Received: 11 May 2023 | Accepted: 17 October 2023

DOI: 10.1002/mar.21931

RESEARCH ARTICLE

Psychology & Marketing WILEY

How to conduct efficient and objective literature reviews using natural language processing: A step-by-step guide for marketing researchers

Serena Pugliese | Verdiana Giannetti | Sourindra Banerjee

Marketing Department, Leeds University Business School, University of Leeds, Leeds, UK

Correspondence
Serena Pugliese, Marketing Department, Leeds University Business School, University of Leeds, Maurice Keyworth Bldg., Moorland Rd., Leeds, LS2 9JT, UK.
Email: s.pugliese@leeds.ac.uk

Abstract

Literature reviews are crucial for attaining a full understanding of the key topics and latest trends in research and instrumental in identifying important research gaps. Unfortunately, conducting literature reviews can be time-consuming, and the outcomes are frequently subjective. Hence, to address such limitations, we detail an alternative, recent approach to conducting literature reviews. In this research, we outline the steps involved in conducting a literature review via natural language processing. Specifically, we illustrate how to (1) select relevant papers using term frequency-inverse document frequency and (2) perform topic modeling analysis through latent Dirichlet allocation to identify key research topics. This study and the associated ready-to-use Python code provide researchers, including those in consumer behavior, with detailed guidance on how to use natural language processing in their literature reviews.

KEYWORDS

literature reviews, marketing methods, marketing research, natural language processing, topic modeling

1 | INTRODUCTION

Literature reviews are crucial for attaining a full understanding of the key topics and latest trends in research and are instrumental in identifying research gaps. Unfortunately, traditional methods of conducting literature reviews are time-consuming and often rely on prior understandings of relevant research, potentially resulting in subjective interpretations.

Against this backdrop, we perceive an opportunity to detail an alternative, recent approach to conducting literature reviews to address these limitations. In this research, we thus outline the steps involved in conducting a literature review using natural language processing. Natural language processing is "a computer-assisted analytical technique aimed at automatically analyzing and comprehending human language (Manning & Schütze, 1999) that allows scholars to easily extract beneficial insights contained in textual datasets while avoiding burdensome computational work (Collobert et al., 2011; Green, 2012)" (Kang et al., 2020, p. 139). Natural language processing enables the efficient and objective identification of key research topics, reducing the tedious work and subjectivity associated with manual coding. In the following sections, we illustrate how to (1) select relevant papers using term frequency-inverse document frequency and (2) conduct topic modeling analysis via latent Dirichlet allocation to identify key research topics in the literature.

Accordingly, we detail the steps involved in this process and provide an illustration based on the field of Culture of Innovation,

which we consider familiar to scholars across various disciplines. Specifically, we apply latent Dirichlet allocation to the full text of 254 papers, automatically selected through term frequency-inverse document frequency, published in 1996–2019 across 15 well-renowned journals.

This study contributes to the marketing literature in two main ways. First, it responds to calls for "utilizing different approaches to reflect on extant knowledge" (Noble et al., 2021, p. 1) by illustrating the approach in conducting a literature review using natural language processing. Second, it offers a step-by-step tutorial detailing how term frequency-inverse document frequency and latent Dirichlet allocation can be employed in literature reviews, with clear benefits for efficiency and objectivity.

We also provide the associated Python code, which can be easily executed with minimal manual input, enabling researchers to implement this approach in conducting their own literature reviews on any subject of interest. We expect this research and the accompanying code to be of particular interest to marketing researchers, including those in consumer behavior, whose work is profoundly influenced by research in adjacent fields, such as psychology, sociology, economics, political science, or law (Clark et al., 2014), by making sparse research *corpora* more easily accessible and comprehensible. Notably, while the code is primarily designed for literature reviews, it (or parts of it) can also be effectively utilized by marketing researchers and practitioners in topic modeling of other textual data. The code is particularly suited to long, structured text, such as corporate or analyst reports, but it can also be adapted to user-generated content, such as customer reviews or social media posts.

2 | METHODOLOGY: OVERVIEW

Our approach builds on well-established natural language processing tools to assist marketing researchers in three tasks: (1) identifying relevant papers, (2) clustering similar papers, and (3) elucidating the topics that define each cluster.

2.1 | Identifying relevant papers

When approaching a literature review, researchers often face the daunting task of sifting through numerous potentially relevant papers to isolate the most relevant ones. To streamline this process without compromising reliability, our approach relies on a widely accepted method of understanding the content of a text, that is, the occurrence of relevant terms (Pennebaker et al., 2001; Rust et al., 2021). Specifically, we propose that the occurrence of terms related to a concept in a paper reflects the concept's significance within the paper. However, we note that when searching for papers on a specific concept, such as "innovation," all the papers retrieved online will likely contain some occurrences of the term "innovation." Consequently, it is necessary to account for the focal term's occurrence in the entire corpus of papers to determine which papers more prominently discuss this concept.

To address this issue, we employ term frequency-inverse document frequency (Spärck Jones, 1972) to automatically identify relevant papers. Term frequency-inverse document frequency is an information retrieval statistic that measures the typicality of a term in a given document relative to a corpus of documents. For each document in a corpus of n documents, the term frequency-inverse document frequency score of term t, TF-IDF(t), can be computed as follows:

$$\mathrm{TF}(t) = \text{number of times } t \text{ occurs in the document}, \tag{1}$$

$$\mathrm{IDF}(t) = \log\left(\frac{1+n}{1+\mathrm{DF}(t)}\right) + 1, \tag{2}$$

$$\text{TF-IDF}(t) = \mathrm{TF}(t) \times \mathrm{IDF}(t), \tag{3}$$

where t is a term, n is the number of documents in the corpus, and DF(t) is the number of documents among the n documents that contain at least one instance of t (Pedregosa et al., 2011). As TF-IDF(t) results from multiplying TF(t) with IDF(t), a higher TF-IDF(t) score can be obtained only if one of the two components, TF(t) or IDF(t), increases. The TF(t) component increases when t is frequent in the focal document, while the IDF(t) component increases when t is infrequent across other documents. Hence, the more t is frequent in the focal document and infrequent across the corpus, that is, typical of the focal document, the higher the TF-IDF(t) score will be for the focal document. In sum, we expect papers relevant to a certain topic to have higher term frequency-inverse document frequency scores for terms related to that topic.

However, it is also important to consider how differently a computer processes terms compared to humans. While humans perceive words such as "innovation" and "innovations" to represent the same concept, computers treat them as distinct. Therefore, to effectively utilize term frequency-inverse document frequency, it is imperative to preprocess papers' text in a manner that enables the computer to recognize that slightly varied terms refer to the same underlying concept. To achieve this, a more comprehensive approach is needed beyond minimal text cleaning, which typically involves removing punctuation, numbers, and figures, converting uppercase instances to lowercase, and eliminating stop-words (common words with little semantic value such as prepositions, conjunctions, and articles). In addition to these steps, one needs to apply lemmatization, stemming, and collocation.

Lemmatization transforms terms to their dictionary form by removing endings such as "-ing", "-ed", or "-s". For instance, "innovations" is reduced to "innovation" through lemmatization.

Stemming reduces terms to their roots. For example, "innovation" becomes "innov" after stemming. These techniques enable the grouping of various forms of the concept "innovation" ("innovate," "innovative," "innovativeness" are all reduced to "innov"), facilitating the evaluation of how frequently the concept appears in a text.

Collocation is used when two terms have a different meaning when they are adjacent compared to their meaning in isolation. For instance, the terms "social" and "media" have a different meaning when they are adjacent, as in "social media," compared to when they are used separately. Collocating these terms implies joining them, typically with "_". This process allows the computer to treat cases such as "social_media" or "new_product" as unique terms, different from their components.

Overall, these three data-cleaning functions modify term frequencies to make them more representative of the actual number of times a certain concept is mentioned.

2.2 | Clustering similar papers and eliciting topics

After identifying relevant papers, researchers must cluster similar papers and elicit topics within each cluster. To automate this stage, we use latent Dirichlet allocation, a well-established topic modeling technique.

Topic modeling is a text-mining approach that reveals abstract topics across a corpus of documents. Topic modeling thus allows the identification of hidden topics across a large, unstructured corpus of documents by clustering terms with similar meanings (Griffiths & Steyvers, 2004; Wang et al., 2012) and provides a list of documents that belong to each topic.

Latent Dirichlet allocation (Blei et al., 2003) has gained traction in marketing research (e.g., Büschken & Allenby, 2016; Tirunillai & Tellis, 2014; Zhong & Schweidel, 2020), particularly for topic modeling of reviews or other user-generated content (Tirunillai & Tellis, 2014) and marketing literature abstracts (e.g., Cano-Marin et al., 2023; Schmitt et al., 2022; Wang et al., 2015).

Formally, latent Dirichlet allocation represents a corpus of documents as a random mixture of latent topics. For a corpus of M documents composed of N terms across K topics, the joint distribution of terms and topics in each document can be described as follows:

$$p(w, z, \theta, \phi, \alpha, \beta) = \prod_{i=1}^{K} p(\phi_i|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \prod_{t=1}^{N} p(z_{j,t}|\theta_j) p(w_{j,t}|\phi_{z_{j,t}}), \quad (4)$$

where w and z are the vectors of all terms and topics; θ is a vector of prior topic probabilities that follows a Dirichlet distribution, ϕ is a vector of term probabilities for a given topic that follows a Poisson distribution, and α and β are the hyperparameters on θ and ϕ, respectively.

Latent Dirichlet allocation offers several advantages over alternative techniques. First, since latent Dirichlet allocation does not rely on assumptions about text structure or the syntactical and grammatical properties of language, it is suitable for extracting latent topics from papers (Tirunillai & Tellis, 2014). It also avoids assumptions about underlying term distributions and relationships among terms (Tirunillai & Tellis, 2014). Second, it employs an unsupervised Bayesian learning algorithm, enabling a clean slate, assumption-free approach to literature reviews (Tirunillai & Tellis, 2014). Finally, it efficiently handles a large number of documents (Tirunillai & Tellis, 2014). Table 1 provides a summary of the natural language processing terms introduced in this research.

2.3 | Comparison to alternative literature review methods

The most commonly used approaches in examining and synthesizing academic literature include bibliographic analyses, meta-analyses, and systematic literature reviews. While all these approaches share

TABLE 1 Definitions.

Term	Definition
Term frequency-inverse document frequency	Information retrieval statistic that captures how typical a term is of a certain document given a certain corpus of documents (Spärck Jones, 1972)
Lemmatization	Reducing each term to its dictionary form
Stemming	Reducing each term to its stem
Collocation	Joining terms that have a different meaning when they are adjacent compared to the meaning they would have if in isolation
Sensitivity	Ability to correctly classify relevant papers among the total number of relevant papers
Specificity	Ability to correctly classify irrelevant papers among the total number of irrelevant papers
Accuracy	Ability to correctly classify relevant and irrelevant papers among the total number of papers
Topic modeling	Text-mining approach that uncovers abstract topics from a corpus of documents
Latent Dirichlet allocation	Topic model that represents a corpus of documents as a random mixture of latent topics (Blei et al., 2003)
Term distinctiveness, saliency, and relevance	How informative a term is for determining a certain topic compared to any other randomly selected term (measured variously)

the goal of extracting insights and understanding published studies, they differ in their methods and scope.

Bibliographic analyses focus on bibliographic data, such as publication titles, authors, keywords, and abstracts, to identify trends, patterns, and relationships within a field (Donthu et al., 2021). Bibliographic analyses are thus primarily aimed at identifying influential papers and authors to provide an overview of research that focuses on field structure and the impact of each paper.

Meta-analyses, conversely, synthesize the results of studies and statistically analyze their data collectively to draw conclusions based on aggregated findings (Paul & Barari, 2022); they provide a more precise estimation of true effect sizes and identify sources of variation across empirical studies.

Systematic literature reviews apply a structured approach to the identification, evaluation, and synthesis of extant research (Palmatier et al., 2018). Systematic literature reviews, therefore, rely on their authors' manual assessment of these papers to offer insights into the current state of knowledge and research gaps, as well as directions for future research.

Our approach combines the objectivity typical of bibliographic analyses and meta-analyses with the ability to consider the full content of papers typical of systematic literature reviews. Table 2 describes how our approach compares to these alternatives.

To conclude, given the ongoing debate concerning artificial intelligence (AI) tools, it is important to clarify the benefits of our method against them. AI tools serve different purposes; to our knowledge, none of them is specifically designed for systematic literature reviews. ChatGPT, currently the most common AI tool for text, focuses primarily on *natural language generation*, a subset of natural language processing that generates text suitable for a given context or task based on extensive training with other data. In contrast, term frequency-inverse document frequency and latent Dirichlet allocation represent *natural language understanding*, a different subset of natural language processing that focuses on extracting information from text. Furthermore, while ChatGPT can provide suggestions for how to conduct a literature review, such suggestions are based on commonly observed patterns within the literature reviews that have been included in ChatGPT's training data, that is, typically, traditional systematic literature reviews. Hence, it may not offer guidance on the specific natural language processing tools applicable to literature reviews. Finally, ChatGPT can generate generic Python codes for term frequency-inverse document frequency or latent Dirichlet allocation. These codes are helpful only to researchers who already understand why these methods are relevant and who explicitly request assistance in how to program them. Additionally, such researchers need to possess sufficient programming skills to integrate and adapt ChatGPT-generated code. In contrast, our approach does not require extensive familiarity with Python programming or prior knowledge of natural language processing tools.

3 | RESEARCH CONTEXT

In the following sections, we use the field of Culture of Innovation as an example to demonstrate the benefits of our approach.

The Culture of Innovation concept has garnered significant attention from both academics and practitioners (Tellis et al., 2009). The intrigue associated with it is evident in quotes such as "culture [of innovation] is a uniquely human product that develops slowly within firms, is tacit and not easily defined, and is not easily transported across firms" (Tellis et al., 2009, p. 7).

Extensive research has examined Culture of Innovation through the lens of corporate- (risk tolerance, willingness to cannibalize, etc.) and national-level (national values, R&D spending, etc.) variables to establish its importance for firms' market performance and financial value. This vast pool of research spans diverse disciplines including marketing, management, and international business. We consider this research field particularly suitable for illustrating the benefits of our approach due to its interdisciplinary nature and its broad appeal to researchers in various marketing areas. Based on the literature, we thus define a Culture of Innovation as a culture (corporate or national) that fosters relentless innovation, ensuring that the focal firm stays on the leading edge of innovation.

4 | METHODOLOGY: IMPLEMENTATION

This section outlines how we used term frequency-inverse document frequency and latent Dirichlet allocation to select relevant papers, cluster similar papers, and elicit key topics. Figure 1 illustrates the position of our approach with regard to the systematic literature review process (Palmatier et al., 2018). Figure 2 depicts an overview of our process.

TABLE 2 Comparison to alternative literature review methods.

Method	Input	Full paper's content	Objective	Efficient	Replicable	Goal
Bibliometric analysis	Authors, institutions, countries, and journals	No	Yes	Yes	Yes	Mapping field structure and impact
Meta-analysis	Results of empirical studies	No	Yes	Yes	Yes	Hypothesis testing
Systematic literature review	Full paper's content	Yes	No	No	No	Organizing knowledge
This method	Full paper's content	Yes	Yes	Yes	Almost	Organizing knowledge

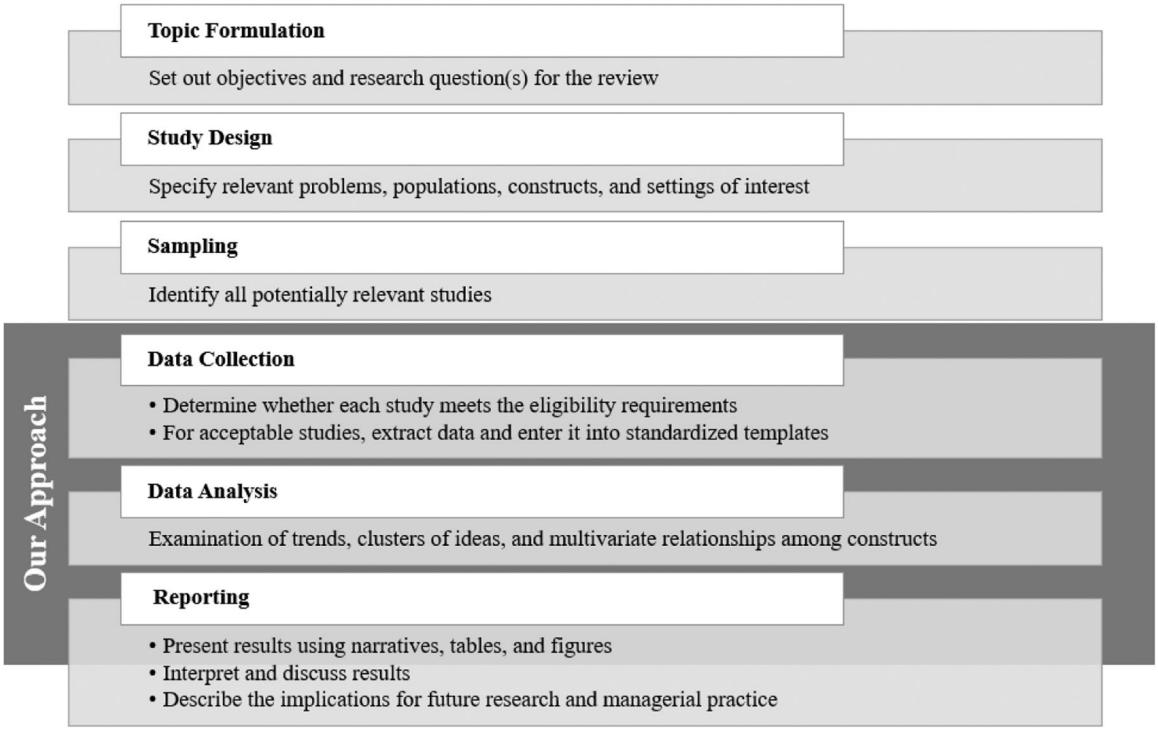

FIGURE 1 Comparison to systematic literature review process (adapted from Palmatier et al., 2018).

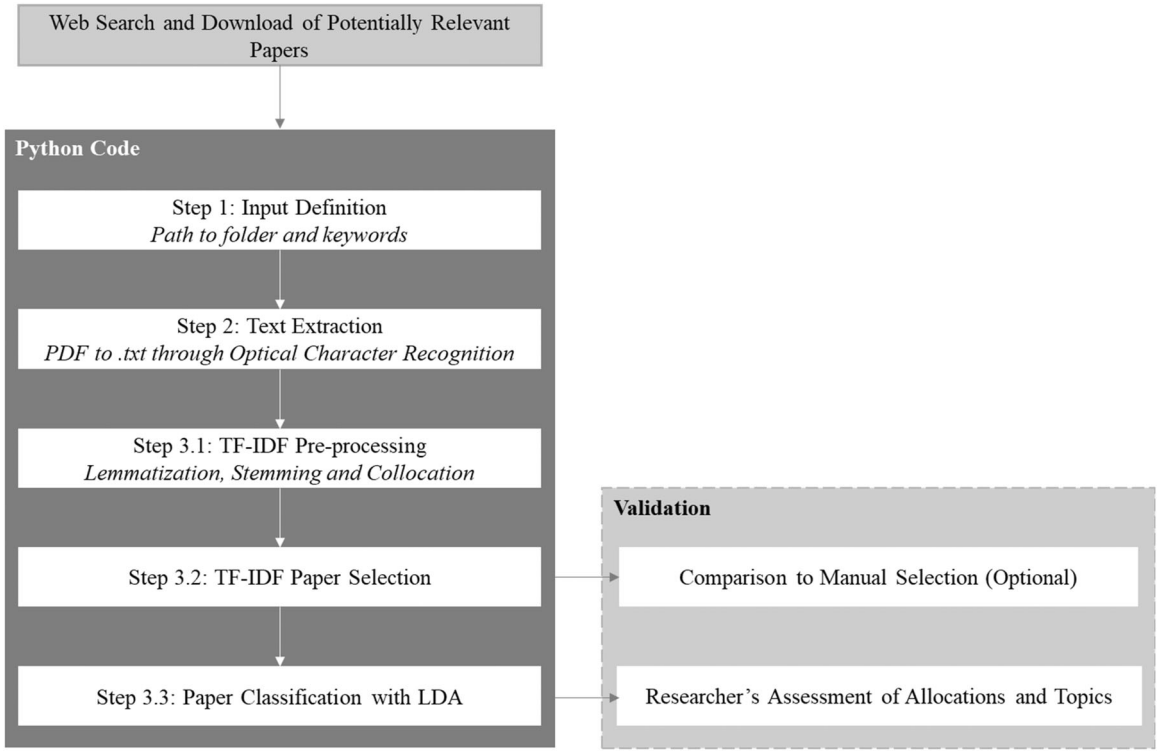

FIGURE 2 Process overview.

The accompanying Python code (at https://osf.io/tud9p/?view_only=7ca1f67541e4442cac3d87b04ab47be1) only requires the path to a folder containing previously downloaded papers in PDF format and a list of relevant search terms to be executed. This code is presented as a Jupyter notebook. To execute it, the user must first install the Jupyter Notebook app (at https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html) and then open the above file within the app. The user must then select the cell they want to run in the code and click the run button at the top of the screen. As the code is modularly structured, different parts can be run independently for the efficient utilization of computing resources.

4.1 | Step 0: Retrieval of potentially relevant papers

To retrieve papers for potential inclusion in the literature review, we used Web of Science. Specifically, we searched for papers whose titles, abstracts, or keywords contained the following two combinations of stem words: *innov** AND *cultur** or *new product**[1] AND *cultur**. The use of stem words, such as "*innov**," "*cultur**," and "*new product**" (instead of "innovation," "culture," and "new product," respectively), allowed us to capture variations such as "innovativeness," "cultural," and "new products" to encompass the entirety of relevant marketing literature on the topic. Following Rubera and Kirca (2012), we queried 15 management, marketing, and international business journals that have extensively covered the focal topic.[2] The initial search yielded 568 papers. To identify recent and impactful papers, we excluded any papers published before 1996 or those that had received, on average, no more than 5 yearly citations since their initial publication.[3] This process resulted in the selection of 207 papers (out of the initially retrieved 568).

Two authors then independently assessed the relevance of each paper and its alignment with the focal topic. While this manual evaluation was conducted to validate the term frequency-inverse document frequency measure, as explained below, it is not necessary for another user to implement the associated code. This process resulted in the selection of 83 relevant papers. Next, we examined the references in these 83 papers to identify other potentially relevant papers that our initial keyword search might have missed. This process yielded an additional 186 potentially relevant papers. Once again, two authors then independently evaluated the relevance of each paper and its alignment with the topic. This process resulted in the selection of 121 additional papers.

Hence, this manual process eventually resulted in a corpus of 204 papers (83 papers from the initial Web of Science search and 121 papers from the reference list search; a flowchart of our manual approach to paper retrieval and selection is provided in Figure 3a).

It is important to reiterate that an evaluation of the relevance of each paper is not needed to utilize the code. We engaged in this process to create a manual benchmark for evaluating the results of the term frequency-inverse document frequency analysis.

Notably, the code we provide does not search or download papers from the internet, as scraping is generally prohibited on platforms such as Web of Science. Instead, the code assumes that users have already downloaded papers and saved them as PDFs in a designated folder.

4.2 | Step 1: Input definition (Code: Block 1)

In this stage, we applied the code to the set of 393 potentially relevant papers (207 papers from the initial Web of Science search and 186 papers from the reference list search). A flowchart outlining our automatic paper selection approach is presented in Figure 3b.

To do so, we placed all the PDFs in a designated folder and then provided the path to that folder (Block 1, line 10) and the terms that were used to search for them (Block 1, line 12), as input in the code. Users should input the path to their folder and their list of search terms in lines 10 and 12, Block 1, respectively.[4]

The first part of the block prints the stems of the provided search terms. Stemming the search terms helps reduce variance in the terms used to identify the concept of interest. These stems are later utilized to distinguish relevant from irrelevant papers.

Before proceeding with the analysis, it is crucial for the user to review this list of stems and determine which stems or combinations thereof best capture their concept of interest. Hence, it is necessary to manually input the final list of relevant stems in the first line of the second code snippet in Block 1.

4.3 | Step 2: Text extraction (Code: Block 2)

Upon obtaining the list of files for analysis via the user-provided path, the code converts each paper to images using the Pdf2Image module for Python. Then, it extracts text from these images using Python Tesseract, an optical character recognition tool. The code treats PDFs as images because older PDFs are often scans, making it challenging to extract text from them with a conventional PDF reading tool. In this stage, the code separates the main content of papers from their reference lists, as including reference lists might inflate word frequencies and bias analyses.

[1]Firms innovate either through product or process innovation. Consistent with prior research (Cillo et al., 2018), we focus on product innovation in defining our keywords for paper retrieval and selection.

[2]Academy of Management Journal, Industrial Marketing Management, International Journal of Research in Marketing, Journal of Business Research, Journal of International Business Studies, Journal of Management, Journal of Management Studies, Journal of Marketing, Journal of Marketing Research, Journal of Product Innovation Management, Journal of the Academy of Marketing Science, Management Science, Marketing Science, Organization Science, and Strategic Management Journal.

[3]We also excluded editorial materials.

[4]To enable line numbers while working with a Jupyter Notebook, click on "View" in the top left part of the screen and then click on "Toggle Line Numbers."

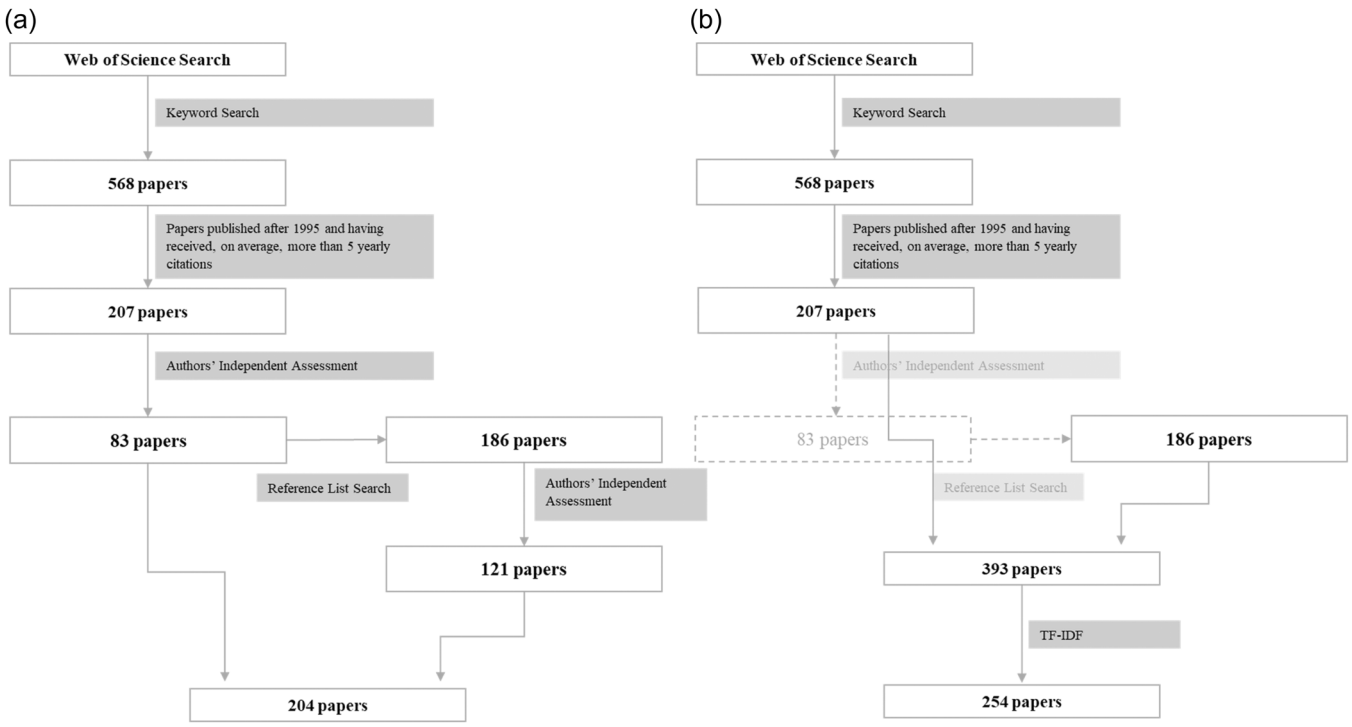

FIGURE 3 Alternative approaches to paper selection: (a) manual versus (b) automated.

4.4 | Step 3.1: Term frequency–inverse document frequency preprocessing (Code: Block 3.1)

To perform textual analysis, the raw text extracted from PDFs must be cleaned, that is, preprocessed. As outlined above, the code preprocesses text by removing punctuation, numbers, figures, uppercases (replaced with lowercases), and stop-words. Then, it reduces each word to its lemma and, subsequently, to its stem. Finally, the code identifies repeated adjacent terms and combines their tokens using "_", allowing for collocation. For instance, "new product" becomes "new_product" during this process.

4.5 | Step 3.2: Term frequency–inverse document frequency paper selection (Code: Block 3.2)

After preprocessing, the code employs term frequency–inverse document frequency to select relevant papers. In our case, term frequency–inverse document frequency was applied to the corpus of 393 papers deemed potentially relevant (Figure 3b). To isolate relevant papers, the code utilizes the term frequency–inverse document frequency scores for the search terms, for example, the stems "innov," "cultur," and "new_product," for each of the potentially relevant papers.

We first focused on each search term in isolation and then labeled papers as relevant if their term frequency–inverse document frequency score for a specific term was above or equal to the mean score in our corpus. For example, Hurley and Hult's (1998) term frequency–inverse document frequency score for "innov" is 0.041. Given that this value was above the mean score for "innov" in our corpus (0.021), Hurley and Hult (1998) was automatically labeled as relevant ("Y," as opposed to "N", in the *tfidf-results* file stored in the Results folder) when using "innov" as the only search term. We alternatively considered search terms in pairs and then labeled papers as relevant if their term frequency–inverse document frequency score was above or equal to the mean score for at least one of them.

4.5.1 | Term frequency–inverse document frequency results evaluation

To assess the effectiveness of this approach, we compared the list of relevant papers suggested by the different term frequency–inverse document frequency criteria to our manual classification. That is, considering our manual classification the correct classification and comparing the outcomes of term frequency–inverse document frequency against it, we assessed the predictive ability of the six alternative term frequency–inverse document frequency criteria detailed above. Based on the match/mismatch of the outcomes of the manual versus automated classification, we computed, for each term frequency–inverse document frequency criterion, the number of

1. True positives, papers classified as relevant by both our manual classification and the automatic term frequency–inverse document frequency classification;

2. True negatives, papers classified as irrelevant by both our manual classification and the automatic term frequency–inverse document frequency classification;
3. False positives, papers classified as irrelevant by our manual classification but relevant by the automatic term frequency–inverse document frequency classification; and
4. False negatives, papers classified as relevant by our manual classification but irrelevant by the automatic term frequency–inverse document frequency classification.

Accordingly, we then calculated the sensitivity, specificity, and accuracy of the six alternative term frequency-inverse document frequency criteria. Specifically, we define term frequency–inverse document frequency sensitivity as its ability to correctly classify relevant papers among the total number of relevant papers (Vassallo et al., 2023) and compute it as follows:

$$\text{Sensitivity} = \frac{\text{Number of True Positives}}{(\text{Number of True Positives} + \text{Number of False Negatives})}. \quad (5)$$

We define term frequency-inverse document frequency specificity as its ability to correctly classify irrelevant papers among the total number of irrelevant papers (Vassallo et al., 2023) and compute it as follows:

$$\text{Specificity} = \frac{\text{Number of True Negatives}}{(\text{Number of True Negatives} + \text{Number of False Positives})}. \quad (6)$$

Finally, we define term frequency-inverse document frequency accuracy as its ability to correctly classify relevant and irrelevant papers among the total number of papers and compute it as follows:

$$\text{Accuracy} = \frac{(\text{Number of True Positives} + \text{Number of True Negatives})}{\text{Total Number}}. \quad (7)$$

In Table 3, we compare the sensitivity, specificity, and accuracy of the six alternative term frequency-inverse document frequency criteria, considering our manual classification the correct one (baseline).

Here, a recurrence equal to or above the mean for at least one notion of innovation ("innov" OR "new_product") yielded the highest accuracy (sensitivity = 87%, specificity = 59%, accuracy = 74%) (Column 6, Table 3). Using this criterion resulted in the selection of 254 papers (Figure 3b) for inclusion in the review. Table 4 details the confusion matrix obtained using the best-performing term frequency-inverse document frequency criterion. Table 4 thus cross-tabulates the manual classification of the 393 potentially relevant papers with the term frequency-inverse document frequency classification and provides the numbers of true positives (177), true negatives (112), false positives (77), and false negatives (27). Notably, among the 77 false positives, papers classified as irrelevant according to the manual classification and relevant according to term frequency-inverse document frequency, 72 were manually excluded for reasons other than not concerning Culture of Innovation, for example, for being case studies with a narrow scope. Furthermore, among the 27 false negatives, papers classified as relevant according to the manual classification but not relevant according to term frequency-inverse document frequency, 14 focus on relevant constructs but employ different terms such as "NPD" instead of "new product development."

Given the high level of accuracy (74%) achieved by the best-performing term frequency-inverse document frequency criterion and our objective of producing an automated literature review, the

TABLE 4 Confusion matrix.

		Automated classification	
		Automatic "relevant"	**Automatic "irrelevant"**
Manual Classification	**Manual "relevant"**	True positives 177 (45%)	False negatives 27 (6.9%)
	Manual "irrelevant"	False positives 77 (19.6%)	True negatives 112 (28.5%)

Note: Sensitivity = 177/(177 + 27)% = 86.765%. Specificity = 112/(112 + 77)% = 59.259%. Accuracy = (177 + 112)/393% = 73.537%.

TABLE 3 Comparison across alternative term frequency-inverse document frequency criteria.

Criterion	1 "innov"[a]	2 "cultur"[a]	3 "new_product"[a]	4 "innov" OR "cultur"[b]	5 "new_product" OR "cultur"[b]	6 "innov" OR "new_product"[b]
Sensitivity	0.716	0.534	0.461	0.858	0.784	**0.868**
Specificity	0.677	0.471	0.836	0.307	0.349	**0.593**
Accuracy	0.697	0.504	0.641	0.593	0.575	**0.735**

Note: Numbers in bold represent the statistics resulting from the selected term frequency-inverse document frequency criterion.

[a]A paper is classified as relevant if its term frequency-inverse document frequency score for the search term (e.g., "innov") is equal to or above the mean in the corpus.

[b]A paper is classified as relevant if its term frequency-inverse document frequency score for at least one of the search terms (e.g., "innov" OR "cultur") is equal to or above the mean in the corpus.

next stages applied latent Dirichlet allocation to the set of 254 automatically selected papers. Notably, we also applied latent Dirichlet allocation to the set of 204 manually selected papers. These results, available upon request, are generally consistent with those obtained using the set of automatically selected papers. In Supporting Information: Appendix A, we report some potential alternative approaches to automated paper selection and explain why we deem term frequency-inverse document frequency the most suitable one.

We note here that the code we provide does not include lines for computing accuracy since we do not expect other users to have a list of manually annotated files to compare these results. However, the *tfidf-results* Excel file, in the Results folder, contains the list of files with the relevance label assigned automatically. If users aim to validate the use of term frequency-inverse document frequency for paper selection in their sample, they can thus add their manual annotations to the file and then easily calculate accuracy using Excel.

4.6 | Step 3.3: Paper classification with latent Dirichlet allocation (Code: Block 3.3)

In this stage, the code focuses only on the papers that were deemed relevant by the term frequency-inverse document frequency criterion selected by the user (manual input required in Block 3.3, line 20 of the first code snippet).

4.6.1 | Preprocessing

The code starts with the raw .txt files and then preprocesses each paper by removing sources of noise and lemmatizing terms. In this stage, the code does not apply stemming due to the different purposes of latent Dirichlet allocation and term frequency-inverse document frequency. To run latent Dirichlet allocation effectively, it is preferable to tag each word with its part of speech (pronoun, adverb, conjunction, etc.) and retain only terms classified as nouns, verbs, adjectives, or adverbs. Stemming terms in text preprocessing for latent Dirichlet allocation renders tagging impossible (as the stem "innov" could correspond to the verb "innovate," the noun "innovation," or the adjective "innovative"). Therefore, we halted preprocessing amid lemmatization.

4.6.2 | Latent Dirichlet allocation estimation

Latent Dirichlet allocation requires the number of topics as an input. However, the optimal number of topics is generally unknown before running the analysis. For this reason, after preprocessing the text (Block 3.3, up to line 45 of the second code snippet), the code runs several latent Dirichlet allocations, setting alternative numbers of topics (between 2 and 20), and automatically selects the best-performing number of topics to represent the corpus (Block 3.3, starting from line 46 of the second code snippet). In machine learning, it is common practice to evaluate the performance of alternative models by dividing the data into two sets, a training set and a testing set. The training set is used to fit the model, and the obtained parameters are then used to predict the output of the testing set (Gareth et al., 2013). To assess the performance of alternative numbers of topics, the code focuses on the log-likelihood and perplexity (normalized log-likelihood) thereof regarding the testing set using 10-fold cross-validation. This involves dividing the sample into 10 random groups and iteratively using each group as the testing set and the remaining nine groups as the training set. In this context, log-likelihood and perplexity essentially capture how probable unseen papers (testing set) are based on what the model has learned from the training set.

4.6.3 | Latent Dirichlet allocation results evaluation

In our case, the results (automatically stored in the *lda.html* file in the Results folder) indicated that the best number of topics according to our corpus is two. Hence, each paper was allocated by latent Dirichlet allocation to one of these two topics. Figure 4 lists these identified topics. Figure 4a represents each topic as a bubble. The larger the bubble is, the higher the number of papers. The further the bubbles are from each other, the more different the topics. Figure 4a shows that the two topics are roughly equivalent in size (132 vs. 122, respectively) and clearly distinct.

Latent Dirichlet allocation generates a list of terms (Figure 4b) that represent the whole corpus. By default, it returns the 30 most *salient* terms, ranked in descending order of saliency. For example, the five most salient terms in our corpus are, in descending order, "orientation," "market," "customer," "performance," and "innovation." The gray bars (originally blue in the output file generated by the code we provide) in Figure 4b, represent overall term frequencies, utilized to compute term saliencies, as we explain below.

Figure 4b shows the most salient terms in our whole corpus. The saliency of terms is determined based on their (a) probability of occurrence in the corpus and (b) distinctiveness, as follows:

$$\text{saliency}(w) = p(w) \times \text{distinctiveness}(w), \tag{8}$$

where $p(w)$ is the observed probability of term w in the corpus (the frequency of term w over the total number of terms) and distinctiveness(w) is its distinctiveness. Following Chuang et al. (2012), each term can be considered more or less distinctive depending on its role in determining a certain topic compared to any other randomly selected term. Distinctiveness(w) is mathematically defined as the Kullback–Leibler divergence (Kullback & Leibler, 1951) between the likelihood that the observed term w was generated by topic t, $p(t|w)$, and the likelihood that any randomly selected term w' was generated by topic t, $p(t)$, as follows:

$$\text{distinctiveness}(w) = \sum_t p(t|w) \log \frac{p(t|w)}{p(t)}. \tag{9}$$

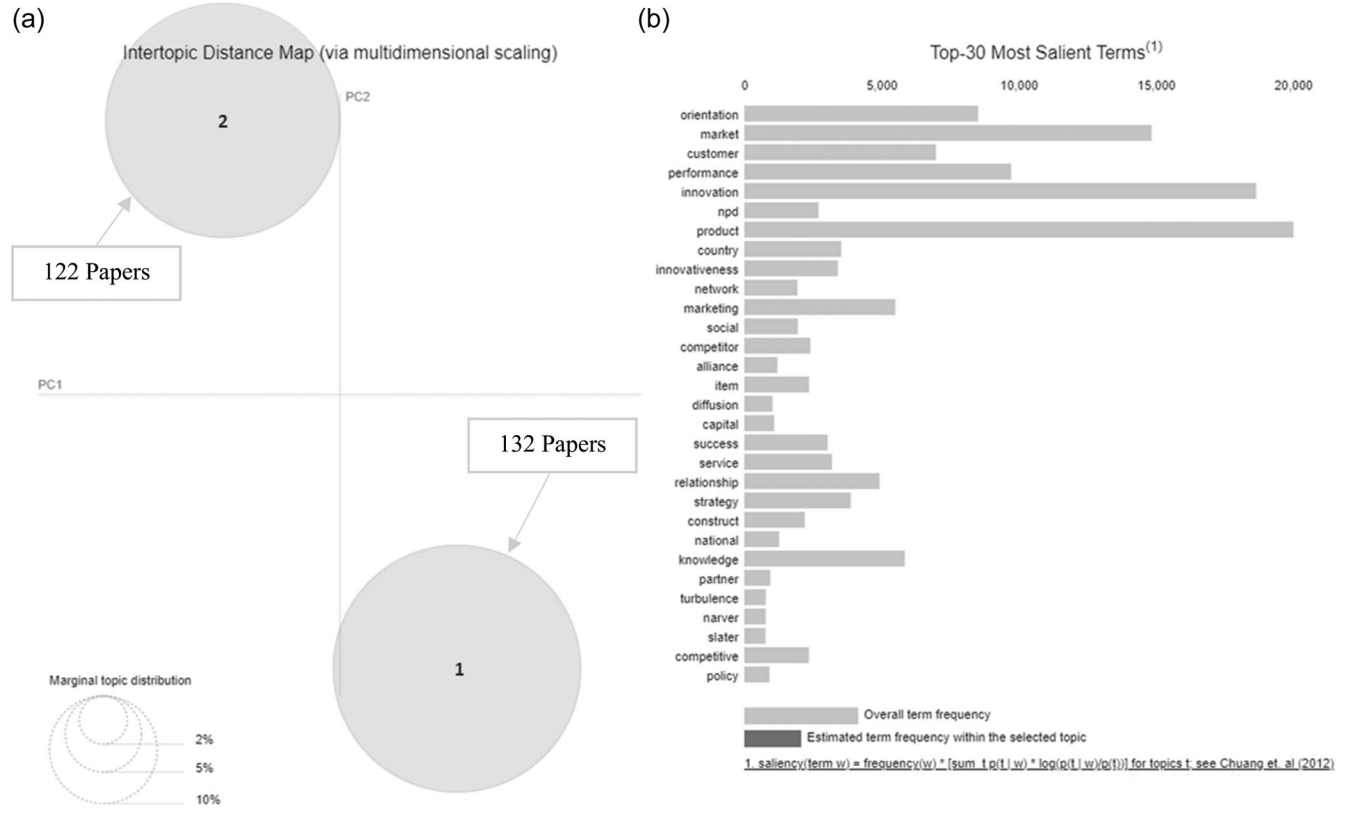

FIGURE 4 Topics and salient terms. (a) Represents each topic as a bubble. The larger the bubble is, the higher the number of papers allocated to that topic. The further the bubbles are from each other, the more different the topics they represent. (b) Represents the most salient terms within the corpus. The gray bars represent overall term frequencies.

Saliency, therefore, provides us with terms that are *potentially* relevant for describing the research topics covered in the corpus. However, salient terms are more informative about the general topic of Culture of Innovation than the distinct topics within it (Sievert & Shirley, 2014). To understand the content of the individual topics, latent Dirichlet allocation returns the most frequent terms per topic. However, as all our papers are within the same field, target similar audiences and are written by authors with similar backgrounds, this may increase the probability that multiple topics have frequent terms in common. Since our aim is to use latent Dirichlet allocation to distinguish different topics, we use *Relevance* (Sievert & Shirley, 2014) to identify the most representative terms in each of these topics.

Relevance (Sievert & Shirley, 2014) is computed as follows:

$$\text{Relevance}(w, t|\lambda) = \lambda \log(\phi_{tw}) + (1 - \lambda)\log\left(\frac{\phi_{tw}}{p_w}\right), \quad (10)$$

where ϕ_{tw} is the probability of term $w \in \{1 \ldots, V\}$ for topic $t \in \{1 \ldots, T\}$, p_w is the marginal probability of term w in the corpus, and λ is a weight parameter ranging between 0 and 1 given to the probability of term w under topic t relative to its lift (the ratio of the probability of term w within topic t over its marginal probability across the corpus). Sievert and Shirley (2014) find the optimal value of λ to be approximately 0.6; hence, we set λ to 0.6 when evaluating term relevance within a topic.

After identifying the different topics based on the relevant terms, we examined the allocation of papers therein. Latent Dirichlet allocation assigns to each paper the probability of belonging to each topic; then, it assigns each paper to the topic with the highest probability. To assess the reliability of the analysis, we reviewed the list of probabilities (automatically stored in the *dominant_topics* file in the Results folder).

Specifically, we filtered the list of papers by topic and probability and then manually examined whether the content of the papers that were assigned a high probability aligned with the topic description derived from the relevant terms. Similarly, we also assessed whether the allocation remained reliable for papers assigned with the lowest probabilities. We found the latent Dirichlet allocation results to be reliable, as the papers assigned with the lowest probabilities still exhibited consistency with their respective topics. Importantly, reviewing the complete list of files and their topic assignments is not necessary when implementing this approach. Notwithstanding this, our evaluation effectively validated the effectiveness of the method in allocating papers to different topics.

4.7 | Step 3.4: Identifying subtopics (Code: Block 3.4, *Optional*)

Finally, in this section, we provide the code for analyzing the papers assigned to each topic separately. Conducting separate latent

Dirichlet allocations on the papers assigned to each topic produces a more detailed picture of the literature, revealing distinct research streams within the previously identified topics. This process follows the same steps described above; the only difference is that the code iterates through different topics and analyses, at each iteration, only the subset of papers assigned to a specific topic. While this step is not strictly necessary, it can be informative by uncovering more nuanced differences in the literature. However, importantly, this step requires at least 10 papers per topic, as the code employs 10-fold cross-validation. It is infeasible to split a set of fewer than 10 papers into 10 parts. Users who seek to identify subtopics within topics containing fewer than 10 papers can adjust the "cv" (cross-validation) parameter in line 54, Block 3.4, to a lower number. Using a lower number of folds is not advisable, however, as it results in less training for the model.

5 | RESULTS

Figure 4a indicates that the two topics are roughly equivalent in size and clearly distinct. The five most salient terms in our corpus, in descending order of saliency, are "orientation," "market," "customer," "performance," and "innovation" (Figure 4b).

Figure 5 displays the 30 most *relevant* terms in Topic 1 (Figure 5a) and Topic 2 (Figure 5b), ranked in descending order. Each term is accompanied by its frequency within the relevant topic (black bar in the figure, red bar in the output file) and its overall frequency across the corpus (gray bar in the figure, blue bar in the output file).

Interpretation of the latent Dirichlet allocation results was performed by the authors based on the most relevant terms assigned to each topic, entailing consideration of the papers until a consensus was reached.

The 10 most relevant terms in the first topic, Topic 1 (132 papers), are "product," "market," "orientation," "performance," "firm," "new," "customer," "marketing," "study," and "relationship," based on which it was possible to label the topic as "Market Orientation and Innovation Performance," driven primarily by terms such as "market," "orientation," "customer," "marketing," and "relationship" (Figure 5a). The 10 most relevant terms in the second topic, Topic 2 (122 papers), are "innovation,""firm,""country,""knowledge,""research,""use,""new,""technology," "model," and "product," based on which it was possible to label the topic as "Cultural Advantages and Innovation Performance," driven primarily by terms such as "firm," "country," "knowledge," "research," and "technology" (Figure 5b).

We subsequently conducted another round of latent Dirichlet allocation to uncover subtopics within the main topics. This process yielded two subtopics for each main topic. Figures 6–8 depict the identified subtopics and their most relevant terms. Specifically, we identified two subtopics within Topic 1, "Market Orientation and Innovation Performance," that is, Subtopic 1.1, "Customer and Competitor Orientation" (83 papers), driven by terms such as "market," "orientation," "customer," and "competitor," and Subtopic 1.2, "Interfunctional Coordination and Cultural Cohesiveness" (49 papers), driven by terms such as "project," "npd," "team," "development," "process," "research," and "integration." Figure 6a shows these two Subtopics, 1.1 and 1.2, which appear to be clearly distinct. The most relevant terms for Subtopics 1.1 and 1.2 are shown in Figure 7a,b, respectively.

Furthermore, we identified two subtopics within Topic 2, "Cultural Advantages and Innovation Performance," that is, Subtopic

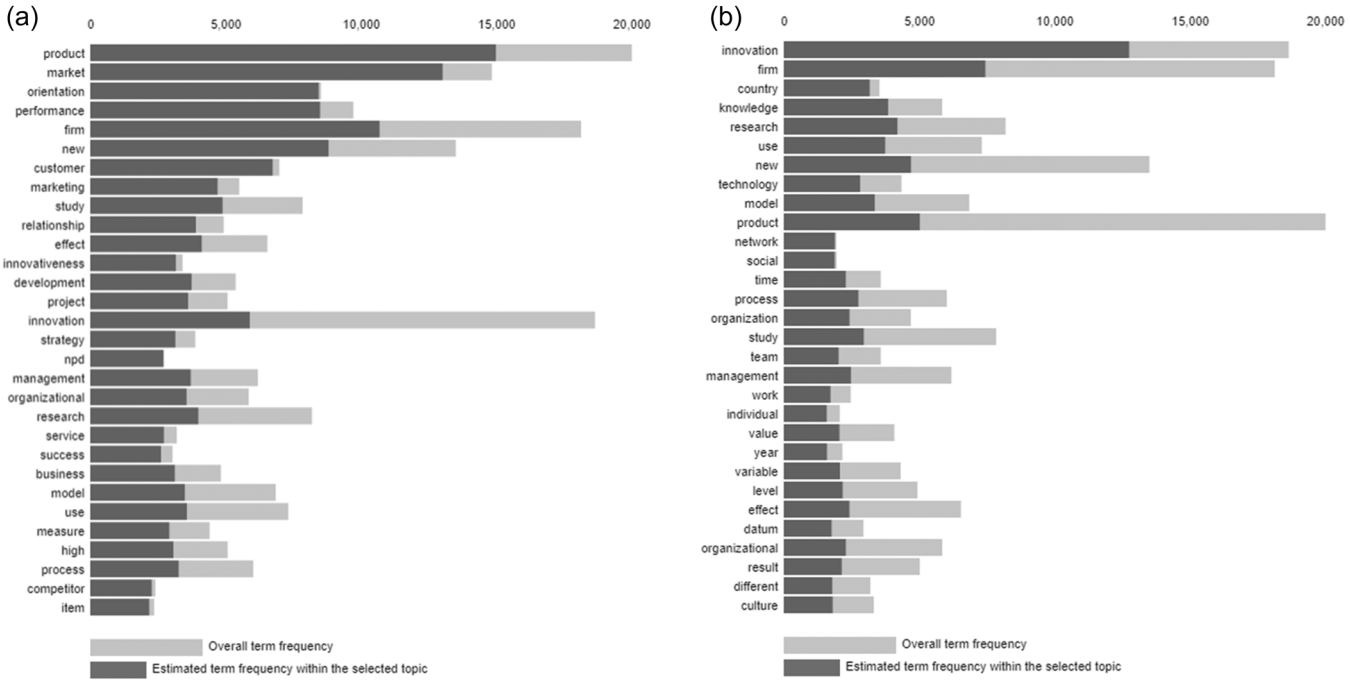

FIGURE 5 Relevant terms in (a) Topic 1 and (b) Topic 2.

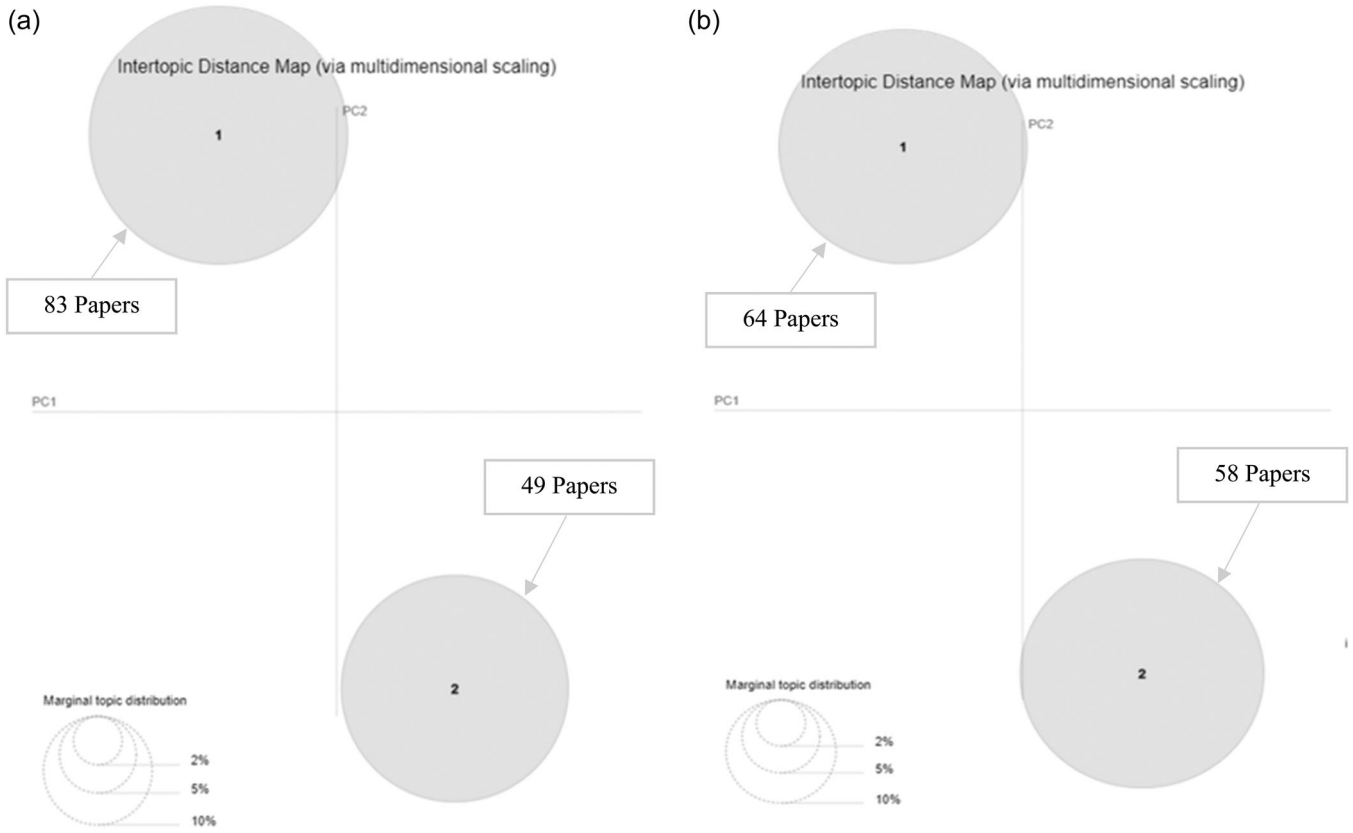

FIGURE 6 Subtopics within (a) Topic 1 and (b) Topic 2.

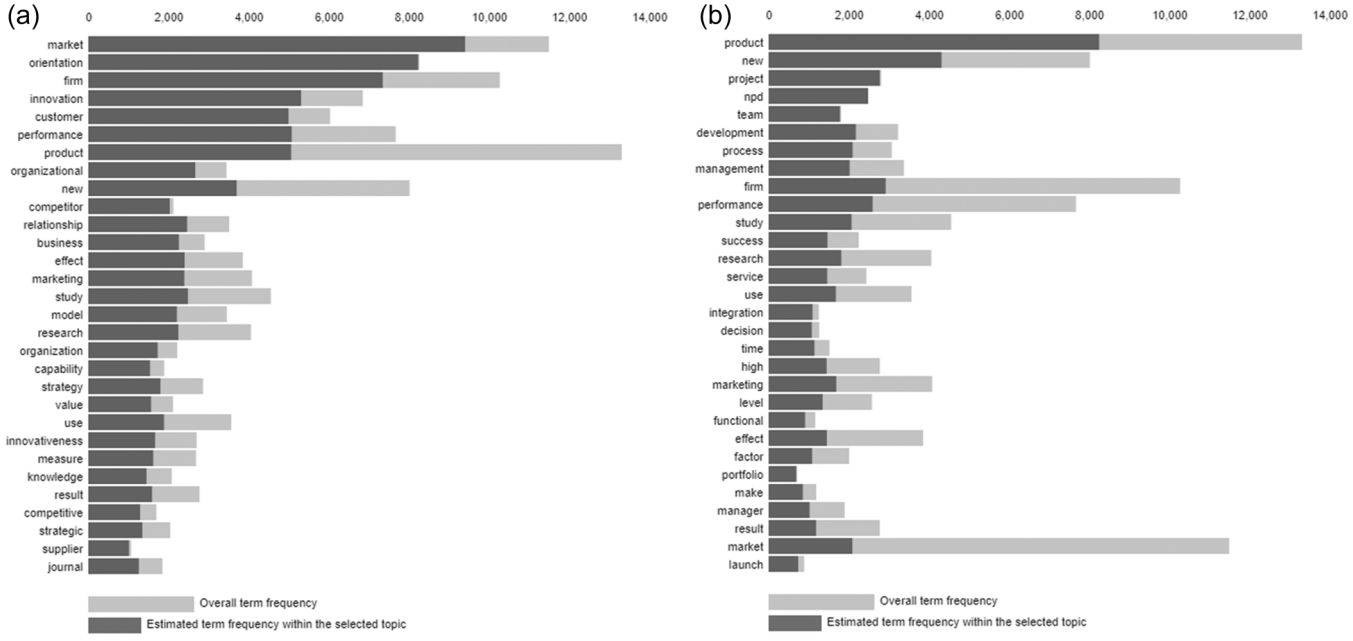

FIGURE 7 Relevant terms in (a) Subtopic 1.1 and (b) Subtopics 1.2.

2.1, "Corporate Culture Advantages" (64 papers), driven by terms such as "innovation," "knowledge," "organization," "capability," and "resource," and Subtopic 2.2, "National Culture Advantages" (58 papers), driven by terms such as "firm," "product," "innovation," "country," "national," and "diffusion." Figure 6b depicts Subtopics 2.1 and 2.2, which similarly appear to be clearly distinct. The most relevant terms for Subtopics 2.1 and 2.2 are shown in Figure 8a,b, respectively.

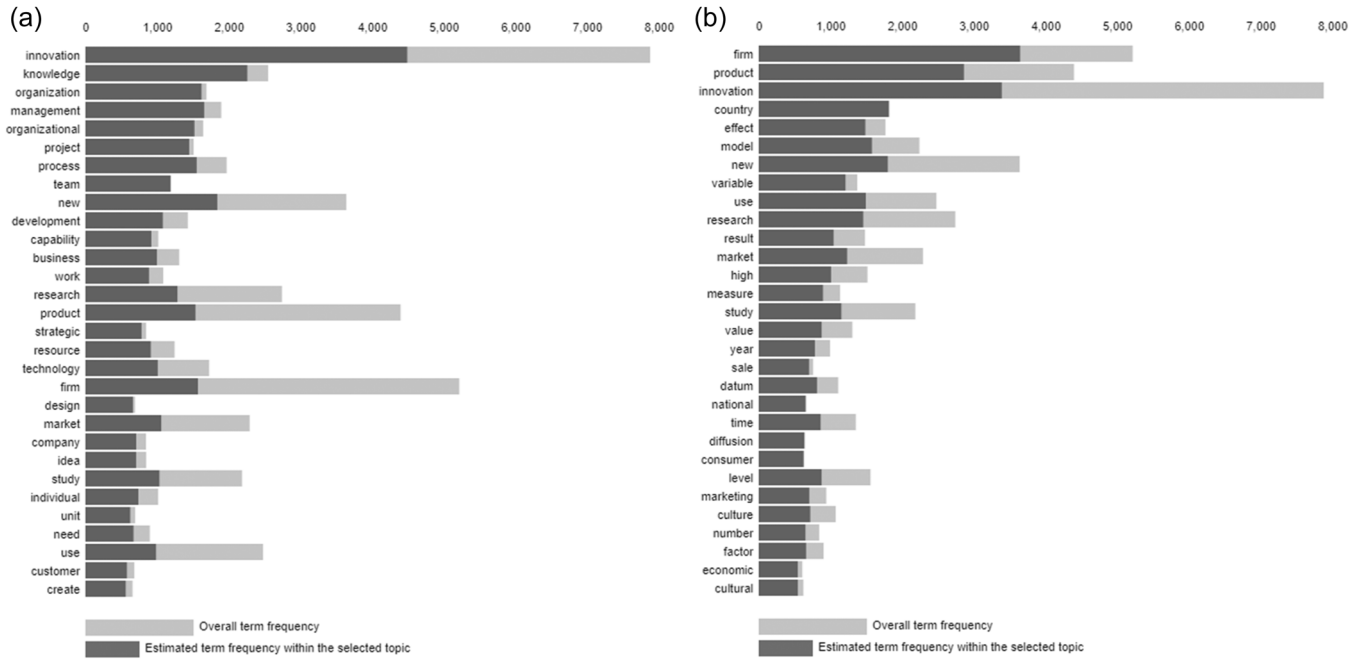

FIGURE 8 Relevant terms in (a) Subtopic 2.1 and (b) Subtopic 2.2.

(a)

	1	2	3	4	5	6
Topic	**1**	**1.1**	**1.2**	**2**	**2.1**	**2.2**
1	product	market	product	innovation	innovation	firm
2	market	orientation	new	firm	knowledge	product
3	orientation	firm	project	country	organization	innovation
4	performance	innovation	npd	knowledge	management	country
5	firm	customer	team	research	organizational	effect
6	new	performance	development	use	project	model
7	customer	product	process	new	process	new
8	marketing	organizational	management	technology	team	variable
9	study	new	firm	model	new	use
10	relationship	competitor	performance	product	development	research
Number of Papers	132	83	49	122	64	58

(b)

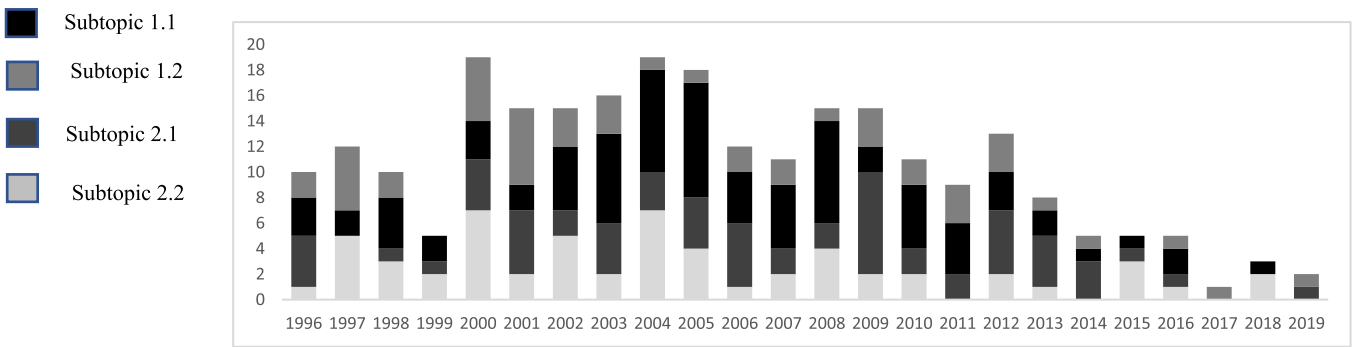

FIGURE 9 Latent Dirichlet allocation: Summary of results. (a) showcases, for each topic/subtopic, the 10 most relevant terms and the corresponding number of papers. (b) showcases research trends.

Figure 9a provides a summary of the results obtained from our latent Dirichlet allocation, showcasing, for each topic/subtopic, the 10 most relevant terms and the corresponding number of papers. We depict research trends over time in Figure 9b. In the interest of space, we do not provide the full list of papers allocated to each topic.

In Supporting Information: Appendix B, we offer a more detailed discussion of these topics and subtopics to highlight the relevance of our results to the Culture of Innovation literature.

6 | DISCUSSION

In this study, we have detailed the steps involved in conducting a literature review using natural language processing. Specifically, we showed how to (1) identify relevant papers using term frequency-inverse document frequency and (2) conduct topic modeling analysis based on latent Dirichlet allocation to identify research topics.

This study makes important contributions to the literature; it is the first, to our knowledge, to provide a step-by-step tutorial on how to conduct a literature review using term frequency-inverse document frequency and latent Dirichlet allocation in business and management scholarship. By utilizing term frequency-inverse document frequency, we were able to automatically select relevant papers; applying latent Dirichlet allocation to the full text of papers enabled us to identify topics and subtopics efficiently and objectively. We have also provided clear definitions and a ready-to-use Python code, enabling future researchers to replicate our approach when conducting their own literature reviews. Notably, our approach allows the processing of numerous papers with minimal effort before any time-costly manual work is to be done, a benefit thus far generally provided only by bibliometric analyses (Gupta et al., 2023; Khan et al., 2020). Importantly, there is practically no limitation on the number of papers the code is able to process. Although it takes approximately 6 hours with an average laptop to analyze approximately 400 papers, as we did, with most of this time spent in text extraction (4/5 hours), due to the modular structure of the code, this step can be skipped when rerunning the analysis on papers already converted to .txt format. Importantly, while we offer the code with a view toward making it possible for researchers to use it for their literature reviews, the code (or parts of it) could also be effectively used in topic modeling of firm- or user-generated content by marketing researchers in general and by consumer behavior researchers in particular.

7 | LIMITATIONS

While our approach enables the attainment of an effective understanding of a large corpus of papers in a few hours, it suffers from some limitations. First, although measures based on the occurrence of meaningful terms are already used in marketing (Rust et al., 2021), they may not perfectly predict the relevance of each paper (achieve 100% accuracy vs. manual classification). Notably, the modular structure of our code allows users to run other parts of the code independently if they prefer to rely on manual paper selection.

Second, latent Dirichlet allocation does not yield fully replicable results due to its Bayesian nature. This inherent variability may lead to slightly different outcomes during each iteration. However, in our experience, the variation across iterations was minimal, and the random component did not significantly impact the usefulness of the results.

Third, importantly, the purpose of topic models, such as latent Dirichlet allocation, is to provide an interpretation of the text that needs to be validated by the user. In other words, it is ultimately the researcher who determines which interpretation of the literature is most informative for them or their readers. The value of this approach lies in its ability to efficiently and objectively cluster papers based on the simultaneous consideration of hundreds of them, a task that would otherwise take days, if not weeks, to accomplish manually.

Despite these limitations, we hope this research can be of help to researchers in marketing academia and beyond.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ORCID

Serena Pugliese http://orcid.org/0009-0008-8309-639X
Verdiana Giannetti http://orcid.org/0000-0001-5703-789X

REFERENCES

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

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953–975.

Cano-Marin, E., Mora-Cantallops, M., & Sánchez-Alonso, S. (2023). Twitter as a predictive system: A systematic literature review. *Journal of Business Research*, *157*, 113561.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 74–77.

Cillo, P., Griffith, D. A., & Rubera, G. (2018). The new product portfolio innovativeness–stock returns relationship: The role of large individual investors' culture. *Journal of Marketing*, *82*(6), 49–70.

Clark, T., Key, T. M., Hodis, M., & Rajaratnam, D. (2014). The intellectual ecology of mainstream marketing research: An inquiry into the place of marketing in the family of business disciplines. *Journal of the Academy of Marketing Science*, *42*, 223–241.

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

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Green, G. M. (2012). *Pragmatics and natural language understanding* (2nd ed.). Psychology Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5228–5235.

Gupta, M., Givi, J., Dey, M., Kent Baker, H., & Das, G. (2023). A bibliometric analysis on gift giving. *Psychology & Marketing*, *40*(4), 629–642.

Hurley, R. F., & Hult, G. T. M. (1998). Innovation, market orientation, and organizational learning: An integration and empirical examination. *Journal of Marketing*, *62*(3), 42–54.

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139–172.

Khan, M. A., Ali, I., & Ashraf, R. (2020). A bibliometric review of the special issues of psychology & marketing: 1984–2020. *Psychology & Marketing*, *37*(9), 1144–1170.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Noble, C. H., Spanjol, J., & Kirca, A. H. (2021). Advancing broad and deep understanding in innovation management: Meta-analyses and literature reviews Accessed July 30, 2021. https://onlinelibrary.wiley.com/pb-assets/assets/15405885/JPIM%20CFP%20-%20Meta%20Analyses%20and%20Reviews-1623685669020.pdf

Palmatier, R. W., Houston, M. B., & Hulland, J. (2018). Review articles: Purpose, process, and structure. *Journal of the Academy of Marketing Science*, *46*, 1–5.

Paul, J., & Barari, M. (2022). Meta-analysis and traditional systematic literature reviews—What, why, when, where, and how. *Psychology & Marketing*, *39*(6), 1099–1115.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*.

Rubera, G., & Kirca, A. H. (2012). Firm innovativeness and its performance outcomes: A meta-analytic review and theoretical integration. *Journal of Marketing*, *76*(3), 130–147.

Rust, R. T., Rand, W., Huang, M. H., Stephen, A. T., Brooks, G., & Chabuk, T. (2021). Real-time brand reputation tracking using social media. *Journal of Marketing*, *85*(4), 21–43.

Schmitt, B., Brakus, J. J., & Biraglia, A. (2022). Consumption ideology. *Journal of Consumer Research*, *49*(1), 74–95.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 63–70.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.

Tellis, G. J., Prabhu, J. C., & Chandy, R. K. (2009). Radical innovation across nations: The preeminence of corporate culture. *Journal of Marketing*, *73*(1), 3–23.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463–479.

Vassallo, J. P., Banerjee, S., Zaman, H., & Prabhu, J. C. (2023). Design thinking and public sector innovation: The divergent effects of risk-taking, cognitive empathy and emotional empathy on individual performance. *Research Policy*, *52*(6), 104768.

Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. *arXiv preprint arXiv*, *1206*, 3298.

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

Zhong, N., & Schweidel, D. A. (2020). Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science*, *39*(4), 827–846.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Pugliese, S., Giannetti, V., & Banerjee, S. (2024). How to conduct efficient and objective literature reviews using natural language processing: A step-by-step guide for marketing researchers. *Psychology & Marketing*, *41*, 427–441. https://doi.org/10.1002/mar.21931