ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

ADVANCING INNOVATION THROUGH TEXTUAL ANALYTICS: A PRACTICAL AND THEORETICAL EXPLORATION OF TOPIC MODELING

Rodríguez Elena María

Department of Innovation and Data Science, ESADE Business School, Ramon Llull University, Barcelona, Spain

DOI: 10.5281/zenodo.19596370

Abstract

Innovation remains a critical driver of economic growth, organizational competitiveness, and societal progress, particularly within an increasingly data-driven digital environment. The exponential growth of unstructured textual data—originating from sources such as social media, customer feedback, academic literature, and patent databases—presents both a challenge and an opportunity for innovation management. Effectively extracting actionable insights from these diverse data streams has become essential for informed decision-making across the innovation lifecycle. This study explores the role of textual analytics, with a specific focus on topic modeling, as a robust methodological approach for uncovering latent knowledge embedded within large text corpora.

Topic modeling, a class of probabilistic and unsupervised machine learning techniques, enables the systematic identification of hidden thematic patterns in textual data. Although not a recent development, its relevance has significantly increased with advancements in natural language processing (NLP) and big data technologies. Positioned within the broader NLP framework, topic modeling provides a transparent, interpretable, and scalable solution for organizing and summarizing complex textual information. Unlike more recent artificial intelligence approaches such as large language models (LLMs), including BERT and GPT, which often face limitations related to generalization, computational cost, and potential inaccuracies, topic modeling offers a more controlled and domain-specific analytical framework.

This paper highlights both the theoretical foundations and practical applications of topic modeling in innovation contexts. It emphasizes how the technique can support strategic decision-making by revealing emerging trends, identifying knowledge gaps, and enhancing organizational learning processes. Furthermore, the study underscores the comparative advantages of topic modeling in terms of interpretability, reproducibility, and adaptability to specific research or business needs. By bridging theoretical insights with practical implementation, this work contributes to a deeper understanding of how textual analytics can be leveraged to drive innovation and sustain competitive advantage in the digital age.

Keywords: Textual Analytics, Topic Modeling, Innovation Management, Natural Language Processing, Unstructured Data.

Introduction

Innovation has long been recognized as a cornerstone of economic growth, corporate competitiveness, and societal advancement. In the digital era, innovation management must contend with an unprecedented influx of unstructured data, ranging from customer reviews and social media content to academic publications and patent filings. These textual sources are replete with insights that, if properly extracted and analyzed, can inform decision-making across the innovation lifecycle. The ability to harness these insights has thus become a strategic

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

imperative for modern organizations. One promising approach to this challenge is topic modeling—a family of probabilistic, unsupervised machine learning techniques that uncover latent thematic structures within text corpora.

Topic modeling is not a new invention, but its relevance has surged in recent years with the proliferation of natural language processing (NLP) tools and big data analytics. Positioned within the broader NLP landscape, topic models provide a mathematically interpretable and relatively transparent mechanism for organizing, summarizing, and deriving insights from large text datasets. While large language models (LLMs), such as BERT and GPT, have captured much of the recent attention in AI research, their general-purpose design, input limitations, and tendency toward hallucination pose limitations in domain-specific applications such as innovation management. In contrast, topic modeling offers a more controlled and customizable approach, with the capacity to generate reproducible and context-sensitive insights.

The core premise of topic modeling is the statistical inference of latent topics—unobserved thematic patterns that recur across documents. Techniques like Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Correlated Topic Models (CTM), and more recent neural models such as BERTopic, allow for scalable and interpretable analysis of unstructured text. These models assign distributions over topics for each document, and distributions over words for each topic, enabling analysts to summarize and visualize large bodies of text with remarkable efficiency.

Organizations are increasingly turning to topic modeling to power their innovation efforts. From ideation to commercialization, topic models help firms navigate dynamic consumer landscapes, track emerging trends, and evaluate internal R&D portfolios. For instance, Procter & Gamble has implemented real-time feedback systems using NLP and topic modeling to enhance product development cycles. DHL leverages voice-of-the-customer analysis to inform service innovation, while L'Oréal monitors social media and fashion blogs to anticipate beauty trends before they become mainstream. At a more strategic level, Siemens applies patent analytics via topic modeling to benchmark and guide innovation investments.

Beyond corporate use cases, topic modeling has also gained traction in academic innovation research. Governments and funding agencies use topic modeling to assess research trends, allocate resources, and evaluate scientific impact. The European Commission, for example, has employed topic modeling to monitor and forecast innovation trajectories in EU-funded projects. In parallel, journals like the *Journal of Product Innovation Management* (JPIM) serve as rich repositories of scholarly insight, and analyzing them with topic models can yield valuable meta-level understandings of the field's evolution.

Despite its growing adoption, the use of topic modeling in innovation management remains fragmented. Different models are often applied without sufficient consideration of their assumptions, suitability, or methodological rigor. This results in inconsistencies and hampers the comparability of findings across studies. Furthermore, while

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

existing literature provides isolated examples of topic modeling applications in innovation, there is a lack of integrative frameworks that classify and evaluate these applications in relation to the stages of the innovation process.

To address this gap, this article makes several important contributions. First, we provide a comprehensive review of five dominant topic modeling approaches, highlighting their theoretical underpinnings, algorithmic characteristics, and practical implications. This review serves as a guide for researchers and practitioners navigating the complex landscape of topic modeling tools. Second, we propose a four-stage framework that maps topic modeling applications onto the core stages of innovation: idea generation, development, commercialization and evaluation, and academic research. This framework offers a structured lens through which to assess the relevance and effectiveness of different topic modeling strategies.

Third, we apply this framework in an empirical analysis of JPIM publications spanning four decades (1984–2023). Using topic modeling, we identify key research themes, track their evolution over time, and highlight underexplored areas that may warrant future investigation. This illustrative case not only demonstrates the practical utility of our framework but also provides a replicable methodology for similar analyses in other domains.

Finally, we conclude by outlining future research directions and practical considerations for effective topic modeling deployment. These include the integration of hybrid models, the use of dynamic topic modeling to capture temporal changes, and the development of domain-specific taxonomies to enhance interpretability.

In sum, this article positions topic modeling as a versatile and powerful tool for innovation management. By systematically categorizing its applications and aligning them with innovation processes, we aim to demystify the technique and empower innovation stakeholders to make informed, data-driven decisions. As organizations increasingly seek to extract value from unstructured data, topic modeling stands out not just as a technical solution, but as a strategic enabler of innovation.

2 | Topic Modeling Approaches

Topic modeling refers to a group of statistical and machine learning methods to identify latent semantic structures and extract meaningful topics within large volumes of text (DiMaggio et al. 2013). Such techniques have the potential to yield important benefits, beyond the capabilities of manual assessments of text documents, such as identifying new concepts that human readers cannot discern or serve as an inductive analysis tool to render new theory from textual data (Hannigan et al. 2019). However, at the same time, practitioners and scholars have an abundance of topic modeling approaches from which to choose, each of which likely was developed in reference to a specific scenario and required methodological decisions that are unfamiliar to users (DiMaggio et al. 2013). When used inaccurately or inappropriately, though, topic modeling can create serious issues regarding optimization, noise sensitivity, and instability that undermine any results it produces (Hannigan et al. 2019).

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

Taking even one step back, the use of topic modeling should be justified given a study's research question and related theory-informed reasoning, rather than mere accessibility of easy-to- use software packages (Chen et al. 2023).

Configuring a topic modeling procedure to serve intended purposes requires a thorough understanding of the similarities and differences among various approaches. To this end, in Figure 1, we offer an overview of the most frequently used probabilistic topic modeling approaches, which could serve as a starting point for deeper explorations. As topic modeling is not a discipline- specific technique, we draw on existing overviews in the neighboring research fields of information systems (e.g., Vayansky and Kumar 2020) and management (e.g., Hannigan et al. 2019) to organize our discussion of the different topic modeling approaches. These fields offer a balanced perspective on topic modeling, including key technical considerations but also advantages and disadvantages in light of translating raw text analysis outcomes to actionable insights; a key focus in line with innovation management (Spanjol, Noble, and Barczak 2024). Broadly speaking, existing topic modeling approaches in innovation- related research can be categorized as basic or advanced, in terms of topic relationships, time dynamics, hierarchical representation, and structural relationships, as we discuss next.[1] In Table 1, for each of these types of topic models, we present the key objective, assumptions, core analytical mechanisms, primary algorithms, and main advantages and disadvantages.

2.1 | Basic Approach

2.1.1 | Latent Dirichlet Allocation

A fundamental, frequently used, generative probabilistic model, called latent Dirichlet allocation (LDA; Blei et al. 2003), defines documents as a random composition of topics, which reflect the words that co- occur in documents, assuming a Dirichlet distribution. This approach makes several pertinent assumptions. Namely, it assumes the optimal topic number is fixed and known. Documents also represent “bags- of- words” whose meaning is relational and emerges from co- occurrence patterns, independent of any particular structure, and each document is independent. Furthermore, it considers topics as a multinomial distribution of words from a fixed vocabulary and assumes topics are unrelated to one another. For inference and learning, LDA primarily draws on variational, expectation propagation, and collapsed Gibbs sampling algorithms.[2]

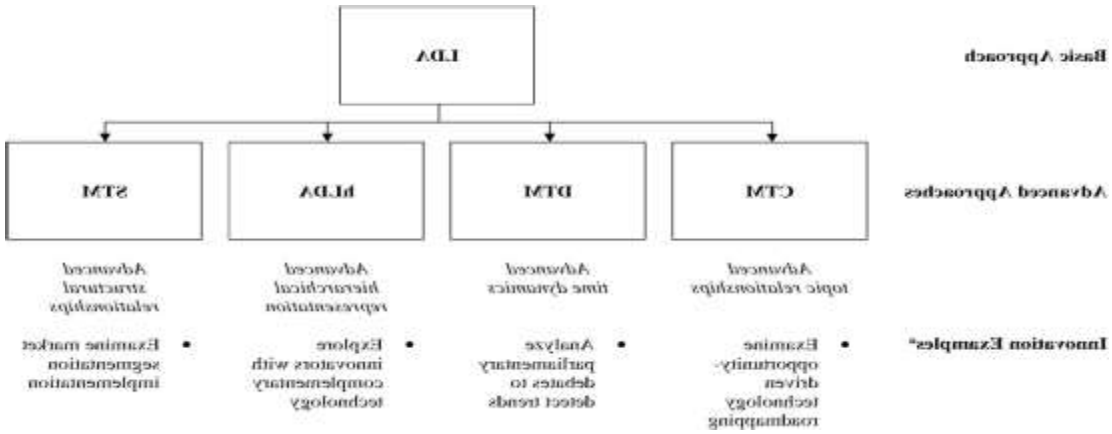
Despite the benefits of LDA for dealing with unseen documents, it suffers from several limitations. The key assumptions often do not hold in real- world applications. For example, predetermining an ideal number of topics (using metrics such as perplexity)[3] tends to be inefficient and imprecise for large data sets, and can cause such metrics to fluctuate, undermining reproducibility. LDA has further exhibited issues related to sparsity for corpora with large vocabulary sizes. In addition, LDA does not account for the order of words in a sentence, so it ignores an important characteristic of real- world textual data. Similarly, not accounting for correlation across topics or structures within a set of

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

FIGURE 1 | Probabilistic topic modeling approaches. [a]For the sake of completeness, we also report an example for LDA in an innovation context; LDA has for example, been used to identify technological



opportunities of emerging technologies.uments limits its ability to deal with various types of data, such as social media data that usually contain such correlations (George and Birla 2018; Vayansky and Kumar 2020).

Regardless of these limitations, LDA can be modified to fit most general analysis tasks, and even multi- model configurations and various types of data. Consequently, many studies use it, gathering input data from online customer reviews, patent descriptions, annual reports, newspaper articles, scientific articles, white papers, and so on, in their efforts to isolate customer needs, conduct business analyses, gauge stakeholder evaluations, or identify breakthrough research (e.g., Antons et al. 2015; Ma et al. 2021).

Advanced Approaches

Correlated Topic Model

In an attempt to overcome the LDA limitation pertaining to independent topics, the correlated topic model (CTM[4]; Lafferty and Blei 2006) follows a virtually identical generative process, except that it samples topic mixtures from a logistic normal distribution, instead of a Dirichlet distribution. This distribution accounts for the pairwise correlation between topics and uses the topic correlations to support its predictions, such that words from related topics are likely to appear within the document. In line with LDA, CTM is built on the assumptions related to the optimal number of topics,[5] "bag- of- words" representation, exchangeable documents, and treats documents as a logistic normal distribution of words from a fixed vocabulary and employs variational algorithms for inference and learning.

Compared with LDA, CTM enables researchers to account for more realistic relationships among topics; it also is more versatile in terms of exploring large corpora, forming predictive distributions, and providing optimized collaborative filtering. Although the outcomes are less sensitive to the assumption that the optimal number of topics is fixed and known, this claim still applies to CTM. As do the limitations related to the "bag- of-words"

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

representation and exchangeable documents assump tions. Even though correlations can only arise between two topics at once, CTM is computationally more expensive than LDA (Vayansky and Kumar 2020) and could result in more general words within topics (George and Birla 2018). Overall, CTM is a suitable approach for many types of data, especially corpora in which strong correlations between topics are to be expected. For example, CTM applied to patent text data has supported opportunity- driven technology roadmapping (Noh et al. 2021).

Dynamic Topic Model

Dynamic topic modeling (DTM; Blei and Lafferty 2006) is designed to overcome another important limitation of LDA, namely, the assumption of the independence of documents within a data set. DTM accounts for changes in topics over time in sequential corpora. In essence, a sequence of models is chained together based on a defined unit of time, within this approach considered a 'slice' of time. In line with LDA and CTM, DTM makes assumptions related to the optimal number of topics,[6] "bag- of- words" representation, and modeling documents as multinomial and logistic normal distributions of words from a fixed vocabulary; and mainly relies on variational Kalman filtering and variational wavelet regression algorithms for inference and learning.

Although it represents an extension of LDA, this approach requires researchers to specify the mentioned discrete unit of time, at the risk of losing specific data characteristics. Furthermore, DTM does not allow for correlation between identified topics. Like LDA and CTM, DTM does not incorporate methods to determine the optimal number of topics. In fact, the variation between data within time slices complicates this process even further (Vayansky and Kumar 2020), and the birth or death of topics are not considered (George and Birla 2018). Similarly, the same limitations with respect to the "bag- of- words" representation and computational requirements apply to DTM.

This probabilistic topic modeling approach is particularly suitable for a filtered or specifically structured corpus, which would benefit from a clear understanding of the evolution of topics, as can be depicted by changes in topic distributions. Innovation research has applied this approach to discover innovation trends and opportunities from sources such as parliamentary speeches, newspaper articles, and crowdsourcing platform data (e.g., Mele et al. 2019; Müller-Hansen et al. 2021).

2.2.3 | Hierarchical Latent Dirichlet Allocation

Another compelling extension to LDA is hierarchical LDA (hLDA; Blei et al. 2010), which moves beyond a single level of underlying topics and instead creates a hierarchy of aggregated topics and subtopics. This hierarchy is established by using a collapsed Gibbs sampling algorithm to identify aggregate topics, generate a set of subtopics and regroup documents accordingly, and finally generate additional subtopics using these groupings. In accordance with the previously discussed approaches, the assumptions of hLDA revolve around "bag- of-

words" representations, exchangeable documents, and considering documents as multinomial distributions of words from a fixed vocabulary. Additionally, the tree's size, shape, and character are not static.

The main advantages of hLDA are that it can locate a tree- like hierarchy of topics, accurately capturing different levels of granularity, and that this approach does not require researchers to specify the optimal number of topics; it can generate this value. Instead, however, it requires other parameters to be defined by the researcher, such as the number of hierarchical levels and the number of terms for each topic. In addition, a document can only follow a single path in the hierarchical tree structure, whereas in reality, documents in distinct subsets often share topics (George and Birla 2018; Hannigan et al. 2019). Following the previous approaches, hLDA also suffers from the limitations related to "bag- of-words" representations, exchangeable documents, and computational intensity.

Existing studies that rely on hLDA's ability to generate a hierarchical representation of internal structures use newspaper articles, crowdsourcing platforms, social media, and patent data to discover innovation trends and opportunities, as well as to examine organizational resources and capabilities (e.g., Mele et al. 2019; Wang et al. 2021).

Structural Topic Modeling

Developed to make fit algorithms more structured and systematic, structural topic modeling (STM; Roberts et al. 2014) also extends LDA; it allows for the inclusion of documents' metadata (e.g., who wrote the text, when the text was written, where the text was published) into the prior document–topic and topic–word distributions. Topical prevalence (i.e., the frequency with which a topic is discussed) and content (i.e., the words used to discuss a topic) are not assumed to be constant across all documents. Rather, metadata can be incorporated as covariates to structure the distributions. STM also assumes that the optimal number of topics is fixed and known,[7] documents are represented by "bags- of- words," documents are exchangeable, and topics can be considered a multinomial distribution of words from a fixed vocabulary. For inference and learning, STM predominantly utilizes partially collapsed variational expectation–maximization algorithms, using Laplace approximation.

Compared to the previous approaches, the inclusion of metadata in STM results in a more comprehensive estimation, in which topics can be correlated. It also supports measures of systematic changes in topical prevalence and content, and estimates of the influence of the included metadata as covariates. However, scaling up in terms of corpus size requires increasing the STM's capacity to process a larger set of documents and more diverse metadata. In such situations, there is substantial room for improvement and efficiency gains related to both theoretical optimality and practical programming applications (Hannigan et al. 2019; Roberts et al. 2014). Moreover, the key added value of STM is dependent on the availability of relevant metadata, and this approach is prone to limitations resulting from the "bag-of- words" representation and exchangeable documents assumptions.

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

Several existing studies have employed STM, using input data related to online customer reviews, online employee reviews, newspaper articles, government policy documents, scientific articles, online business idea competitions, online crowdsourcing platforms, or company profile data, with goals of sourcing, evaluating, and selecting ideas; screening regulations and policies; segmenting markets; and detecting emerging research topics, innovation trends, and opportunities (e.g., Guenduez and Mettler 2023; Nathan and Rosso 2022).

Topic Modeling Applications in Innovation

Topic modeling has the potential to reveal underlying topic structures that can aid innovation efforts in several ways (see Figure 2). First, because it provides evaluations of customer needs and sourcing ideas, it can facilitate idea generation. Second, topic modeling can support the development of testing, marketing, and operational strategies, as it provides a way to interpret stakeholder feedback, conduct business analyses, and screen the regulations and policies that are critical for the introduction of new products. Third, it can evaluate the likely effectiveness of an innovation and its impact on business performance. Fourth, topic modeling can offer an overview of innovation research in scientific outlets, which can highlight breakthrough research. In the following sections, we detail the uses of topic modeling in innovation according to this framework and discuss its various applications for innovation.

Topic Modeling Uses Framework

To structure various uses of topic modeling in innovation management, we develop a comprehensive framework of application areas within innovation management that draws on previous work on innovation processes. We distinguish the generated insights according to whether they are more relevant to innovation practitioners or innovation scholars. That is, topic modeling can help innovation practitioners understand and contribute to key stages of innovation processes. Innovation scholars, across each stage of the innovation process, can review and build innovation management knowledge and detect trends in previous research.

In Figure 2, we simplify the innovation process into three stages: idea generation, development, and commercialization and evaluation. These stages reflect theories about new product development processes (Grönlund et al. 2010), as outlined in the stage-gate and agile models by Cooper (1990, 2008) and colleagues (Cooper and Sommer 2016). During the ideation stage, topic modeling can help innovators source and select ideas, identify special interest groups and thought leaders, and assess customer needs. In the development stage, detailed testing, operations planning, and marketing activities take place (Cooper 1990). Topic modeling can help innovation practitioners conduct analyses, enhance quality controls, and screen regulation policies. Finally, in the last stage of the innovation process, the commercialization and evaluation stage, topic modeling can assist in financial projections, evaluations of the marketing launch plan, and post- audits of the project (Cooper 1990).

To structure the framework, we collected articles on innovation management and topic modeling.[8] From an initial set of 8301 articles, we identified only journals with an impact factor greater than 4. We further manually reduced

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

the articles and removed those that ultimately were not relevant to innovation management and topic modeling, leading to a final set of 1099 articles. We manually mapped each article to match it with one of the three topic modeling uses outlined in Figure 2 (idea generation: 61.1%, development: 22.5%, and commercialization and evaluation: 20.0%), and articles primarily focused on topic modeling of academic innovation research (25.8%).[9] Table 2 illustrates some topic modeling research examples, the data and topic modeling approach used, and the application area.

3.2 | Idea Generation

Topic modeling is a valuable tool for informing the idea generation phase, during which markets and customer preferences are evaluated, trends are analyzed, and ideas are selected.

For existing review papers in academic innovation research, see Table 3.

3.2.1 | Isolate and Examine Customer Needs, Preferences and Requirements

In detail, topic modeling can isolate customer needs, preferences, and requirements to develop product innovation ideas (Bernier et al. 2023). Furthermore, Zhang et al. (2022) use LDA, combined with sentiment classifiers of customer reviews for laptops and smartphones, and thereby identify customers' needs more accurately. In turn, their analysis produces product attribute categories to inform product design strategies (Zhang et al. 2022).

3.2.2 | Discover Innovation Trends and Opportunities

Another prevalent use of topic modeling involves uncovering trends and opportunities by analyzing patent data (Ma et al. 2021; Park et al. 2019) or technical documents (Choi and Kwon 2023). Sun et al. (2021) review full patent texts and thereby identify photovoltaic technologies according to their novelty and relevance. They developed a list of 98 patents that offer potential breakthroughs for product innovation and development policy. That is, topic modeling can discover innovation opportunities based on customer needs and preferences, revealed in consumer- generated data, as well as through analyses of researcher and patent data.

3.2.3 | Source, Assess and Select Ideas

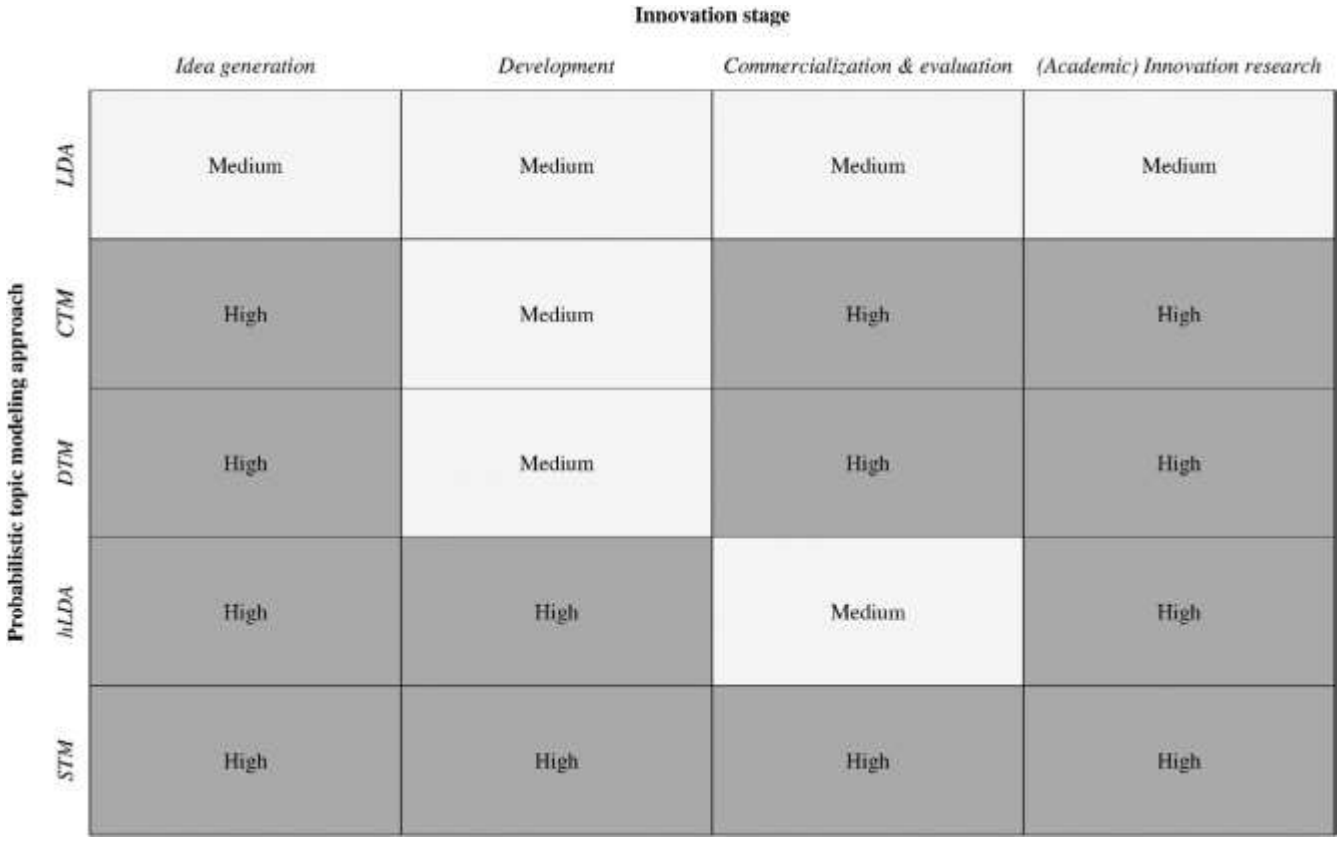
Idea generation requires efforts to source, evaluate, and select ideas, which can be supported through topic modeling. Lee and Sohn (2019) use LDA to analyze Kickstarter descriptions of software- related projects to discover innovative topics hidden in these project sets. The authors combine these results with a conjoint analysis to determine the most preferred topics, with respect to the amount of funding received. The results recommend businesses to invest in smart assistant services in certain domains (e.g., finding job opportunities).

Service improvement ideas also can be derived from user- generated content. Yin et al. (2023) use topic modeling of user- generated posts on the largest auto product website in China to identify consumer pain points. They validate the analysis results and confirm that their classification model yields breakthrough, useful, feasible, and adoptable innovation ideas (Yin et al. 2023). Thus, topic modeling is useful and important for the process of idea generation.

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article



3.2.4 | Identify and Detect Insights From Special Interest Groups and Thought Leaders

Not only do customers and patents provide useful insights that can stimulate product innovation, but special interest groups and thought leaders also can provide meaningful, targeted insights accessible through text analysis to generate new ideas. As such, topic modeling has been employed to identify thought leaders in knowledge celebrity communities who might anticipate trends (Chen et al. 2022).

Analyzing special interest groups, Zeng (2018) examines renewable online communities to anticipate future business environments and guide innovation decisions, by applying LDA to

FIGURE 3 | Future potential of topic modeling approaches across innovation stages. The future potential of the topic modeling approaches across the innovation stages is based on our own recommendations and we would like to highlight that certain research questions can very well be answered by deploying approaches that score relatively lower. Medium potential indicates that while these methodscould offer valuable insights, their impact may be limited due to a suboptimal fit with the respectiveinnovation stage. High potential indicates strong promise for generating valuable insights due to excellent fitbetween their unique capabilities and the respective innovation

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

stage. As we discuss in the followingsections, tailored approaches could possibly possess even higher future potential by combining theadvantages of multiple topic modeling approaches.

chatter in an online community. The identified topics, “finding the right wind turbines” and “wind turbines and their characteristics,” provide insights into what customers are prioritizing and which features should be developed next by investigating the comments included among those topics. As these examples show, topic modeling can be applied to platforms and online groups consisting of special interest groups and thought leaders that have specialized insights or drive product and market trends.

Overall, existing research has deployed LDA, DTM, hLDA, hierarchical Dirichlet process (HDP)[10] and STM in enquiries within the idea generation stage (e.g., Bernier et al. 2023; Choi and Kwon 2023; Li et al. 2020; Mele et al. 2019; Müller-Hansen et al. 2021). Whereas the flexibility of LDA (Blei et al. 2003) inherently makes it suitable for all highlighted uses in this stage, DTM's focus on evolution over time (Blei and Lafferty 2006) explains its dominant use in discovering innovation trends and opportunities. hLDA (Blei et al. 2010), in line with DTM, is mostly relied on for this purpose, reflecting the complexity of discovering such trends. While STM is also applied toward this end, the possibility to include covariates (Roberts et al. 2014) also makes it an attractive approach for sourcing, assessing, and selecting ideas. Figure 3 illustrates our recommendation on the future potential of the topic modeling approaches, based on our assessment of the fit between their unique capabilities and the respective innovation stage.[11] However, we hasten to add that certain research questions could still be effectively addressed using approaches with relatively lower scores. We conclude that idea generation can greatly benefit from the advanced capabilities of CTM to deal with large corpora in which topics are expected to be correlated (e.g., obtaining thought leader insights from social media data), DTM to capture time dynamics (e.g., discovering trends in patent data), hLDA to expose hierarchical structures (e.g., isolating customer needs and preferences and requirements in customer review data), and STM to model the structural influences by incorporating covariates (e.g., sourcing ideas from crowdfunding platform(s) data).

Development

In the development stage, topic modeling could aid in conducting business analyses, ensuring quality management, examining organizational resources and capabilities, and developing roadmaps.

| Conduct Business Analysis

By examining online reviews, researchers can identify a set of competitors that might be benchmarked for useful innovations. With the help of topic modeling, an LDA analysis of more than 8 million customer reviews of hotels has provided insights into which hotel attributes offer the most competitive advantage for different segments (Ye et al. 2020). Such an automated text approach provides managers with a systematic, efficient way to conduct business analysis (e.g., identify key competitors and adjust product and service development efforts accordingly).

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

Ensure Quality Management in New Product Development

Existing research has also relied on topic modeling for quality management. For instance, Barravecchia et al. (2023) deploy STM on online customer review data to demonstrate how such types of data, that is, digital voice-of- customers data, can be drawn on to track and monitor product quality perceptions over time.

Examine Organizational Resources and Capabilities

Moreover, topic modeling has assisted evaluations of how teams should be structured to achieve breakthrough innovations. Vakili and Kaplan (2021) use it to review patent data from four distinct technological domains (RFID, MRI, nanotube, and stem cell technologies). With topic modeling, they derive a measure of cognitive novelty, according to which topics feature vocabulary that describes ideas, and thus can assess novel breakthroughs and innovative outcomes. In turn, they demonstrate that the selection of teams to foster innovation depends on the type of technology.

Screen Regulation and Policies

Topic modeling has been employed to analyze policy announcements, providing insights into variations in state responses during pandemics (Capano et al. 2020) and to assess the impact of policies on sales (Li et al. 2021). For example, Guenduez and Mettler (2023) utilized STM in conjunction with qualitative narrative analysis to identify six predominant policy narratives in AI policies from 33 different countries. These narratives included enhancing national frameworks for AI development and deployment, and promoting strategic AI collaboration at both national and international levels. This methodology enabled a comprehensive comparison of policy narratives across diverse governmental settings.

Roadmap Development

The development of (technological) roadmaps has also been supported by topic modeling. In particular, Kim and Geum (2021) use LDA to develop data-driven technology roadmaps by iden tifying underlying topics related to technology and markets. Similarly, Noh et al. (2021) rely on CTM to extract technological topics from patent text data in an effort to integrate technological opportunities with market opportunities.

For the development stage, previous research tends to draw on LDA, CTM, hLDA, and STM (e.g., Guenduez and Mettler 2023; Kim and Geum 2021; Noh et al. 2021; Wang et al. 2021). LDA's flexibility, again, has earned this approach an application in all uses within this stage (Blei et al. 2003). A more specialized application in roadmap development is observed for CTM, owing to its ability to account for more realistic relationships between topics (Lafferty and Blei 2006) in underlying patent data. Similarly, extant research has solely deployed hLDA for examining organizational resources and capabilities, benefiting from identifying hierarchical structures (Blei et al. 2010) in sets of patent texts. The improved model fit of STM (Roberts et al. 2014) has resulted in applications geared toward conducting business analyses, ensuring quality management in new product development, and screening regulations and policies. The future potential of the topic modeling approaches in this stage appears to

be most promising for hLDA and STM (see Figure 3). That is, development can be enhanced through the advanced capacity of hLDA to establish hierarchical representations (e.g., screening regulation and policy documentation data) and STM to capture structural influences (e.g., roadmap development based on patents).

Commercialization and Evaluation

During the commercialization and evaluation stages of the innovation process, it is essential to execute the marketing launch plan and operations plan, followed by a review of project and product performance.

Segment Market

Extant studies have relied on topic modeling for segmentation purposes. For instance, Schröder et al. (2019) segment online users based on the latent interest underlying their browsing behavior. Deploying LDA on clickstream panel data, the authors demonstrate how users' browsing baskets can be used as a basis for segmentation.

3.4.2 | Predict and Measure Innovation Performance

Topic modeling can also provide new measures of innovation activity and innovation performance. Choi et al. (2021) develop a new multidimensional measure of diversification on the basis of topic modeling of annual reports; they find that diversification leads to higher firm value. Thus, topic modeling can help build measures of innovation activity or diversification based on language data. In particular, the combination of such data with performance and business-relevant measures supports assessments of innovation effectiveness.

3.4.3 | Evaluate Launch

Post- launch research could evaluate the overall performance and impact of diversification further to decide whether to roll out innovations across industries or geographic markets. Nathan and Rosso (2022) use STM to build a new measure of innovative activity by analyzing product and service launch news articles; with topic modeling, they cluster text fragments pertaining to the same real- world launch event. In this case, STM helped reduce the original array of observations to a condensed number of launch events. These data, in combination with other data sets pertaining to firm performance and business insights, indicate positive relationships between prior patents and launches, as well as between launches and the business performance of small enterprises (Nathan and Rosso 2022).

Previous research zooming in on the commercialization and evaluation stage has mostly used LDA and STM (Fresneda et al. 2022; Slof et al. 2021). The flexibility of LDA (Blei et al. 2003), also in this stage, makes it a suitable approach for all uses within this stage. STM, on the other hand, has been applied for segmenting markets and evaluating launches, underscoring the relevance of incorporating covariates (Roberts et al. 2014) toward these ends. As outlined in Figure 3, continued applications could likely benefit from alternative approaches for such tasks, relying on the sophisticated abilities of CTM to optimally process large datasets and account for more realistic relationships between topics (e.g., evaluating launch on the basis of social media and/or news platform

data), DTM to capture temporal dynamics (e.g., measuring (innovation) performance based on annual report data), and STM to examine influences of external factors (e.g., segmenting markets by using social media data, complemented with other user- related data, as input).

3.5 | Innovation Management Research

In Table 3, we summarize representative articles providing topic model reviews of innovation research domains. Most of them focus on specific topics within innovation research. For example, Hopp et al. (2018) investigate disruption research using LDA in combination with network, cluster, and linear regression analysis. Li et al. (2022) identify potential breakthrough research in the domain of solar cell technologies using HDP. Rakas and Hain (2019) use LDA to examine innovation systems research. Antons et al. (2015) use an LDA approach to review innovation research published in *JPIM* articles up to 2013. Future endeavors could move beyond LDA and conventional STM approaches to develop tailored configurations with the goal of capturing advanced topic relationships (Lafferty and Blei 2006), temporal (Blei and Lafferty 2006), hierarchical (Blei et al. 2010), and structural relationships (Roberts et al. 2014) dynamics (see Figure 3). Such research efforts could push boundaries regarding the comprehensiveness of reviewing and building innovation management knowledge. Our review study discussed next extends and complements past work by examining articles published in *JPIM* until 2023, employing a multi- level STM approach combined with bibliometric and regression analysis.

4 | Illustrative Topic Modeling Application

Zooming in on the last application area (Figure 2), innovation research, we offer a finer-grained perspective on how topic mod eling can be used to spur innovation. We use *JPIM* research as input data to demonstrate the use of topic modeling to review and build innovation management knowledge (i.e., uncover *JPIM*'s topical structure), detect emerging research topics (i.e., based on degree of research attention), and identify impactful research (i.e., based on academic citations).

Topic modeling in innovation research can yield valuable insights that build on existing knowledge of innovation management, detect emerging research topics, and identify groundbreaking research. This type of topic modeling typically involves a selection of an entire cohort of scientific articles, including their abstracts, titles, and keywords. Thus, by scrutinizing published research, innovation scholars can generate insights for the stages of the innovation process, depending on the mined articles' focus and selection criteria.

4.1 | Methodology

To establish these insights, we systematically collect the required data, which we subject to a mixed- method approach consisting of (multi- level) STM and bibliometric, and regression analyses. We gather all *JPIM* articles available at the time of data collection. By the end of May 2023, *JPIM*'s online archives contained 2237 records, published between the first *JPIM* issue in 1984 and the third issue in 2023 (222 issues in total). Applying screening and eligibility criteria to this set resulted in a final sample of 1444 *JPIM* publications.[12] We collected the full-

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

length PDF versions and accompanying metadata using a combination of *JPIM*'s online archives and the Web of Science and Dimensions databases, such that we could ensure cross-validation of the compiled metadata.

We use the corpus of 1444 full-length *JPIM* articles[13] and their accompanying metadata as input for a multi-level STM. In es sence, we seek to combine the advantages of the approaches that we reviewed previously into one configuration, such that we (1) use STM (stm package in R) for advanced structural relationships and to allow for topic correlations; (2) incorporate time dynamics in the covariates; and (3) deploy STM on two levels, with the goal of uncovering the underlying (hierarchical) topical structure of *JPIM* research, in terms of both research themes (i.e., global level) and subtopics within the identified research themes (i.e., granular level).

We subject the topics identified in the STM analyses to ordinary least squares (OLS) regression analyses. That is, we take a topic's average prevalence and regress it on the articles' year of publication to reveal how the research topic's relative share in each publication year has evolved over time. At the global level, we depict how the research themes have evolved over *JPIM*'s 40- year history, on which we overlay some of the influential *JPIM* publications per research theme.

To isolate the most influential *JPIM* publications per topic on both levels, we use descriptive bibliometrics. With this approach, we can quantify, aggregate, and rank the relative contributions of *JPIM* articles. Descriptive bibliometric analyses on the publication level indicate the most academically influential publications (i.e., most cited *JPIM* publications) within each identified *JPIM* topic. In turn, the outcomes of the global- level

STM analysis serve as input for our integrated depiction of the evolution of *JPIM* research over time, for which we overlay the separate *JPIM* research themes with their most academically influential publications.

To examine the academic impact of the *JPIM* research themes, we follow existing impact studies, with academic impact proxied by citations (Warren et al. 2021). We obtained citation data about articles that cited any of the 1255 focal *JPIM* articles from the Dimensions database. The 106,296 citations (mean = 84.698; median = 47.000; SD = 125.276) prompted by the *JPIM* articles in our corpus are highly skewed, though, such that many articles prompted no citations. Guided by model fit indices (see Supporting Information: Appendix A, models A–D), we subject the resulting measures to a zero- inflated negative binomial regression, which is appropriate given the skew and number of null observations. Considering the timespan of our data set, we apply an age correction and obtain age- weighted citation counts (number of citations divided by the article's age; Jin et al. 2007) to correct for age- related effects.

4.2 | Findings

4.2.1 | Underlying Topical Structure of *JPIM* Research

We uncover *JPIM*'s underlying topical structure using a multi- level STM approach, with publication year and document length as topical prevalence covariates for the model estimation to capture any systematic changes. The degree to which a document is associated with a certain topic can vary as a function of when the article was

published and its length. For our model estimation, on both levels, we use spectral initialization (i.e., non-negative matrix factorization of the word co- occurrence matrix; Roberts et al. 2014) and the default for priors.

Because our aim is to look beyond a document's textual content and consider the bibliometric focus of our mixed-method approach, we rely on the most influential publications with a topic proportion of at least 0.500 (i.e., the majority of the respective *JPIM* publication relates to the respective topic), in combination with the most frequently used terms (Wang et al. 2015), to label the 14 underlying themes. This exercise resulted in the following labels for the identified *JPIM* research themes: (1) digital transformation, (2) responsible innovation and value, (3) innovation adoption and resistance, (4) idea generation, (5) new product design, (6) disruptive innovation, (7) radical innovation (management), (8) environmental turbulence and learning, (9) service innovation and new service design, (10) new product design speed/time performance, (11) cross- functional integration, (12) new product launch and market entry, (13) concept testing and innovation diffusion, and (14) new product development success and failure.[14] Supporting Information: Appendix B offers a comprehensive overview of these research themes, along with the most frequently used terms, academically most influential publications, and received research attention per theme.

Organizing research into research themes greatly improves the accessibility of findings, which are typically hidden in lengthy and complex academic articles (Andersen and Hackos 2018), reducing the need for innovation researchers and practitioners to navigate the full breadth of academic literature. Extending related previous endeavors (e.g., Antons et al. 2015; Mertens et al. 2023), our illustrative multi-level STM approach adds an additional layer of accessibility to the underlying topical structure by further organizing subtopics into research themes. Given our focus on supporting innovation processes, the resulting research themes are then categorized and contextualized along the innovation process stages (Cooper 1990) in Table 5 below. Uncovering such hierarchical dynamics takes structuring research in the field to the next level, helping both innovation researchers and practitioners quickly identify areas of interest for deeper exploration.

4.2.2 | Evolution of *JPIM* Research Themes Over Time to Detect Emerging Themes

Innovation management researchers exhibit considerable interest in the development of research themes in *JPIM* (e.g., Antons et al. 2015; Guo 2008). Contributing to this ongoing conversation, we offer an integral overview of *JPIM*'s 40-year evolution, drawing on a sophisticated topic modeling configuration. Figure 4 illustrates the evolution of *JPIM* research themes for the period 1984–2023 in terms of research attention received, including the most impactful publications per identified theme. Seven

were added based on publication year); the size of the nodes is based on the age- weighted number of citations. In the online version of the present article, using Adobe Acrobat, hovering over nodes will display the respective record's details (in the offline version, this information can be found in Supporting Information: Appendix B). Hyperlinks to full publication details are included with each of the nodes. The standardized coefficients reflect

how a respective research theme's relative share, in terms of research attention, has evolved over time, per publication year (based on online publication date). Significant positive coefficients indicate *emerging* research themes, significant negative coefficients represent *mature* research themes, and insignificant coefficients refer to *stable* research themes. To facilitate comparison across different studies, both independent and dependent variables are standardized (Darlington and Hayes 2017). *$p < 0.05$, **$p < 0.01$,

pertinent research themes emerge over time: digital transformation, responsible innovation and value, innovation adoption and resistance, idea generation, new product design, disruptive innovation, and radical innovation (management). These emerging research themes reflect the widespread implications of digital transformation (Appio et al. 2021), the growing prominence of different forms of responsible innovation and value creation (Bstieler et al. 2015; Prahalad 2012; Sjödin et al. 2020), acknowledgments of both innovation adoption and resistance (Lee and Coughlin 2015), the fundamental importance of idea generation (Micheli et al. 2019) and new product design (Veryzer and Borja de Mozota 2005), and the relevance of disruptive innovation (Markides 2006) and radical innovation (management) (Ritala and Hurmelinna- Laukkanen 2013). Notably, technological developments and open innovation are also featured within these research themes (with the exception of open innovation in the research theme related to innovation adoption and resistance).

The identified *JPIM* research themes also contain topics for which research attention is declining, including cross-functional integration (Griffin and Hauser 1996), new product launch and market entry (Lee and O'Connor 2003), concept testing and innovation diffusion (Fuchs and Schreier 2011), and new product development success and failure (Narver et al. 2004). These research themes fulfilled a vital role during the growth and formative years of *JPIM* (1984–2008) but have started to fade as the journal reaches maturity. These developments coincide with the increasing dominance of emerging research themes at the end of the journal's formative years. While undeniably challenging, important new knowledge could be generated by advancing these more traditional innovation management research themes (Guo 2008).

Other themes exhibit a stable time trend, also referred to as “evergreen” or “wallflower” topics (Antons et al. 2015). These themes are the very core of the innovation management domain: environmental turbulence and learning (e.g., Calantone et al. 2003), service innovation and new service development (e.g., Storey et al. 2016) , and new product development speed/ time performance (e.g., Cooper and Kleinschmidt 1994). The

importance of advancing core innovation management research

TABLE 5 | Analysis of *JPIM*'s academic impact, 1984–2023.

themes has been recognized in prior *JPIM* overview articles (e.g., Antons et al. 2015).

As an illustration of the proposed multi- level STM approach, we use research theme 2 and use STM to identify subtopics within this research theme on responsible innovation and value. We

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

TABLE 5 | (Continued)

Lexical variation: diversity	−0.136	(0.050)	**	0.873
Quantifiers	−0.079	(0.035)	*	0.924
Readability: Flesch score	−0.050	(0.034)		0.951
3. Controls related to presentation and author properties				
Title: length	−0.173	(0.034)	***	0.842
Title: question	0.010	(0.102)		1.010
Title: colon	0.146	(0.032)	***	1.157
Appendix	0.045	(0.064)		1.046
Number of pages	−0.226	(0.085)	**	0.798
Number of authors	−0.014	(0.034)		0.986
4. Controls related to time properties				
Article age	−0.169	(0.061)	**	0.845
Zero- inflation model[f]				
5. Controls related to time properties				
Article age	3.433	(0.808)	***	30.982
Log-likelihood		−3443		
Wald's χ2Change (df)		$41.142_{(6)}$***		

[a]See Supporting Information: Appendices A and C for an overview of the regression models of *JPIM*'s academic impact, in which the present model corresponds to Model D and Model 3, respectively.

[b]Incidence rate ratio.

[c]Negative binomial with log link. [d]New service development. [e]New product development. [f]Binomial with logit link. [g]Research themes relating to stage 1: Idea generation (Figure 2). [h]Research themes relating to stage 2: Development (Figure 2).

[i]Research themes relating to stage 3: Commercialization and Evaluation (Figure 2).

cross-functional integration (β = 0.107, IRR = 1.112, p = 0.048). Second, with respect to writing property controls, *JPIM* articles with less lexical diversity (β = −0.136, IRR = 0.873, p = 0.007) or quantifiers (β = −0.079, IRR = 0.924, p = 0.023) prompt more age- weighted citations. Third, among controls related to presentation properties, we find that longer *JPIM* articles (β = −0.226, IRR = 0.798, p = 0.008) and longer titles (β = −0.173, IRR = 0.842, p < 0.001) have decreasing effects on the age-weighted number of citations, whereas the use of colons in the title (β = 0.146, IRR = 1.157, p < 0.001) tends to increase them. Fourth, regarding time property

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

controls, relatively older *JPIM* articles seem to attract fewer age- weighted citations ($\beta = -0.169$, IRR = 0.845, $p = 0.006$).

Factoring in the academic impact of the identified research themes (Mertens et al. 2023), could even further inform the discussion on the future development of the field (Wetzels et al. 2023). After all, receiving research attention does not equal impact (Antons et al. 2015), as emerging research themes could be little impactful, just as themes receiving declining research attention could very well be among the most impactful themes, for example, new product development success and failure. Bringing together macro- level insights on research attention received and garnered age-weighted number of academic citations could offer a solid and structured foundation for the future development of the field— making informed decisions on orchestrating continued research efforts toward building on emerging, refreshing stable, or revitalizing mature innovation management research areas.

5 | Future Outlook

After demystifying the various types of topic modeling, organizing their use in innovation research, and illustrating the benefits such techniques can deliver, we leverage these insights to look forward and outline potential applications of topic modeling for the components of the proposed framework (summarized in Table 6).

5.1 | Innovation Research (And Researchers)

Building novel contributions requires a state- of-the- art understanding of the existing body of innovation literature (Barczak 2017). The unprecedented, rapid expansion of extant knowledge makes it ever more difficult for innovation researchers to stay abreast of relevant findings. As a further advancement of the multi- level topic modeling suggested herein, LLMs might be built into topic models to help develop such state- of-the- art understanding. Uncovering the underlying topical structure and their respective developments over time, on multiple levels, could serve this purpose.

Both macro-level developments on research themes and sub topics within these research themes can be generated to help alleviate the bottleneck created by the sheer size of available research. Continued research efforts to move the innovation research field forward thus could be orchestrated more effectively. Taking the incorporation of LLM models even further, beyond performing the actual task of topic modeling, LLM-based gen erative AI could add significant value at various stages of such configurations (McCloskey et al. 2024). For example, such models can be leveraged to search for relevant innovation literature and then label topics in the resulting topic models.

Current topic modeling applications predominantly rely on text data as input. However, there is a surge in other forms of unstructured data, such as voice, image, and video (Balducci and Marinova 2018). For example, important insights for innovation might be harvested from aspects beyond mere customer review text, such as

image and video data (Wang et al. 2022). In such configurations, computer vision can be used to extract so-called "visual words" from images or (each image within) videos, therewith creating visual word documents and thus,

TABLE 6 | Future research agenda.

selection that will provide innovation capabilities and allow for the implementation of required modifications, and (5) collecting white papers to optimize the firm's own investor value. Such endeavors could greatly benefit by complementing topic modeling with other techniques, for example, classification and cluster analysis (Ye et al. 2020), multiple correspondence analysis (Stamolampros et al. 2020), and network analysis (Wang et al. 2021). All these efforts could broaden insights and encompass relevant stakeholder groups to offer a more comprehensive perspective on development. In line with the previous discussion on bringing together topic models and LLMs, rather than solely relying on either of these techniques for such tasks, the infusion of LVMs and other generative models could be fruitful as part of mixed configuration (Spanjol, Noble, Baer, et al. 2024). In addition, generative AI could be used to merge collected insights from the multiple stakeholders involved to better aid firms' development activities.

5.4 | Commercialization and Evaluation

A more comprehensive perspective can be offered for commercialization and evaluation as well. In terms of predicting and measuring innovation performance, existing topic modeling applications tend to be limited to one source (e.g., Bongini et al. 2022), with some initial attempts to expand further (e.g., Slof et al. 2021). New applications should be set up to provide a multi-source perspective. For example, com bining white papers with annual (ESG) reports to determine critical success factors, newspaper articles for legitimacy assessments, app store information to evaluate existing strategies, and customer social media data to gauge satisfaction all could deepen insights with regard to measuring and predicting innovation performance. Extracting meaningful insights from such combinations of input data likely requires tailored methodological configurations, leveraging topic modeling's complementarity with, for example, competing regression analysis (Slof et al. 2021), sentiment analysis and qualitative techniques (Gregoriades and Pampaka 2020), and cluster analysis (Fresneda et al. 2022). Also, within this stage, generative AI could help integrate various data sources to better plan and assess commercialization efforts (Spanjol, Noble, Baer, et al. 2024).

6 | Discussion

6.1 | Theoretical Implications

Existing research has acknowledged the unique value of topic modeling for extracting meaningful insights from the vast amount of unstructured text data to improve innovation processes (Choi and Kwon 2023; Guenduez and Mettler 2023). The abundance of available models and serious risks of inadequate use, however, make it ever more difficult for researchers and practitioners to choose and configure topic modeling procedures (Hannigan et al. 2019), especially in fragmented fields such as innovation management (Antons et al. 2015; Banville and

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

Landry 1989). In response, this study creates a contextualized common ground for effectively using topic modeling for existing and continued inquiries related to the (stages of the) innovation process, thereby offering important implications for innovation research.

First, isolating and structuring the mostly used probabilistic modeling approaches provides innovation researchers and practitioners a solid starting point to develop an understanding of the similarities of, and differences between, these established approaches (Hannigan et al. 2019; Vayansky and Kumar 2020). Moreover, drawing on research outside the traditional realm of innovation management could inspire discipline- spanning theoretical integration (Spanjol, Noble, Baer, et al. 2024), or the overhaul of theoretical paradigms rooted in other research disciplines.

Second, by presenting a comprehensive framework and illustrative application, the present study structures and contextualizes the suitability and future potential of the most frequently used topic modeling approaches along the stages of the innovation process (Cooper 1990). As a result, innovation researchers can readily spot promising unexplored applications, which could facilitate more robust theory development via validating extant or igniting fresh theorizing.

Third, this study lays out pertinent future research avenues for configuring topic modeling procedures along the components of the proposed framework (McCloskey et al. 2024; Spanjol, Noble, Baer, et al. 2024), enabling innovation researchers to address existing and future challenges in understanding and contributing to the innovation stages. Similarly, the illustrative application could serve as a blueprint for continued topic modeling efforts geared toward capturing a variety of dynamics buried in complex datasets (Antons et al. 2015). The suggested configurations could better equip researchers to refine and complement existing theoretical frameworks, thereby capitalizing on opportunities for new theoretical development in this research field.

6.2 | Managerial Implications

Successfully integrating topic modeling in innovation management holds significant potential to help managers optimize innovation processes, an important driver of revenue, growth, and organizational health (Banholzer et al. 2023; Manly et al. 2023). Against this backdrop, it is important to carefully allocate resources to enhance results for the focal firm and other relevant stakeholder groups. This study can help innovation managers improve these processes and unlock such mutually beneficial outcomes.

First, organizing the space of available topic modeling approaches offers innovation practitioners an accessible starting point for developing a deeper understanding of key approaches, including main strengths and limitations. With this knowledge, managers would be able to recognize when (and which specific configuration of) topic modeling might be appropriate, greatly benefiting the translation of unstructured text data into actionable insights (Spanjol, Noble, and Barczak 2024), in turn, serving as input for making informed decisions on addressing operational hurdles or

strategic questions related to streamlining innovation processes (de Backer and Rinaudo 2019; De Jong et al. 2024).

Second, the comprehensive framework put forward in this study, including an overview of future potential and an illustrative application, contextualizes topic modeling applications along the fundamental stages of the innovation process (Cooper 1990). For interested innovation practitioners, the adoption of topic modeling approaches based on the aggregated patterns of suggested, validated research practices presented herein (Spanjol, Noble, and Barczak 2024) expands managers' toolbox to shed additional light on key aspects of innovation processes, and thus reduces the risk of misguided decision making.

6.3 | Limitations and Conclusion

With this comprehensive review and illustrative application, we aim to aid innovation management researchers and practitioners in leveraging unstructured text data to enhance innovation processes. By (a) outlining prominent topic modeling approaches in innovation management, (b) organizing existing topic modeling applications along the stages of new product development into a comprehensive four- part framework, (c) zooming in on one component and concretely showcasing how topic modeling can be relied on to support innovation processes, and (d) offering a future outlook along the components of the proposed framework, this article demonstrates the unique capabilities of topic modeling to spur innovation management.

In spite of making the use of topic modeling more accessible to both innovation management scholars and practitioners, the present study also comes with limitations. First, despite the comprehensive and systematic search underlying the review of the most frequently used topic modeling approaches in innovation management research, some articles might have been overlooked as a result of variations in terminology. Relatedly, by organizing the discussion of the most frequently used approaches according to carefully selected existing overviews in neighboring research fields, some less prominent but promising approaches may have been excluded from our discussion. Therefore, it would be insightful to extend our review by explicitly focusing on the applicability of such additional topic modeling approaches in innovation management.

Second, beyond overlooking topic modeling applications, selecting and mapping articles to the components of the proposed framework remains a subjective process. To minimize researcher bias, and in line with common practice (see e.g., Spieth et al. 2025), we specify selection criteria, protocoled the matching of articles with our framework's components and resolve any discrepancies by consensus. Continued research could apply other selection criteria or focus on more specific innovation management processes to provide additional insights into the added value of topic modeling in innovation management.

Third, our illustrative topic modeling application examines *JPIM*'s underlying topical structure, along with the evolution of the constituent research themes and most impactful publications.

ISSN: 3065-0593

Journal of Marketing and Digital Media

Research Article

While research themes might be classified as emerging, stable or mature within the journal boundaries of *JPIM*, it could very well be the case that these themes have received diverging research attention within the broader landscape of innovation research (Antons et al. 2015; Page and Schirr 2008). Zooming out to include other relevant innovation journals, such as *R&D Management*, *Research Policy*, *Technological Forecasting and Social Change*, and *Technovation* (Sarstedt et al. 2024) could provide such complementary insights. In a similar vein, in this study, academic impact is operationalized as age-weighted num ber of citations. While not the specific focus of the present study, recent research has adopted a broader perspective on impact and has included additional impact measures related to general public uptake (Gonsalves et al. 2021; Thelwall et al. 2023), more specifically focused on public policy documents (Fang et al. 2020), or patents (Gazni 2020). Incorporating such complementary impact measures could offer a more holistic perspective on the impact of *JPIM*'s research themes. In conclusion, understanding how to effectively apply topic modeling could greatly benefit both innovation management researchers and practitioners. We hope that this study empowers these stakeholder groups to leverage the unique capabilities of topic modeling for extracting meaningful insights from the wealth of available unstructured data toward further optimizing innovation processes.

Endnotes

Short text- optimized topic models, such as the biterm topic model (BTM; Yan et al. 2013) and dual sparse topic model (DSTM; Lin et al. 2014), represent another category of advanced topic models, albeit less frequently relied upon in innovation research than the advanced topic model categories discussed in the present study.

For more technical details, interested readers are referred to the original works of the discussed topic modeling approaches.

Frequently used quality indicators to determine the optimal number of topics in innovation research are topic coherence, divergence, perplexity and log- likelihood (e.g., Chunmian et al. 2022; Gregoriades and Pampaka 2020).

The Pachinko allocation model (PAM) is another topic correlation model (Li and McCallum 2006) that is considerably less prominent in innovation research than the other approaches covered herein.

Existing innovation research has relied on marginal corpus likelihood to determine the optimal number of topics (e.g., Noh et al. 2021).

Existing innovation research has used topic coherence, log-likelihood, perplexity, exclusivity and coherence to determine the optimal number of topics (e.g., Mele et al. 2019; Müller- Hansen et al. 2021).

Existing innovation research has employed semantic coherence and exclusivity to determine the optimal number of topics (e.g., Dehler- Holland et al. 2022).

Using the Dimensions database (Wetzels et al. 2023), we carried out (a) an unrestricted search using (different variants of) the search terms “topic model” in combination with “innovation”, (b) an unrestricted search using

(different variants of) the search terms "topic model" in combination with either "new product development," "new service development," "idea generation," "ideation," "post- launch," "product launch," "product evaluation," "service evaluation," "customer feedback," "lead user," "thought leader" or "opinion leader," and, (c) a restricted search using (different variants of) the search terms "topic model" in innovation management-dedicated journals, that is, *JPIM*, *Journal of Innovation and Knowledge*, *Technological Forecasting and Social Change*, and *Technovation*.

Some articles are matched with multiple topic modeling uses. In line with common practice (e.g., Lee et al. 2024), two authors independently coded the list of 4274 articles—hits on the employed search terms in journals with an impact factor greater than 4—using as selection criteria that articles should be relevant to both innovation management and topic modeling. Similarly, a different combination of authors mapped the resulting 1099 articles to the components of the proposed framework, using protocols based on Cooper (1990)'s definition of the stages of the innovation process. For both exercises, any discrepancies were resolved by consensus.

In HDP, "hierarchical" applies to the addition of another level to the Dirichlet process and not to organizing topics hierarchically, so topic organization remains flat, similar to LDA (Krasnov and Sen 2019).

We thank the anonymous reviewer for the helpful suggestion to include this table.

After screening, we omit records of abstracts, book reviews, calls for papers, commentaries, corrigendum, editorials, errata, indexes, notes, special issue introductions, spotlight articles, rejoinders, and retractions.

Following standard preprocessing procedures, 189 articles were dropped from the analyses due to sparsity.

Examining the topic correlations reveals rather independent research themes overall. A positive correlation was found only between the two research themes on cross- functional integration and NPD success and failure, indicating that these research themes are likely to be discussed together within a document (Roberts et al. 2014).

References

Andersen, R., & Hackos, J. A. (2018). Increasing the value and accessibility of academic research: Perspectives from industry. In *Proceedings of the 36th ACM International Conference on the Design of Communication* (pp. 1–10). Association for Computing Machinery.

Antons, D., Kleer, R., & Salge, T. O. (2015). 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.

Appio, F. P., Frattini, F., Petruzzelli, A. M., & Neirotti, P. (2021). Digital transformation and innovation management: A synthesis of existing research and an agenda for future studies. *Journal of Product Innovation Management, 38*(1), 4–20.

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science, 46*(4), 557–590.

Banholzer, M., Doherty, R., Morris, A., & Schwaitzberg, S. (2023). Innovative growers: A view from the top. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/innovative-growers-a-view-from-the-top

Banville, C., & Landry, M. (1989). Can the field of MIS be disciplined? *Communications of the ACM, 32*(1), 48–60.

Barczak, G. (2017). Writing a review paper. *Journal of Product Innovation Management, 34*(2), 120–121.

Barravecchia, F., Mastrogiacomo, L., & Franceschini, F. (2023). Product quality tracking based on digital voice-of-customers. *Total Quality Management & Business Excellence, 34*(11–12), 1386–1409.

Becker, L., Coussement, K., Büttgen, M., & Weber, E. (2022). Leadership in innovation communities: The impact of transformational leadership language on member participation. *Journal of Product Innovation Management, 39*(3), 371–393.

Bernier, C., DiMaggio, P., & Heckscher, C. (2023). When content is king: Using topic models to analyze online innovation crowdsourcing. *Innovation: Organization & Management, 25*(2), 177–200.

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

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM, 57*(2), 1–30.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120). Association for Computing Machinery.

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

Bongini, P., Osborne, F., Pedrazzoli, A., & Rossolini, M. (2022). A topic modeling analysis of white papers in security token offerings: Which topic matters for funding? *Technological Forecasting and Social Change, 184*, 122005.

Bstieler, L., Hemmert, M., & Barczak, G. (2015). Trust formation in university–industry collaborations in the U.S. biotechnology industry: IP policies, shared governance, and champions. *Journal of Product Innovation Management, 32*(1), 111–121.

Calantone, R., Garcia, R., & Dröge, C. (2003). The effects of environmental turbulence on new product development strategy planning. *Journal of Product Innovation Management, 20*(2), 334–347.

Capano, G., Howlett, M., Jarvis, D. S., Ramesh, M., & Goyal, N. (2020). Mobilizing policy (in) capacity to fight COVID-19: Understanding variations in state responses. *Policy and Society, 39*(3), 285–308.

Chen, X., Chua, A. Y., & Pee, L. G. (2022). Who sells knowledge online? An exploratory study of knowledge celebrities in China. *Internet Research, 32*(3), 916–942.

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures, 17*(2), 111–130.

Choi, J., Menon, A., & Tabakovic, H. (2021). Using machine learning to revisit the diversification–performance relationship. *Strategic Management Journal, 42*(9), 1632–1661.

Choi, K. H., & Kwon, G. H. (2023). Strategies for sensing innovation opportunities in smart grids: In the perspective of interactive relationships between science, technology, and business. *Technological Forecasting and Social Change, 187*, 122210.

Chunmian, G., Shi, H., Jiang, J., & Xu, X. (2022). Investigating the demand for blockchain talents in the recruitment market: Evidence from topic modeling analysis on job postings. *Information & Management, 59*(7), 103513.

Claudy, M. C., Peterson, M., & Pagell, M. (2016). The roles of sustainability orientation and market knowledge competence in new product development success. *Journal of Product Innovation Management, 33*(S1), 72–85.

Cooper, R. G. (1990). Stage-gate systems: A new tool for managing new products. *Business Horizons, 33*(3), 44–54.

Cooper, R. G. (2008). Perspective: The stage-gate idea-to-launch process—Update, what's new, and nexgen systems. *Journal of Product Innovation Management, 25*(3), 213–232.

Cooper, R. G. (2021). Accelerating innovation: Some lessons from the pandemic. *Journal of Product Innovation Management, 38*(2), 221–232.

Cooper, R. G., & Kleinschmidt, E. J. (1994). Determinants of timeliness in product development. *Journal of Product Innovation Management, 11*(5), 381–396.

Cooper, R. G., & Sommer, A. F. (2016). The agile–stage-gate hybrid model: A promising new approach and a new research opportunity. *Journal of Product Innovation Management, 33*(5), 513–526.

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models*. Guilford Press.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.