performance of the model. The choice of metric is dependent

on the task and application domain. However, it should be kept in mind that conclusions generated are conditioned on the data; that is, a model is good only insofar as the data are representative of the population. Second, there are other criteria to judge the merit of a model, such as the time it takes to build the model and its interpretability.

Practical Recommendations. In topic modeling, one can use the aggregate topic probabilities of unseen documents (Wallach, Murray, Salakhutdinov, & Mimno, 2009) as an evaluation metric. In clustering, internal and external evaluation criteria are used. External criteria use previous knowledge about the data (i.e., prior information) and internal criteria only use the data. We already mentioned two criteria in the clustering section, which are Dunn's index and the silhouette coefficient (Rendón et al., 2011).

Both dimensionality reduction and distance and similarity computing are usually evaluated on their impact on the text classification and text clustering performance (Forman, 2003). That is, an effective dimensionality reduction technique must contribute to the improvement of classification or clustering performance. An analogous comment can be made for distance and similarity computing, since these measures often serve as input to the clustering (e.g., k-means) and classification task (e.g., nearest neighbor), although there are applications where distance and similarity measures are used as a standalone method (Houvardas & Stamatatos, 2006; Lewis, 1992b; Mihalcea et al., 2006). An example of the latter is comparing (parts of) leader and subordinate resumes to operationalize person-supervisor similarity. In information retrieval, where the task is to match queries to document content, performance metrics for distance or similarity measures are precision, recall, and the F-measure.

Postprocessing

The postprocessing step may involve domain experts to assist in determining how the output of the models can be used to improve existing processes, theory, and/or frameworks. Two major issues are usually addressed here. The first is to find out whether the extracted patterns are real and not just random occurrences due to the sheer size of the data (e.g., by applying Bonferroni's principle). The second is, as with all empirical research, whether data and results are valid. Establishing the reliability, validity (e.g., content, predictive, and discriminant validity), and credibility of the output of TM models is particularly important for TM to gain legitimacy in organizational research. It is important to note here that it is not the TM procedures that need to be validated but the output (in the same manner that we do not validate factor analysis), for example, the predictions of a TM-based classifier.

Prior to being applied to support decision making and knowledge generation, the validity of TM-based findings will need to be established. When TM is used to identify and operationalize constructs, using different forms of data triangulation will help generate construct validity evidence. For example, in our job analysis example of TM application, which follows below, we enlisted the help of job analysts and subject matter experts (SMEs) in evaluating the output of the TM of vacancy texts. In other cases, TM outcomes could be compared to survey data, such as for the aforementioned study on the role of personality in language use (Yarkoni, 2010). More generally, TM-based models will require a comparative evaluation in which (part of) the TM output is correlated with independent data sources or other "standards" (such as the aforementioned survey or expert data). Though it is easy to view TM as a mechanistic means of extracting information from data, the input of domain experts is critically important. Finally, there is no reason why validity assessment procedures, such as those outlined by Binning and Barrett (1989) to establish the validity of personnel decisions, cannot be applied to TM output.

Practical Recommendations. A straightforward practice for construct validation is to have independent experts validate TM output. For example, in text classification, SMEs may be consulted from

time to time to assess whether the resulting classifications of text are correct or not. A high agreement between the experts and the model provides an indication of the content-related validity of the model. The agreement is usually quantified using measures such as the Cohen's kappa or intraclass correlation coefficient.

Another way to validate TM output is through replication, data triangulation, and through an indirect inferential routing (Binning & Barrett, 1989). The standard can be established by obtaining external data using accepted measures or instruments that may provide theory based operationalizations that should or should not be correlated to the model. Such correlations give an indication of validity. For example, to validate experience requirements extracted from job vacancies, one can administer questionnaires to job incumbents asking them about their experience. Validity is then ascertained through the correlation between both operationalizations. This can be replicated on various types of text to assess if the TM model consistently generates valid experience requirements for a particular occupation. In theory, one could even compute full multitrait multimethod correlation matrices (Campbell & Fiske, 1959) to compare the measurements obtained from TM with established instruments, although in practice it may be difficult to obtain the fully crossed dataset that it requires.

Text Mining Applied to Job Analysis

To illustrate the key steps in TM we provide an example from job analysis. Job analysis aims to collect and analyze any type of job-related information to describe and understand a job in terms of behaviors necessary for performing the job (Sanchez & Levine, 2012; Voskuijl, 2005). Job analytic data are traditionally collected through interviews, observations, and surveys among SMEs, including job holders, supervisors, and job analysts (Morgeson & Dierdorff, 2011). Here, we apply a TM approach to automatically classify job information from vacancies and assess whether the worker attributes necessary for effective job performance emerge from the vacancies to show that TM might be useful tool to job analysts.

Job analysis may be a fertile ground for TM due to the abundance of textual sources of job information, a prime example being online job vacancies. The ability to analyze a big corpus of job vacancies addresses several limitations of existing job analysis data collection strategies. Job vacancies provide up-to-date job information, offer the potential to capture the dynamism of jobs in the contemporary workplace (Sanchez & Levine, 2012), and may be used to reduce bias inherent in existing data collection strategies for job analysis (Dierdorff & Morgeson, 2009; Morgeson & Campion, 1997, 2000; Morgeson, Delaney-Klinger, Mayfield, Ferrara, & Campion, 2004; Sanchez & Levine, 2000). Also, job vacancies are inexpensive and relative easy to obtain and since TM techniques are based on algorithms that are optimized for performance, job information extraction can be made efficient and reliable (McEntire, Dailey, Osburn, & Mumford, 2006).

Previous studies in the field of information technology have extracted skills from vacancies by counting the frequency of preselected keywords related to computer programming (e.g., Java, Python, etc.; Smith & Ali, 2014; Sodhi & Son, 2010). One limitation of this approach is that it may not be effective in detecting emergent skills because researchers may fail to specify the appropriate keywords. Here we assess whether TM is of use in analyzing the content of vacancies. In line with the work of Sackett and Laczo (2003), it seems useful to be able to disentangle information referring to worker attributes on one hand and work activities on the other. Also, worker attribute requirements may differ across job professions and/or job industries. Determining key worker attributes (e.g., technical skills) for specific jobs and how these worker attributes compare and contrast across jobs could be of use in job classification (Harvey, 1986), training needs analysis (Arthur, Bennett, Edens, & Bell, 2003), compensation (Verwaeren, Van Hoye, & Baeten, 2016), recruitment (Abdessalem & Amdouni, 2011), and the generation of synthetic validity evidence (Scherbaum, 2005).

Classification of Job Information Types

Preprocessing. We partnered with two organizations, namely, Monsterboard and Textkernel, which provided access to vacancy data from various employment websites. Of the different fields that vacancies usually contain, our focus is on the *job description* field, which usually lists activities associated with the job and the attributes required from the applicants.

Since our analysis operates at the sentence level and some of our variables are derived from the words, we started by applying sentence and word segmentation. Once words and sentences were identified, we converted letters to lowercase and removed stop words. The criteria used to determine whether a word is a stop word or not were based on the standard English language (RANKS NL, n.d.) stop word list and our own inductive identification of words that did not appear to be associated with the types of job information we were interested in detecting. Hence, conjunctions, articles, and prepositions were deleted. We retained the following stop words, "to," "have," "has," "had," "must," "can," "could," "may," "might," "shall," "should," "will," "would," because these were useful for the classification task. Specifically, sentences containing "to" and "will" often contain job activities, whereas, "have," "has," "had," "should," and "must" are suggestive of worker attributes. Having deleted the irrelevant stop words, we removed punctuation except for intraword dashes to avoid separating words which together express a single meaning (e.g., problem-solving, customer-oriented, pro-active, etc.). Finally, we stripped the extra whitespaces that resulted from the deletion of particular characters. The output was a collection of sentences in which all letters were in lowercase from which the irrelevant stop words and punctuation had been stripped.

For the transformation step we deviated from the approach of using solely words as variables. We generated a list of variables that would potentially be able to predict the category membership of sentences, that is, into either work activities (e.g., tasks) or worker attributes (e.g., skills). We used knowledge from the job analysis field and eye-balling coupled with statistical tests to preselect these variables. Based on definitions of tasks, for example, we deduced that often, these are indicated by sentences that consist of an action verb, the object of the action, the source of information or instruction, and the results (Morgeson & Dierdorff, 2011; Voskuijl, 2005). We also expected verbs to be more prevalent in activity sentences than in attribute sentences. Using part-of-speech tagging, we computed features such as the percentage of verbs in a sentence and the part-of-speech of the first word. For the POS labels, we based the tags on the "Penn part of speech tags" (Penn Part of Speech Tags, n.d.).

For the purposes of our analysis, we put all noun, verb, adjective, and adverb related tags together. We grouped related tags under one general derived tag, since we did not require detailed information about each tag. For example, singular or mass noun (NN), plural noun (NNs), singular proper noun (NNP), plural proper noun (NNPS) were all subsumed under the "noun" tag. Other noteworthy tags are TO (to), CD (cardinal number) and MD (modal), these tags appeared important for discriminating between work activities and worker attributes. The TO tag is indicative of job activity (e.g., "*to* ensure project stays on track for assigned client projects"), because it reflects either the results of an action or the indefinite form of a verb in a task. The presence of a CD tag most often points to the years of education or work experience required from job applicants, and hence is indicative of the worker attribute category. The complete list of variables can be found in Table 4.

A total of 168 variables were constructed. In the future, to create finer (sub)classifications of job information types, additional features will likely be needed. We computed the 168 variables for each sentence (our unit of analysis) and constructed vectors that represent each sentence. We then collected the vectors in a data matrix.

Application of classification techniques. The data matrix served as the input data for the classification of job information. To construct models, we needed labeled training data. We examined each sentence

Table 4. The 168 Variables for the Vacancy Mining task.

Feature Type	Number of Derived Features	Variable Type
Part of speech (POS) tag of the first word	1	Categorical (actual POS)
Is the first word in this sentence unique in work activity sentences (based on the labeled data)?	1	Numeric
Is the first word in this sentence unique in worker attribute sentences (based on the labeled data)?	1	Numeric
Is the last word in this sentence unique in work activity sentences (based on the labeled data)?	1	Numeric
Is the last in this sentence unique in worker attribute sentences (based on the labeled data)?	1	Numeric
Proportion of adjectives	1	Numeric
Proportion of verbs	1	Numeric
Proportion of the word "to"	1	Numeric
Proportion of modal verbs	1	Numeric
Proportion of numbers	1	Numeric
Proportion of adverbs	1	Numeric
Proportion of nouns	1	Numeric
Proportion of nouns, verbs, adjectives, adverbs, and other part of speech tags followed by another verb	5	
Proportion of unique words found only in work activity sentences (based on the labeled data)	1	Numeric
Proportion of unique words found only in worker attributes sentences (based on the labeled data)	1	Numeric
Frequency of keywords for work activity and worker attributes sentences	149	Numeric

and employed standard definitions from the job analysis literature (these were accumulated in a coding manual that is available in the Supplemental Materials) to label each sentence as either a work activity or a worker attribute. In establishing the labeled training data, mixed sentences containing both activity and attribute information were split and buffer sentences not containing any relevant information were dropped. For the construction of the classification model we added a 169th column to the data matrix. This column contained the classification of sentences into either job attribute (0) or job activity (1) as derived from the manually labeled sentences.

For the classifier, three techniques were tested, namely naive Bayes (NB), support vector machine (SVM), and random forest (RF). We chose these as they are purportedly the most effective classifiers for text classification (Aggarwal & Zhai, 2012). We built each classifier and assessed its performance through 10-fold cross-validation using accuracy and F-measure as performance metrics. These performance metrics reflect our objective of creating an accurate classifier that favors neither one of the categories (attribute or activity).

The parameter set for each technique and the classification results are summarized in Table 5. The mean of the two metrics from the 10-fold cross validation suggests that SVM and RF perform better than NB. A comparison of the mean accuracies using a one-way ANOVA found that at least one mean accuracy was different from the rest, $F(2.27) = 15.94$, $p = .000$. A post hoc analysis using Tukey's honestly significant difference (HSD) method revealed that the mean accuracy of NB is significantly different from the other two techniques (RF, $p = .001$; SVM, $p = .000$), whereas SVM and RF did not significantly differ from one another ($p = .988$). These high accuracies can be explained by the appropriateness of the extracted variables and the suitability of these classifiers for

Table 5. Parameters and Performance Metrics for the Three Classifiers.

	Parameter	Accuracy (%)	F-measure for Job Activity	F-measure for Job Attribute
Support vector machine	Dot product kernel Cost of misclassification = 1	97.30	.9703	.9751
Random forest	Number of trees grown = 500 Number of variables sampled at each split = 4	97.31	.9700	.9750
Naive Bayes	Laplace smoothing = 0.01	96.60	.9463	.9554

studying text data. To make predictions even more valid, one can aggregate them (e.g., by means of majority voting).

We then ran the classifier on over a million sentences and obtained an additional 270,000 work activity sentences and 317,000 worker attribute sentences. These are the sentences in which all three classifiers agree and have high confidence on their predictions.

Postprocessing. Since it is difficult to find job experts that have expertise across job professions, the following discussion of validity is based solely on nursing jobs and experts in those. Specifically, we wanted to assess whether the extracted work activities for nurses correspond to actual nursing tasks. We validated the TM application to job analysis in two ways. First, we asked a nursing expert (i.e., training coordinator) to examine the condensed list of 76 nursing tasks that we extracted from the nursing vacancies for consistency with the actual tasks executed in practice. The 76 nursing tasks were obtained by first extracting task sentences from vacancies, and then applying clustering to group similar tasks together. Hence, we only presented core nursing tasks to the expert.

The SME classified 93.3% of the extracted tasks as representative of actual nursing jobs. The expert validation provided initial support for the content validity of the TM model as the collected information from the vacancies appears to accurately reflect the job. Second, we compared the TM results with traditional job analysis, namely a task inventory, to validate our results by data triangulation. The task inventory consisted of four interviews and a two-day observation with SMEs (i.e., nurses and head nurse) from two German hospitals. More information about the task inventory is available in the Supplemental Materials). Tasks from both lists were rated as synonyms (i.e., exactly the same), similar (i.e., different wording, same meaning), or dissimilar (i.e., different wording and meaning) based on the decision rules of Tett, Guterman, Bleier, and Murphy (2005). Based on this comparison 55.6% of all tasks were found in both lists, whereas 29.1% were unique to the task inventory and 15.2% to the online vacancies. The relatively high correspondence ($\geq$50%) between the list of task collected by TM and the list of tasks collected in the task inventory further established convergent validity.

Topic Modeling on Worker Attributes

We now proceed with our second of aim of analyzing all of the extracted worker attributes (i.e., not restricted to solely those of the nurses). Our goal is to summarize the worker attributes and find worker attribute constructs and use these to cluster jobs. For this purpose we applied topic modeling using LDA to the extracted worker attribute sentences. We set the number of topics equal to 140 based on two criteria. One criterion is based on topic distances as discussed in the article of Cao, Xia, Li, Zhang, and Tang (2009) and the other is based on the idea that LDA is a matrix factorization mechanism and the quality of the factorization depends on choosing the right number of topics (for additional information we refer the reader to the article of Arun, Suresh, Madhavan, & Murthy,

Table 6. Some of the Topics Obtained From Applying LDA to Worker Attribute Sentences.

Topic 100	Topic 86	Topic 132	Topic 75
development	new	travel	communication
software	learn	willingness	written
agile	quickly	willing	oral
methodologies	willingness	work	verbal
application	adapt	time	interpersonal
scrum	technologies	needed	presentation
design	internet	internationally	effective
life	desire	international	listening
Topic 18	**Topic 45**	**Topic 20**	**Topic 105**
highly	detail	sales	results
motivated	attention	selling	leadership
oriented	oriented	salesforcecom	others
self	organizational	outside	goals
driven	accuracy	crm	achieve
organized	multitask	success	influence
starter	follow	account	motivate
selfstarter	details	inside	deliver
Topic 60	**Topic 15**	**Topic 55**	**Topic 108**
scripting	attitude	design	problem
python	positive	adobe	solving
linux	can	creative	analytical
programming	energetic	photoshop	solver
java	team	user	troubleshooting
perl	flexible	illustrator	approach
languages	enthusiastic	graphic	abilities
unix	professional	production	capabilities
Topic 61	**Topic 129**	**Topic 81**	**Topic 16**
license	work	management	data
valid	team	time	analysis
drivers	independently	project	quantitative
driving	part	organizational	research
record	environment	change	statistics
transportation	pressure	people	economics
reliable	members	planning	statistical
vehicle	member	pm	modeling

2010). We use variational expectation maximization to estimate the parameters of the LDA model. For the interest of space and purpose of illustration we show in Table 5 a subset of twelve topics generated from LDA. Looking at the top 8 words, Topics 75, 18, 45, 108, and 129 appear to point to behavioral/personal qualities. Topic 75 could be interpreted as interpersonal communication skills, Topic 18 as self-motivation, Topic 45 seems to pertain to attention to detail, Topic 108 seems to be about analytical and problem-solving skills, and Topic 129 about team-working. Topics 132 and 16 are attributes that were seldom considered in job analysis studies (e.g., Harvey, 1986) and may as well reflect new worker attributes sought by contemporary organizations. Topic 132 seems to be about willingness to travel and the ability to operate on a flexible work schedule and Topic 16 about data analytical skills. The rest of the Topics seem to be about technical skills specific to certain professions such as sales for Topic 20 and software/programming for Topics 100 and 60. Topic 61 pertains to a specific requirement and is about having a valid driving license. Interestingly, even without giving LDA prior information about which worker attributes to expect it still appears to

recover both technical and soft skill requirements. Though it is a bit difficult to interpret Topics 86, 105 and 15 they seem to be topics pertaining to generic personal qualities such as the ability to learn new things quickly (86), goal-setting and leadership (105), and possessing a positive, energetic, and enthusiastic attitude (15). We can visualize the correlations among words within each topic to aid interpretation. We show this in Figure 3 where an edge between words indicate a correlation of at least 0.1 and the thickness of an edge indicates the strength of correlation. The word-networks are in line with our interpretations and show that a topic could capture more than 2 worker attributes; the model put them in one topic because they tend to co-occur. From the topics we can generate hypotheses about which behavioral/personal characteristics are actually required to carry out a particular job, which could then be tested in an empirical study.

Investigating the relationship between topics provides a way to assess the convergent/divergent validity of the topical content. Here we cannot directly use correlation since topics are assumed to be uncorrelated, however, we can use the "distance" between topics. To get a better idea about how to judge whether an association is low or high, we suggest using simulation techniques such as Monte Carlo or permutation tests. In this case, magnitude is always application dependent. To compute distance, we use the Jensen-Shannon divergence which measures the distance between probability distributions. Here we focus the discussion on Topic 75, which we previously interpreted as interpersonal communication skills. Topic 75 is closer to Topics 13, 30, 51, 88, 111, 129, and 103 (please refer to the Supplemental Materials for the complete list of topics). Topics 13 (effective oral and written communication), 30 (professional demeanor), 129 (team work), and 103 (analytical and problem solving skills) all relate to interpersonal skills hence these qualities are expected to relate to interpersonal communication. A noteworthy similarity exists between Topics 132, 77, and 119 which are willingness to travel, ability to work on a flexible schedule, and work relocation, respectively. We can further explore this relationship by performing a more inference driven investigation by comparing the findings here to the results obtained by interviewing SMEs or job holders, which will further help in establishing construct validity. Aside from similar topics there are also less similar ones, for example Topic 75 (interpersonal communication skills) is least similar to Topics 31 (finance) and 89 (programming languages). Possible interpretations include range restriction (that is, if job incumbents in a position do not vary on certain characteristics these characteristics may not be mentioned in the vacancies), but it could also mean that interpersonal communication is not essential to perform jobs requiring those specific technical skills, or that incumbents who excel in those jobs have low interpersonal communication skills.

To examine the relationship among all topics simultaneously, we applied multidimensional scaling and projected the topics on 2 dimensions. Figure 4a shows the projections of topics on 2 dimensions. Topics 7, 8, 9, 6, 25, and 35 (bottom rightmost, fourth quadrant) are close together because they all relate to programming or software skills. This also holds for Topics 123, 124, 128, 107, and 133 (bottom leftmost, third quadrant) which are about written and oral communication skills. Topics 46, 52, 50, 83, and 31 (upper between first and second quadrants) are about how someone should work (fast paced and dynamic) and the qualities needed to perform the work (adaptable, able to multitask, and can work independently or in a team).

The output from LDA allows us to determine the most likely topic for each document. Here we want to find the most likely worker attribute for each job. Consider Topics 16 and Topic 18. Most jobs under Topic 16 are quantitatively oriented jobs such as data scientist, statistician, and financial analyst. On the other hand jobs under Topic 18 appear to pertain mostly to sales, marketing, and customer management. Note that in LDA, each document can have more than one topic (each document is actually a mixture of topics), we can utilize all topic probabilities for each document and construct a hierarchical clustering of jobs. In Figure 4b we show part of the cluster dendrogram highlighting medically related jobs.

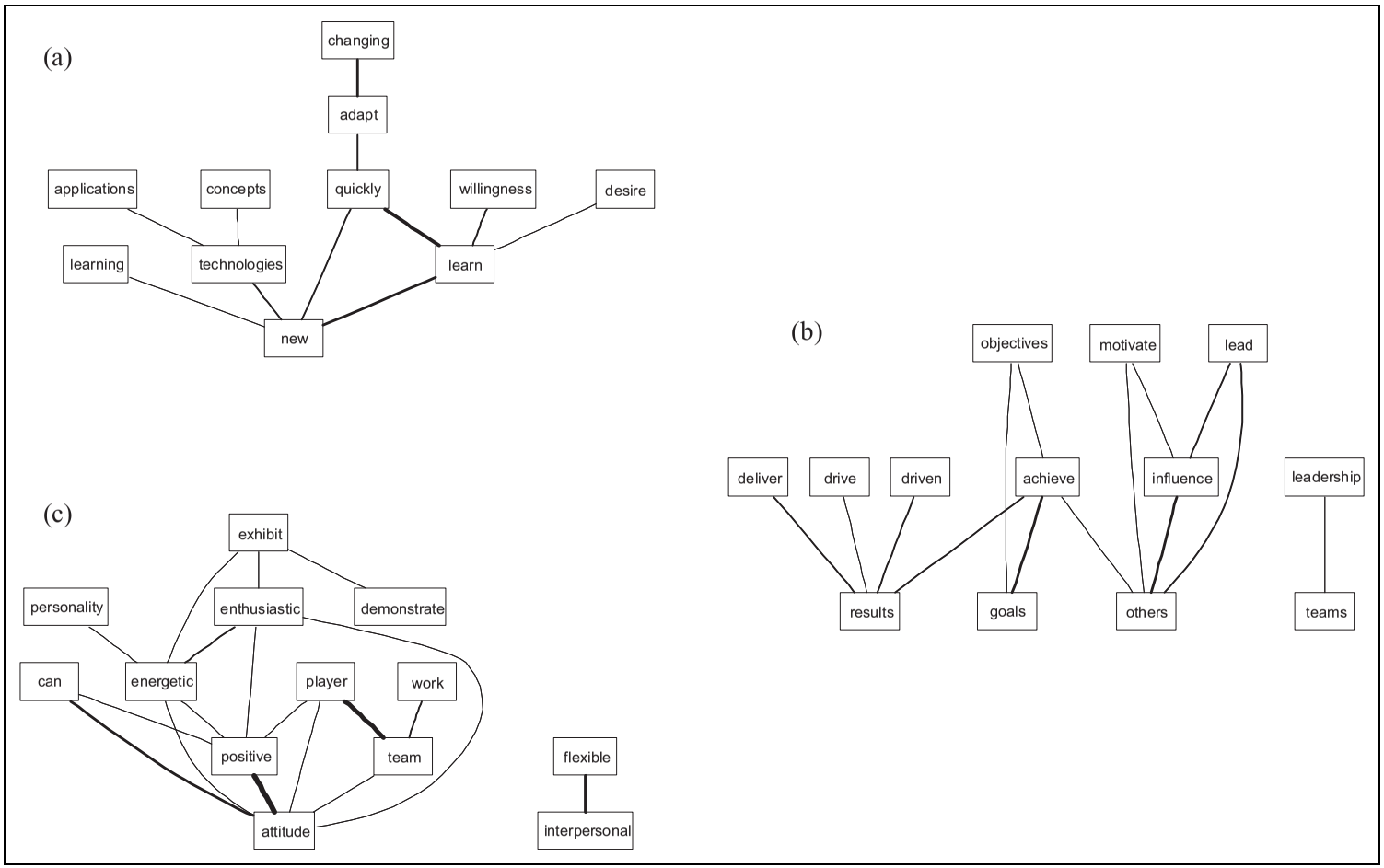

Figure 3. Word correlation networks for (a) Topic 86, (b) Topic 105, and (c) Topic 15.

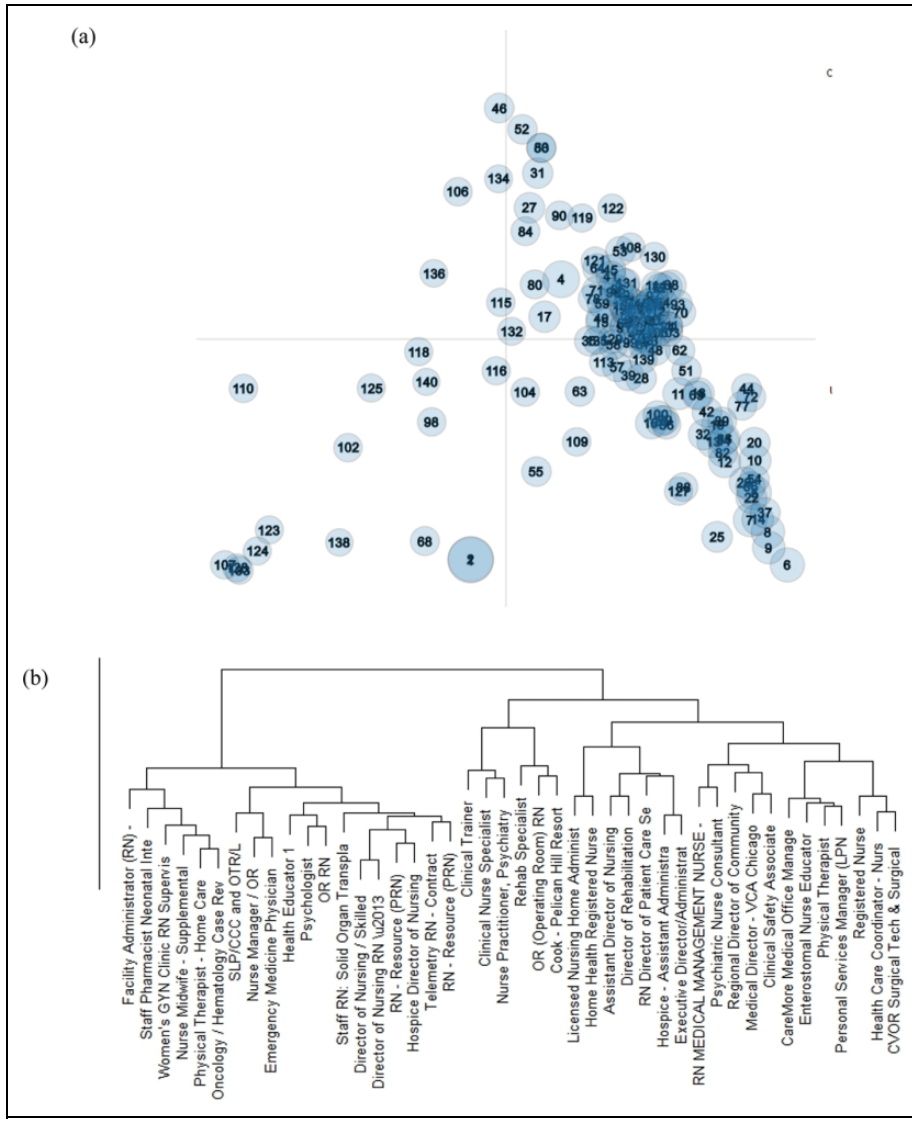

Figure 4. (a) Intertopic distance map and (b) cluster dendrogram of medically related jobs.

Terms associated to topics give us an idea about the possible interpretation of topics, however, we need to examine the relationship graph to help us surmise the context in which these words are used. Also, topic modeling showed that it is not only possible to accurately classify job information from vacancies but that we can also derive behavioral characteristics that are valued or required by employers from potential or existing job holders. We further made use of the extracted job information by summarizing the worker attributes on 140 dimensions, defining "job similarity" based on topic mixtures, and then clustering the jobs. Further analysis can be performed such as analyzing trends of worker attributes required by organizations across time, occupations, companies, and geographical regions given that these types of information are generally provided in the vacancies. Also, one can build a network of work activities to examine relationship among tasks.

Data collection, through TM, is faster, cheaper, and more reliable than traditional job analytic methods (McEntire et al., 2006). For our work on nursing tasks extraction, data triangulation showed that a substantial amount of the extracted tasks may be characterized as context-specific (e.g., caring for patients with spine surgery, caring for mentally ill patients) and that not all nurses perform these tasks. These tasks reflect idiosyncrasies in jobs that may be overlooked with data collection from SMEs because it would be impossible to interview, observe, and/or survey all nurses. Due to context-specificity, traditional ways of data collection have compromised the reliability of job-analytic data, causing bias (Dierdorff & Morgeson, 2009; Morgeson & Campion, 1997, 2000; Morgeson et al., 2004; Sanchez & Levine, 2000). Our application of TM, however, showed that this information can be extracted automatically from vacancies to complement, enrich, and strengthen traditional methods of job analysis.

Of course there are also validity concerns associated with online vacancies as a data source. First, there are noticeable differences in the quality of the information across sources. For example vacancies posted by recruitment agencies are often lower in quality (e.g., level of detail, clarity of information) compared to vacancies posted by organizations. Data triangulation for the nurses also showed that specificity varied a lot between TM and task inventory data. There are for example five tasks about medication (i.e., prepare medication, arrange medication new patients, check medication, and hand out medication), all with extensive descriptions in the task inventory, whereas the TM counterpart is only "administration of medication." Thus the level of detail is much lower there. Second, online data, as all secondary data, is often produced with very different purposes than the research purpose it may subsequently be repurposed for, in this case job analysis. For example, online vacancies are aimed at recruiting employees, which means that the included information might be biased through advertising only certain, mainly positive, aspects of the job and/or not mentioning very mundane tasks. Tasks unique to the traditional task inventory included, for example more mundane and less positive, but very frequently occurring tasks in the nursing profession (e.g., washing patients, changing patients, cleaning beds, checking temperature). Third, not all jobs are advertised online (Sodhi & Son, 2007), potentially leaving out relevant information and jobs. Our recommendation to further validate the relationships is to compare the results we obtained with alternative sources of information such as interviewing SMEs or job incumbents, and computing measures traditionally used in interrater reliability as what we did with nursing tasks

Discussion and Summary

This article presented TM steps and associated methodologies to provide a sense of the applicability of TM methodologies within the field of organizational research. When confronted with a large volume of text data, TM can reduce personnel and cost constraints (i.e., hiring manual coders). Besides discussing steps and techniques, the practical recommendations sections offered tips on how to start TM and which tools to use, and we illustrated how TM can be applied in the field of job analysis.

When incorporating TM in organizational research, domain knowledge or theory can help supplement the more inductive approach often followed in TM and we tried to illustrate the role and importance of such knowledge and theory to a number of TM steps (see Figure 1). TM also allows flexibility and opportunity to recover potentially useful patterns which have previously been inaccessible from large amounts of text. Yet, for the expansion of TM to areas where research goals are not only to classify or to cluster but also to explain, using existing knowledge or theory and incorporating this into the analysis from the start is vital (see also George, Haas, & Pentland, 2014).

In establishing or evaluating the reliability and validity of a given study using TM, a key question is whether we should adopt our evaluative criteria from the qualitative research tradition (cf. Yu, Jannasch-Pennell, & DiGangi, 2011), the quantitative research tradition, or perhaps even both. Of

course, insofar as a TM study can withstand scrutiny from both methodological perspectives, this only serves to increase its credibility. Yet, the relevance of specific quality measures is likely to be contingent upon the epistemological orientation and specific objectives of the researcher. That is, for more exploratory or descriptive studies, such as those relying on topic modeling and clustering (see Table 1), it is not mandatory to impose strategies designed for establishing the validity of inferences. "By definition 'inference' is an act of expanding the conclusion from a smaller subset to a broader set (e.g., from the sample statistics to the population parameter), but most qualitative studies do not aim to make 'valid inferences'" (Yu et al., 2011, p. 736). Krippendorf (2012) echoed this for CATA stating that "deductive and inductive inferences are not central to content analysis" (p. 36). Nevertheless, TM output can also provide a starting point for studies aiming to take an inferential route.

TM also has limitations and constraints. TM requires specific expertise and resources. Not all organizations/researchers have the computing resources to develop massive TM applications or the necessary expertise to execute these appropriately. The expertise and computing resources constraint could be addressed by outsourcing the task to companies and people who specialize in TM. Another limitation is the question of the representativeness of the information found in text data. The quality of the data will matter for the outcomes as with any type of data. The limitation of text data as an incomplete source of information could be mitigated by supplementing the analysis with additional types of data. For instance, in our job vacancy analysis we could triangulate our findings against the Occupational Information Network (Jeanneret & Strong, 2003), or other data sources that provide rich job information.

The different legal and ethical considerations that come with using particular forms of text data form a final limitation. Some text data are proprietary or contain privacy sensitive information that may be difficult to anonymize. The difficulty of obtaining permission to use text data can be addressed in part by implementing safeguards to protect the confidentiality of the data and to perform the analysis securely. Wider ethical concerns (Van Wel & Royakkers, 2004) on the use of "big" data, urgently need further and wider development and discussion.

We hope our discussion of TM helps foster dialogue and collaboration between organizational researchers and data scientists, particularly text miners. Though most discussions here have centered on how TM can help organizational research, TM as a field also has something to gain from organizational research. The richness of problems that organizational research is trying to analyze can stimulate the creation of novel TM methodologies, thereby, contributing to its advancement. In sum, the deluge of text data, the need to combine qualitative approaches with their quantitative counterparts, and the resulting progress for the two fields (organizational research and TM) brought by the interplay of theory and methods make the inclusion of TM methods ever more relevant to organizational research.

Appendix

Glossary of Terms as Used by the Text Mining Community

Corpus: A collection of documents.

Document: A sequence of characters or a string. In this context, it is better understood as a file containing words, punctuations and special characters in a particular language. It is synonymous with the word *text*. Examples of documents are hiring offers, email messages, company mission statements, responses to open-ended survey questions, journal articles, and books.

Feature: A variable used to capture a characteristic of text data. The word feature is usually synonymous with (input) variable (i.e., terms), although at times input variables may be preprocessed to compute a feature (Guyon & Elisseeff, 2003), somewhat akin to the process by means of

which survey items may be preprocessed to yield a score for a construct. The application of feature selection techniques yields a subset of features that is then used to construct a classification model.
Labeled data: In classification, labeled data refer to documents whose category membership is known. For purposes of constructing and evaluating the algorithm, these are respectively split into training and test data.
Term: A unit in a document. It can be a word, a phrase, or a sentence. Punctuation marks can also be considered as terms.
Test data: Human-labeled data used to evaluate the performance of a model.
Training data: Human-labeled data used to construct the classification model.
Vector: An arranged array of numbers that represent the scores on features for a particular document
Vocabulary or **lexicon**: The set of all unique terms in a corpus.

Acknowledgments

Vladimer B. Kobayashi is on study leave from the University of the Philippines Mindanao. An earlier version of this article was presented as Kobayashi, V. B., Berkers, H. A., Mol, S. T., Kismihók, G., & Den Hartog, D. N. (2015, August). Augmenting organizational research with the text mining toolkit: All aboard! In J. M. LeBreton (Chair), *Big Data: Implications for Organizational Research*, showcase symposium at the 75th Annual Meeting of the Academy of Management, Vancouver, BC, Canada. The authors would like to acknowledge the anonymous reviewers at *Organizational Research Methods* and the Academy of Management for their constructive comments on earlier drafts of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the European Commission through the Marie-Curie Initial Training Network EDUWORKS (Grant PITN-GA-2013-608311) and by the Society of Industrial and Organizational Psychology Sidney A. Fine Grant for Research on Job Analysis, for the Big Data Based Job Analytics Project.

Supplemental Material

Supplementary material for this article is available online at http://journals.sagepub.com/doi/suppl/10.1177/1094428117722619.

References

Abdessalem, W. K. B., & Amdouni, S. (2011). E-recruiting support system based on text mining methods. *International Journal of Knowledge and Learning*, *7*(3), 220-232. https://doi.org/10.1504/IJKL.2011.044542

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 163-222). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-3223-4_6

Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge, MA: MIT Press.

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computer Review*, *27*(4), 509-523. https://doi.org/10.1177/0894439309332293

Arthur, W., Jr., , Bennett, W., Jr., , Edens, P. S., & Bell, S. T. (2003). *Effectiveness of training in organizations: A meta-analysis of design and evaluation features*. Washington, DC: American Psychological Association. Retrieved from http://psycnet.apa.org/journals/apl/88/2/234/

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.),

Advances in knowledge discovery and data mining (pp. 391-402). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-642-13657-3_43

Associated Press. (2013, July 29). *AP, Meltwater settle copyright dispute*. Retrieved from http://www.ap.org/Content/AP-In-The-News/2013/AP-Meltwater-settle-copyright-dispute

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 245-250). New York, NY: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=502546

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*(3), 478-494. https://doi.org/10.1037/0021-9010.74.3.478

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics*, *1*, 17-35.

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

Borovikov, E. (2014). *A survey of modern optical character recognition techniques* (arXiv:1412.4183 [Cs]). Retrieved from http://arxiv.org/abs/1412.4183

Bsoul, Q., Salim, J., & Zakaria, L. Q. (2013). An intelligent document clustering approach to detect crime patterns. *Procedia Technology*, *11*, 1181-1187. https://doi.org/10.1016/j.protcy.2013.12.311

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the whole in parts: Text summarization for web browsing on handheld devices. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 652-662). New York, NY: ACM. https://doi.org/10.1145/371920.372178

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105. https://doi.org/10.1037/h0046016

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Journal of Neurocomputing*, *72*(7-9), 1775-1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with naïve Bayes. *Expert Systems with Applications*, *36*(3, pt. 1), 5432-5435. https://doi.org/10.1016/j.eswa.2008.06.054

Cohen Priva, U., & Austerweil, J. L. (2015). Analyzing the history of cognition using topic models. *Cognition*, *135*, 4-9. https://doi.org/10.1016/j.cognition.2014.11.006

Conrad, J. G., Al-Kofahi, K., Zhao, Y., & Karypis, G. (2005). Effective document clustering for large heterogeneous law firm collections. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law* (pp. 177-187). New York, NY: ACM. https://doi.org/10.1145/1165485.1165513

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web* (pp. 519-528). New York, NY: ACM. https://doi.org/10.1145/775152.775226

Derpanis, K. G. (2006). *K-means clustering*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.217.5155

Dierdorff, E. C., & Morgeson, F. P. (2009). Effects of descriptor specificity and observability on incumbent work analysis ratings. *Personnel Psychology*, *62*(3), 601-628. https://doi.org/10.1111/j.1744-6570.2009.01151.x

El-Hamdouchi, A., & Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *Computer Journal*, *32*(3), 220-227. https://doi.org/10.1093/comjnl/32.3.220

Faguo, Z., Fan, Z., Bingru, Y., & Xingang, Y. (2010). Research on short text classification algorithm based on statistics and rules. In *2010 Third International Symposium on Electronic Commerce and Security (ISECS)* (pp. 3-7). New York, NY: IEEE. https://doi.org/10.1109/ISECS.2010.9

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*(2), 291-314.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289-1305.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information retrieval: Data structures and algorithms*. Upper Saddle River, NJ: Prentice Hall.

George, G., Haas, M., & Pentland, A. (2014). From the editors big data and management. *Academy of Management Journal*, *57*(2), 321-326.

Ghani, R., Probst, K., Liu, Y., Krema, M., & Fano, A. (2006). Text mining for product attribute extraction. *SIGKDD Explorations Newsletter*, *8*(1), 41-48. https://doi.org/10.1145/1147234.1147241

Ghosh, S., Roy, S., & Bandyopadhyay, S. K. (2012). A tutorial review on text mining algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(4). Retrieved from http://ijarcce.com/upload/june/6-A%20tutorial%20review%20on%20Text%20Mining%20Algorithms.pdf

Grimes, S. (2008, August 1). *Unstructured data and the 80 percent rule*. Retrieved from https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

Guo, Y., Li, Y., & Shao, Z. (2015). An ant colony-based text clustering system with cognitive situation dimensions. *International Journal of Computational Intelligence Systems*, *8*(1), 138-157. https://doi.org/10.1080/18756891.2014.963986

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157-1182.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447-457. https://doi.org/10.1037/met0000120

Harvey, R. J. (1986). Quantitative approaches to job classification: A review and critique. *Personnel Psychology*, *39*(2), 267-289.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191-205.

Holton, C. (2009). Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems*, *46*(4), 853-864. https://doi.org/10.1016/j.dss.2008.11.013

House, R. J., Spangler, W. D., & Woycke, J. (1991). Personality and charisma in the US presidency: A psychological theory of leader effectiveness. *Administrative Science Quarterly*, *36*, 364-396.

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. In *Artificial intelligence: Methodology, systems, and applications* (pp. 77-86). New York, NY: Springer. Retrieved from http://link.springer.com/chapter/10.1007/11861461_10

Hu, J., Sun, X., Lo, D., & Li, B. (2015). Modeling the evolution of development topics using dynamic topic models. In *2015 IEEE 22nd International Conference on Software Analysis, Evolution and Reengineering (SANER)* (pp. 3-12). New York, NY: IEEE. https://doi.org/10.1109/SANER.2015.7081810

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168-177). New York, NY: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1014073

Huang, H., & Zhang, B. (2009). Text segmentation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 3072-3075). New York, NY: Springer. Retrieved from http://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_421

Inmon, W. H. (1996). The data warehouse and data mining. *Communications of the ACM*, *39*(11), 49-50.

Inmon, W. H., & Nesavich, A. (2007). *Tapping into unstructured data: Integrating unstructured data and textual analytics into business intelligence*. Upper Saddle River, NJ: Prentice Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264-323.

Jang, H., Song, S. K., & Myaeng, S. H. (2006). Text mining for medical documents using a hidden Markov model. In *Proceedings of the Third Asia Conference on Information Retrieval Technology* (pp. 553-559). Berlin, Germany: Springer-Verlag. https://doi.org/10.1007/11880592_45

Jeanneret, P. R., & Strong, M. H. (2003). Linking O*net job analysis information to job requirement predictors: An O*net application. *Personnel Psychology*, *56*(2), 465-492. https://doi.org/10.1111/j.1744-6570.2003.tb00159.x

Jolliffe, I. (2005). *Principal component analysis*. New York, NY: Wiley. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0470013192.bsa501/full

Jonsson, H., Nugues, P., Bach, C., & Gunnarsson, J. (2010). Text mining of personal communication. In *2010 14th International Conference on Intelligence in Next Generation Networks (ICIN)* (pp. 1-5). New York, NY: IEEE. https://doi.org/10.1109/ICIN.2010.5640938

Kabanoff, B. (1997). Computers can read as well as count: Computer-aided text analysis in organizational research. *Journal of Organizational Behavior*, *18*(S1), 507-511. https://doi.org/10.1002/(SICI)1099-1379(199711)18:1+<507::AID-JOB904>3.0.CO;2-0

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. New York, NY: Springer.

Kirkpatrick, S. A., Wofford, J. C., & Baum, J. R. (2002). Measuring motive imagery contained in the vision statement. *Leadership Quarterly*, *13*(2), 139-150. https://doi.org/10.1016/S1048-9843(02)00096-6

Klema, V., & Laub, A. J. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, *25*(2), 164-176.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text classification for organizational researchers: A tutorial. *Organizational Research*, *21*(3), 766-799.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (*Vol. 14*, pp. 1137-1145). Retrieved from http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf

Korkontzelos, I., Mu, T., Restificar, A., & Ananiadou, S. (2011). Text mining for efficient search and assisted creation of clinical trials. In *Proceedings of the ACM Fifth International Workshop on Data and Text Mining in Biomedical Informatics* (pp. 43-50). New York, NY: ACM. https://doi.org/10.1145/2064696.2064706

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 721-735. https://doi.org/10.1109/TPAMI.2008.110

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259-284. https://doi.org/10.1080/01638539809545028

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2013). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, *21*. Retrieved from http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/

Lee, J., & Hong, Y. S. (2013, August). *Business model mining: Analyzing a firm's business model with text mining of annual report*. Paper presented at the 19th International Conference on Engineering Design, Seoul, Korea.

Lee, S., Baker, J., Song, J., & Wetherbe, J. C. (2010). An empirical comparison of four text mining methods. In *43rd Hawaii International Conference on System Sciences (HICSS)* (pp. 1-10). https://doi.org/10.1109/HICSS.2010.48

Lewis, D. D. (1992a). Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language* (pp. 212-217). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1075527.1075574

Lewis, D. D. (1992b). *Representation and learning in information retrieval*. Amherst: University of Massachusetts. Retrieved from http://ciir.cs.umass.edu/pubfiles/UM-CS-1991-093.pdf

McEntire, L. E., Dailey, L. R., Osburn, H. K., & Mumford, M. D. (2006). Innovations in job analysis: Development and application of metrics to analyze job data. *Human Resource Management Review*, *16*(3), 310-323. https://doi.org/10.1016/j.hrmr.2006.05.004

McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods*, *16*(1), 152-184. https://doi.org/10.1177/1094428112459910

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI* (*Vol. 6*, pp. 775-780). Retrieved from http://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf

Mitchell, T. M. (1997). *Machine learning*. Burr Ridge, IL: McGraw-Hill.

Montanelli, R. G., Jr., , & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika*, *41*(3), 341-348.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*, *82*(5), 627-655.

Morgeson, F. P., & Campion, M. A. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior*, *21*(7), 819-827.

Morgeson, F. P., Delaney-Klinger, K., Mayfield, M. S., Ferrara, P., & Campion, M. A. (2004). Self-presentation processes in job analysis: A field experiment investigating inflation in abilities, tasks, and competencies. *Journal of Applied Psychology*, *89*(4), 674-686.

Morgeson, F. P., & Dierdorff, E. C. (2011). Work analysis: From technique to theory. In *APA handbook of industrial and organizational psychology, vol. 2: Selecting and developing members for the organization* (pp. 3-41). Washington, DC: American Psychological Association. https://doi.org/10.1037/12170-001

Nenkova, A., & Bagga, A. (2003). Email classification for contact centers. In *Proceedings of the 2003 ACM Symposium on Applied Computing* (pp. 789-792). New York, NY: ACM. https://doi.org/10.1145/952532.952689

Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends in Information Retrieval*, *4*(3), 175-246.

Osinski, S., & Weiss, D. (2005). A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, *20*(3), 48-54. https://doi.org/10.1109/MIS.2005.38

Palmer, D. (2010). Text preprocessing. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd ed., pp. 9-30). London, UK: Chapman & Hall/CRC.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. https://doi.org/10.1561/1500000011

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program*. Retrieved from http://liwc.wpengine.com

Penn Part of Speech Tags. (n.d.). Retrieved from http://cs.nyu.edu/grishman/jet/guide/PennPOS.html

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 91-100). New York, NY: ACM. https://doi.org/10.1145/1367497.1367510

Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In A. Kao & S. R. Poteet (Eds.), *Natural language processing and text mining* (pp. 9-28). London, UK: Springer. https://doi.org/10.1007/978-1-84628-754-1_2

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 569-577). New York, NY: ACM. Retrieved from http://dl.acm.org/citation.cfm? id=1401960

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137. https://doi.org/10.1108/eb046814

ProgrammableWeb. (n.d.). *API directory*. Retrieved from http://www.programmableweb.com/apis/directory

Radev, D., Otterbacher, J., Winkel, A., & Blair-Goldensohn, S. (2005). NewsInEssence: Summarizing online news topics. *Communications of the ACM*, *48*(10), 95-98.

RANKS NL. (n.d.). *Stopwords*. Retrieved from http://www.ranks.nl/stopwords

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of Computers and Communications*, *5*(1), 27-34.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064-1082.

R Programming/Text Processing. (2014, June 26). Retrieved from http://en.wikibooks.org/wiki/R_Programming/Text_Processing

Sackett, P. R., & Laczo, R. M. (2003). Job and work analysis. In I. B. Weiner (Ed.), *Handbook of psychology* (pp. 48-87). New York, NY: John Wiley. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0471264385.wei1202/abstract

Saggion, H., & Gaizauskas, R. (2005). Experiments on statistical and pattern-based biographical summarization. In C. Bento, A. Cardoso, & G. Dias (Eds.), *Progress in artificial intelligence* (pp. 611-621). Berlin, Germany: Springer. Retrieved from http://link.springer.com/chapter/10.1007/11595014_60

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513-523.

Sanchez, J. I., & Levine, E. L. (2000). Accuracy or consequential validity: Which is the better standard for job analysis data? *Journal of Organizational Behavior*, *21*(7), 809-818.

Sanchez, J. I., & Levine, E. L. (2012). The rise and fall of job analysis and the future of work analysis. *Annual Review of Psychology*, *63*(1), 397-425. https://doi.org/10.1146/annurev-psych-120710-100401

Scherbaum, C. A. (2005). Synthetic validity: Past, present, and future. *Personnel Psychology*, *58*(2), 481-515.

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. In *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 379-388). San Francisco, CA: Morgan Kaufmann. Retrieved from http://dl.acm.org/citation.cfm? id=645528.657484

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*(2), 320-347. https://doi.org/10.1177/1094428109335949

Singh, N., Hu, C., & Roehl, W. S. (2007). Text mining a decade of progress in hospitality human resource management research: Identifying emerging thematic development. *International Journal of Hospitality Management*, *26*(1), 131-147. https://doi.org/10.1016/j.ijhm.2005.10.002

Smith, D., & Ali, A. (2014). Analyzing computer programming job trend using web data mining. *Issues in Informing Science and Information Technology*, *11*. Retrieved from http://iisit.org/Vol11/IISITv11p203-214Smith0494.pdf

Sodhi, M. S., & Son, B.-G. (2007). *Industry requirements of operations research skills based on statistical content analysis of job ads* (SSRN No. 1011468). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1011468

Sodhi, M. S., & Son, B.-G. (2010). Content analysis of OR job advertisements to infer required skills. *Journal of the Operational Research Society*, *61*(9), 1315-1327. https://doi.org/10.1057/jors.2009.80

Solka, J. L. (2008). Text data mining: Theory and methods. *Statistics Surveys*, *2*, 94-112. https://doi.org/10.1214/07-SS016

Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern Analysis and Applications*, *8*(1-2), 199-209. https://doi.org/10.1007/s10044-005-0256-3

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining* (*Vol. 400*, pp. 525-526). Retrieved from https://wwws.cs.umn.edu/tech_reports_upload/tr2000/00-034.ps

Tett, R. P., Guterman, H. A., Bleier, A., & Murphy, P. J. (2005). Development and content validation of a hyperdimensional taxonomy of managerial competence. *Human Performance*, *13*(3), 205-251.

Theeboom, T., Van Vianen, A. E. M., Beersma, B., Zwitser, R., & Kobayashi, V. (in press). A practitioner's perspective on coaching effectiveness. In L. Nota & S. Soresi (Eds.), *Counseling and coaching in times of crisis and transitions: From research to practice*. London, UK: Routledge.

Van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, *6*(2), 129-140.

Verwaeren, B., Van Hoye, G., & Baeten, X. (2016). Getting bang for your buck: The specificity of compensation and benefits information in job advertisements. *International Journal of Human Resource Management*. Advance online publication.

Vo, D.-T., & Ock, C.-Y. (2015). Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications*, *42*(3), 1684-1698. https://doi.org/10.1016/j.eswa.2014.09.031

Voskuijl, O. (2005). Job analysis: Current and future perspectives. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 27-47). Malden, MA: Blackwell.

Waal, A. de, Venter, J., & Barnard, E. (2008). Applying topic modeling to forensic data. In I. Ray & S. Shenoi (Eds.), *Advances in digital forensics IV* (pp. 115-126). New York, NY: Springer. https://doi.org/10.1007/978-0-387-84927-0_10

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1105-1112). New York, NY: ACM. Retrieved from http://dl.acm.org/citation.cfm? id=1553515

Wang, L., & Chen, Y. (2008). Conversation extraction in dynamic text message stream. *Journal of Computers*, *3*(10), 86-93.

Willett, P. (2006). The Porter stemming algorithm: Then and now. *Program*, *40*(3), 219-223. https://doi.org/10.1108/00330330610681295

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 412-420). San Francisco, CA: Morgan Kaufmann. Retrieved from http://dl.acm.org/citation.cfm? id=645526.657137

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363-373. https://doi.org/10.1016/j.jrp.2010.04.001

Youn, S., & McLeod, D. (2007). A comparative study for email classification. In K. Elleithy (Ed.), *Advances and innovations in systems, computing sciences and software engineering* (pp. 387-391). New York, NY: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4020-6264-3_67

Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *Qualitative Report*, *16*(3), 730-744.

Zhang, W., Yoshida, T., & Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, *21*(8), 879-886. https://doi.org/10.1016/j.knosys.2008.03.044

Zhang, Y., Chen, M., & Liu, L. (2015). A review on text mining. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (pp. 681-685). New York, NY: IEEE. https://doi.org/10.1109/ICSESS.2015.7339149.

Author Biographies

Vladimer B. Kobayashi is a PhD student at the University of Amsterdam and on study leave from the University of the Philippines Mindanao. His current research interest is in labor market driven learning analytics. Specifically, he uses text mining techniques and machine learning to automatically extract information from job vacancies, to understand education-to-labor market transition, to study job mobility, to determine success in the labor market, and to match education and labor market needs. He plans to continue this line of research by applying his training and knowledge in the industry context.

Stefan T. Mol is an assistant professor of organizational behavior at the Leadership and Management Section of the Amsterdam Business School of the University of Amsterdam, the Netherlands. His research interests center on a variety of topics relating to the broad fields of organizational behavior and research methods, including but not limited to refugee integration in the labor market, employability, job crafting, calling, work

identity, psychological contracts, expatriate management, text mining (with a focus on vacancies), learning analytics, and meta-analysis.

Hannah A. Berkers is a PhD candidate in organizational behavior at the Leadership and Management Section of the Amsterdam Business School of the University of Amsterdam, the Netherlands. Her research interests include work and professional identity, calling, employee well-being, meaningfulness, job crafting, and a task-level perspective on work.

Gábor Kismihók is a postdoc of knowledge management at the Leadership and Management Section of the Amsterdam Business School of the University of Amsterdam, the Netherlands. His research focuses on the bridge between education and the labor market, and entails topics such as learning analytics, refugee integration in the labor market, and employability.

Deanne N. Den Hartog is a professor of organizational behavior, head of the Leadership and Management Section, and director of the Amsterdam Business School Research Institute at the University of Amsterdam, the Netherlands and visiting research professor at the University of Southern Australia. Her research interests include leadership, HRM, proactive work behavior, and trust. She has published widely on these topics and serves on several editorial boards.