Shahgholian, A, Odacioglu, EC, Zhang, L and Allmendinger, R

Big Textual Data Research for Operations Management: Topic Modeling with Grounded Theory

Big Textual Data Research for Operations Management: Topic Modeling with Grounded Theory

Eyyub C. Odacioglu

Department of Engineering Management, The University of Manchester, Manchester, UK

Lihong Zhang

Department of Engineering Management, The University of Manchester, Manchester, UK

Richard Allmendinger

Alliance Manchester Business School, The University of Manchester, Manchester, UK

Azar Shahgholian

Liverpool Business School, Liverpool John Moores University, Liverpool, UK

Corresponding author:

Eyyub C. Odacioglu can be contacted at: eyyub.odacioglu@manchester.ac.uk

Abstract

Purpose

There is a growing need for methodological plurality in advancing Operations Management (OM) research, especially with the emergence of Machine Learning (ML) techniques for analysing extensive textual data. To bridge this knowledge gap, this paper introduces a new methodology that combines ML techniques with traditional qualitative approaches, aiming to reconstruct knowledge from existing publications.

Design/methodology/approach

In this pragmatist-rooted abductive method where human-machine interactions analyse big data, we employ Topic Modeling (TM), an ML technique, to enable Constructivist Grounded Theory (CGT). A four-step coding process (Raw Coding, Expert Coding, Focused Coding, and Theoretical Coding) is deployed to strive for procedural and interpretive rigour. To demonstrate the approach, we collected data from an open-source professional Project Management (PM) website and illustrated our research design and data analysis leading to theory development.

Findings

Our results show that TM significantly improves the ability of researchers to systematically investigate and interpret codes generated from large textual data, thus contributing to theory building.

Originality/value

This paper presents a novel approach that integrates an ML-based technique with human hermeneutic methods for empirical studies in OM research. Using Grounded Theory (GT), this method reconstructs latent knowledge from massive textual data and uncovers management phenomena hidden in published data, offering a new way for academics to develop potential theories for business and management studies.

Keywords: Big Data, Grounded Theory, Machine Learning, Topic Modeling
Paper Type: Research Paper

1. Introduction

Advancements in digital technologies have opened new opportunities to collect and analyse large data sets. Big data can be collected from diverse sources, such as documents, web, social media, ERP and CRM systems, and cloud platforms, in various formats. Big data analysis techniques cover a broad spectrum, including statistics, machine learning, data mining, and optimisation (Choi et al., 2018; Hassani et al., 2020). This enables organisations to make informed decisions and optimise their processes (Feng and Shanthikumar, 2018; Matthias et al., 2017). A significant facet of big data is large textual data, from which organisations can derive insights about their operations and understand user perspectives and intentions through opinion and sentiment analysis (Beheshti-Kashi et al., 2018; Hassani et al., 2020), assess their service quality (Mejia et al., 2021), and comprehend people's behaviour during crises such as the COVID-19 pandemic and its implications on supply chains (Wilk et al., 2023). For the Operations Management (OM) research community, scholars underscore the exploration of Machine Learning (ML) methods and big textual data techniques to foster theory development, as highlighted in the literature (Bansal et al., 2020; Chou et al., 2023).

One such method that OM researchers have utilised to contribute to theory is Grounded Theory (GT) (Chenger and Pettigrew, 2023; Guo et al., 2022; Lu et al., 2023). However, scholars approach GT cautiously due to the challenges inherent in qualitative research, such as the lack of quality criteria, specific skill requirements, time and effort, generalisation from limited samples, and the need to develop rigorous research designs (Tracy, 2010). Potential misuse of qualitative methods raises concerns among them, as it may lead to weak and questionable research outcomes (Taylor and Taylor, 2009). To address these concerns, scholars such as Gioia et al. (2013), Strauss and Corbin (1997), and Binder and Edwards (2010) provided guidelines for conducting rigorous studies and ensuring methodological rigour and research quality. However, the introduction of big data into research has imposed a re-evaluation of conventional methods heavily reliant on human effort, as their effectiveness has been diminished. Although theory abstraction by itself is traditionally challenging with conventional methods (Bak, 2005), the amount of data in big data research exceeds a researcher's data analysis ability (Lesnikowski et al., 2019). Consequently, manual annotation (coding) or NVivo coding presents significant challenges when dealing with large data sets.

In this study, we proposed to combine a big data technique with a GT method to address this issue. Big data techniques have the potential to facilitate research processes involving vast amounts of data (Baumer et al., 2017; Schmiedel et al., 2019), and among these techniques, Topic Modeling (TM) has emerged as a promising approach due to its inductive and exploratory nature (Ignatow and Mihalcea, 2017). TM simplifies the text exploration process by generating a set of interpretable codes that allow researchers to code a vast amount of data (Dimaggio et al., 2013). Over the past two decades, we have witnessed the emergence of TM in different management areas (Hannigan et al., 2019), such as tourism (Lau et al., 2005), information systems (Muresan and Harper, 2004), organisational research (Schmiedel et al., 2019), and marketing (Lawrence et al., 2010). However, the application of ML methods to support OM research is limited, and there is a call for the application of these methods (Chou et al., 2023), particularly a call for the application of TM in OM (Bansal et al., 2020). In this work, we aim to contribute to the field by combining TM with a GT method to address the aforementioned gaps. We contend that this integration has the potential to advance both OM research and practices, allowing scholars to make significant contributions to theory in future research.

Therefore, we address the following question: How can a big data technique, i.e., TM, be embedded in a conventional analytical process, i.e., GT, so as to generate new knowledge in OM research?

Recently, Baumer et al. (2017) have compared the processes of a GT approach with TM and have highlighted potential contributions that researchers can make to the development of new knowledge using these two methods. Bryant (2017) also discussed the opportunities that big data approaches bring for GT approaches. More recently, Croidieu and Kim (2018) and Odacioglu et al. (2022) utilised such a method. We have considered integrating the TM algorithm of Blei et al. (2003) into the Constructivist GT (CGT) method of Charmaz (2006) to enhance interpretation through the analytic process and establish procedural rigour with clearly designed steps to research.

This study primarily aims to illustrate the method's process with an example for OM researchers and to increase awareness of ML applications for OM researchers and practitioners. We illustrated the method by providing an example in the Project Management (PM) field as a specific area of OM where organisations dynamically generate a vast amount of textual content, making it challenging to analyse these documents in real-time. However, the proposed methodology empowers practitioners to overcome this hurdle and effectively analyse the ever-flowing textual data. Using this method, practitioners can effectively identify hidden meanings within textual data. Through the sense-making process of these meanings and building links between them, they gain the ability to make well-informed decisions promptly and with ease, enabling them to respond efficiently.

The following section places the proposed methodology in context with existing research related to GT and TM. Section 3 introduces the proposed methodology with the data collection and preparation processes. The demonstration of the application of the methodology is presented in Section 4. A discussion of the findings is presented in Section 5, and the paper is concluded in Section 6.

2. Literature Review

Experience has shown that new research methods are best introduced using examples and empirical studies (Bryant, 2017; Feagin et al., 2016; Jacobs and Tschötschel, 2019; Kotzab et al., 2006; Morse et al., 2021). This approach not only facilitates audience comprehension and their understanding of the methodology's steps with the underlying assumptions but also provides researchers with a comprehensive learning experience while promoting transparency and reproducibility in research. Hence, researchers in the field of OM have produced a variety of guidelines focusing on different research methodologies. For instance, Will M. Bertrand and Fransoo (2002) have provided guidelines for quantitative models, Forza (2002) for survey research, Coughlan and Coghlan (2002) for Action Research, Stuart et al. (2002) and Voss et al. (2002) for Case Study, and Binder and Edwards (2010) for GT. Despite these efforts to operationalise different methods and encourage methodological contributions alongside theoretical and practical advancements (Shang and Rönkkö, 2022), OM researchers seem cautious regarding new methodologies; TM could be an example of that.

Our analysis of the literature revealed that a scarce number of OM scholars used TM in their work with limited transparency in the research method (Dominguez-Péry et al., 2021; Kinra et al., 2020; Ko et al., 2019; Xiao et al., 2021). Additionally, these studies are mainly limited to historical or archival textual data analysis. The latest trend in TM approaches is to extract insights (Hannigan et al., 2019), and the future is to reconstruct knowledge and develop a theory. Recognising the need for increased awareness and understanding of theorising with big textual data using TM, particularly within the OM context, Bansal et al. (2020) shed light on this topic and emphasised critical considerations for conducting research with TM. There is a growing call for utilising ML-based big textual data analysis techniques, including TM, in OM research (Bansal et al., 2020; Chou et al., 2023). To address the mentioned gaps and demands, we aim to bridge this gap and provide methodological guidance by employing a bricolage method, supplemented with an illustrative example, which has been conducted in a controlled medium.

To structure our discussion, we will begin by providing an overview of GT and Big Data. Next, we will delve into prior research on theory building with TM.

2.1. Grounded Theory and Brief Historical Evolution

Glaser and Strauss (1967) introduced GT as a method for systematically analysing data to derive theory. It relies on inductive data coding and applies constant comparison to build categories to reach abstraction. Charmaz (2017a) described this method as an iterative process and added that the method involves coding, simultaneous data collection, and data analysis to generate analytic categories based on codes. Three GT schools of thought (Straussian, Glaserian, and Charmazian) are widely accepted and used in the literature (Díaz et al., 2022; Morse et al., 2021). The first school of thought is Straussian (Strauss and Corbin, 1990), which is the widely utilised version of GT. However, Glaser (1992) criticised the prescriptive nature of the Straussian approach and argued that the procedural steps force theory onto data rather than theory emerging from data. Glaser (1992) presented the Classical GT, staying true to the positivist nature of the original version. Classical GT highlights the significant role of data and emphasises that GT is all about data, and anything can be data (Morse et al., 2021). Afterwards, Charmaz (2006) combined the two schools to form CGT by adding her own interpretation. Table I illustrates a comparison of these schools.

Table I. Comparison of Three Schools of Thought in GT

	Glaserian	**Straussian**	**Charmazian**
Ontology	Positivism	Interpretivist	Constructivist
Epistemology	Objective	Pragmatic	Subjective
Philosophical Influence	Free from influence	Interpretivism	Constructivism with Pragmatism
Reasoning	Purley inductive	Inductive but allows abductive reasoning	Inductive coding with abductive reasoning
Researchers Role	Unbiased researcher, observer, distant and detached	Engaged and actively interprets data	Constructs data and the theory
Literature Review	Allowed after data analysis	Allowed before and during the data collection process	Depends on the preference of the researcher
Data Coding and Analysis	I. Open Coding, II. Selective Coding, III. Theoretical Coding	I. Open Coding, II. Axial Coding, III. Selective Coding	I. Initial Coding, II. Focused Coding, III. Theoretical Coding
Adapted From: McCall and Edwards (2021) and Sebastian (2019)			

For the proposed methodology, we have selected CGT due to its ability to combine the strengths of two schools of thought, promoting active involvement of participants and the contribution of collaborative insights. CGT enables a nuanced exploration of research topics by facilitating the co-construction of knowledge between the researcher and the participants. Acknowledging the inherent subjectivity of human involvement, we believe that the idea of an unbiased researcher and objective approach is ideal but not possible. The researcher engages in and actively interprets data and constructs data and theory, i.e., labelling as initial coding and categorising as focused coding and building links between these categories to build a theory. For everything that a human is involved in, we talk about biases, e.g., human biases influence AI. A challenge we need to address is finding a solution to reduce the biases of researchers.

2.1.1. Constructivist Grounded Theory (CGT)

The CGT originated as a product of two existing schools of thought and combined the strengths of its predecessors through a distinct epistemological orientation, a nuanced conception of the researcher's roles and background, recognition of diverse realities and values, emphasis on the importance of data, and contextualisation of research (Charmaz, 2017a). Although it rejected the positivist orientation (the idea of an ideal unbiased researcher) by acknowledging the researcher’s role in construction, it stressed methodical flexibility and recognised the possibilities of emergent data interpretation (Morse et al., 2021). CGT emphasises that scholars should clarify their biases, beliefs, and positions (i.e., reflexivity) while conducting research (Thornberg and Dunne, 2019). Although its constructivist nature assumes that multiple realities exist ontologically, it recognises the interactive and interpretive nature of the theory construction process and emphasises the researcher's role during the process (Inaba and Kakai, 2019). The researcher's role, position, and reflexivity form the basis of the CGT, with its roots in pragmatism during the study (Charmaz and Keller, 2016; Clarke and Charmaz, 2019). These key characteristics are vital for the proposed methodology and form the basis of it because the researcher has a significant role during the analysis process.

CGT is an emergent abductive method that starts with empirical data and develops an inductive understanding of a phenomenon (Charmaz, 2017b). It evolves from an inductive inquiry to a pragmatic abductive analysis in two ways: The researcher first casts doubt into the theory abstraction process and, second, reshapes the research questions that will trigger new data collection and analysis (Morgan, 2020). The method relies on the analytic tools of GT and reflexivity to support the researcher's reflective thinking, such as constant comparison (of data, codes, and categories), iteration, memo writing, and reasoning (Charmaz, 2017a; Morse et al., 2021). It starts with coding and completes with developing a substantive theory, i.e., a real-life problem concerning theories in a particular context grounded in data (Charmaz, 2006; Morse et al., 2021). It consists of three coding steps: open, simple, and provisional *Initial Coding*; directed, selective, and categorised *Focused Coding*; and substantive constructs *Theoretical Coding*, and its strength comes from process flexibility and interpretation (Charmaz, 2006). These steps also influenced the proposed method.

2.2. Big (textual) Data; Opportunities and Challenges

Although scholars have been discussing big data for a while, Hu et al. (2014) emphasised that there is no consensus on a single definition of big data. However, they found that the definitions are centred around 3+1 Vs, which are volume (amount of data), variety (diverse types of data), velocity (data generation speed), and value (potential insights that can be derived from data). By adding veracity (trustworthiness of data), Mills (2019) (p.10) defined big data as *"rapidly generated, digitally encoded information of significant volume, velocity, variety, value, and veracity"*. Although this understanding aligns with our understanding of big data in general, these definitions fall short of reflecting our understanding of big textual data. The closest definition for big textual data can be the definition of Mills (2018) (p. 591), which is *"digitally encoded (qualitative) information of unprecedented scope or scale about a phenomenon"*. Nevertheless, this definition overlooks a crucial aspect, which is capability. Hu et al. (2014) presented another definition, describing big data as data size that exceeds the capturing, storing, managing, and analysing capability of typical software tools. Although this definition lacks clarity regarding the specific characteristics of a typical software tool, it is evident that textual data requires human involvement. To address this, we propose enhancing the Mills (2018) definition by adding the following clarification: "data size that exceeds a human's analysis and interpretation capability in effectively handling and making sense of textual data under constraints such as time".

Management researchers swiftly adapt to new techniques, including big data approaches, to make an impact in the field (Arora et al., 2016; George et al., 2014). However, big data brings challenges and opportunities for both practitioners (Kache and Seuring, 2017) and researchers (Favaretto et al., 2020). For example, for practice in relation to emerging Industry 4.0 technologies such as the Internet of Things (IoT), big data applications improve productivity, enhance operational efficiency, facilitate process optimisation with real-time data processing, and enable organisations to make better predictions for the future (Chase Jr., 2013; Mithas et al., 2022). Although organisations have recognised the competitive advantage that big data can provide, many still lack the necessary knowledge to leverage it effectively (Guha and Kumar, 2018). Despite the availability of vast amounts of data, the main challenge for organisations is not in acquiring the required technologies or collecting the data (Chase Jr., 2013). Instead, the difficulty lies in transforming this abundance of data into practical insights that can significantly improve decision-making processes (Feng and Shanthikumar, 2018).

For researchers, the realm of big data presents a distinct narrative, and the literature has shown that although extensive research has been conducted on the technologies and analytical methods required to unlock the potential value of big data, empirical research on big data is not yet widespread in the OM community (Matthias et al., 2017). While it requires a new skill set and extensive effort to learn these skills, granularity and the amount of data provide opportunities to solve daily life business problems and identify better answers to further research questions (George et al., 2016). It challenges conventional research methods. Quantitative studies make use of big data through three data analytic clusters: descriptive, predictive, and prescriptive(Chong and Shi, 2015), whereas qualitative studies employ big textual data approaches, including TM, text classification, and text clustering (Kobayashi et al., 2018). With its inductive nature, TM is the most suggested (if not the most utilised) method for exploratory studies (Hannigan et al., 2019; Ignatow and Mihalcea, 2017). Its inductive nature makes it particularly suitable for uncovering latent patterns and themes within big textual data.

2.2.1. Topic Modeling (TM)

Recent advances in AI and its widespread applications have dramatically reshaped our perception of AI, ML, and big data, resulting in increased public awareness and a surge of curiosity. Among these applications, TM can explore large textual data sets and inductively generate codes among large corpora (Dimaggio, 2015; Ignatow and Mihalcea, 2017; Lee et al., 2020). It is not a standalone method to apply and develop a theory. It can explore new patterns and hidden topics in a large corpus that a researcher might miss with hand-coding (Dimaggio et al., 2013). However, in regard to sense-making and theorising in management research, human interpretation is required to understand the results of TM (Choi and Song, 2018; Han et al., 2021; Moro et al., 2020). While the possibility of AI replacing human interpretation cannot be ruled out, currently, AI and computer-based tools are increasingly used for supportive purposes, emphasising supplementing rather than replacing human judgment (Van Eck and Waltman, 2014).

For management research, TM is a complementary methodological approach to assist researchers in analysing a management phenomenon (Schmiedel et al., 2019). The current trend in TM research is an algorithm that mines text to generate topics, and researchers render knowledge with pattern recognition-based qualitative analysis methods (Bansal et al., 2020) as in coding-based approaches (Gioia, 2021). For example, researchers utilised TM with thematic analysis (Piepenbrink and Gaur, 2017), discourse analysis (Aranda et al., 2021), and content analysis (Chuang et al., 2014) in their studies. However, these analysis methods are lacking behind GT for theory building (Fairhurst and Putnam, 2019; Hsieh and Shannon, 2005; Nowell et al., 2017). There is a need for a method that can analyse big textual data and support theorising with a robust process to overcome the shortcomings of existing template-based methodologies and highlight interpretive rigour with abductive reasoning.

2.3. Prior Research for Rendering Theory with TM

The potential of TM for theory contribution to management research is being accepted in the literature, and the growing number of publications proves that TM can facilitate knowledge (re-)construction for management research. Hannigan et al. (2019) reviewed the literature and illustrated novel knowledge identification applications that used TM. Nelson (2020) built a computational GT for a sociology study and used TM and other computer-based pattern recognition techniques to identify linguistic relations to build their theory. In a researcher-driven study, Croidieu and Kim (2018) combined the Straussian GT with TM. Baumer et al. (2017) conducted a comparison of GT and TM in the context of management research. They simultaneously applied both GT and TM, along with content analysis, to the same dataset. This approach aimed to highlight similarities in data collection, analysis procedures, and outcomes. The authors concluded that a novel combination of GT and TM is needed. This study also has proven that these methods dramatically shorten the analysis time compared to conventional methods. This call is meaningful because qualitative researchers (especially grounded theorists) are looking for new methodologies that combine the theory building power of GT with big data approaches (Walsh et al., 2015). These studies influenced us to develop this new methodology.

The similarities between CGT and TM made it reasonable to integrate one into another. As argued in previous sections, CGT and TM share the significance of data, pragmatic roots, methodological steps to explore the (latent/hidden) meaning in data, and the requirement of a high level of human interpretation (Baumer et al., 2017; Cintron and Montrosse-Moorhead, 2021; Li et al., 2021). Both methods start with inductive coding for exploratory purposes, and the human interpretation takes the study toward abductive reasoning and theory building (Charmaz, 2006; Hannigan et al., 2019). The technical, epistemological, and ontological fit makes it feasible to combine TM and CGT (Table II).

Table II. Comparisons between CGT and TM

Constructivist Grounded Theory (Fairhurst and Putnam, 2019)	**Topic Modeling**
Emergent and unfolding from coding and abstracting to categories	Emergent and unfolding from coding and abstracting to patterns or themes (Isoaho et al., 2021; Schmiedel et al., 2019)
Iterates between data and analysis	An Iterative algorithm to build topics and human iterates within codes (Blei et al., 2003; Croidieu and Kim, 2018)
Inductive coding / Abductive with the human hermeneutic.	Inductive topic modeling / Abductive human labelling and hermeneutic (Kaplan and Vakili, 2015; Shrestha et al., 2021)
Language as representing phenomena or categories	Language as representing topics and meanings (Jacobs and Tschötschel, 2019)
Aim to develop a "mid-range (substantive) theory"	Allow researchers to make sense of topics to develop a theory (Hannigan et al., 2019; Kobayashi et al., 2018)
Follows systematic coding steps	Follows systematic data processing and qualitative analysis steps (Hannigan et al., 2019; Schmiedel et al., 2019)
Requires contextual knowledge to understand meanings	Requires contextual knowledge to interpret topics (Baumer et al., 2017)
Pragmatist Roots focus on meanings and actions	Interpretivist Roots focus on latent meanings, and Pragmatic Roots focus on real-life implications (Cintron and Montrosse-Moorhead, 2021; Li et al., 2021)
Relies on rich data, and the researcher needs to stay close to the data	Relies on big data and staying close to data is preferred but not necessary (Baumer et al., 2017; Hannigan et al., 2019)

3. Research Design

This section presents the methodology that embeds TM into CGT to overcome the issues associated with conventional GT. The proposed methodology aims to identify the latent meanings and knowledge in big textual data and interpret them for knowledge reconstruction (Figure 1).

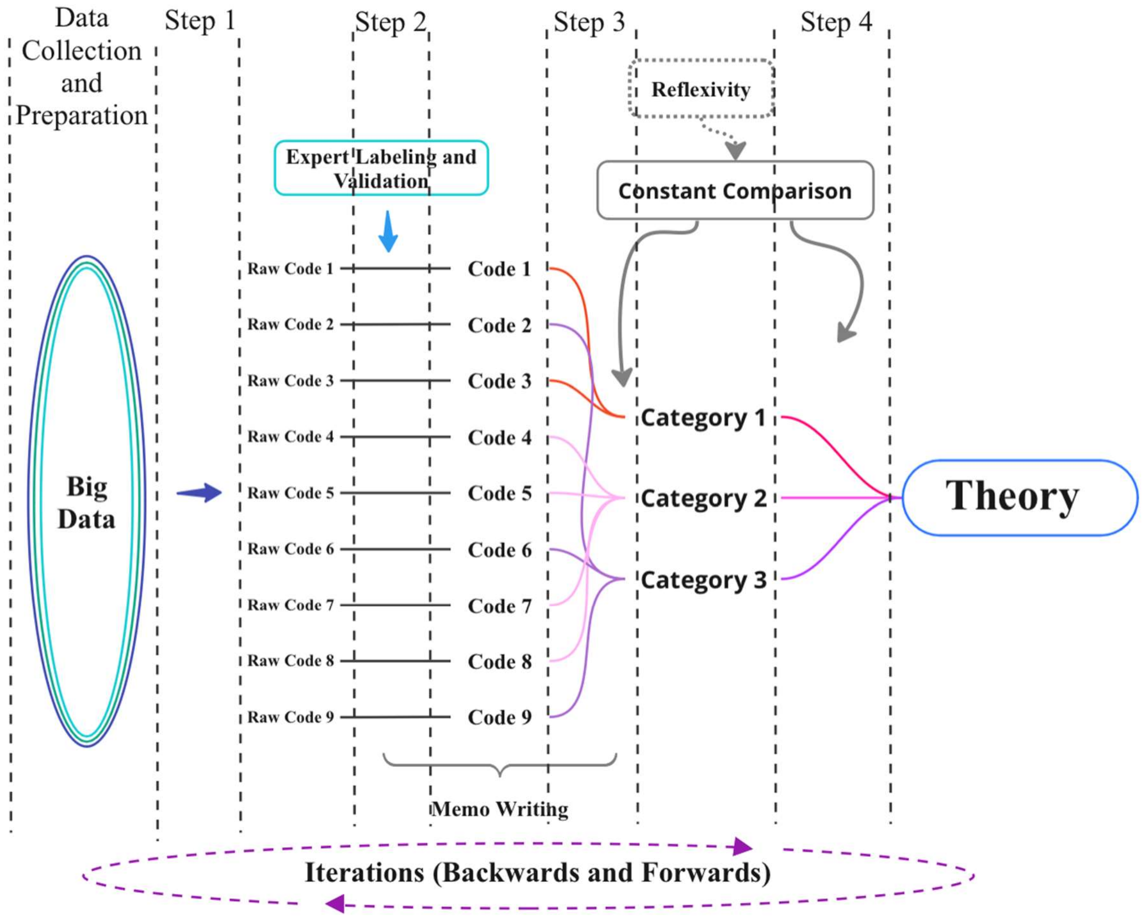

Figure 1. Steps of the proposed methodology

3.1. Data Collection and Preparation

TM utilises diverse data from various sources, both online and offline. Online sources include websites of organisations (Maier et al., 2018), customer products or service reviews (Schmiedel et al., 2019), Google reviews (Lee and Yu, 2018), and social media content (Baumer et al., 2017). Offline sources include archives or historical documents (Croidieu and Kim, 2018), open-ended survey responses (Cintron and Montrosse-Moorhead, 2021), and even existing research data collected for other studies (Nelson, 2020). In this digital era, the availability of data has increased significantly, enabling access from publicly available online platforms to a wide range of textual content on OM-related topics. For instance, reports on supply chain management and circular economy from open-source government or organisation databases (e.g., European Commission, OECD, and World Bank) are valuable resources. Social media data can also be utilised to assess public opinion on a provided service. Additionally, companies' annual reports or market announcements can provide valuable insights into operational performance. In their study, Bansal et al. (2020) provided even more specific sources for further exploration.

For our example, we have compiled data from the Project Management Institute (PMI) official website (www.pmi.org). The PMI, a non-profit organisation established in 1969, stands as a paramount global professional association dedicated to the advancement of the project management discipline. PMI aims to enhance the status of project management as a well-recognised profession. Their contribution includes the cultivation of industry benchmarks, the encouragement of professional growth, and the provision of a wide range of resources aimed at helping professionals improve their project management skills and abilities. PMI provides the most current and comprehensive understanding at their official website. This website serves as a vital hub for professionals, facilitating networking and community building. It offers guidance on certifications and professional development while acting as a central source for industry standards and best practices – essential for project managers worldwide. It provides a wealth of publications and resources, including articles, webinars, and industry insights that improve knowledge and skills in PM.

Once the corpus is ready, any text analysis method requires pre-processing before applying the technique itself for data preparation (Kobayashi et al., 2018). The pre-processing consists of cleaning up the corpus, such as removing stop words, prefixes and suffixes, and extra white spaces, to make it ready for the algorithm to process (Nelson, 2020; Schmiedel et al., 2019). Following this pre-processing is the Raw Coding.

3.2. Step 1 Raw Coding

The pre-processing is followed by TM processing. We used the Python programming language (Version 3.9.13) on the Anaconda-Spyder platform (Version 5.2.2) and the Gensim topic-modelling Python library (Version 4.1.2). We called TM algorithm-generated topics as raw codes. To produce these raw codes, we chose a widely used algorithm of Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Dimaggio et al., 2013; Hannigan et al., 2019; Lancichinetti et al., 2015). Before running the algorithm, a decision must be made on the hyperparameter settings: α (determines the distribution of topics in a document) and β (influences the distribution of words in a topic), and k (represents the number of topics that the model will attempt to identify in the given dataset). Further details on the algorithm and settings are provided in Appendix A.

3.3. Step 2 Expert Coding

When the algorithm generates raw codes (i.e., topics), the next step involves labelling for validation (Dimaggio et al., 2013; Gioia, 2021). Raw codes may seem like "*gibberish*", so they require human interpretation to make sense of them (Bansal et al., 2020). However, this step is prone to researchers' biases, given that researchers themselves mainly carry it out. To mitigate potential bias and ensure the validity of the generated topics, we adopted the approach of Kaplan and Vakili (2015), inviting domain experts to label the raw codes based on their own contextual knowledge and understanding. Experts observed the words within a given topic, tried to make sense of them collectively with links, labelled them, and semantically validated them by rating their relevancy (Grimmer and Stewart, 2013).

Although experts possess sound domain knowledge, they are not necessarily familiar with the research method, scope, aims, and objectives. To mitigate this, we proposed introducing the study to the experts beforehand, providing background information, and explaining the labelling and rating process. Upon completion, the raw codes turn into validated codes with labels.

3.4. Step 3 Focused Coding

The first two steps earmark the TM to provide procedural rigour, while the next two steps follow the CGT principle to qualify interpretive rigour. In CGT, the researcher focuses on the most frequently occurring codes within the data (Charmaz, 2006). TM performs this automatically and extracts frequently occurring words from big textual data, identifying hidden meanings as topics that frequently repeat patterns of co-occurring words (Brett, 2012). As there are an infinite number of word combinations and topic alternatives, the algorithm selects frequently repeated words based on the chosen hyperparameters (Dimaggio et al., 2013). Expert labelling adds human sense to topics and validates them by rating relevance (Grimmer and Stewart, 2013).

The researcher engages with codes to identify themes, patterns, and similarities within the code pool, constructing categories. This process isn't a straightforward pattern recognition; it involves answering questions about why these codes occurred and why experts labelled them in a particular way. The approach is iterative, requiring a constant back-and-forth between data, labels, and underlying reasons. Additionally, seeking further clarification from experts, if needed, helps to comprehend the broader significance of the codes and facilitates the construction of meaningful categories.

During this process, the researcher should always keep in mind the research's aim and objectives, as well as the identified knowledge gap from the literature. The researcher's own knowledge, position, and reflexivity become more apparent in this step during category development (Charmaz, 2020). Researchers should apply GT's analytic tools in this step (Morse et al., 2021), engaging in constant comparison and iteration to identify similarities and differences (Charmaz, 2006). Eventually, the researcher will discover significant categories that emerge with links between codes (Thornberg and Charmaz, 2014). Before moving to the next stage, researchers may remove unassigned codes or categories formed by the least number of codes.

3.5. Step 4 Theoretical Coding

The final step, Theoretical Coding, involves developing theoretical constructs from categories. Up until now, this method has been expected to generate data-driven and empirical codes and categories by conducting constant comparisons of data, codes, and categories (Charmaz, 2017c; Thornberg and Charmaz, 2014). This enables researchers to discover the meaning and reasoning of categories (Charmaz, 2017c). Researchers can aim to develop features and dimensions of constructs (i.e., categories) (Charmaz, 2015; Gioia et al., 2013). This involves juxtaposing categories to build meaningful links through constant comparison and identifying the ideas or meanings underlining these categories to reach analytical abstraction (Charmaz, 2006).

Identifying links and relations between categories requires critical thinking. Although the proposed methodology aims to work with big data and reduce the biases of the researcher in the initial steps, the researcher may delve deeper in this step, performing multiple zoom-ins and zoom-outs on each category. This involves revisiting categories, codes, and even the original data, making observations and reasoning between categories to search for similarities, differences, and theoretical connections, and building a meaningful link between pertinent categories. Hence, the abstraction process should result in a real-life problem concerning substantive theory (Easterby-Smith et al., 2018; Morse et al., 2021).

4. Methodology Demonstration

The previous section introduced the steps of the proposed methodology. This section illustrates each step with an example. As applied in the literature (Croidieu and Kim, 2018; Jacobs and Tschötschel, 2019; Kaplan and Vakili, 2015; Nelson, 2020), we selected a particular subject for the demonstration. Hence, we chose collaboration and innovation intensive Complex Innovation Projects (CIPs) to exemplify the process due to our vested interest in this area.

4.1. Complex Innovation Projects

The positive impact of collaborative efforts of various internal and external stakeholders in innovation, such as clients, different levels of suppliers, and internal teams, has been long discussed and proven in OM (Bahemia et al., 2017). For example, scholars highlighted the benefits of collaborative innovation, particularly when incorporating new technological tools, as it can enhance the innovation process and lead to operational synergy (Esposito De Falco et al., 2017). Furthermore, another study emphasised the significance of formal and informal collaboration mechanisms (socialisation) to foster strong relationships between partners during a project (Aaltonen and Turkulainen, 2018). It is important to note that the first study originates from the OM stream, while the second study comes from a newer stream that combines PM and OM. Although OM researchers have recognised PM as a distinct form of work due to its unique and temporary nature, they acknowledge that OM encompasses a wide range of organisational activities, from novel to routine and from variety to volume. Projects typically fall towards the novel and varied end of this spectrum, demanding different perspectives, approaches, methods, tools, and techniques compared to ongoing, repetitive operations while still being an integral part of OM (Maylor et al., 2018). In this context, our focus is on a specific type of project, known as CIPs, which entail high risk and require extensive collaboration within an ecosystem (Johnson et al., 2021), where a limited number of studies have been conducted.

CIPs deliver bespoke products, systems, technologies, and services in accordance with the requirements of a (industrial or governmental) client (Acha et al., 2004; Davies and Brady, 2000). By nature, these projects involve high cost and risk (financial and technical), require intensive engineering and technological knowledge and capabilities, and take place in project-based structures or organisations (Hobday, 1998). Aircraft, air traffic control systems, airports, mass transportation systems, manufacturing equipment, large software systems like e-government, and telecommunication networks are typical outcomes of these projects. Each project involves incremental or radical innovation (Jesus et al., 2021). These projects have a long lifecycle, and the output of these projects is structures that consist of interdependent subsystems and components (Galati et al., 2019; Park and Ji, 2015). Organisations engage in simultaneous collaborative activities with various stakeholders to reach complementary resources or share the costs and risks that occurred (Chakkol et al., 2018; Lee and Yoon, 2015). Chakkol et al. (2018) highlighted three key characteristics that make collaboration even more challenging for these projects. Firstly, collaboration is constrained by the duration of the projects (Davies and Hobday, 2005), and the time limit makes it challenging since it is not easy to build cooperative norms and mutual trust in this limited time. Secondly, the high-cost and risky nature and coordination among multiple organisations cause ambiguity in project prediction and planning (Brady et al., 2005). Third, organisational structures can be vague since numerous companies, teams, and individuals are involved in these projects. Several scholars have made significant contributions to the field of managing CIPs (see, e.g., Davies et al. (2011); Roehrich et al. (2019)), while more recent research has focused on exploring governance mechanisms for such projects (see, e.g., Aaltonen and Turkulainen (2022); Chakkol et al. (2018)). However, despite these valuable individual contributions, a comprehensive view that integrates both management and governance aspects has not yet been provided.

In fact, the existing OM literature has extensively examined the PM and innovation management faces of these projects (Crespin-Mazet et al., 2019; Davies et al., 2011; Davies et al., 2016; Hobday, 2000; Johnson et al., 2021), as well as the governance and formal and informal mechanisms associated with them (Aaltonen and Turkulainen, 2022; Chakkol et al., 2018; Johnson et al., 2021). Notably, these two aspects have often been discussed separately due to their relevance to different levels of management (Müller et al., 2014; Müller et al., 2015). By leveraging big data, we obtained a comprehensive view and highlighted crucial categories essential for CIPs through the showcase. This also validates the potential of our methodology to contribute to the theory.

4.2. Data Collection and Preparation

To ensure data validity and reliability in accordance with the objectives of the study, we used the PMI website for data collection in this example. The website provides access to a variety of documents, including academic papers from the official journal (the Project Management Journal "PMJ") and practical publications on PM. The PMJ follows a blind peer-review process, while other publications on the website undergo an editorial check. The website's built-in search function in the "Standards & Publications" section facilitated consistent and valid data retrieval. We deliberately selected this website due to its strong connection with PM and its community, as well as its extensive and up-to-date content dating back to 1970. The website comprises more than 15,000 textual publications categorised into around 45 subjects (e.g., risk management, quality management, portfolio management) and 22 industries (e.g., aerospace, infrastructure). Focusing on "collaboration" and "innovation", we used the on-site search function to retrieve relevant case studies and white papers, which serve as valuable sources of real-life knowledge, insights, and best practices.

Our search resulted in 37 case studies and 38 white papers, which were then refined by removing duplications, literature reviews, conceptual papers, training manuals, and documents that were not in textual format. This left us with 68 publications for analysis, containing sections such as abstract, introduction, literature review, methodology, findings and discussions, and conclusions. To concentrate on practical aspects, we specifically included sections tied to real-life experiences, which are findings, discussions, and conclusions. The employed algorithm compares documents to generate topics for processing. To address potential length imbalances and preserve the interpretability of results, we made a natural division and separated sections into subsections of articles. The final corpus comprises 620 documents (the unit of analysis) spanning over 480 pages and 325,000 words.

4.3. Raw Coding

We started by generating raw codes. As discussed in Section 3.2, decisions on hyperparameters and the best number of topics are vital for the TM. Since the library used learns from the corpus and automatically calculates hyperparameters, we focused on selecting the optimal number of topics. To determine *this number*, we followed heuristic approaches, as preferred by many social science scholars, which consider factors such as interpretability, stability, and scalability rather than relying solely on statistical coherence measures (Baumer et al., 2017; Croidieu and Kim, 2018; Dimaggio et al., 2013; Kaplan and Vakili, 2015; Shadrova, 2021).

Applying the heuristic approach proposed by Dimaggio et al. (2013), we chose 24 topics. Table III illustrates some of these topics. Additionally, Figure 2 illustrates the intertopic distance map (a visual representation that illustrates the similarity (or dissimilarity) between topics in a model based on their word distributions and relationships) (Sievert and Shirley, 2014). Further information about the selection process and the intertopic distance map can be found in Appendix A.

Our initial analysis suggests that Topic 4 and Topic 13 are generic codes that are represented in most documents and hinder their interpretability, so we decided to remove them.

Table III. Part of the results for 24 topics

Topic Number	Words
Topic 1	Agency, regulation, process, technology, practice, increase, development, work, compliance, develop, allow, government, impact, address, improve
Topic 2	Autonomy, stakeholder, parent, network, center, environment, type, indicator, report, appear, resource, use, success, influence, informant
Topic 3	Innovation, work, business, solution, team, activity, people, process, time, help, support, idea, requirement, mentor, customer
Topic 4	Change, process, sponsor, program, strategy, business, plan, implementation, management, benefit, define, support, state, require, assessment
… …	
Topic 24	People, quotient, course, dimension, know, intelligence, thing, level, week, code, culture, management, work, capacity, acquisition

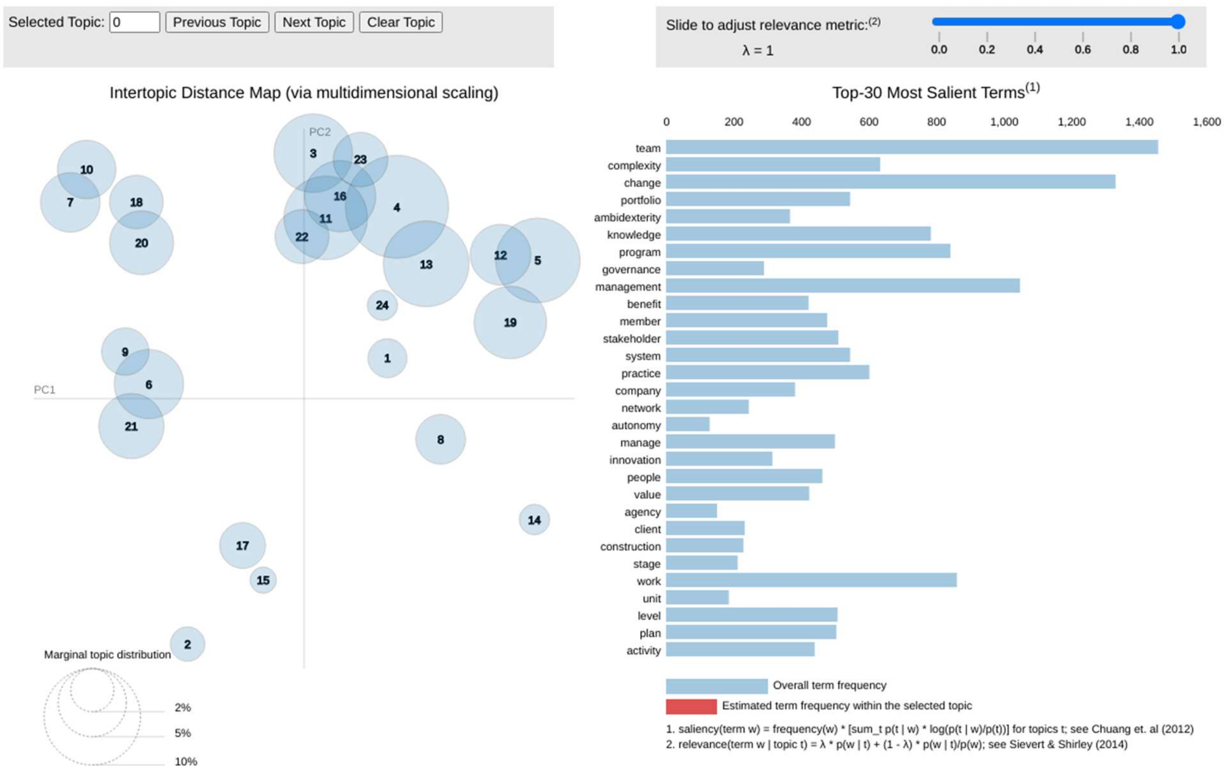

Figure 2. The topic distance map (left) and top-30 salient words (right)

TM relies on the nature of linguistics because words used together are related to a topic discussed in the content. The algorithm works based on words' co-occurrence to build a topic. However, the output, a bag of words, lacks the underlying reasons and cannot be converted into a meaningful sentence. The process of (re-)constructing knowledge from these words is a human job, best illustrated by a playground metaphor. For instance, a topic with words like "park, play, green, child, dog, ball" can be directly labelled as "playing in the park", but it raises questions about the players, the game, the purpose, and the manner of play. Only experts in the domain can interpret and draw meaningful connections between words. Interpretation can vary according to the specialisation of the expert, adding value and insights into the research. Moreover, expert labelling replaces researcher labelling to avoid biases in the validation process.

While raw codes can help to identify the data's content (i.e., PM, collaboration and innovation), they do not convey the deeper meaning behind the words and require content and context knowledge. Given the large amount of data and limited processing time, it is not possible to fully familiarise yourself with the content of the documents. However, experts with contextual knowledge play a crucial role in obtaining insights.

4.4. Expert Coding

We engaged PhD candidates as experts (will be called participants), following the approach of Kaplan and Vakili (2015). We used a snowballing method to select four PhD candidates from the Management of Projects PhD Programme who had real-life experience. At a workshop, we introduced the research and coding process, providing an example of common understanding. During the workshop, participants labelled and rated the topics, aided by sample documents showcasing topic distributions.

Participants first engage in topics to form an initial understanding before proceeding to read the given documents to build their insights. They asserted that they had utilised their intellectual creativity and conducted online searches to enhance their understanding of specific terms or concepts to provide labelled topics, rated for semantic validation (the topic's relation to sample documents: 1 unrelated, 2 somewhat related, 3 quite related, 4 highly related, 5 related), and provided a brief reason for labels.

The next step involves constructing labels for each topic, essentially finalising the topic-labelling process to convert them into codes. To begin, we analysed the topics, considering the given labels and calculating the average semantic validation ratings. Initially, we planned to remove topics below the threshold of 2 (Topics 3, 21, and 22), but we decided to include them in the analysis and asked participants to determine whether to keep or remove them.

During the analysis, we observed variations in semantic validation rates yet observed similarities in some labels, with four results being nearly identical. Accordingly, we synthesised a single label for those four results and two alternative labels for the rest.

Subsequently, we allowed participants to choose an alternative label and rate their confidence level for that label on a scale of 1 to 5. We also allowed them to retain their initial labels or suggest a new one if they were not confident in constructed labels. Furthermore, we sought their input on whether to keep or remove topics with low semantic validation ratings. Participants approved the labels and recommended keeping two out of three topics with low semantic validation rates. With this process, the Expert Coding stage was completed, yielding 20 codes for the Focused Coding process.

4.5. Focused Coding

Focused Coding is a category-building process that eventually forms a set of constructs for a conceptual/theoretical framework. There are two contrasting approaches at the edge of the category-constructing spectrum: the statistical and subjective approaches. Researchers using the statistical approach develop categories by grouping related or overlapped bubbles on the intertopic distance map (Section 4.3). This method begins with code grouping as suggested on the map and progresses to understanding, identifying, and defining the links between the codes. The other end is the subjective approach rooted in the interpretivist nature of humans (Glaser and Strauss, 1967). It starts from scratch, juxtaposing and connecting the codes based on their attributes and potential links. The conclusions and findings derived from these two data approaches may, therefore, vary based on research questions and researchers' interests.

It could be argued that the statistical relevance may ignore the hidden links between the theoretical constructs, while the subjective representation would also carry the residue of biases. Therefore, to address these above drawbacks, we consider the heuristic approach between the two, which is based on experts' context knowledge and analytical preference (i.e., hermeneutic orientation) (Mees-Buss et al., 2022). We initially used the statistical categories as tentative categories and then engaged in an iterative process involving constant comparison of codes. Throughout this process, we consulted with participants, participant-given labels, participant reasons, and scholarly works to build our analytical understanding. This iterative approach allowed us to break down the tentative categories and identify new ones, resulting in a robust and nuanced analytical understanding.

In our case, there exist two pathways to abstract codes into categories. Firstly, one code can be used/shared by multiple categories. This data strategy is well supported by the conventional thematic analysis, where an interviewee's response (i.e., code) can be used for different themes (i.e., categories). For example, we allocated Topic 15 into two categories (*Collaborative Project Planning and Control, Selecting Right Projects*). Second, the fact that multiple codes can be analysed in one category is the same. For example, the *Collaborative Governance category* consists of codes of Topics 1, 2, 16, and 17.

Our experience shows that this is not as straightforward as we initially thought. For example, we originally planned to remove one code during the initial iterations (the code from Topic 14, "Effective Stakeholder Communication") because the words seemed very specific to the contents as opposed to being generic, as in Topics 4 and 13. However, our further analysis revealed that the algorithm generated this topic intersection of two distinctive contents. One was from a university project, and another was from the medicine/drug industry. We focused on its capability dimension, which can build trust, transparency, and commitment among stakeholders with communication, so categorised with other capability-related codes. We applied these two data strategies and ended up with seven categories relevant to innovation and collaboration in PM. Table IV illustrates these categories alongside the aggregated dimensions.

4.6. Theoretical Coding

Theoretical Coding is the final step we use to merge categories into dimensions or constructs of a substantive theory (Charmaz, 2006). We continued a constant comparison between categories and examined their higher-level links to build constructs (i.e., aggregated dimensions). This abstraction process involved zoom-ins/-outs to finally form three tentative dimensions (i.e., *Collaborative Decision-Making*, *Collaborative Execution*, and *Collaboration for Innovation*). For example, Table IV shows that *Collaborative Decision-Making* consists of two categories, which are *Collaborative Governance* and *Selecting the Right Project*. Overall, our dataset strongly supports this theoretical coding. The aggregated dimensions gathered around the collaborative, managerial, and governance activities of CIP. In our data set, 53 out of 68 publications used the term "complex," and 30 of those used "complexity".

We are very cautious about theory building in this paper, as these three dimensions emerged based on our understanding, interest, reflexivity, and interpretation. We believe that these reconstructed elements can lead to a framework as a substantive theory for the management of CIPs. Nevertheless, researchers can keep constant self-debating to seek stronger links between categories and establish better mechanisms.

Table IV. Evolvement process of Raw Codes to Aggregated Dimensions

<table>
<tr><th>Codes
Raw Codes - Label</th><th>Categories</th><th>Aggregated Dimensions</th></tr>
<tr><td>Topic 1 - Collaborative Regulation Development</td><td rowspan="4">Collaborative Governance</td><td rowspan="7">Collaborative Decision-Making</td></tr>
<tr><td>Topic 2 - Autonomy and Success in Collaboration</td></tr>
<tr><td>Topic 16 - Project Governance for Complexity</td></tr>
<tr><td>Topic 17 - Governance of Inter-Organisational Networks</td></tr>
<tr><td>Topic 8 - Management Role in Project Selection</td><td rowspan="3">Selecting the Right Project</td></tr>
<tr><td>Topic 15 - Strategic Project Prioritisation</td></tr>
<tr><td>Topic 19 - Dynamics of Project Portfolio Management</td></tr>
<tr><td>Topic 9 - Project Re-planning Process</td><td rowspan="2">Collaborative Project Planning and Control</td><td rowspan="4">Collaborative Execution</td></tr>
<tr><td>Topic 15 - Strategic Project Prioritisation</td></tr>
<tr><td>Topic 5 - Collaboration for Change Management</td><td rowspan="2">Execute, Monitor & Control</td></tr>
<tr><td>Topic 10 - Collaboration in/for Addressing complexity</td></tr>
<tr><td>Topic 3 - Fostering Innovation with Intra-Organisational Collaboration</td><td rowspan="4">Capabilities for Innovation</td><td rowspan="12">Collaboration for Innovation</td></tr>
<tr><td>Topic 6 - Project Manager’s Capability for Ambidexterity</td></tr>
<tr><td>Topic 7 - Value Co-creation</td></tr>
<tr><td>Topic 14 - Effective Stakeholder Communication</td></tr>
<tr><td>Topic 12 - Fostering Collaboration Through the Contracts</td><td rowspan="5">Collaboration Boosting Tools</td></tr>
<tr><td>Topic 18 - The influence of cultural Habits on Project Team Behaviour</td></tr>
<tr><td>Topic 20 - Distributed and Virtual Team Collaboration</td></tr>
<tr><td>Topic 22 - Uniting Teams for Collaboration</td></tr>
<tr><td>Topic 24 - Cultural Adaptation for Better Collaboration</td></tr>
<tr><td>Topic 11 - Collaborative Knowledge Ecosystems</td><td rowspan="3">Knowledge Management</td></tr>
<tr><td>Topic 18 - The influence of cultural Habits on Project Team Behaviour</td></tr>
<tr><td>Topic 23 - Knowledge Management in Collaborative Innovation</td></tr>
</table>

5. Discussion

This paper has introduced and validated a new methodology for integrating TM with GT to advance OM research. The proposed methodology opens new perspectives for OM research and has the potential to contribute to the theory of OM research.

5.1. Methodological Implications

The methodological contributions of this paper are positioned in the research stream in proposing a new methodology for OM (Alfaro-Tanco et al., 2021; Oliva, 2019; Van Aken et al., 2016). Moreover, scholars have recently called for innovative approaches to enhance the OM theory (Choi et al., 2016; Tang, 2016; Wickert et al., 2021). Our study contributes to this area by bridging conventional methodologies with big data techniques. As data is considered the new oil, the vast amount of textual data being generated necessitates OM research to keep pace and employ such techniques (Bansal et al., 2020; Chou et al., 2023). By addressing this need, we aim to drive the conversation further and encourage more studies that utilise ML and big (textual) data to contribute to the theory.

First, for TM analysis, there is no consensus on the ideal document length (whether it is a tweet, an online review, an article, or a novel). Longer texts are more likely to contain words from different and disconnected topics that TM struggles to distinguish, so scholars suggested splitting texts into segments (Shadrova, 2021; Silva et al., 2021). Our showcase revealed that even when data are collected from the same source, the length of documents can vary. Tang et al. (2014) stressed the critical role of document length, noting that the documents should ideally be sufficiently long but not excessively lengthy. We found that balancing the word count among documents is crucial for performance, as relatively shorter documents may dominate the algorithm due to the sampling processes (Attias, 1999; Blei et al., 2003), leading to the potential loss of topics in longer documents.

Second, our proposed methodology acknowledges the role of the researcher and allows insight from the participants. It allows researchers to participate in the co-creation process with pragmatic underpinnings, following abductive reasoning where researchers bring doubt to the research (Chandrasekaran et al., 2023; Van Aken et al., 2016). But more importantly, it allows and encourages participants to be involved in the research and share their insights in reasoning. The main approach in these studies is reliant on the researchers' secondary data analysis. Although two previous studies (Kaplan and Vakili, 2015; Rinke et al., 2021), engaged with experts, their involvement was limited to providing labels and semantic ratings for their study. In contrast, we contributed to these methods by seeking insights from participants by asking their reasoning for enhanced analysis CGT. This approach is particularly valuable for OM research since it predominantly deals with real-life problems and necessitates collaboration with practitioners as active participants in the research process.

5.2. Managerial Implications

The proposed methodology also holds practical implications for OM practitioners, who would benefit from this approach for their decision-making processes. As highlighted both above and in the introduction, we are all generating a large amount of textual data, including organisations, such as bid documents, contracts, procedures and policies, reports, and other related documents. By utilising this approach, practitioners can streamline their processes and make informed decisions based on insights derived from textual data.

Moreover, with a pragmatic mindset, practitioners see such big data technologies as a black box where they can put their input to generate the final output automatically. This study opens this black box to them so they can be actively involved in the process with their expert opinion, leading to precise and tailored outcomes based on their operational and research context. In other words, this enhances the effectiveness of the decision-making process for their operations and projects. Furthermore, working on these large textual data, providing expert-informed opinions, and analysing outcomes with the CGT approach enables practitioners to extract valuable latent information about their operations. This, in turn, facilitates the generation of improved documents, procedures, and policies for processes such as risk management, bid management, and organisational memory.

Notably, this methodology also benefits Small and Medium Enterprises (SMEs). While incumbents have established marketing operations, SMEs often lack the required capabilities and resources. However, with this methodology, SMEs can leverage online platforms, such as social media, to gain insight into their products and services. On social media, they are generally concerned with their number of followers but do not use them to get ideas about their services and products (Zhan et al., 2020). By gathering and analysing them in a timely manner, SMEs can gauge public perceptions, expectations, and preferences, allowing them to reshape their projects accordingly.

6. Conclusion and Future Work

In this paper, we presented a novel and robust methodology that integrates TM with GT to advance OM research. Through the application of this novel methodology to published case studies of project collaboration and innovation, we effectively identified latent topics by using the LDA algorithm and transformed them into expert codes by engaging with domain experts. We then devised a heuristic method to categorise the expert codes into focused codes and finally aggregate the categories into tentative dimensions for managing CIPs that potentially inform new theories.

While we focused on CIPs, future studies can seek to study other areas of OM, such as supply chain management, operations strategy, manufacturing, and quality management. By applying our approach to diverse OM domains, researchers can make valuable contributions to the OM theory and gain insights from different data sources, benefiting the wider OM research community. Furthermore, future studies can involve more stakeholders in research. The inclusion of a wide range of stakeholders has the potential to yield richer insights, in this manner, enhancing our understanding and contributing to the advancement of CIPs. This collaborative approach has the capacity to drive improvements and foster enhancements in the OM research domain.

Reference

Aaltonen, K. & Turkulainen, V. (2018). 'Creating relational capital through socialization in project alliances' *International Journal of Operations & Production Management*, 38 (6), pp. 1387-1421.

Aaltonen, K. & Turkulainen, V. (2022). 'Institutionalization of a collaborative governance model to deliver large, inter-organizational projects' *International Journal of Operations & Production Management*, 42 (8), pp. 1294-1328.

Acha, V., Davies, A., Hobday, M. & Salter, A. (2004). 'Exploring the capital goods economy: Complex product systems in the uk' *Industrial and Corporate Change*, 13 (3), pp. 505-529.

Alfaro-Tanco, J. A., Avella, L., Moscoso, P. & Näslund, D. (2021). 'An evaluation framework for the dual contribution of action research: Opportunities and challenges in the field of operations management' *International Journal of Qualitative Methods*, 20.

Aranda, A. M., et al. (2021). 'From big data to rich theory: Integrating critical discourse analysis with structural topic modeling' *European Management Review*, 18 (3), pp. 197-214.

Arora, A., et al. (2016). 'Question-based innovations in strategy research methods' *Strategic Management Journal*, 37 (1), pp. 3-9.

Attias, H. (1999). 'A variational baysian framework for graphical models' *Advances in neural information processing systems*, 12.

Bahemia, H., Squire, B. & Cousins, P. (2017). 'A multi-dimensional approach for managing open innovation in npd' *International Journal of Operations & Production Management*, 37 (10), pp. 1366-1385.

Bak, O. (2005). 'Towards triangulation—blending techniques in supply chain management context' *Research Methodologies in Supply Chain Management: In Collaboration with Magnus Westhaus*, pp. 331-346.

Bansal, P., Gualandris, J. & Kim, N. (2020). 'Theorizing supply chains with qualitative big data and topic modeling' *Journal of Supply Chain Management*, 56 (2), pp. 7-18.

Baumer, E. P. S., et al. (2017). 'Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?' *Journal of the Association for Information Science and Technology*, 68 (6), pp. 1397-1410.

Beheshti-Kashi, S., Buch, R., Lachaize, M. & Kinra, A. (2018). 'Big textual data in transportation: An exploration of relevant text sources': Springer International Publishing, pp. 395-399.

Binder, M. & Edwards, J. S. (2010). 'Using grounded theory method for theory building in operations management research' *International Journal of Operations & Production Management*, 30 (3), pp. 232-259.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). 'Latent dirichlet allocation' *Journal of machine Learning research*, 3 (Jan), pp. 993-1022.

Brady, T., Davies, A. & Gann, D. M. (2005). 'Creating value by delivering integrated solutions' *International Journal of Project Management*, 23 (5), pp. 360-365.

Brett, M. R. (2012). 'Topic modeling: A basic introduction' *Journal of digital humanities*, 2 (1), pp. 2-1.

Bryant, A. (2017). *Grounded theory and grounded theorizing: Pragmatism in research practice*. Oxford University Press.

Chakkol, M., Selviaridis, K. & Finne, M. (2018). 'The governance of collaboration in complex projects' *International Journal of Operations & Production Management*, 38 (4), pp. 997-1019.

Chandrasekaran, A., Oliva, R. & Salvador, F. (2023). 'Intervention-based research in operations management' *Foundations and Trends® in Technology, Information and Operations Management (Forthcoming)*.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

Charmaz, K. (2015). 'Grounded theory', in *Qualitative psychology: A practical guide to research methods.* pp. 53-84 [Online]. Version.

Charmaz, K. (2017a). 'Constructivist grounded theory' *The Journal of Positive Psychology*, 12 (3), pp. 299-300.

Charmaz, K. (2017b). 'The power of constructivist grounded theory for critical inquiry' *Qualitative Inquiry*, 23 (1), pp. 34-45.

Charmaz, K. (2017c). 'Special invited paper' *International Journal of Qualitative Methods*, 16 (1).

Charmaz, K. (2020). '"With constructivist grounded theory you can't hide": Social justice research and critical inquiry in the public sphere' *Qualitative Inquiry*, 26 (2), pp. 165-176.

Charmaz, K. & Keller, R. (2016). 'A personal journey with grounded theory methodology. Kathy charmaz in conversation with reiner keller' *Forum : Qualitative Social Research*, 17 (1).

Chase Jr., C. W. (2013). 'Using big data to enhance demand-driven forecasting and planning' *The Journal of Business Forecasting*, 32 (2), p. 27.

Chenger, D. & Pettigrew, R. N. (2023). 'Leveraging data-driven decisions: A framework for building intracompany capability for supply chain optimization and resilience' *Supply Chain Management: An International Journal*.

Choi, D. & Song, B. (2018). 'Exploring technological trends in logistics: Topic modeling-based patent analysis' *Sustainability*, 10 (8), p. 2810.

Choi, T.-M., Cheng, T. C. E. & Zhao, X. (2016). 'Multi-methodological research in operations management' *Production and Operations Management*, 25 (3), pp. 379-389.

Choi, T. M., Wallace, S. W. & Wang, Y. (2018). 'Big data analytics in operations management' *Production and Operations Management*, 27 (10), pp. 1868-1883. DOI: 10.1111/poms.12838.

Chong, D. & Shi, H. (2015). 'Big data analytics: A literature review' *Journal of Management Analytics*, 2 (3), pp. 175-201.

Chou, Y. C., Chuang, H. H. C., Chou, P. & Oliva, R. (2023). 'Supervised machine learning for theory building and testing: Opportunities in operations management' *Journal of Operations Management*, 69 (4), pp. 643-675.

Chuang, J., et al. (2014). 'Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations', *Advances in Neural Information Processing Systems workshop on human-propelled machine learning*, 2014. pp. 1-9.

Cintron, D. W. & Montrosse-Moorhead, B. (2021). 'Integrating big data into evaluation: R code for topic identification and modeling' *American Journal of Evaluation*.

Clarke, A. E. & Charmaz, K. (2019). 'Grounded theory and situational analysis', in *SAGE Research Methods Foundations* [Online]. Version.

Coughlan, P. & Coghlan, D. (2002). 'Action research for operations management' *International Journal of Operations & Production Management*, 22 (2), pp. 220-240.

Crespin-Mazet, F., Romestant, F. & Salle, R. (2019). 'The co-development of innovative projects in cops activities' *Industrial Marketing Management*, 79 pp. 71-83.

Croidieu, G. & Kim, P. H. (2018). 'Labor of love: Amateurs and lay-expertise legitimation in the early u.S. Radio field' *Administrative Science Quarterly*, 63 (1), pp. 1-42.

Davies, A. & Brady, T. (2000). 'Organisational capabilities and learning in complex product systems: Towards repeatable solutions' *Research policy*, 29 (7-8), pp. 931-953.

Davies, A., Brady, T., Prencipe, A. & Hobday, M. (2011). 'Innovation in complex products and systems: Implications for project-based organizing' *Advances in Strategic Management*, 28 pp. 3-26.

Davies, A., Dodgson, M. & Gann, D. (2016). 'Dynamic capabilities in complex projects: The case of london heathrow terminal 5' *Project management journal*, 47 (2), pp. 26-46.

Davies, A. & Hobday, M. (2005). *The business of projects: Managing innovation in complex products and systems*. Cambridge: Cambridge University Press.

Díaz, J., Pérez-Martínez, J., Pérez, C. & González-Prieto, Á. (2022). 'Applying inter-rater reliability and agreement in collaborative grounded theory studies in software engineering' *SSRN Electronic Journal*.

Dimaggio, P. (2015). 'Adapting computational text analysis to social science (and vice versa)' *Big Data & Society*, 2 (2).
Dimaggio, P., Nag, M. & Blei, D. (2013). 'Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.S. Government arts funding' *Poetics*, 41 (6), pp. 570-606.
Dominguez-Péry, C., Tassabehji, R., Vuddaraju, L. N. R. & Duffour, V. K. (2021). 'Improving emergency response operations in maritime accidents using social media with big data analytics: A case study of the mv wakashio disaster' *International Journal of Operations & Production Management*, 41 (9), pp. 1544-1567. DOI: 10.1108/ijopm-12-2020-0900.
Easterby-Smith, M., Thorpe, R., Jackson, P. & Jaspersen, L. J. (2018). *Management & business research*. *Management and business research* 6th edn. Los Angeles: SAGE.
Esposito De Falco, S., Renzi, A., Orlando, B. & Cucari, N. (2017). 'Open collaborative innovation and digital platforms' *Production Planning & Control*, 28 (16), pp. 1344-1353.
Fairhurst, G. T. & Putnam, L. L. (2019). 'An integrative methodology for organizational oppositions: Aligning grounded theory and discourse analysis' *Organizational Research Methods*, 22 (4), pp. 917-940.
Favaretto, M., De Clercq, E., Schneble, C. O. & Elger, B. S. (2020). 'What is your definition of big data? Researchers' understanding of the phenomenon of the decade' *PLOS ONE*, 15 (2).
Feagin, J. R., Orum, A. M. & Sjoberg, G. (2016). *A case for the case study*. UNC Press Books.
Feng, Q. & Shanthikumar, J. G. (2018). 'How research in production and operations management may evolve in the era of big data' *Production and Operations Management*, 27 (9), pp. 1670-1684.
Forza, C. (2002). 'Survey research in operations management: A process-based perspective' *International Journal of Operations & Production Management*, 22 (2), pp. 152-194.
Galati, F., Bigliardi, B., Galati, R. & Petroni, G. (2019). 'Managing structural inter-organizational tensions in complex product systems projects: Lessons from the metis case' *Journal of Business Research*.
George, G., Haas, M. R. & Pentland, A. (2014). 'Big data and management' *Academy of Management Journal*, 57 (2), pp. 321-326.
George, G., Osinga, E. C., Lavie, D. & Scott, B. A. (2016). Big data and data science methods for management research. Academy of Management Briarcliff Manor, NY.
Gioia, D. (2021). 'A systematic methodology for doing qualitative research' *The Journal of Applied Behavioral Science*, 57 (1), pp. 20-29.
Gioia, D. A., Corley, K. G. & Hamilton, A. L. (2013). 'Seeking qualitative rigor in inductive research' *Organizational Research Methods*, 16 (1), pp. 15-31.
Glaser, B. G. (1992). *Basics of grounded theory analysis: Emergence vs forcing*. Sociology press.
Glaser, B. G. & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.
Griffiths, T. L. & Steyvers, M. (2004). 'Finding scientific topics' *Proceedings of the National Academy of Sciences*, 101 (suppl_1), pp. 5228-5235.
Grimmer, J. & Stewart, B. M. (2013). 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts' *Political Analysis*, 21 (3), pp. 267-297.
Guha, S. & Kumar, S. (2018). 'Emergence of big data research in operations management, information systems, and healthcare: Past contributions and future roadmap' *Production and Operations Management*, 27 (9), pp. 1724-1735.
Guo, X., et al. (2022). 'Supply chain transformation and technology management challenges in developing regions: Inductive theory building from rural chinese nanostores' *Journal of Operations Management*, 68 (5), pp. 454-486.
Han, X., Zhu, D., Lei, M. & Daim, T. (2021). 'R&d trend analysis based on patent mining: An integrated use of patent applications and invalidation data' *Technological Forecasting and Social Change*, 167 p. 120691.
Hannigan, T. R., et al. (2019). 'Topic modeling in management research: Rendering new theory from textual data' *Academy of Management Annals*, 13 (2), pp. 586-632.

Hassani, H., et al. (2020). ‘Text mining in big data analytics’ *Big Data and Cognitive Computing*, 4 (1), p. 1. DOI: 10.3390/bdcc4010001.
Hobday, M. (1998). ‘Product complexity, innovation and industrial organisation’ *Research policy*, 26 (6), pp. 689-710.
Hobday, M. (2000). ‘The project-based organisation: An ideal form for managing complex products and systems?’ *Research Policy*, 29 (7-8), pp. 871-893.
Hsieh, H.-F. & Shannon, S. E. (2005). ‘Three approaches to qualitative content analysis’ *Qualitative health research*, 15 (9), pp. 1277-1288.
Hu, H., Wen, Y., Chua, T.-S. & Li, X. (2014). ‘Toward scalable systems for big data analytics: A technology tutorial’ *IEEE Access*, 2 pp. 652-687.
Ignatow, G. & Mihalcea, R. (2017). 'Topic models', in *Text Mining: A Guidebook for the Social Sciences* [Online]. Version.
Inaba, M. & Kakai, H. (2019). 'Grounded text mining approach: A synergy between grounded theory and text mining approaches', in *The SAGE Handbook of Current Developments in Grounded Theory* [Online]. Version.
Isoaho, K., Gritsenko, D. & Mäkelä, E. (2021). ‘Topic modeling and text analysis for qualitative policy research’ *Policy Studies Journal*, 49 (1), pp. 300-324.
Jacobs, T. & Tschötschel, R. (2019). ‘Topic models meet discourse analysis: A quantitative tool for a qualitative approach’ *International Journal of Social Research Methodology*, 22 (5), pp. 469-485.
Jesus, G. T., Itami, S. N., Segantine, T. Y. F. & Chagas Junior, M. F. (2021). ‘Innovation path and contingencies in the china-brazil earth resources satellite program’ *Acta Astronautica*, 178 pp. 382-391.
Johnson, M., Roehrich, J. K., Chakkol, M. & Davies, A. (2021). ‘Reconciling and reconceptualising servitization research: Drawing on modularity, platforms, ecosystems, risk and governance to develop mid-range theory’ *International Journal of Operations & Production Management*, 41 (5), pp. 465-493.
Kache, F. & Seuring, S. (2017). ‘Challenges and opportunities of digital information at the intersection of big data analytics and supply chain management’ *International Journal of Operations & Production Management*, 37 (1), pp. 10-36.
Kaplan, S. & Vakili, K. (2015). ‘The double-edged sword of recombination in breakthrough innovation’ *Strategic Management Journal*, 36 (10), pp. 1435-1457.
Kinra, A., Hald, K. S., Mukkamala, R. R. & Vatrapu, R. (2020). ‘An unstructured big data approach for country logistics performance assessment in global supply chains’ *International Journal of Operations & Production Management*, 40 (4), pp. 439-458.
Ko, D. G., Mai, F., Shan, Z. & Zhang, D. (2019). ‘Operational efficiency and patient-centered health care: A view from online physician reviews’ *Journal of Operations Management*, 65 (4), pp. 353-379.
Kobayashi, V. B., et al. (2018). ‘Text mining in organizational research’ *Organizational Research Methods*, 21 (3), pp. 733-765.
Kotzab, H., Seuring, S., Müller, M. & Reiner, G. (2006). *Research methodologies in supply chain management*. Springer Science & Business Media.
Lancichinetti, A., et al. (2015). ‘High-reproducibility and high-accuracy method for automated topic classification’ *Physical Review X*, 5 (1).
Lau, K.-N., Lee, K.-H. & Ho, Y. (2005). ‘Text mining for the hotel industry’ *Cornell Hotel and Restaurant Administration Quarterly*, 46 (3), pp. 344-362.
Lawrence, R., et al. (2010). ‘Social media analytics: The next generation of analytics-based marketing seeks insights from blogs’ *OR/MS Today*, 37 (1), pp. 26-31.
Lee, J. J. & Yoon, H. (2015). ‘A comparative study of technological learning and organizational capability development in complex products systems: Distinctive paths of three latecomers in military aircraft industry’ *Research Policy*, 44 (7), pp. 1296-1313.

Lee, K. & Yu, C. (2018). ‘Assessment of airport service quality: A complementary approach to measure perceived service quality based on google reviews’ *Journal of Air Transport Management*, 71 pp. 28-44.
Lee, L. W., Dabirian, A., McCarthy, I. P. & Kietzmann, J. (2020). ‘Making sense of text: Artificial intelligence-enabled content analysis’ *European Journal of Marketing*, 54 (3), pp. 615-644.
Lesnikowski, A., et al. (2019). ‘Frontiers in data analytics for adaptation research: Topic modeling’ *WIREs Climate Change*, 10 (3), p. e576.
Li, M., et al. (2021). 'Applying bayesian hyperparameter optimization towards accurate and efficient topic modeling in clinical notes', 2021 2021. IEEE.
Lu, Q., Zhou, Y., Luan, Z. & Song, H. (2023). ‘The effect of smes' ambidextrous innovations on supply chain financing performance: Balancing effect and moderating effect’ *International Journal of Operations & Production Management*.
Maier, D., et al. (2018). ‘Applying lda topic modeling in communication research: Toward a valid and reliable methodology’ *Communication Methods and Measures*, 12 (2-3), pp. 93-118.
Matthias, O., Fouweather, I., Gregory, I. & Vernon, A. (2017). ‘Making sense of big data – can it transform operations management?’ *International Journal of Operations & Production Management*, 37 (1), pp. 37-55.
Maylor, H., Meredith, J. R., Söderlund, J. & Browning, T. (2018). ‘Old theories, new contexts: Extending operations management theories to projects’ *International Journal of Operations & Production Management*, 38 (6), pp. 1274-1288.
McCall, C. & Edwards, C. (2021). ‘New perspectives for implementing grounded theory’ *Studies in Engineering Education*, 1 (2), p. 93.
Mees-Buss, J., Welch, C. & Piekkari, R. (2022). ‘From templates to heuristics: How and why to move beyond the gioia methodology’ *Organizational Research Methods*, 25 (2), pp. 405-429.
Mejia, J., Mankad, S. & Gopal, A. (2021). ‘Service quality using text mining: Measurement and consequences’ *Manufacturing & Service Operations Management*, 23 (6), pp. 1354-1372.
Mills, K. A. (2018). ‘What are the threats and potentials of big data for qualitative research?’ *Qualitative Research*, 18 (6), pp. 591-603.
Mills, K. A. (2019). ‘Big data for qualitative research’.
Mithas, S., Chen, Z. L., Saldanha, T. J. V. & De Oliveira Silveira, A. (2022). ‘How will artificial intelligence and industry 4.0 emerging technologies transform operations management?’ *Production and Operations Management*, 31 (12), pp. 4475-4487.
Morgan, D. L. (2020). ‘Pragmatism as a basis for grounded theory’ *The Qualitative Report*, 25 (1), p. 64.
Moro, A., Joanny, G. & Moretti, C. (2020). ‘Emerging technologies in the renewable energy sector: A comparison of expert review with a text mining software’ *Futures*, 117 p. 102511.
Morse, J. M., et al. (2021). *Developing grounded theory: The second generation*. 2nd Edition edn.: Routledge.
Müller, R., Pemsel, S. & Shao, J. (2014). ‘Organizational enablers for governance and governmentality of projects: A literature review’ *International Journal of Project Management*, 32 (8), pp. 1309-1320.
Müller, R., Pemsel, S. & Shao, J. (2015). ‘Organizational enablers for project governance and governmentality in project-based organizations’ *International Journal of Project Management*, 33 (4), pp. 839-851.
Muresan, G. & Harper, D. J. (2004). ‘Topic modeling for mediated access to very large document collections’ *Journal of the American Society for Information Science and Technology*, 55 (10), pp. 892-910.
Nelson, L. K. (2020). ‘Computational grounded theory: A methodological framework’ *Sociological Methods & Research*, 49 (1), pp. 3-42.
Nowell, L. S., Norris, J. M., White, D. E. & Moules, N. J. (2017). ‘Thematic analysis’ *International Journal of Qualitative Methods*, 16 (1).

Odacioglu, E., Zhang, L. & Allmendinger, R. (2022). ‘Combining topic modeling with grounded theory: Case studies of project collaboration’ *arXiv pre-print server*. DOI: arxiv:2207.02212.

Oliva, R. (2019). ‘Intervention as a research strategy’ *Journal of Operations Management*, 65 (7), pp. 710-724.

Park, T. Y. & Ji, I. (2015). ‘From mass production to complex production: Case of the korean telecom equipment sector’ *Asia-Pacific Journal of Accounting and Economics*, 22 (1), pp. 78-102.

Piepenbrink, A. & Gaur, A. S. (2017). 'Topic models as a novel approach to identify themes in content analysis', *Academy of Management Proceedings*, 2017. Academy of Management Briarcliff Manor, NY 10510. p. 11335.

Rinke, E. M., et al. (2021). ‘Expert-informed topic models for document set discovery’ *Communication Methods and Measures*, pp. 1-20.

Roehrich, J. K., Davies, A., Frederiksen, L. & Sergeeeva, N. (2019). ‘Management innovation in complex products and systems: The case of integrated project teams’ *Industrial Marketing Management*, 79 pp. 84-93.

Schmiedel, T., Müller, O. & Vom Brocke, J. (2019). ‘Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture’ *Organizational Research Methods*, 22 (4), pp. 941-968.

Schwarz, C. (2018). ‘Ldagibbs: A command for topic modeling in stata using latent dirichlet allocation’ *The Stata Journal*, 18 (1), pp. 101-117.

Sebastian, K. (2019). ‘Distinguishing between the strains grounded theory: Classical, interpretive and constructivist’ *Journal for Social Thought*, 3 (1).

Shadrova, A. (2021). ‘Topic models do not model topics: Epistemological remarks and steps towards best practices’ *Journal of Data Mining & Digital Humanities*, 2021. DOI: 10.46298/jdmdh.7595.

Shang, G. & Rönkkö, M. (2022). ‘Empirical research methods department: Mission, learnings, and future plans’ *Journal of Operations Management*, 68 (2), pp. 114-129.

Shrestha, Y. R., He, V. F., Puranam, P. & von Krogh, G. (2021). ‘Algorithm supported induction for building theory: How can we use prediction models to theorize?(forthcoming in organization science)’.

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

Silva, C. C., Galster, M. & Gilson, F. (2021). ‘Topic modeling in software engineering research’ *Empirical Software Engineering*, 26 (6). DOI: 10.1007/s10664-021-10026-0.

Strauss, A. & Corbin, J. (1990). ‘Basics of grounded theory methods’ *Beverly Hills: Sage*.

Strauss, A. & Corbin, J. M. (1997). *Grounded theory in practice*. Sage.

Stuart, I., et al. (2002). ‘Effective case research in operations management: A process perspective’ *Journal of Operations Management*, 20 (5), pp. 419-433.

Tang, C. (2016). ‘Innovative om research: Why? What? And how’ *Manufacturing Service Oper. Management*, 18 (2), pp. 178-183.

Tang, J., et al. (2014). 'Understanding the limiting factors of topic modeling via posterior contraction analysis', *International conference on machine learning*, 2014. PMLR. pp. 190-198.

Taylor, A. & Taylor, M. (2009). ‘Operations management research: Contemporary themes, trends and potential future directions’ *International Journal of Operations & Production Management*, 29 (12), pp. 1316-1340.

Thornberg, R. & Charmaz, K. (2014). 'The sage handbook of qualitative data analysis', [Online]. Version.

Thornberg, R. & Dunne, C. (2019). 'Literature review in grounded theory', in *The SAGE Handbook of Current Developments in Grounded Theory* [Online]. Version.

Tracy, S. J. (2010). ‘Qualitative quality: Eight “big-tent” criteria for excellent qualitative research’ *Qualitative inquiry*, 16 (10), pp. 837-851.

Van Aken, J., Chandrasekaran, A. & Halman, J. (2016). ‘Conducting and publishing design science research’ *Journal of Operations Management*, 47-48 (1), pp. 1-8.

Van Eck, N. J. & Waltman, L. (2014). 'Visualizing bibliometric networks': Springer International Publishing, pp. 285-320.

Voss, C., Tsikriktsis, N. & Frohlich, M. (2002). ‘Case research in operations management’ *International Journal of Operations & Production Management*, 22 (2), pp. 195-219.

Walsh, I., et al. (2015). ‘What grounded theory is…a critically reflective conversation among scholars’ *Organizational Research Methods*, 18 (4), pp. 581-599.

Wickert, C., et al. (2021). ‘Management research that makes a difference: Broadening the meaning of impact’ *Journal of Management Studies*, 58 (2), pp. 297-320.

Wilk, V., Mat Roni, S. & Jie, F. (2023). ‘Supply chain insights from social media users’ responses to panic buying during covid-19: The herd mentality’ *Asia Pacific Journal of Marketing and Logistics*, 35 (2), pp. 290-306. DOI: 10.1108/apjml-06-2021-0400.

Will M. Bertrand, J. & Fransoo, J. C. (2002). ‘Operations management research methodologies using quantitative modeling’ *International Journal of Operations & Production Management*, 22 (2), pp. 241-264.

Xiao, S., Ho, Y. C. & Che, H. (2021). ‘Building the momentum: Information disclosure and herding in online crowdfunding’ *Production and Operations Management*, 30 (9), pp. 3213-3230. DOI: 10.1111/poms.13425.

Zhan, Y., et al. (2020). ‘Leveraging social media in new product development: Organisational learning processes, mechanisms and evidence from china’ *International Journal of Operations & Production Management*, 40 (5), pp. 671-695. DOI: 10.1108/ijopm-04-2019-0318.

Appendix A

Selecting the Right Topic Model

As a Bayesian Model, LDA has become the most popular TM algorithm (Hannigan et al., 2019). Functioning as a generative probabilistic model, LDA facilitates the analysis of unstructured text data to uncover latent topics within a given corpus (Blei et al., 2003). It assumes that documents consist of a mixture of several topics, each represented by a probability distribution over words. Its goal is to estimate the topic mixture across a document collection and the word distribution within each topic.

The LDA process starts with document collection and pre-processing to generate a corpus. Once prepared, LDA randomly assigns words in each document to topics and builds bags of words based on a selective value of *k*. Iteratively, the model updates topic probability distributions until its convergence is achieved, resulting in stable topic allocations. LDA calculates probabilities for word-topic relations and document-topic generation. The final output includes sets of topics, each represented by a word distribution, and documents represented by topic distributions indicating topic proportions.

In their study, Dimaggio et al. (2013) asserted that LDA does not yield a single solution, so researchers have to explore different models and choose the best-serving model for the "*analytic purpose*". The choice of hyperparameters α and β significantly influences results. α controls topic distribution across documents, with higher values yielding more diverse topics and lower values producing fewer topics. β determines word sharing between topics, affecting topic distinctiveness, with higher values leading to more overlapping words and lower values lead to distinct topics with fewer shared words.

Selecting α and β remains an open challenge, with scholars adopting optimisation, fixed values, defaults, or variable parameters (Silva et al., 2021). Default values (0.25 and 0.1, respectively) (Schwarz, 2018), or $\alpha = 50 / k$ with $\beta = 0.1$ when $k>50$ (Griffiths and Steyvers, 2004) are common approaches. In our study, we utilised the Gensim Topic Modeling Python Library's built-in α and β optimisation features.

The most significant parameter, *k*, often relies on heuristic approaches, particularly when interpretation is a focus (Dimaggio et al., 2013). Initially, we optimised our corpus and code with the trial-error method, iterating over different *k* values and producing topic groups for analysis, then learned from the analysis results and made improvements accordingly. We refined our approach based on generated topic groups and analysis outcomes, making code adjustments (e.g., modifying the stop-word list, altering iteration numbers, and addressing data set inconsistencies). Each iteration introduced enhancements.

Upon reaching code and corpus maturity, we generated solutions with *k* ranging from 5 to 50. We evaluated solutions and their interpretability in line with the suggestion of Dimaggio et al. (2013). Eventually, 24 topics demonstrated superior promise for our study compared to other outcomes.

The intertopic distance map (Sievert and Shirley, 2014) is a visual representation that helps to elucidate relationships and proximities between topics. This graphical tool offers an initial understanding of topic distinctiveness or similarity based on their word content and probabilities. Each circle or point on the map represents a topic, and its proximity indicates how closely related they are in terms of word usage and distribution. The distance between circles signifies the statistical dissimilarity or separation between topics; shorter distances indicate more similarity, while longer distances suggest dissimilarity. By examining the positions and closeness of the circles, an analyst can gain insights into topic relations, facilitating the identification of initial topic clusters or themes within the data. One would choose to develop themes based on this map. However, since it relies solely on documents and does not consider topic labels, we suggest using this map as a guiding tool for developing themes.