R&D MANAGEMENT

The application of text mining methods in innovation research: current state, evolution patterns, and development priorities

David Antons, Eduard Grünwald, Patrick Cichy and Torsten Oliver Salge

School of Business and Economics, RWTH Aachen University, Aachen, Germany. antons@time.rwth-aachen.de, gruenwald@time.rwth-aachen.de, cichy@time.rwth-aachen.de; salge@time.rwth-aachen.de

Unstructured data in the form of digitized text is rapidly increasing in volume, accessibility, and relevance for research on innovation and beyond. While traditional attempts to analyze text (i.e., qualitative analysis) are limited in processing large amounts of data, text mining presents a set of approaches that allow researchers to explore large-scale collections of texts in an efficient manner. Given the potential of text mining as a method of inquiry, the primary purpose of this manuscript is to enable both novice and more experienced innovation researchers to select, specify, document, and interpret text mining techniques in a way that generates valid and reliable knowledge for the innovation management community. This involved taking stock of text mining applications in the field of innovation research to date by means of a systematic review of 124 journal articles employing text mining techniques and are published in a basket of the 10 premier innovation management and 8 top general management journals. The results of the systematic manual and computational analysis of these articles do not only illustrate the state and evolution of text mining applications in our field, but also allow for evidence-based recommendations regarding their future use. Here, our paper presents methodological, conceptual, and contextual development priorities that will contribute to establishing higher methodological standards in text mining and enhance the methodological richness in our field.

1. Introduction

Growing quantities of digitized text are available for researchers today. Such texts comprise among others online content from newspapers and social media, company press releases, user reviews about experiences and products as well as scientific articles and discourse. In fact, profound digitization efforts in virtually all industries will certainly continue to increase the variety and quantity of such unstructured data. There is a long tradition of drawing on text from a variety of sources in management research in general and studies on innovation in particular. Scholars have used qualitative approaches to text analysis including manual coding, discourse analytical methods, or grounded theory (Duriau et al., 2007). These manual procedures, however, are labor intensive and seem to have reached their natural

limits when it comes to analyzing increasingly large amounts of text material (Jamiy et al., 2015; Kobayashi et al., 2018). Consequently, researchers have started to explore the opportunities that the computer-aided, or automated, analysis of textual data offers (Janasik et al., 2009; Wiedemann, 2013). In simple terms, the case for text mining becomes stronger the larger the text corpus and the less accessible it is to manual content analytical techniques. Whenever the text corpus is amenable to manual analysis, manual coding remains the gold standard. Studies on R&D- and innovation-related phenomena have already begun to take advantage of the potential that the computational exploration of large-scale collections of texts as a method of inquiry promises. As such, innovation scholars developed tools for technology forecasting and road mapping based on full-text analyses of patents (e.g., Lee et al., 2008a; Choi et al., 2013), found that industries converge using newspaper articles (Kim et al., 2015), investigated the infringement risk of patents (Bergmann et al., 2008), developed patent-based profiles of inventors (Moehrle et al., 2005), and reviewed the innovation research landscape using journal publications (e.g., Antons et al., 2016).

Text mining – a field located at the intersection of computer and information science, mathematics, and (computational) linguistics – promises not only to analyze large text corpora efficiently, but also to do so in a transparent and reproducible manner (Humphreys and Wang, 2018). As such, text mining is said to fuel advances in theoretical as well as phenomenological knowledge relevant to managerial practice (Müller et al., 2016). With regard to innovation research in particular, Nambisan et al. (2017) note that the pervasive diffusion of digital technologies fundamentally changes not only the very nature of innovation, but also how we study innovation processes and outcomes. Calls from different sub-fields of business-related research share the excitement of applying computational methods to better exploit unstructured data (e.g., Agarwal and Dhar, 2014; George et al., 2014, 2016; Chintagunta et al., 2016; Antons and Breidbach, 2018).

It is against this background that the present study seeks to provide the first systematic review of text mining applications in innovation research. We identify and review a set of 124 articles that have been published on an innovation-related topic in a basket of the 10 premier innovation management journals and the top 8 general management journals. Grounded in the systematic manual and computational analysis of these 124 articles, we document the state and evolution of text mining applications in innovation research and derive conceptual, methodological, and contextual priorities for future text mining applications. From a *conceptual* perspective, we argue that text mining applications in innovation research have now reached a state of maturity where the need for demonstrations and case studies of text mining in innovation research is rapidly declining. We provide scholars with the conceptual background needed to develop research designs that take text mining applications in innovation research to the next level. From a *methodological* perspective, we identify issues that warrant attention in future studies using large-scale and automated text analysis. Therefore, we provide actionable knowledge on how to employ text mining in innovation research by proposing a set of best practices. Finally from a *contextual* perspective, we identify areas in the wider innovation research landscape where text mining techniques could help shed light on challenging research questions.

Overall, we argue that innovation scholars have much to benefit from these evidence-based insights into the current state, evolution patterns, and development priorities of text mining applications in innovation research. This will enable both novice and more experienced researchers to select, specify, document, and interpret text mining techniques in a way that generates valid and reliable knowledge for academic research and managerial practice alike. In particular, this paper is meant to empower innovation scholars with no or little previous knowledge in computer-aided text analysis to employ text mining in ways that help them accomplish their respective research objectives. Moreover, our paper can stimulate critical reflection among innovation scholars already experienced in using text mining in view of establishing even stronger methodological standards in future research. Our review of 124 text mining applications demonstrates the value of text mining as a rapidly diffusing methodological approach whose meaningful application can contribute to advancing the field. We believe that our review and guidelines can further enhance the methodological richness of innovation management research and will allow our community to take advantage of the opportunities that the digital transformation of our field offers.

2. A brief overview of text mining

Text mining has developed over decades and across scientific disciplines to form a now-substantial and diverse body of literature on the computer-aided analysis of textual data. This includes, among others, fields such as statistics, computer science, (computational) linguistics, library science, and computer science (Miner et al., 2012). Moreover, different

terminology has evolved including text mining, computational content analysis, and natural language processing. Despite their different emphasis, they share a clear focus on automated, computer-supported processing and analysis of text as a form of natural language. Due to the different traditions and underlying lines of thought of these disciplines, a unifying definition of text timing remains absent. However, there is broad agreement on the general process of analyzing large-scale text data, following the general process of knowledge discovery in databases (Fayyad et al., 1996). We adopt this logic and present Figure 1 as a simplified overview of the process underlying a typical text mining study.

Here, we distinguish between the process phases of data gathering, data preprocessing, content analysis, and integration of the text mining findings and results into the study. Data gathering, i.e., the collection of data from databases and archives or scraping data from websites or social media, is beyond the scope of our review on text mining techniques.[1] We, thus, focus on the steps of data preprocessing and content analysis. Please also note that Figure 1 implies that these activities might be executed in iterations to improve results by refining text preprocessing. As part of our discussion, we then discuss how innovation scholars can integrate text mining findings into their research design, for instance, by combining it with complementary qualitative or statistical analysis.

In what follows, we provide an overview of prevalent techniques used for (1) text preprocessing and natural language processing, (2) dictionary-based techniques to classify words into categories, and (3) algorithmic techniques to classify texts or textual units like sentences or paragraphs into predefined categories (so-called supervised algorithms) or to cluster texts or textual units into homogeneous groups without prior knowledge of those groups (so-called unsupervised algorithms).

2.1. *Text preprocessing and natural language processing*

The unifying theme linking the various text mining techniques is the idea of turning unstructured data in form of text into structured data in form of numbers so that mathematical and statistical algorithms can be applied (Miner et al., 2012). A common approach is representing text documents in a matrix, where each column represents a document and each row represents a specific term. The cells, then, contain the frequency of the term in the respective document. This approach comes with, at least, two major issues: (1) One of the essential characteristics and a potential shortcoming of a representation of textual data in such a document-term matrix is that it neglects the linguistic structure within a document treating it as a 'bag-of-words'. (2) Moreover, large collections of documents can result in a document-term matrix of very large size or, in mathematical terms, a large number of dimensions. To begin to address both limitations, various text preprocessing techniques have been developed in fields such as computational linguistics and computer and information science (Rüdiger et al., 2017). We introduce the underlying logic of these techniques below and refer to the literature in order to guide the interested reader on how to implement these techniques.

Text preprocessing usually starts with tokenization, which is the process of converting a sequence of characters into tokens. For most purposes, scholars use words as tokens and separate strings at white spaces. But one could also use sentences or

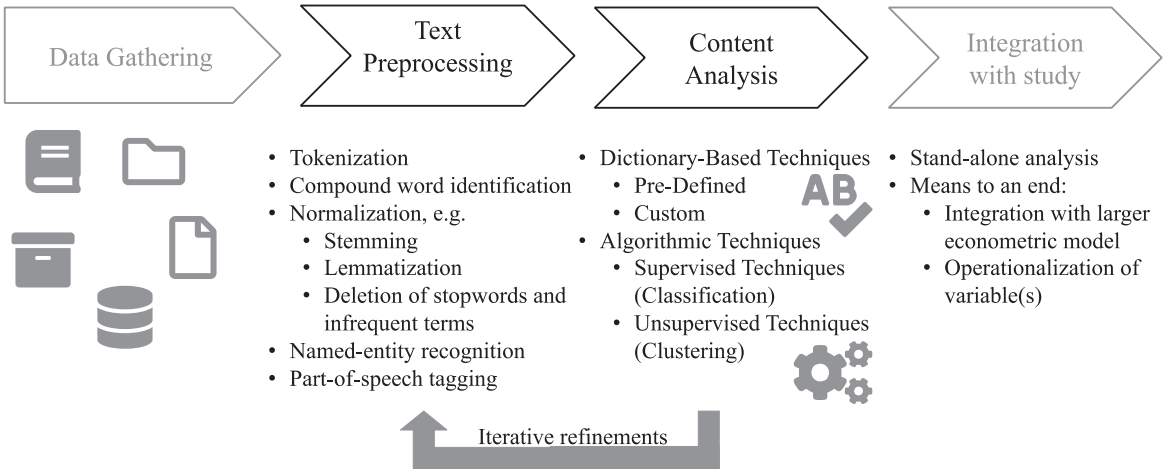

Figure 1. Text mining process and techniques.

paragraphs as tokens. Using word tokens and converting tokens into a document-term matrix yields the above-mentioned 'bag-of-words'. This simple approach, however, fails to account for the linguistic structure of compound words such as 'innovation culture' or 'open innovation' that jointly have a specific meaning. Specific algorithms are now available (e.g., Wang et al., 2007) to detect such word combinations, also known as n-grams, in text and mark them.[2] Usually, n-grams have to be replaced before constructing a document-term matrix so that the meaning of the compound is not lost. This text preprocessing technique, therefore, addresses the first limitation introduced above.

The second major issue of large, that is, high-dimensional document-term matrices results from the considerable variability of human language. For instance, we use language and words in temporal, plural, or otherwise inflected forms. In automated text analysis, this inflates the dimensionality of a document-term matrix, since each term would be added as a separate column (or row) to the matrix. Hence, preprocessing techniques have been developed to reduce such variability while preserving the word meaning. Stemming or lemmatization are techniques that reduce word variability by reducing words to their stem or their dictionary form (the lemma) (Rüdiger et al., 2017). Other techniques involve converting all text to lower cases and removing all punctuation (e.g., Antons et al., 2016; Chae and Olson, 2018), dropping stop words such as pronouns and function words that do not carry any meaning (e.g., 'the', 'and', or 'it') (e.g., Blei et al., 2003). Other scholars further reduce the document-term matrix by removing infrequent words (e.g., Hornik and Grün, 2011), or weigh words using measures such as the term frequency – inverse document frequency (tf-idf) measure (e.g., Wang et al., 2007; Antons and Breidbach, 2018; Chae and Olson, 2018). The tf-idf measure is used to spot those words that discriminate documents from each other. Terms appearing only in a focused set of documents are seen as discriminating while words that are shared across many or all documents are defined as not discriminating. Finally, scholars replace abbreviations in text with fully spelled terminology.

Beyond handling these issues, preprocessing may also add information in form of tags to words or phrases, which might be helpful to choose certain text elements or to structure texts. For instance, *named entity recognition* identifies those documents from a larger collection that contain names of a certain person, organization, or location. With *part-of-speech (POS) tagging,* scholars may tag words based on their grammatical function (e.g., noun, verb, or adjective).

Below, we move from text preprocessing phase to the actual phase of computer-aided content analysis employing either *dictionary-based techniques* originating from linguistics and psychology or *algorithmic techniques* from statistics and computer science.

2.2. Computational content analysis with dictionary-based techniques

Text mining techniques that rely on word frequency counts to measure contextual, psychological, linguistic, or semantic concepts and constructs are among the most widely adopted approaches for computer-aided analysis of textual data in management-related research so far (Duriau et al., 2007; Short et al., 2010).

Scholars may either employ existing dictionaries that were developed in previous research or create new dictionaries that would match their needs. From an epistemological point of view, dictionaries can be developed deductively based on existing theory (e.g., Gamache et al., 2015 used theory, construct definitions, and survey items to build a dictionary for the construct of regulatory focus), inductively generated from the corpus at hand (e.g., Henry, 2008 developed a list of positive and negative words used in earnings press releases to measure their tenor), or by combining these two approaches (e.g., Short et al., 2010 for the construct of entrepreneurial orientation). Especially to measure the general positivity or specific emotions in text, validated dictionaries already exist (e.g., Pennebaker et al., 2015). Dictionary-based text mining is sometimes referred to as *sentiment analysis* or *automated content analysis* depending on the objective pursued.

2.3. Computational content analysis with algorithmic techniques

Algorithmic content analytical techniques originating from computer science and statistics are typically categorized into ***classification*** *or supervised techniques* and *clustering or unsupervised techniques*.

2.3.1. Classification or supervised techniques

Researchers often seek to assign textual objects like documents or words to predefined categories. For example, Li (2010) was interested in understanding the content of forward-looking statements in the management discussion and analysis section of annual reports. After manually coding 30,000 sentences from forward-looking statements, he used a classification algorithm to categorize the content of 13 million additional forward-looking statements automatically. From a machine-learning perspective,

this means that the researcher wants to generate tags. Tags represent a kind of metadata summarizing the category the text belongs in. Prior text mining literature has developed classification procedures that are able to generate tags from pre-categorized textual content. They do so by observing existing tags and textual content to learn the classification scheme that had been applied to the pre-categorized textual content. Based on this learned classification scheme, they suggest tags for new and untagged textual content. Classifiers share the fact that they are supervised, that is, they are trained on pre-categorized data.

A classifier is able to distinguish categories producing a binary outcome, very much like a yes/no decision (Tong and Koller, 2001). A well-known example of such a binary classifier is a spam filter implemented in most of the existing email services. However, other algorithms exist that are able to distinguish multiple outcomes or categories. Broadening the scope of binary classifiers, multiple binary classifiers compare input data against each class (Fung and Mangasarian, 2005). Depending on the algorithm settings, the outcome will be a classification to the class that the new document is most likely to belong to or number of weights per class indicating the likelihood of class membership per class for the new document.

2.3.2. Clustering or unsupervised techniques

The classification procedures described above all rely on so-called supervised algorithms, which require additional information from the researcher (e.g., a pre-classified training data set) as input. Clustering procedures, in contrast, do not require any prior human knowledge when employed for the computer-aided analysis of textual data. They rely on so-called unsupervised algorithms that group textual content based on similarity (Jain, 2010). Clustering procedures are part of the broader category of dimensionality-reduction techniques – just like exploratory factor analysis, a well-known procedure among innovation scholars developing for instance new measurement scales.

In order to meaningfully group documents in a text corpus, most approaches draw on similarity as a measure of distance between documents. When documents are represented in a document-term matrix, distance between documents understood as vectors may then be calculated using common measures such as the Euclidean, Cosine, or Manhattan distance (Ingersoll et al., 2013). Other, non-distance-based algorithms use probabilistic modeling to determine similarity. Here, similarity is calculated as a probability of membership in a cluster.

To work with the clustering result, the clusters obtained typically have to be labeled (Ingersoll et al., 2013). This usually involves a manual process that is based on inspecting documents that are at the core of the final clusters or inspecting most frequent/most distinctive terms and phrases of a cluster. Obviously, these two approaches can also be combined during the labeling process. Clustering of documents or other textual content may be done in a more fine-grained way by adding another layer of analysis. That layer takes words into account and clusters words, rather than documents, based on the assumption that words that co-occur across documents are used to express a certain latent topic. Based on those topics, document similarity can be quantified. This approach has recently been used to review and map published research (Wang et al., 2015a; Antons et al., 2016; Brust et al., 2017; Antons and Breidbach, 2018; Hopp et al., 2018), to summarize patents (Kaplan and Vakili, 2015), and to judge topical newness of research articles (Antons et al., 2019).

Table 1 provides an overview of the text mining techniques discussed above with special emphasis on their value proposition to researchers. This typology will serve as an important structuring device during our subsequent systematic review of text mining applications in innovation research. In particular, it will allow us to assign each article identified to one or more of the types of analytical techniques summarized in Table 1.

3. A review of text mining applications in innovation research

To provide scholars with a structured overview of the state and evolution of text mining application in the field of innovation research, we conducted a systematic literature review (Mulrow, 1994; Tranfield et al., 2003; Denyer and Tranfield, 2009). We proceeded in the following way.

3.1. Search strategy and article selection

First, we defined a set of journals that can be considered the top journals publishing innovation research. To do so, we reviewed prior research discussing impactful innovation management journals (e.g., Linton and Thongpapanl, 2004; Thieme, 2007; Goffing et al., 2019) and selected the most impactful from the respective lists. We included the following 10 premier innovation management journals: (1) *Research Policy,* (2) *Journal of Product*

Table 1. Overview of text mining techniques

Text mining process phase	Type of algorithm	Selected question	Purpose	Selected technique(s)	Disciplinary origin(s)
Text preprocessing	Tokenization	How can I transform a text, i.e., string, into words or other textual entities?	Transferring strings into single textual entities or tokens (e.g., words, sentences)	White space separation	Computer Science
	Compound word identification	How can I identify words that have a joint meaning?	Identifying words with a joint meaning that gets lost in a bag-of-words	n-grams	Computer Science, Statistics
	Normalization and noise reduction	How can I cope with too many variables in my Document-Term-Matrix?	Reducing dimensionality of Document-Term-Matrix	Stemming, Lemmatization, Deletion of stop words, infrequent term and tf-idf weighing	Computer Science, Data Science, Statistics, Computational Linguistics, Natural Language Processing
	Linguistic analysis	How can I identify words with a special meaning or grammatical function?	Tagging of words	Named-entity recognition, Part-of-speech tagging	Statistics, Computational Linguistics, Natural Language Processing
Content analysis	Dictionary-based techniques	How can I identify how latent sociological or psychological traits and states are reflected in natural language?	Measuring contextual, psychological, linguistic, or semantic concepts and constructs	Pre-defined dictionaries	Computational Linguistics, Psychology, Natural Language Processing
				Customized dictionaries	Computational Linguistics, Psychology, Natural Language Processing
	Algorithmic techniques	How can I assign texts to predefined classes?	Classifying of textual entities into predefined categories	Supervised learning techniques such as binary or multi-class classifiers	Computer Science, Data Science, Statistics
		How can I group together similar documents?	Clustering of textual entities into formerly undefined and unknown groups	Unsupervised learning techniques such as LDA, k-means or non-negative matrix factorization	Computer Science, Data Science, Statistics

Innovation Management, (3) *R&D Management,* (4) *Technological Forecasting and Social Change,* (5) *Research-Technology Management,* (6) *Technovation,* (7) *Industrial and Corporate Change,* (8) *IEEE Transactions on Engineering Management,* (9) the *Journal of Technology Transfer,* and (10) the *Journal of Engineering and Technology Management*. Moreover, as research on innovation is also published in leading management journals, we also included innovation research published in the following eight general and strategic management journals that are part of the Financial Times list of the 50 top journals: (1) the *Academy of Management Journal,* (2) *Administrative Science Quarterly,* (3) the *Journal of Management,* (4) the *Journal of Management Studies,* (5) *Management Science,* (6) *Organization Science,* (7) *Organization Studies,* and (8) the *Strategic Management Journal.*[3]

As our second step, we composed a comprehensive list of keywords authors regularly use to indicate the application of a text mining technique in their article. This list comprised expressions that related to various alternative notations of the methodological domain of text mining (e.g., 'NLP'; 'computer-assisted text analysis'; 'textual data mining'), some of the most common approaches (e.g., 'topic models'; 'sentiment analysis'), and specific aspects of their implementation (e.g., 'bag-of-words'; 'text preprocessing'). Table S1 in the appendix provides a full list of our search terms. We then compiled a search query based on these terms and searched the journal archives directly via their official websites. Doing so ensured that our query was performed not only on titles and abstracts but on the actual full texts of all research articles published in the 18 journals identified above.

Third, we downloaded all respective search results and the second and third authors manually inspected them. We selected only those research articles into our final sample that explicitly addressed an innovation-related topic and made substantial use of text mining broadly defined. This selection procedure resulted in 124 innovation research articles that employed text mining as a method of inquiry and were published in one of the 18 journals identified above. For each of the 124 articles, we extracted both meta-data and the full texts for analysis.

3.2. Data analysis

We applied a combination of bibliometric analysis and manual coding.[4] We extracted not only the full texts, but also meta-data for all 124 research articles from the Web of Science. Meta-data included author names, titles, keywords, abstracts, and references.

As for the *bibliometric analyses*, the R-package 'bibliometrix' was used. The package provides several tools for quantitative research models in bibliometrics and scientometrics. The functional scope includes importing and formatting of raw data, the actual bibliometric analyses, as well as the creation of matrices and networks for the visualization of co-citations, couplings, collaborations, and co-work analyses. After importing the data from the Web Of Science in BibTex format, we used the individual meta-information to extract the author keywords and to build a co-occurrence network. Subsequently, adopting the Louvain-clustering method, we visualized the distribution and grouping of the author keywords.

As for the *manual analysis*, two of the authors and a research assistant independently read the 124 papers and coded them using a standardized coding scheme informed by our overview of text mining techniques presented above. This team extracted from the full texts the type and amount of text analyzed as well as the software or algorithms used for analysis and whether it was proprietary. The team then engaged in manual coding and investigated the type of text mining approach used. Here, we distinguished between techniques of text preprocessing and natural language processing, dictionary, classification, and clustering, which we described above. For dictionaries, we also distinguished between articles using pre-defined dictionaries and those developing their own dictionaries. Some articles included in our sample were coded as using other approaches. These papers, for instance, used simple keyword extraction and illustrated the usage of keywords by means of word clouds or similar techniques. To provide an indication of the content of the papers, the authors coded the articles using the subject areas of the Product Development and Management Association (PDMA) *Body of Knowledge* (Griffin and Somermeyer, 2007). Many articles related to more than one subject area, so we allowed for multiple assignments. Some articles like reviews even related to all subject areas (e.g., Antons et al., 2016). Finally, we coded the type of usage of text mining using the categories of *demonstration* (i.e., showcasing text mining or a certain algorithm or software as a research approach), *case study* (i.e., using text mining to study a certain technological field), *review* (i.e., using text mining to review a strand of research), and *variable creation* (i.e., using to operationalize a variable for an econometric model). In case of disagreement during the

coding procedure, a third author was consulted. The full coding results for all 124 articles are available in the appendix to this paper (see Table S3).

4. State and evolution of text mining applications in innovation research

The bibliometric and manual analyses conducted on our corpus of 124 articles allow us to document the state and evolution of text mining applications in innovation research below. In particular, we zoom in into (1) the *journal outlets* publishing innovation research informed by text mining approaches, (2) the type and amount of *textual data* processed, (3) the *thematic content* of the substantive research, (4) the *text mining algorithms*, (5) the *specific outcome* of text mining for the research performed, and (6) the *reporting quality* of the text mining application.

4.1. Journal outlets

Although a broad range of journals has already published innovation research using text mining, almost half of the articles (61) have been published in *Technological Forecasting and Social Change*.[5] *R&D Management* published 10 articles, and the *Journal of Product Innovation Management*, *Research Policy*, and *Technovation* have published eight articles each. *Technovation* (2001), *R&D Management* (2002), and *Technological Forecasting and Social Change* (2003) started publishing text mining applications already in the early 2000s and jointly account for 75% of all articles published before 2012. In contrast, most other journals in our basket have only recently added to the rapidly growing body of innovation research applying text mining. Figure 2 depicts the sharp increase in text mining applications in innovation research especially since 2015.

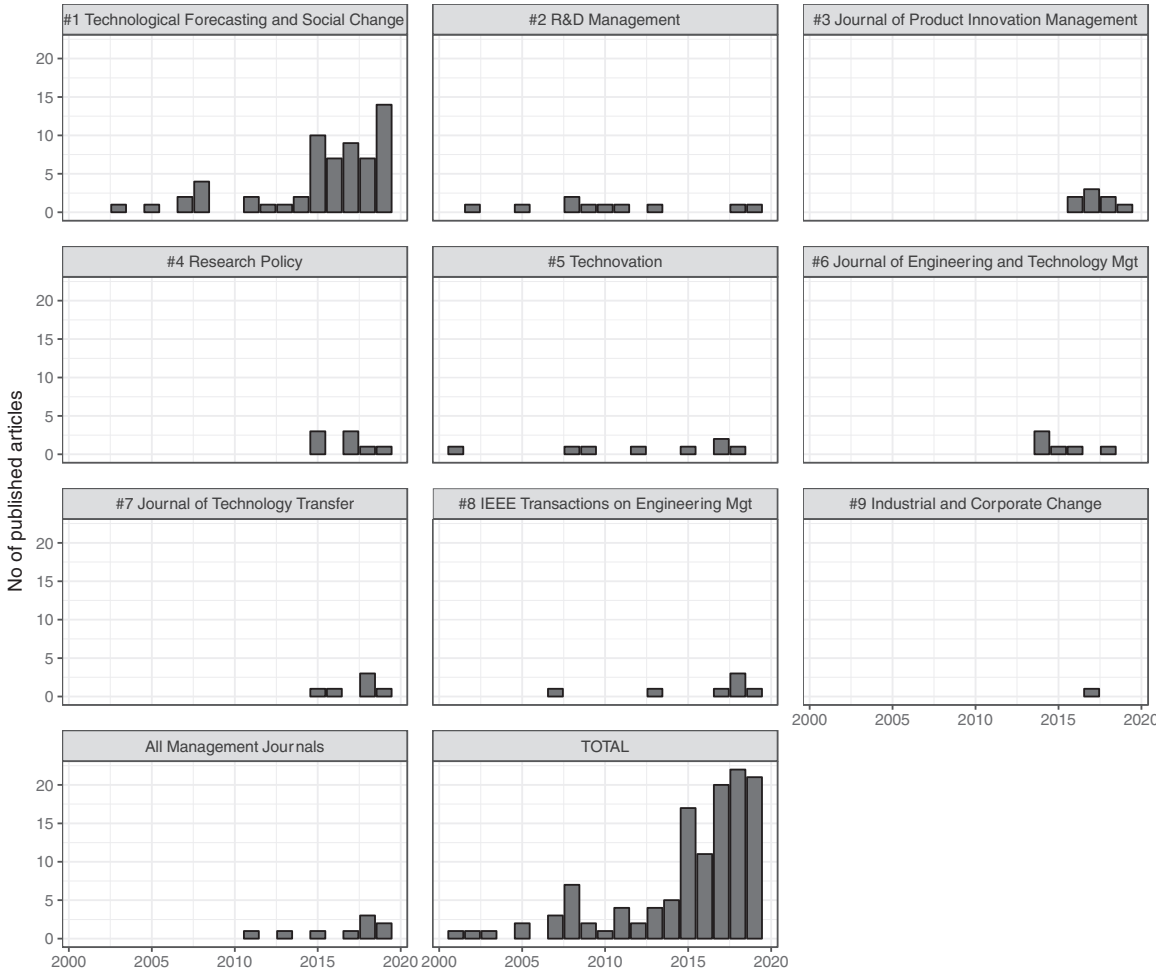

Figure 2. Text mining Articles per journal and year.

4.2. Textual data

With regard to the type of textual data analyzed, we found that 48 articles investigate research papers and 44 analyze patents. Less frequently studied are texts like blog posts and forum entries (12), newspaper articles (7), tweets and other social media entries (5), and product reviews (3). Other texts like annual reports, emails, press releases, interviews have each only been analyzed once. Interestingly, we find a broad variability with regard to the amount of texts that have been used. Of the 124 articles, 58.9% use less than 10,000 texts, 22.6% use more than 10,000 but less than 100,000 documents, 4.8% analyze more than 100,000 but less than 1,000,000 texts, and 13.7% investigate even more than 1,000,000 texts.

4.3. Thematic content

To understand the themes the 124 articles sought to examine with the help of text mining, we first used the author-generated keywords of each research article to compute a keyword–network graph illustrated in Figure 3. Here, nodes indicate keywords, node size the number of occurrences in our corpus, and links showing whether terms appear together in scientific articles.

Not surprisingly, we observe that text mining is the most central and important keyword tying together the network. On the left, we see two clusters that are not connected to the other nodes. Here, the light-gray nodes refer to topics from entrepreneurship and opportunity recognition. The purple nodes refer to meaning, cohesion, and semantic concepts. The main cluster composed of the blue-colored nodes mainly covers technical terms like text mining, text clustering, topic modeling, and bibliometric analysis. Similarly, green nodes are also more methodological referring to information retrieval from scientific texts. The light-brown nodes refer to themes like content and sentiment analysis and texts such as user-generated content, product reviews, and mass media. The red nodes seem to cover themes from technology management such as technology intelligence, technology road mapping, patent analysis, and technology monitoring. Finally, the orange nodes represent topics like technological change, adoption, dominant design, and life cycle. The red and green clusters are the largest clusters not composed of merely technical terms. They refer to technology management and retrieval of information from scientific texts and patents. This is in-line with the finding that most text mining applications in innovation research nowadays take advantage of academic publications and patents.

In addition to the bibliometric analysis of author keywords, we relied on our manual coding of the article content against the PDMA subject areas to map the topic landscape of text mining applications in innovation research. Figure 4 reveals that most studies are part of the larger subject area ‘strategy, planning, and decision making’. As we have seen, many studies analyze patents to monitor technological improvements and development or to develop technological road maps informed by academic literature. This clearly falls into the category of supporting strategic technology management and decision-making. It is noteworthy that only few studies are categorized as ‘co-developments and alliances’ or ‘people, teams, and culture’.

4.4. Text mining algorithms

Although a broad range of algorithms for automated text analysis exists, most articles in our dataset apply some clustering technique. Even though the number of articles in our corpus sharply increased in recent years, the increase in clustering articles was disproportionately stronger. Figure 5 depicts this trend. Interestingly, we can also see that innovation scholars have started relatively early to develop custom dictionaries and are doing so until today to a greater extent than using predefined dictionaries.

4.5. Text mining outcomes

Text mining algorithms have been used for all four outcome categories defined in our coding scheme (demonstration, case study, review, variable). Figure 6 tracks the evolution over time. First applications focused on demonstrating the applicability and value of text mining algorithms or software. Surprisingly, until today, most articles still fall in this category of demonstrations. In 2007, scholars started using text mining techniques for reviewing academic literature as well as for informing domain-specific case studies, for instance on a particular field of technology. While the number of reviews only recently started to increase, case studies are nowadays the second most frequent type of text mining application. Only since 2011, innovation scholars have started to use text mining techniques to measure specific constructs and variables. This category has experienced a sharp increase in recent years. Interestingly, of the 18 articles using text mining techniques to measure a variable being part of a larger (econometric) model, six appeared in the premier general management journals included in the prestigious Financial Times 50-journal list. In addition, two studies appeared in *Research Policy* and four in the *Journal of*

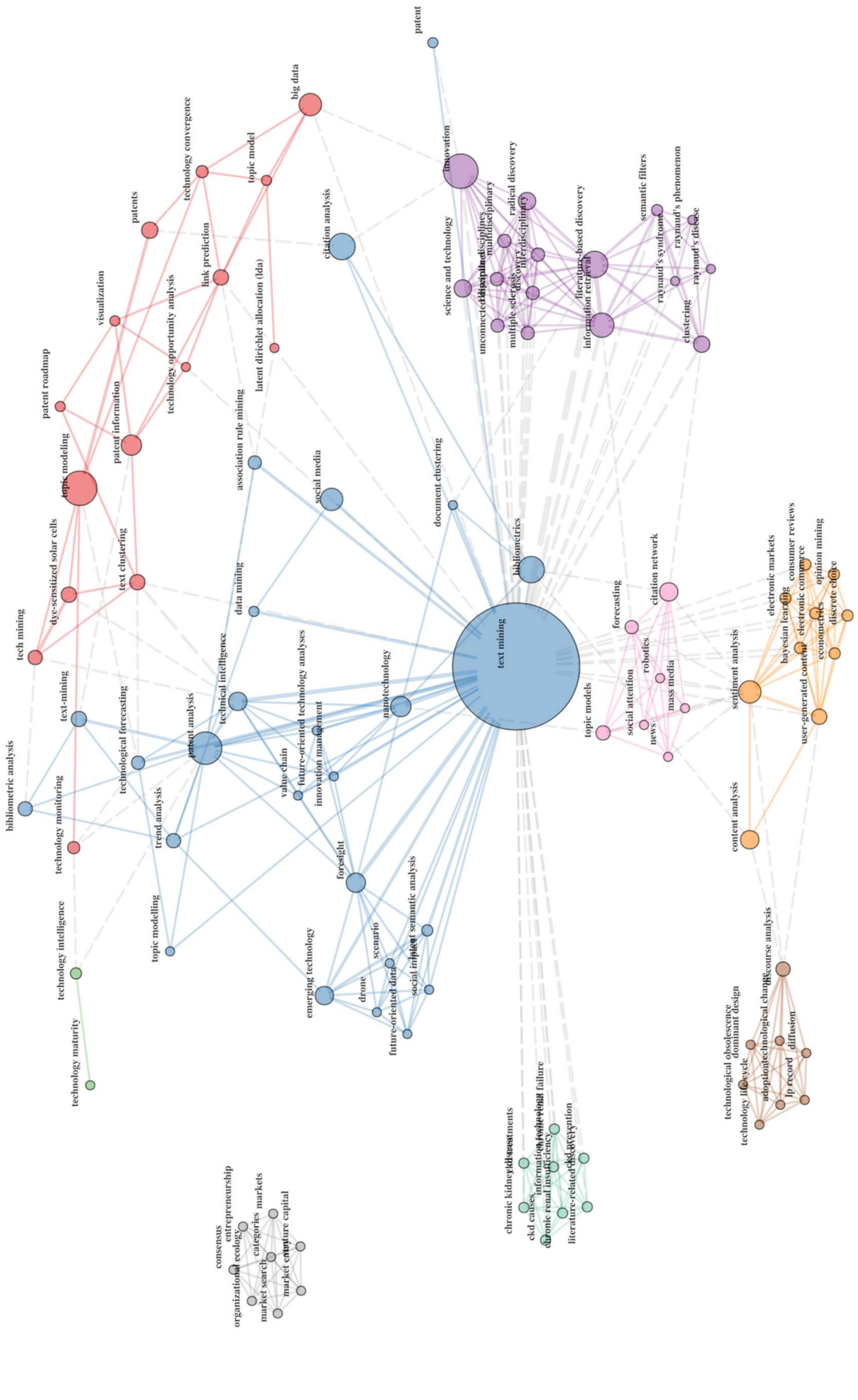

Figure 3. Network composed of author-given article keywords. [Colour figure can be viewed at wileyonlinelibrary.com]
Note: We generated this illustration by means of the R-package bibliometrix and using Louvain-clustering.

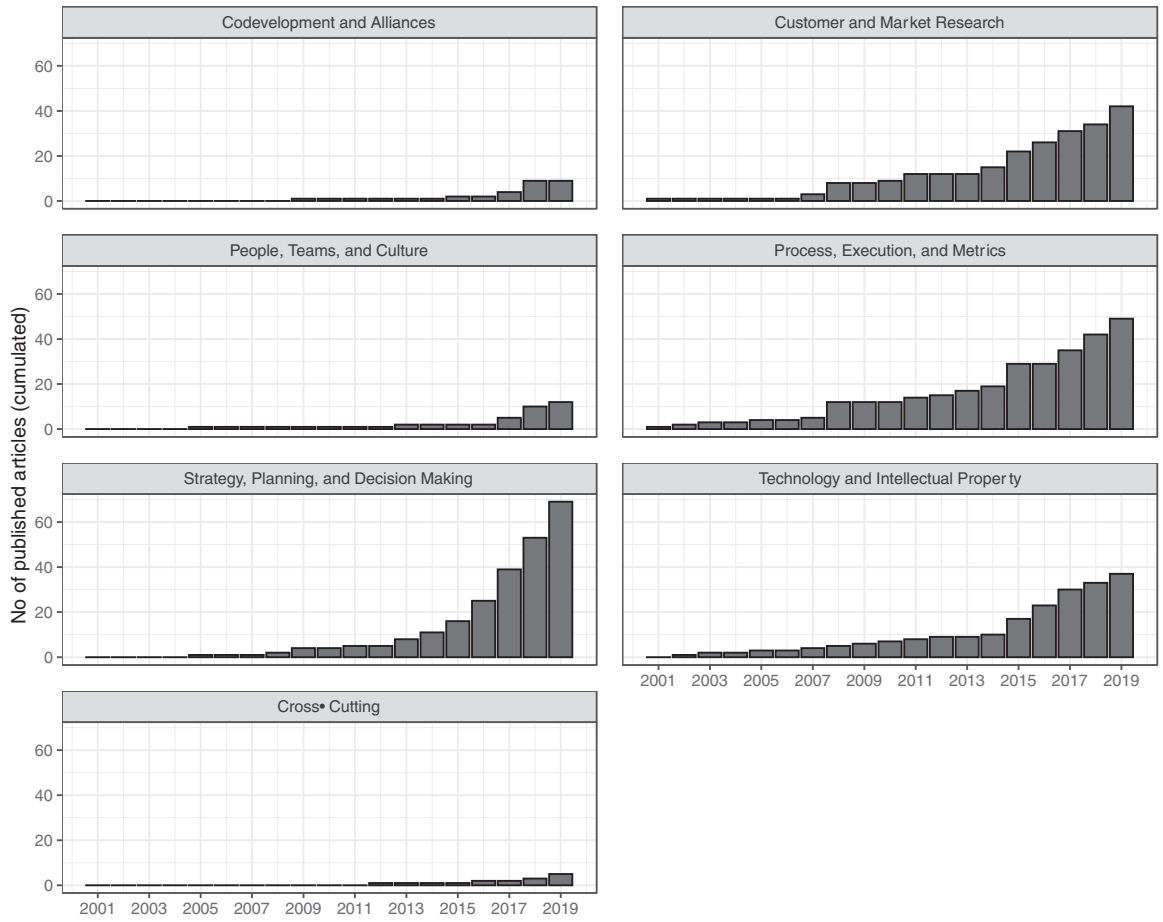

Figure 4. The subject focus of text mining applications over time.

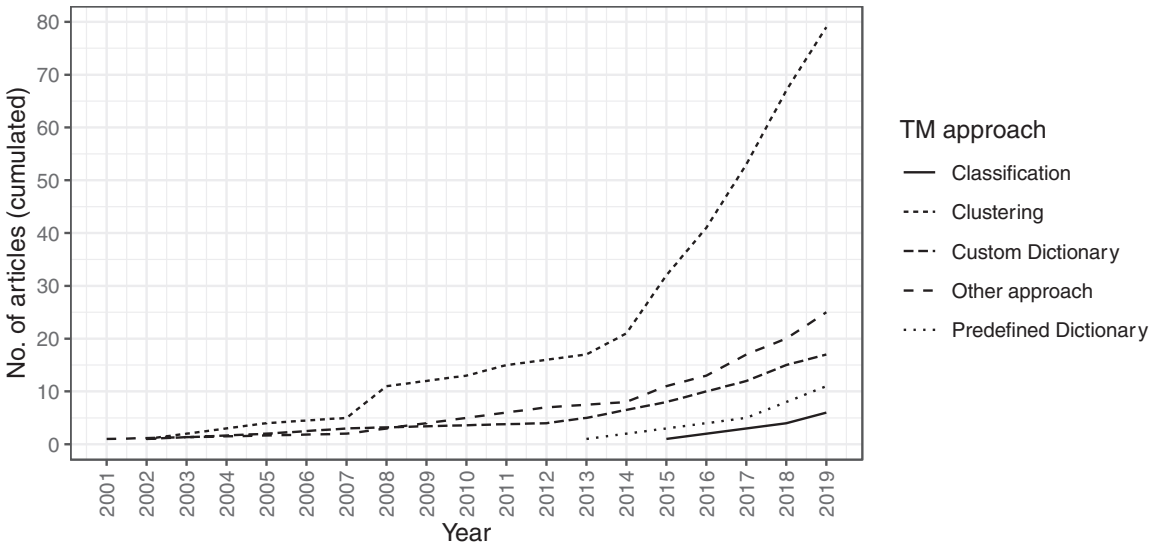

Figure 5. The application of text mining approaches in innovation research over time.

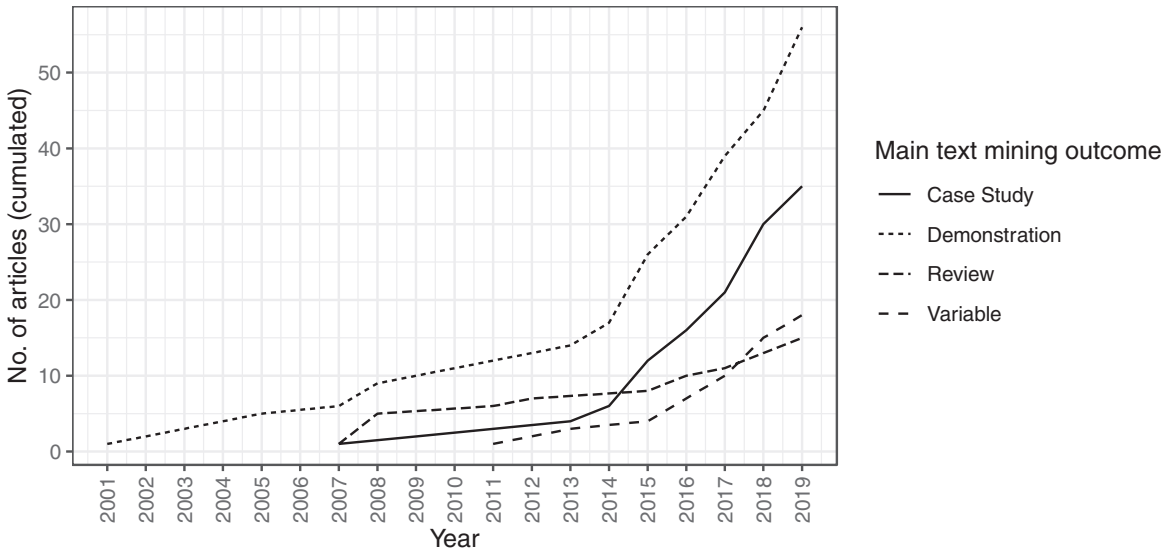

Figure 6. The type of usage of text mining outcomes over time.

Table 2. Types of software used and extent to which preprocessing was described

	Software			
Preprocessing	Not indicated	Non-proprietary	Proprietary	Total
Not indicated	19	9	28	56
Indicated	3	2	8	13
Detailed	13	21	21	55
Total	35	32	57	124

Product Innovation Management, arguably two of the most impactful innovation management journals.

4.6. Reporting quality

For text mining to meaningfully enhance the methodological repertoire of innovation research, its application has to be documented in a detailed manner to ensure transparency and replicability of studies. Against this backdrop, we found that of the 124 articles, roughly 10% do not report the size of the dataset they are using and 8% do not indicate the source of their data. We also found that 35 articles do not specify the software they used. Of the 89 articles reporting the software employed, 57 apply a proprietary software package and 32 use free software or packages and functions available for free programming languages like Python or R. As described above, text preprocessing is a major step in text mining preparing the text for analysis.

As any data cleaning and alteration might affect outcomes of analysis, scholars are well advised to document these steps in detail. We find that 56 articles do not describe the data preprocessing conducted. Only 55 articles describe their preprocessing in detail and 13 at least reported the preprocessing techniques employed. As Table 2 indicates, 19 studies neither report the software used nor the preprocessing techniques applied. This limits the replicability of these studies. Interestingly, we find that of the 37 studies that report software but do not report preprocessing, 28 apply proprietary software. It seems that some proprietary packages impede introspection into their procedures limiting their scientific value. While 42 studies report both software as well text preprocessing in detail, only 37 of those also report the source and size of the dataset used. This means that only 29.8% of all articles can be regarded as fully transparent to the reader. Figure 7 depicts the cumulative number of fully transparent articles published over time. Although it is good to see that the articles that are fully transparent to the audience have increased sharply in recent years, still the majority of articles published in recent years continues to lack in transparency and, thereby, replicability. As Figure 8 shows, the standards of reporting preprocessing have increased in recent years. That said, studies continue to be published with insufficient transparency

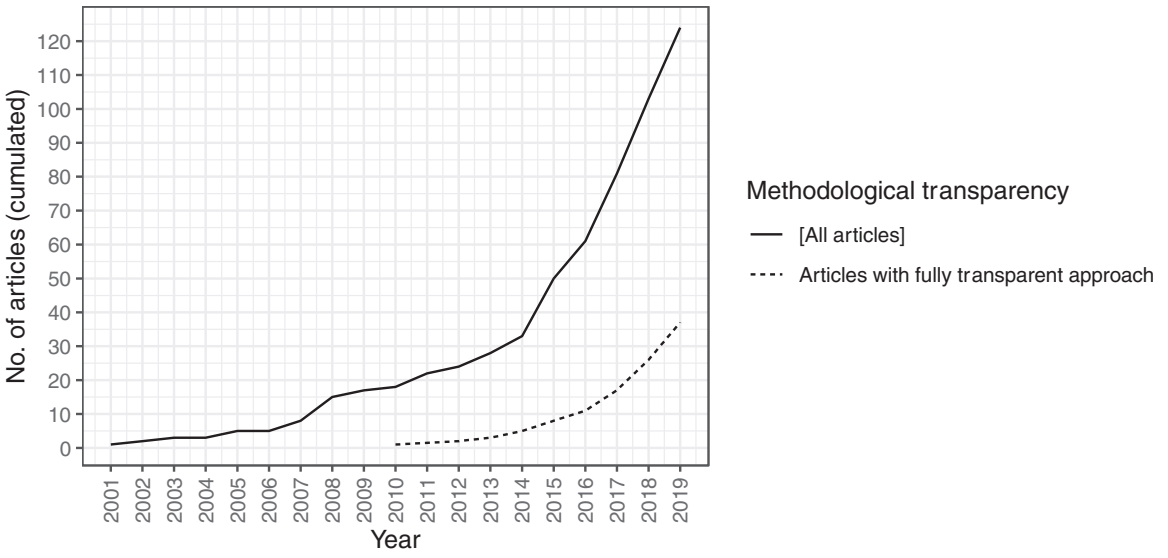

Figure 7. Trajectory of methodological transparency of text mining applications.

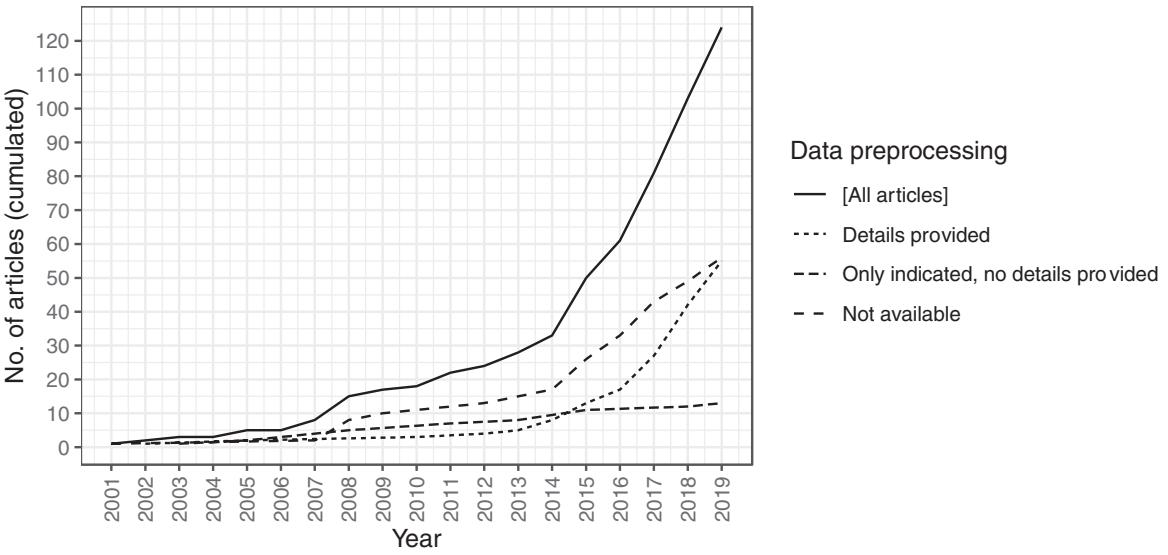

Figure 8. Trajectories of transparency of data preprocessing.

on data preprocessing. Finally, we see in Figure 9 that scholars are increasingly using non-proprietary software to conduct text mining. Here, programming languages like R and Python are mainly used offering the advantage of making the program code fully available to reviewers and fellow scholars.

5. Priorities for future text mining applications in innovation research

Based on our systematic review of the state and evolution of text mining applications in innovation research, we delineate conceptual, methodological, and contextual priorities for future innovation research applying text mining. We believe that these research priorities are of interest to both novices to text mining as well as experienced users. Novices may become acquainted with the state-of-the-art and readily implementable algorithms to kick-start their research projects. Experienced users, in turn, might draw on insights into the main limitations and challenges of current text mining applications to help raise the standards for rigorous and reliable text mining applications in innovation research and beyond.

From a *conceptual* point of view, we argue that innovation scholars now have suitable exemplars at hand that showcase the meaningful application and

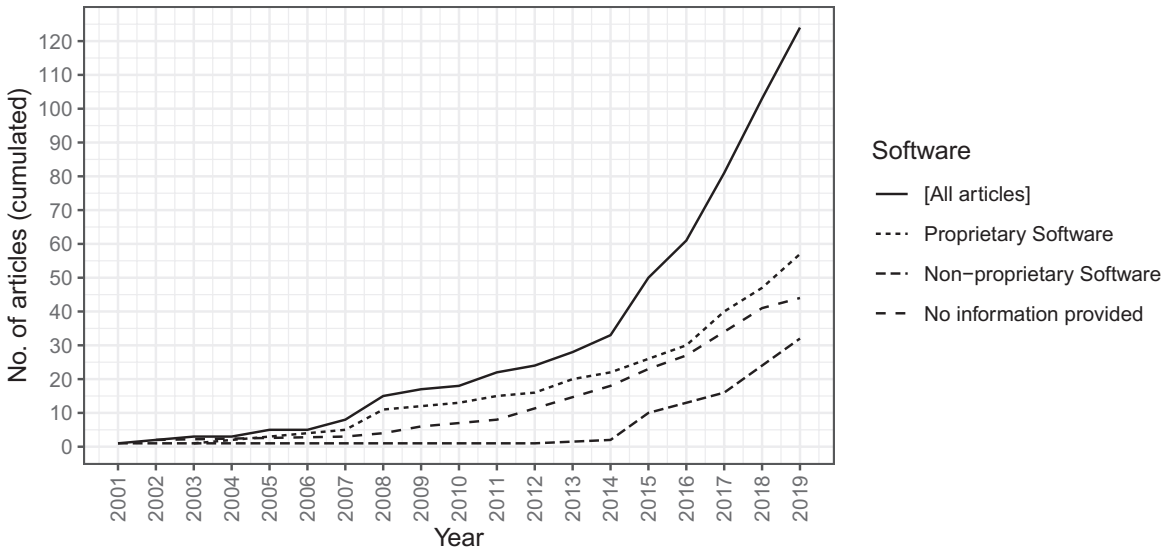

Figure 9. Trajectories of software usage.

value of text mining in innovation research. As we have seen in recent years, the number of studies that use text mining as a means to the end of generating variables for inclusion in subsequent statistical models sharply increased. For instance, Kaplan and Vakili (2015) used a topic model to generate a text-based measure of knowledge recombination that they subsequently incorporate as an independent variable into their econometric model. Similarly, Antons et al. (2019) applied LDA to generate measures of topic newness to quantitate a possible citation premium of topic newness. Angus (2019) computed document similarity measures to investigate the link between search distance and firm performance. Finally, Bednar et al. (2013) investigated how media coverage of a firm affects strategic change. We argue that this trend towards applying text mining techniques to operationalize new, previously unexplored variables or to improve the measurement of existing variables instead of simply demonstrating the techniques' usage or providing a case study can be interpreted as a sign of methodological maturity. This argument is further supported by the fact that 12 out of 18 articles operationalizing variables by means of text mining appeared in the most prestigious journals, as we have shown before. Even though the application of text mining as an innovative measurement approach is still in its early days in innovation research, we argue that it has much to offer to innovation scholars and their research endeavors.

We propose that adding text mining to an innovation scholar's toolbox allows them not only to harness the value of larger bodies of textual data, but also to increase the objectivity and reproducibility of empirical findings based on textual analysis. Text mining somehow bridges qualitative and quantitative research traditions by structuring, analyzing, and understanding textual data at scale. This link to both qualitative and quantitative research traditions is also evident in the discussion on quality criteria of text mining (e.g., Yu et al., 2011). Those text mining algorithms not including probability measures will perform in a consistent and repeatable manner on the same text as well as across texts. When the corpus and the settings of the algorithms are held stable, re-running the algorithms should yield the same results. Hence, text mining techniques are likely to produce more consistent results than human coders, especially with a growing amount of data (Yu et al., 2011). Moreover, it is possible to check the consistency of different algorithms that were designed for the same task in order to see whether they produce comparable results (McKenny et al., 2016). Going beyond reliability, it is necessary to ensure construct and criterion validity. This is of particular importance in studies testing theory and hypotheses in order to avoid both Type I errors of rejecting true hypotheses, and Type II errors of failing to reject incorrect hypotheses. Especially for those techniques that quantitate the content of text documents and whose measures are used as variables in quantitative models, it is necessary to establish construct validity. Prior research has shown that it is possible to do this by using measures derived from text mining (Short et al., 2010).

From a *methodological* perspective, we argue that future scholars are well advised to follow the best practice of studies that are fully transparent to readers. For us, this means that scholars need to disclose (1) the

Table 3. Employing text mining in research on innovation (examples)

	Innovation-related theme	Research focus	Data	Potentially useful text mining operations
#1	Novelty	Novelty detection	Scientific articles, patents, grant applications	• Clustering to group documents covering similar topics • Compare word distributions and usage across documents that constitute a new topic and documents that have been published in similar years on similar topics → Derive insights into if, how, and to what extent language differs that describes truly novel and breakthrough technological developments.
#2	Technology hypes	Innovation and media hypes	Newspaper articles	• Custom dictionaries on selected technologies and their application (e.g., Genome editing, cloud computing, or artificial intelligence) to measure media attention ('Does the article refer to selected technology?') • Pre-defined dictionaries to examine specific emotions and time references ('To what extent does an article include language that expresses fear, excitement, or uncertainty? Does the article contain language that refer to the past, present, or future?') → Derive insights into what characteristics of an innovation attract media attention and how journalists document its diffusion.
#3	Creativity	Exploration vs. exploitation strategies in crowdsourcing contests	Multiple technology proposals from same solver (team)	• Classifier to classify a proposal as close or distant to a solver's knowledge base ('Is a technology proposal similar to all previous ones of the same solver?') → Use classification of proposals as proxy for an exploration vs. exploitation strategy. Use it as explanatory variable in a regression model explaining success in contests while controlling for call characteristics.
#4	New product announcements	The effect of positive tenor of announcements on later sales	Full texts of new product announcements	• Sentiment analysis to account for the amount of positive words ('Is this a positively written new product announcement?') → Utilize the number of positive words or the share of positive words related to total words as an independent variable in regression models explaining sales volumes of various new products.
#5	Lifecycle management	The sweet spot of novelty in software updates	Customer reviews (e.g., Google Play Store)	• Clustering techniques to identify topics in reviews ('What is a review about?') • Pre-defined dictionaries to examine reviews' sentiment ('Is a review positively or negatively connotated?) → Combine both approaches to discover not only what parts of the software update raised particular attention of the user base, but whether the updates were evaluated positively or negatively.
#6	Innovation performance	The effect of team vs. product characteristics for crowdfunding success	Transcripts of crowdfunding campaigns (e.g., Kickstarter campaigns)	• Clustering techniques to find product categories of Kickstarter campaigns. • Sentiment analysis to find the positive sentiment for covered topics. • Pre-defined dictionaries to examine the share of technical and colloquial language. → Combine the approaches to find out how the topics and the language of Kickstarter campaign descriptions are related to reaching funding goals.

type of texts used, (2) the source from which this data were drawn, (3) the amount of data analyzed, (4) the software used to run preprocessing as well as analyses, (5) the techniques that were used to preprocess the data including detailed descriptions of how this affected the textual basis, and (6) the kind of algorithm(s) used to analyze the text including, if applicable, choices made to fine-tune the algorithm by means of setting methodological parameters. Only such fully transparent studies (e.g., Wang et al., 2010; Kaplan and Vakili, 2015; Antons et al., 2016, 2018; Hopp et al., 2018) enable future replication studies and the cumulative development of knowledge in our field.

As part of our review, we found that scholars are increasingly using free software like R and Python and use available packages and functions to run text preprocessing and text mining algorithms. We see this as particularly advantageous, since this puts the researcher into the driver's seat. Compared to some proprietary software, this allows for full introspection of how data are handled and how analytic algorithms are used. Moreover, it enables researchers to be fully transparent to fellow scholars and reviewers by making data and code of the analysis available by means of online repositories. However, it comes at the cost of learning the required programming skills and of becoming familiar with the specific text mining packages available. To ease the process for innovation scholars, we prepared an appendix that lists popular text mining packages for R and Python by type of analysis to be performed (see Table S2).

This leads us to the *contextual* question of identifying the fields of innovation research for which text mining applications for variable measurement and inclusion in subsequent models appear particularly promising. Texts like product reviews, patents, customer interviews, and product preannouncements in form of press releases have always been a vital source for innovation research. As it is almost impossible to craft an all-embracing and coherent picture of research questions that can be answered by using text mining to analyze such texts, we outline fields of research that are based both on our experience of using text mining in innovation research and are particularly salient or trending in the innovation research landscape (e.g., Antons et al., 2016; Antons and Breidbach, 2018).

Table 3 provides a non-exhaustive list of six exemplary research themes at the heart of our field of innovation research that we argue could benefit from the meaningful use of text mining techniques. This includes (#1) the development of novel measures for central, yet so far difficult to capture innovation concepts such as novelty using clustering techniques; (#2) the granular mapping of innovation trajectories in scientific and public discourse using custom dictionaries applied to media data; (#3) the scalable categorization of proposals (e.g., for grants, technical solutions, patents, articles) as exploratory or exploitative using classification techniques based on text-based distance measures; (#4) the fine-grained assessment of text content and sentiment (e.g., in new product announcements) as predictors of subsequent adoption using pre-defined dictionaries for positivity and/or negativity; (#5) the analysis of customer reviews on product updates using clustering and dictionary-based techniques, and (#6) the link between topic and language characteristics in product descriptions (e.g., on kick-starter) and the subsequent product success.

Overall, many of these examples illustrate distinct computational techniques can be combined to generate promising answers to discovery-oriented or theory-guided research questions. An analytical strategy can hence consist in a combination of multiple text mining techniques conducted in sequence (see #2, #4) and/or a combination of text mining techniques with econometric analyses (see #3, #6).

To find additional sweet spots for research applying text mining methodologies, we also refer the reader to Figure 4. Here, we see that some fields in innovation research, as represented by the PDMA subject areas, have not applied text mining as other areas or do not show a strong trajectory. As a case in point, research on networks, alliances, and ecosystems could benefit from using text mining. Scholars could use text mining techniques to mirror capabilities of firms active in an ecosystem by analyzing changes in their patent stocks and scientific publications to see whether these changes are associated with competitive advantages of these ecosystems.

6. Conclusion

Text mining, that is computer-aided analysis of textual data, offers a great opportunity to advance scholarship in innovation management. This motivated us to provide innovation scholars with an overview of available text mining methods, a systematic review of the state and evolution of text mining applications in innovation research as well as a set of priorities for those considering to apply text mining techniques in their own research. This included guidance on available R and Python libraries for data preprocessing, dictionaries, classifying and clustering. Our intent was to inform and to empower scholars to implement text mining in their research in a rigorous way, even if they are new to this methodological field. Overall, this review and the research priorities we delineate can facilitate the meaningful use of text mining by both novices and experienced scholars and

contribute to the methodological richness in our field of innovation research and beyond.

Acknowledgements

We are thankful to input that we received from several friendly reviewers and colleagues with whom we discussed our text mining-related research. In particular, we thank Michael Barrett, Christoph Breidbach, Amol Joshi, Rajiv Kohli, Dirk Lüttgens, Frank Piller, and Matthias Rüdiger. We would also like to thank Leona Brust and Julian Mertes, who supported our text mining research projects as student assistants.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site: Supporting information.

References

Agarwal, R. and Dhar, V. (2014) Editorial—big data, data science, and analytics: the opportunity and challenge for is research. *Information Systems Research*, **25**, 3, 443–448.

Albert, T., Moehrle, M.G., and Meyer, S. (2015) Technology maturity assessment based on blog analysis. *Technological Forecasting and Social Change*, **92**, 196–209.

Alencar, M.S.M., Porter, A.L., and Antunes, A.M.S. (2007) Nanopatenting patterns in relation to product life cycle. *Technological Forecasting and Social Change*, **74**, 9, 1661–1680.

Al-Hasan, A., Yim, D. and Lucas, H.C. (2019) A tale of two movements: Egypt during the Arab spring and occupy wall street. *IEEE Transactions on Engineering Management*, **66**, 1, 84–97.

Angus, R.W. (2019) Problemistic search distance and entrepreneurial performance. *Strategic Management Journal*, **40**, 2011–2023.

Antons, D. and Breidbach, C. (2018) Big data, big insights? Advancing service innovation and design with machine learning. *Journal of Service Research*, **21**, 1, 17–39.

Antons, D., Kleer, R., and Salge, T.O. (2016) Mapping the topic landscape of JPIM, 1984–2013: in search of hidden structures and development trajectories. *Journal of Product Innovation Management*, **33**, 6, 726–749.

Antons, D., Joshi, A.M., and Salge, T.O. (2019) Content, contribution, and knowledge consumption: uncovering hidden topic structure and rhetorical signals in scientific texts. *Journal of Management*, **45**, 7, 3035–3076.

Archak, N., Ghose, A., and Ipeirotis, P.G. (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Science*, **57**, 8, 1485–1509.

Arts, S., Cassiman, B., and Gomez, J.C. (2018) Text matching to measure patent similarity. *Strategic Management Journal*, **39**, 1, 62–84.

Basole, R.C., Park, H., and Chao, R.O. (2018) Visual analysis of venture similarity in entrepreneurial ecosystems. *IEEE Transactions on Engineering Management*, **66**, 4, 568–582.

Bednar, M.K. (2012) Watchdog or lapdog? A behavioral view of the media as a corporate governance mechanism. *Academy of Management Journal*, **55**, 131–150.

Bednar, M.K., Boivie, S., and Prince, N.R. (2013) Burr under the saddle: how media coverage influences strategic change. *Organization Science*, **24**, 3, 910–925.

Beretta, M. (2019) Idea selection in web-enabled ideation systems. *Journal of Product Innovation Management*, **36**, 1, 5–23.

Berger, J. and Milkman, K.L. (2012) What makes online content viral? *Journal of Marketing Research*, **49**, 2, 192–205.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J.P., Moehrle, M.G., and Erdmann, V.A. (2008) Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips. *R&D Management*, **38**, 5, 550–562.

Bharadwaj, N., Noble, C.H., Tower, A., Smith, L.M., and Dong, Y. (2017) Predicting innovation success in the motion picture industry: the influence of multiple quality signals. *Journal of Product Innovation Management*, **34**, 5, 659–680.

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

Broniatowski, D.A. and Magee, C.L. (2017) The emergence and collapse of knowledge boundaries. *IEEE Transactions on Engineering Management*, **64**, 3, 337–350.

Brust, L., Breidbach, C., Antons, D., and Salge, T.O. (2017) Service-dominant logic and information systems research: a review and analysis using topic modeling. Proceedings of the International Conference on Information Systems (ICIS), Seoul.

Chae, B. and Olson, D. (2018) A topical exploration of the intellectual development of decision sciences 1975–2016: intellectual development of decision sciences 1975–2016. *Decision Sciences*.

Chen, S.-H., Huang, M.-H., and Chen, D.-Z. (2012) Identifying and visualizing technology evolution: a case study of smart grid technology. *Technological Forecasting and Social Change*, **79**, 6, 1099–1110.

Chen, H., Zhang, G., Zhu, D., and Lu, J. (2017) Topic-based technological forecasting based on patent data: a case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change*, **119**, 39–52.

Chen, H., Wang, X., Pan, S., and Xiong, F. (2019) Identify topic relations in scientific literature using topic modeling. *IEEE Transactions on Engineering Management*, 1–13.

Chintagunta, P., Hanssens, D.M., and Hauser, J.R. (2016) Editorial—marketing science and big data. *Marketing Science*, **35**, 3, 341–342.

Choi, D.G., Lee, Y., Jung, M., and Lee, H. (2012) National characteristics and competitiveness in MOT research: a comparative analysis of ten specialty journals, 2000–2009. *Technovation*, **32**, 1, 9–18.

Choi, S., Kim, H., Yoon, J., Kim, K., and Lee, J.Y. (2013) An SAO-based text-mining approach for technology roadmapping using patent information. *R&D Management*, **43**, 1, 52–74.

Conti, A., Denas, O., and Visentin, F. (2014) Knowledge specialization in Ph.D. student groups. *IEEE Transactions on Engineering Management*, **61**, 1, 52–67.

Coussement, K., Debaere, S., and De Ruyck, T. (2017) Inferior member participation identification in innovation communities: the signaling role of linguistic style use. *Journal of Product Innovation Management*, **34**, 5, 565–579.

Cunningham, S.W. and Kwakkel, J.H. (2014) Tipping points in science: a catastrophe model of scientific change. *Journal of Engineering and Technology Management*, **32**, 185–205.

Deephouse, D.L. and Carter, S.M. (2005) An examination of differences between organizational legitimacy and organizational reputation. *Journal of Management Studies*, **42**, 2, 329–360.

Denyer, D. and Tranfield, D. (2009) Producing a systematic review. *The Sage Handbook of Organizational Research Methods*, 671–689.

Dernis, H., Squicciarini, M., and de Pinho, R. (2016) Detecting the emergence of technologies and the evolution and co-development trajectories in science (DETECTS): a 'burst' analysis-based approach. *The Journal of Technology Transfer*, **41**, 5, 930–960.

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

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996) The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**, 11, 27–34.

Fung, G.M. and Mangasarian, O.L. (2005) Multicategory proximal support vector machine classifiers. *Machine Learning*, **59**, 1–2, 77–97.

Furukawa, T., Mori, K., Arino, K., Hayashi, K., and Shirakawa, N. (2015) Identifying the evolutionary process of emerging technologies: a chronological network analysis of world wide web conference sessions. *Technological Forecasting and Social Change*, **91**, 280–294.

Gamache, D.L., McNamara, G., Mannor, M.J., and Johnson, R.E. (2015) Motivated to acquire? The impact of CEO regulatory focus on firm acquisitions. *Academy of Management Journal*, **58**, 4, 1261–1282.

George, G., Haas, M., and Pentland, A. (2014) Big data and management. *Academy of Management Journal*, **57**, 2, 231–326.

George, G., Osinega, E.C., Lavie, D., and Scott, B.A. (2016) From the editors – big data and data science methods for management research. *Academy of Management Journal*, **59**, 5, 1493–1507.

Geum, Y., Lee, H., Lee, Y., and Park, Y. (2015) Development of data-driven technology roadmap considering dependency: an ARM-based technology roadmapping. *Technological Forecasting and Social Change*, **91**, 264–279.

Ghazinoory, S., Ameri, F., and Farnoodi, S. (2013) An application of the text mining approach to select technology centers of excellence. *Technological Forecasting and Social Change*, **80**, 5, 918–931.

Goffin, K., Ahlström, P., Bianchi, M. and Richtnér, A. (2019) Perpsective: state-of-the-art: the quality of case study research in innovation management. *Journal of Product Innovation Management*, **36**, 5, 586–615.

Griffin, A. and Somermeyer, S. (2007) *The PDMA Toolbook 3 for New Product Development*. Hoboken, NJ: Wiley.

Grover, P., Kar, A.K., Dwivedi, Y.K., and Janssen, M. (2019) Polarization and acculturation in US election 2016 outcomes – can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, **145**, 438–460.

Guo, J., Wang, X., Li, Q., and Zhu, D. (2016) Subject–action–object-based morphology analysis for determining the direction of technological change. *Technological Forecasting and Social Change*, **105**, 27–40.

Han, E.J., and Sohn, S.Y. (2015) Patent valuation based on text mining and survival analysis. *The Journal of Technology Transfer*, **40**, 5, 821–839.

Henry, E. (2008) Are investors influenced by how earnings press releases are written? *The Journal of Business Communication*, **45**, 4, 363–407.

Höllerer, M.A., Jancsary, D., Barberio, V., and Meyer, R.E. (2019) The interlinking theorization of management concepts: cohesion and semantic equivalence in management knowledge. *Organization Studies*.

Hoornaert, S., Ballings, M., Malthouse, E.C., and Van den Poel, D. (2017) Identifying new product ideas: waiting for the wisdom of the crowd or screening ideas in real time. *Journal of Product Innovation Management*, **34**, 5, 580–597.

Hopp, C., Antons, D., Karminski, J., and Salge, T.O. (2018) The topic landscape of disruption research—a call for consolidation, reconciliation, and generalization. *Journal of Product Innovation Management*, **35**, 3, 458–487.

Hornik, K. and Grün, B. (2011) topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, **13**, 1–30.

Huang, J.-Y. (2016) Patent portfolio analysis of the cloud computing industry. *Journal of Engineering and Technology Management*, **39**, 45–64.

Humphreys, A. and Wang, R.J.-H. (2018) Automated text analysis for consumer research. *Journal of Consumer Research*, **44**, 6, 1274–1306.

Hwang, S. and Shin, J. (2019) Extending technological trajectories to latest technological changes by overcoming time lags. *Technological Forecasting and Social Change*, **143**, 142–153.

Ingersoll, G.S., Morton, T.S. and Farris, A.L. (2013) *Taming Text: How to Find, Organize, and Manipulate It*. Shelter Island, NY: Manning Publications Co.

Ittipanuvat, V., Fujita, K., Sakata, I., and Kajikawa, Y. (2014) Finding linkage between technology and social issue: a literature based discovery approach. *Journal of Engineering and Technology Management*, **32**, 160–184.

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 8, 651–666.

Jamiy, F.E., Daif, A., Azouazi, M., and Marzak, A. (2015) The potential and challenges of big data – recommendation systems next level application. arXiv preprint arXiv:1501.03424.

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

Jeong, Y., Lee, K., Yoon, B., and Phaal, R. (2015) Development of a patent roadmap through the generative topographic mapping and bass diffusion model. *Journal of Engineering and Technology Management*, **38**, 53–70.

Jeong, Y., Park, I., and Yoon, B. (2019) Identifying emerging research and business development R&D areas based on topic modeling and visualization with intellectual property right data. *Technological Forecasting and Social Change*, **146**, 655–672.

Jiang, H., Qiang, M., Fan, Q., and Zhang, M. (2018) Scientific research driven by large-scale infrastructure projects: a case study of the three gorges project in China. *Technological Forecasting and Social Change*, **134**, 61–71.

Johnson, S.L., Safadi, H., and Faraj, S. (2015) The emergence of online community leadership. *Information Systems Research*, **26**, 1, 165–187.

Jun, C.N. and Chung, C.J. (2016) Big data analysis of local government 3.0: focusing on Gyeongsangbuk-Do in Korea. *Technological Forecasting and Social Change*, **110**, 3–12.

Kaplan, S. and Vakili, K. (2015) The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, **36**, 10, 1435–1457.

Kayser, V. (2017) Comparing public and scientific discourse in the context of innovation systems. *Technological Forecasting and Social Change*, **115**, 348–357.

Kim, J. and Lee, C. (2017) Novelty-focused weak signal detection in futuristic data: assessing the rarity and paradigm unrelatedness of signals. *Technological Forecasting and Social Change*, **120**, 59–76.

Kim, J. and Park, Y. (2017) Leveraging ideas from user innovation communities: using text-mining and case-based reasoning. *R&D Management*, **49**, 2, 155–167.

Kim, N., Lee, H., Kim, W., Lee, H., and Suh, J.H. (2015) Dynamic patterns of industry convergence: evidence from a large amount of unstructured data. *Research Policy*, **44**, 9, 1734–1748.

Kim, D., Lee, H., and Kwak, J. (2017a) Standards as a driving force that influences emerging technological trajectories in the converging world of the internet and things: an investigation of the M2M/IoT patent network. *Research Policy*, **46**, 7, 1234–1254.

Kim, H., Hong, S., Kwon, O., and Lee, C. (2017b) Concentric diversification based on technological capabilities: link analysis of products and technologies. *Technological Forecasting and Social Change*, **118**, 246–257.

Kim, H., Ahn, S.-J., and Jung, W.-S. (2019a) Horizon scanning in policy research database with a probabilistic topic model. *Technological Forecasting and Social Change*, **146**, 588–594.

Kim, P.H., Kotha, R., Fourné, S.P.L., and Coussement, K. (2019b) Taking leaps of faith: evaluation criteria and resource commitments for early-stage inventions. *Research Policy*, **48**, 6, 1429–1444.

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

Kostoff, R.N. (2008) Literature-related discovery (LRD): potential treatments for cataracts. *Technological Forecasting and Social Change*, **75**, 2, 215–225.

Kostoff, R.N. (2011) Literature-related discovery: potential treatments and preventatives for SARS. *Technological Forecasting and Social Change*, **78**, 7, 1164–1173.

Kostoff, R.N. and Briggs, M.B. (2008) Literature-related discovery (LRD): potential treatments for Parkinson's disease. *Technological Forecasting and Social Change*, **75**, 2, 226–238.

Kostoff, R.N., Koytcheff, R.G., and Lau, C.G.Y. (2007) Global nanotechnology research literature overview. *Technological Forecasting and Social Change*, **74**, 9, 1733–1747.

Kostoff, R.N., Block, J.A., Stump, J.A., and Johnson, D. (2008a) Literature-related discovery (LRD): potential treatments for Raynaud's phenomenon. *Technological Forecasting and Social Change*, **75**, 2, 203–214.

Kostoff, R.N., Briggs, M.B., and Lyons, T.J. (2008b) Literature-related discovery (LRD): potential treatments for multiple sclerosis. *Technological Forecasting and Social Change*, **75**, 2, 239–255.

Kwon, H. and Park, Y. (2018) Proactive development of emerging technology in a socially responsible manner: data-driven problem solving process using latent semantic analysis. *Journal of Engineering and Technology Management*, **50**, 45–60.

Kwon, H., Kim, J., and Park, Y. (2017) Applying LSA text mining technique in envisioning social impacts of emerging technologies: the case of drone technology. *Technovation*, **60–61**, 15–28.

Kwon, H., Park, Y., and Geum, Y. (2018) Toward data-driven idea generation: application of wikipedia to morphological analysis. *Technological Forecasting and Social Change*, **132**, 56–80.

Landers, R.N., Brusso, R.C., Cavanaugh, K.J., and Collmus, A.B. (2016) A primer on theory-driven web scraping: automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, **21**, 4, 475–492.

Lee, H. and Kang, P. (2018) Identifying core topics in technology and innovation management studies: a topic model approach. *The Journal of Technology Transfer*, **43**, 5, 1291–1317.

Lee, S., Lee, S., Seol, H., and Park, Y. (2008a) Using patent information for designing new product and technology: keyword based technology roadmapping. *R&D Management*, **38**, 2, 169–188.

Lee, S., Lee, S., Hyeonju, S., and Park, Y. (2008b) Using patent information technology: keyword based technology roadmapping. *R&D Management*, **38**, 2, 169–188.

Lee, S., Yoon, B., and Park, Y. (2009) An approach to discovering new technology opportunities: keyword-based patent map approach. *Technovation*, **29**, 6–7, 481–497.

Lee, W.S., Han, E.J., and Sohn, S.Y. (2015) Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. *Technological Forecasting and Social Change*, **100**, 317–329.

Lee, J., Kim, C., and Shin, J. (2017) Technology opportunity discovery to R&D planning: key technological performance analysis. *Technological Forecasting and Social Change*, **119**, 53–63.

Li, F. (2010) The information content of forward-looking statements in corporate filings: a naive Bayesian machine learning approach. *Journal of Accounting Research*, **48**, 5, 1049–1102.

Li, Y., Arora, S., Youtie, J., and Shapira, P. (2018) Using web mining to explore Triple Helix influences on growth in small and mid-size firms. *Technovation*, **76–77**, 3–14.

Li, X., Xie, Q., Daim, T., and Huang, L. (2019a) Forecasting technology trends using text mining of the gaps between science and technology: the case of perovskite solar cell technology. *Technological Forecasting and Social Change*, **146**, 432–449.

Li, X., Xie, Q., Jiang, J., Zhou, Y., and Huang, L. (2019b) Identifying and monitoring the development trends of emerging technologies using patent analysis and twitter data mining: the case of perovskite solar cell technology. *Technological Forecasting and Social Change*, **146**, 687–705.

Linton, J.D. and Thongpapanl, N.T. (2004) Perspective: ranking the technology Innovation Management Journals. *Journal of Product Innovation Management*, **21**, 123–139.

Lipizzi, C., Iandoli, L., and Marquez, J.E.R. (2016) Combining structure, content and meaning in online social networks: the analysis of public's early reaction in social media to newly launched movies. *Technological Forecasting and Social Change*, **109**, 35–49.

Lu, L.Y.Y. and Liu, J.S. (2016) A novel approach to identify the major research themes and development trajectory: the case of patenting research. *Technological Forecasting and Social Change*, **103**, 71–82.

Luo, C., Kumar, S., Mallick, D.N., and Luo, B. (2018) Impacts of exploration and exploitation on firm's performance and the moderating effects of slack: a panel data analysis. *IEEE Transactions on Engineering Management*, **66**, 4, 613–620.

Ma, T., Zhang, Y., Huang, L., Shang, L., Wang, K., Yu, H., and Zhu, D. (2017) Text mining to gain technical intelligence for acquired target selection: a case study for China's computer numerical control machine tools industry. *Technological Forecasting and Social Change*, **116**, 162–180.

Ma, J., Abrams, N.F., Porter, A.L., Zhu, D., and Farrell, D. (2019) Identifying translational indicators and technology opportunities for nanomedical research using tech mining: the case of gold nanostructures. *Technological Forecasting and Social Change*, **146**, 767–775.

Magerman, T., Looy, B.V., and Debackere, K. (2015) Does involvement in patenting jeopardize one's academic footprint? An analysis of patent-paper pairs in biotechnology. *Research Policy*, **44**, 9, 1702–1713.

McKenny, A.F., Aguinis, H., Short, J.C., and Anglin, A.H. (2016) What doesn't get measured does exist: improving the accuracy of computer-aided text analysis. *Journal of Management*, **44**, 7, 2909–2933.

Mejía, C. and Kajikawa, Y. (2019) Technology news and their linkage to production of knowledge in robotics research. *Technological Forecasting and Social Change*, **143**, 114–124.

Miner, G., Elder, J., IV, Fast, A., Hill, T., Nisbet, R., and Dursun, D. (2012) *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Waltham, MA: Academic Press.

Moehrle, M.G. and Caferoglu, H. (2019) Technological speciation as a source for emerging technologies. Using semantic patent analysis for the case of camera technology. *Technological Forecasting and Social Change*, **146**, 776–784.

Moehrle, M.G., Walter, L., Geritz, A., and Müller, S. (2005) Patent-based inventor profiles as a basis for human resource decisions in research and development. *R&D Management*, **35**, 5, 513–524.

Moehrle, M.G., Wustmans, M., and Gerken, J.M. (2017) How business methods accompany technological innovations – a case study using semantic patent analysis and a novel informetric measure. *R&D Management*, **48**, 3, 331–342.

Momeni, A. and Rost, K. (2016) Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. *Technological Forecasting and Social Change*, **104**, 16–29.

Mora, L., Deakin, M., and Reid, A. (2019) Combining co-citation clustering and text-based analysis to reveal the main development paths of smart cities. *Technological Forecasting and Social Change*, **142**, 56–69.

Mora-Valentín, E.-M., Ortiz-de-Urbina-Criado, M., and Nájera-Sánchez, J.-J. (2018) Mapping the conceptual structure of science and technology parks. *The Journal of Technology Transfer*, **43**, 5, 1410–1435.

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

Mulrow, C.D. (1994) Systematic reviews: rationale for systematic reviews. *British Medical Journal*, **309**, 6954, 597–599.

Nambisan, S., Lyytinen, K., Majchrzak, A., and Song, M. (2017) Digital innovation management: reinventing

innovation management research in a digital world. *MIS Quarterly*, **41**, 1, 223–238.

Nathan, M. and Rosso, A. (2015) Mapping digital businesses with big data: some early findings from the UK. *Research Policy*, **44**, 9, 1714–1733.

Newman, N.C., Porter, A.L., Newman, D., Trumbach, C.C., and Bolan, S.D. (2014) Comparing methods to extract technical content for technological intelligence. *Journal of Engineering and Technology Management*, **32**, 97–109.

Noh, H. and Lee, S. (2019) Where technology transfer research originated and where it is going: a quantitative analysis of literature published between 1980 and 2015. *The Journal of Technology Transfer*, **44**, 3, 700–740.

Nokelainen, T. and Dedehayir, O. (2015) Technological adoption and use after mass market displacement: the case of the LP record. *Technovation*, **36–37**, 65–76.

Ogawa, T. and Kajikawa, Y. (2017) Generating novel research ideas using computational intelligence: a case study involving fuel cells and ammonia synthesis. *Technological Forecasting and Social Change*, **120**, 41–47.

Olmedilla, M., Send, H., and Toral, S.L. (2019) Identification of the unique attributes and topics within smart things open innovation communities. *Technological Forecasting and Social Change*, **146**, 133–147.

Páez-Avilés, C., Van Rijnsoever, F.J., Juanola-Feliu, E., and Samitier, J. (2018) Multi-disciplinarity breeds diversity: the influence of innovation project characteristics on diversity creation in nanotechnology. *The Journal of Technology Transfer*, **43**, 2, 458–481.

Pennebaker, J.W., Boyd, R.L., Jordan, K., and Blackburn, K. (2015) *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Porter, A.L. and Pei, R. (2011) Profiling leading scientists in nanobiomedical science: interdisciplinarity and potential leading indicators of research directions. *R&D Management*, **41**, 3, 288–306.

Porter, A.L., Garner, J., Carley, S.F., and Newman, N.C. (2019) Emergence scoring to identify Frontier R&D topics and key players. *Technological Forecasting and Social Change*, **146**, 628–643.

Randhawa, K., Wilden, R., and Hohberger, J. (2016) A bibliometric review of open innovation: setting a research agenda. *Journal of Product Innovation Management*, **33**, 6, 750–772.

Randhawa, K., Wilden, R., and Gudergan, S. (2018) Open service innovation: the role of intermediary capabilities. *Journal of Product Innovation Management*, **35**, 5, 808–838.

Rezaeian, M., Montazeri, H., and Loonen, R.C.G.M. (2017) Science foresight using life-cycle analysis, text mining and clustering: a case study on natural ventilation. *Technological Forecasting and Social Change*, **118**, 270–280.

Riccaboni, M. and Moliterni, R. (2009) Managing technological transitions through R&D alliances. *R&D Management*, **29**, 2, 124–135.

Ruckman, K. and McCarthy, I. (2017) Why do some patents get licensed while others do not? *Industrial and Corporate Change*, **26**, 4, 667–688.

Rüdiger, M., Antons, D., and Salge, T.O. (2017) From text to data: on the role and effect of text pre-processing in text mining research. *Academy of Management Proceedings*, **2017**, 1, 16353.

Seo, W., Yoon, J., Park, H., Coh, B., Lee, J.-M., and Kwon, O.-J. (2016) Product opportunity identification based on internal capabilities using text mining and association rule mining. *Technological Forecasting and Social Change*, **105**, 94–104.

Shapira, P., Gök, A., Klochikhin, E., and Sensier, M. (2014) Probing 'green' industry enterprises in the UK: a new identification approach. *Technological Forecasting and Social Change*, **85**, 93–104.

Shibata, N., Kajikawa, Y., Takeda, Y., and Matsushima, K. (2008) Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, **28**, 11, 758–775.

Short, J.C., Broberg, C., Cogliser, C.C., and Brigham, K. (2010) Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods*, **13**, 2, 320–347.

Smalheiser, N.R. (2001) Predicting emerging technologies with the aid of text-based data mining: the micro approach. *Technovation*, **21**, 689–693.

Sohrabi, B. and Khalilijafarabad, A. (2018) Systematic method for finding emergence research areas as data quality. *Technological Forecasting and Social Change*, **137**, 280–287.

Song, B. and Suh, Y. (2019) Identifying convergence fields and technologies for industrial safety: LDA-based network analysis. *Technological Forecasting and Social Change*, **138**, 115–126.

Song, K., Kim, K.S., and Lee, S. (2017) Discovering new technology opportunities based on patents: text-mining and F-term analysis. *Technovation*, **60–61**, 1–14.

Suh, J.H. (2015) Forecasting the daily outbreak of topic-level political risk from social media using hidden Markov model-based techniques. *Technological Forecasting and Social Change*, **94**, 115–132.

Suominen, A., Toivanen, H., and Seppänen, M. (2017) Firms' knowledge profiles: mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, **115**, 131–142.

Tanskanen, K., Ahola, T., Aminoff, A., Bragge, J., Kaipia, R., and Kauppi, K. (2017) Towards evidence-based management of external resources: developing design propositions and future research avenues through research synthesis. *Research Policy*, **46**, 6, 1087–1105.

Teso, E., Olmedilla, M., Martínez-Torres, M.R., and Toral, S.L. (2018) Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective. *Technological Forecasting and Social Change*, **129**, 131–142.

Thieme, J. (2007) Perspective: the world's top innovation management scholars and their social capital. *Journal of Product Innovation Management*, **24**, 214–229.

Titus, V., House, J.M., and Covin, J.G. (2017) The influence of exploration on External corporate venturing activity. *Journal of Management*, **43**, 5, 1609–1630.

Tong, S. and Koller, D. (2001) Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, **2**, 45–66.

Tranfield, D., Denyer, D., and Smart, P. (2003) Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, **14**, 3, 207–222.

Uotila, J., Maula, M., Keil, T., and Zahra, S.A. (2009) Exploration, exploitation, and financial performance: analysis of S&P 500 corporations. *Strategic Management Journal*, **30**, 2, 221–231.

Venugopalan, S. and Rai, V. (2015) Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change*, **94**, 236–250.

Vogel, R., Hattke, F., and Petersen, J. (2017) Journal rankings in management and business studies: what rules do we play by? *Research Policy*, **46**, 10, 1707–1722.

Wang, X., Mccallum, A., and Wei, X. (2007) Topical N-grams: phrase and topic discovery, with an application to information retrieval. *Proceedings – IEEE International Conference on Data Mining, ICDM*, 697–702.

Wang, M.Y., Chang, D.S., and Kao, C.H. (2010) Identifying technology trends for R&D planning using TRIZ and text mining. *R&D Management*, **40**, 5, 491–509.

Wang, X., Bendle, N.T., Mai, F., and Cotte, J. (2015a) The journal of consumer research at 40: a historical analysis. *Journal of Consumer Research*, **42**, 1, 5–18.

Wang, M.-Y., Fang, S.-C., and Chang, Y.-H. (2015b) Exploring technological opportunities by mining the gaps between science and technology: microalgal biofuels. *Technological Forecasting and Social Change*, **92**, 182–195.

Wang, X., Qiu, P., Zhu, D., Mitkova, L., Lei, M., and Porter, A.L. (2015c) Identification of technology development trends based on subject–action–object analysis: the case of dye-sensitized solar cells. *Technological Forecasting and Social Change*, **98**, 24–46.

Watts, R.J. and Porter, A.L. (2003) R&D Cluster Quality Measures and Technology Maturity. *Technological Forecasting and Social Change*, **70**, 8, 735–758.

Webb, L.M., Gibson, D.M., Wang, Y., Chang, H.C., and Thompson-Hayes, M. (2015) *Selecting, Scraping, and Sampling Big Data Sets from the Internet: Fan Blogs as Exemplar*. SAGE.

Whalen, R. (2018) Boundary spanning innovation and the patent system: interdisciplinary challenges for a specialized examination system. *Research Policy*, **47**, 7, 1334–1343.

Wiedemann, G. (2013) Opening up to big data: computer-assisted analysis of textual data in social sciences. *Historical Social Research/Historische Sozialforschung*, **38**, 4, 332–357.

Woo, J., Lee, M.J., Ku, Y., and Chen, H. (2015) Modeling the dynamics of medical information through web forums in medical industry. *Technological Forecasting and Social Change*, **97**, 77–90.

Wu, F.-S., Hsu, C.-C., Lee, P.-C., and Su, H.-N. (2011) A systematic approach for integrated trend analysis—the case of etching. *Technological Forecasting and Social Change*, **78**, 3, 386–407.

Yoon, B. and Magee, C.L. (2018) Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. *Technological Forecasting and Social Change*, **132**, 105–117.

Yoon, B. and Park, Y. (2005) A systematic approach for identifying technology opportunities: keyword-based morphology analysis. *Technological Forecasting and Social Change*, **72**, 2, 145–160.

Yoon, B. and Park, Y. (2007) Development of new technology forecasting algorithm: hybrid approach for morphology analysis and conjoint analysis of patent information. *IEEE Transactions on Engineering Management*, **54**, 3, 588–599.

Yoon, B.U., Yoon, B.C., and Park, Y.T. (2002) On the development and application of a self-organizing feature map-based patent map. *R&D Management*, **32**, 4, 291–300.

Yu, X. and Zhang, B. (2019) Obtaining advantages from technology revolution: a patent roadmap for competition analysis and strategy planning. *Technological Forecasting and Social Change*, **145**, 273–283.

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

Zachary, M.A., McKenny, A.F., Short, J.C., and Payne, G.T. (2011) Family business and market orientation: construct validation and comparative analysis. *Family Business Review*, **24**, 3, 233–251.

Zeng, M.A. (2018) Foresight by online communities – the case of renewable energies. *Technological Forecasting and Social Change*, **129**, 27–42.

Zhang, Y., Porter, A.L., Hu, Z., Guo, Y., and Newman, N.C. (2014) Term clumping' for technical intelligence: a case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, **85**, 26–39.

Zhang, Y., Zhang, G., Chen, H., Porter, A.L., Zhu, D., and Lu, J. (2016) Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, **105**, 179–191.

Zhou, Y., Dong, F., Kong, D., and Liu, Y. (2019) Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies. *Technological Forecasting and Social Change*, **144**, 205–220.

Notes

[1]We refer interested readers to introductions to data gathering and web scraping for social scientists like Landers et al. (2016), Kobayashi et al. (2018), or Webb et al. (2015).

[2]Much research is focused on overcoming the limitations of presenting text in a document-term matrix. Other techniques that help overcoming the 'bag-or-words'

issue are parsing, the later discussed part-of-speech tagging, or synonym recognition. Parsing techniques based on pretrained neuronal networks reach high accuracy in predicting word relations and part-of-speech. Recognition of synonyms can be done via lexical databases such as WordNet, which links words into semantic relations including synonyms, hyponyms, and meronyms. As of today, implementing many of these techniques, however, requires advanced programming skills.

[3]Intrigued by a reviewer comment, we also inspected journals beyond that list that frequently publish text mining studies on innovation-related themes. As a case in point, the journal *Scientometrics* published more than 240 articles where the title, abstract, or keywords included terms like innovation, innovativeness, or R&D along with text mining vocabulary. However, we found that these articles applied text mining techniques primarily as a means to compile science, R&D, and innovation indicators often using patent and journal data. Innovation management research articles, in contrast, employed text mining techniques on a much broader array of texts often in an effort to unpack the content of salient innovation management concepts and phenomena including knowledge consumption (Antons et al., 2019), knowledge recombination (Kaplan and Vakili, 2015), idea selection (Beretta, 2019), user communities (Kim and Park, 2017), technology evolution and roadmaps (Choi et al., 2013), or disruptive innovation (Hopp et al., 2018).

[4]Our own use of manual coding (coupled with bibliometric analyses) of the 124 journal articles identified for our systematic review is consistent with a view of diligent manual analysis by multiple human coders as the gold standard for content analysis, whenever the size of the text corpus is such that is accessible to manual analysis.

[5]This might be due to the fact that the journal's scope is focused on technology evolution, technology trends, and technology road maps – all subjects particularly suitable for text mining applications. These studies often rely on patent or publication data and apply clustering techniques. One should note that most studies tend to have the form of a literature review, a demonstration of a technique, or a case study.

Dr David Antons is a professor and co-director of the Institute for Technology and Innovation Management (TIM) in the TIME Research Area at RWTH Aachen University, Germany. He received his PhD from the same university working in close cooperation with the RWTH Psychological Institute. He held visiting appointments at the universities of Cambridge and Melbourne. His research interests include knowledge sharing across organizational, spatial, and disciplinary boundaries; psychological influences on decision-making; individual learning from feedback; and text mining approaches. Recent contributions have been published in journals such as Academy of Management Review, Journal of Management, Academy of Management Perspectives, Journal of Service Research, and Journal of Product Innovation Management.

Mr Eduard Grünwald is a research associate at the Institute for Technology and Innovation Management (TIM) in the TIME Research Area at RWTH Aachen University, Germany. He received his master's degree in business administration from the same university. His research focuses on the integration of text mining methods in management research.

Dr Patrick Cichy is a postdoctoral researcher at the Institute for Technology and Innovation Management (TIM) in the TIME Research Area at RWTH Aachen University, Germany. He received his PhD from the same university and conducted research at the National University of Singapore. His research focuses on digital business models and services; privacy and cyber security; as well as on institutional change. He published research on these topics in the Proceedings of the International Conference on Information Systems.

Dr Torsten Oliver Salge is a professor and co-director of the Institute for Technology and Innovation Management (TIM) within the TIME Research Area at RWTH Aachen University, Germany. He received his PhD from the University of Cambridge and has held (visiting) appointments at universities in Auckland, Buenos Aires, Bochum, Cambridge, Duisburg, Oxford, and Philadelphia. His current research interests include collaborative innovation, organizational search, and learning from performance feedback. Recent contributions have been published in journals such as Academy of Management Review, Journal of Applied Psychology, Journal of Management, Journal of Service Research, MIS Quarterly, Journal of Product Innovation Management, and Research Policy.

