X SINGEP
Simpósio Internacional de Gestão, Projetos, Inovação e Sustentabilidade
International Symposium on Management, Project, Innovation and Sustainability
ISSN: 2317-8302
CYRUS Institute of Knowledge MAKE A DIFFERENCE
CIK 10th INTERNATIONAL CONFERENCE

Analisando a literatura atual em Gestão do Conhecimento - Uma abordagem com Processamento de Linguagem Natural

Analyzing the extant research on Knowledge Management - a Natural Language Processing approach

DEBORAH FORONI
UNINOVE – UNIVERSIDADE NOVE DE JULHO

FELLIPE SILVA MARTINS
UNINOVE – UNIVERSIDADE NOVE DE JULHO

Nota de esclarecimento:
O X SINGEP e a 10ª Conferência Internacional do CIK (CYRUS Institute of Knowledge) foram realizados de forma remota, nos dias 26, 27 e 28 de outubro de 2022.

Agradecimento à orgão de fomento:
The authors would like to thank CAPES for the scholarship and funding throughout the development of this paper.

Analisando a literatura atual em Gestão do Conhecimento - Uma abordagem com Processamento de Linguagem Natural

Objetivo do estudo
Neste artigo, pretendemos analisar a pesquisa científica recente sobre Gestão do Conhecimento (GC) usando uma abordagem automatizada de agrupamento de tópicos e comparar com o agrupamento encontrado em revisões de literatura de GC feitas manualmente.

Relevância/originalidade
Nossos resultados mostram que é possível extrair e agrupar tópicos de pesquisa de um corpus de artigos científicos em Gestão do Conhecimento usando Processamento de Linguagem Natural (PLN) de última geração.

Metodologia/abordagem
Usamos um novo algoritmo de Processamento de Linguagem Natural (PLN) chamado BERTopic (que gera agrupamentos profundos de tópicos interpretáveis) para determinar a composição interna da pesquisa existente em Gestão do Conhecimento.

Principais resultados
Os resultados apontam para 58 tópicos organizados em quatro agrupamentos bem estabelecidos - três dos quais correspondem aproximadamente a domínios conceituais e um grande cluster de aplicações de domínio metodológico-substantivo.

Contribuições teóricas/metodológicas
Considerando o estado da arte em PLN para extração e agrupamento de tópicos teóricos (e BERTopic como método principal), acreditamos que os resultados obtidos são promissores, considerando as vantagens de reprodutibilidade, transparência, código aberto, automação e baixo viés.

Contribuições sociais/para a gestão
Estamos confiantes nas estratégias baseadas em PLN para servir como uma abordagem complementar aos procedimentos tradicionais na avaliação de aspectos estratégicas.

Palavras-chave: Knowledge Management, Literature Review, Natural Language Processing, Topic Modeling, BERTopic

Analyzing the extant research on Knowledge Management - a Natural Language Processing approach

Study purpose
In this paper, we intend to analyze the recent scientific research on Knowledge Management (KM) using an automated topic clustering approach to compare with the clustering found in current human-made KM literature reviews.

Relevance / originality
Our results show that it is possible to extract and cluster research topics from a corpus of scientific papers in Knowledge Management using state-of-the-art Natural Language Processing (NLP).

Methodology / approach
We have used a novel Natural Language Processing (NLP) algorithm called BERTopic (that generates deep clusters of interpretable topics) to ascertain the inner composition of the extant Knowledge Management research.

Main results
Results point to 58 topics organized in four well-established clusters - three of which roughly correspond to conceptual domains and one large cluster of methodological-substantive domain applications.

Theoretical / methodological contributions
Considering the state of the art in NLP for theory topic extraction and clustering (and BERTopic as a main method), we believe that while the results obtained are promising, considering the advantages of reproducibility, transparency, open source code, automation and low bias.

Social / management contributions
We are confident in NLP-based strategies to serve as a complementary approach to traditional procedures in evaluating strategic aspects.

Keywords: Gestão do conhecimento, Revisão de Literatura, Processamento de Linguagem Natural, Modelagem de Tópics, BERTopic.

1 Introduction

Knowledge Management (KM) is a theoretical area of enquiry that has amassed a huge volume of past research (Kakabadse et al., 2003; Wallace et al., 2011; Ramy et al., 2018). Consequently, it has become a wide and deep research field from which many reviews of literature - systematic and otherwise - have emerged (Al-Emran et al., 2018). As such, considering the KM field as a vast source of potential research avenues, we intend in this paper to analyze the current scientific production on KM through an automated, transparent, and reproducible process by employing Natural Language Processing (NLP) clustering tools. Thus, our aim in this paper is to compare an automated analysis and clustering of KM topics stemming from the literature with the state-of-the-art literature reviews available.

As a subset of management, Knowledge Management focuses on valuable yet intangible assets that impact organizations. While considered abstract, subjective or even elusive in its definition (Alavi and Leidner 2001; Gaviria-Marin et al., 2019), knowledge is seen as a viable asset, particularly when confined to procedures, models, systems and tools and, thus, prone to be managed (Sambamurthy et al., 2003; Elia et al., 2020). As such, it has become an ubiquitous presence in research, spreading from the management core towards a plethora of different areas of application.

On the other hand, KM alone is not a silver bullet to all organizational problems. For instance, knowledge management initiatives fail due to a vast array of reasons (Weber, 2007). Consequently, the extant literature reflects the many adaptations of KM tenets to every nook and cranny of scientific endeavor as well as multi- and interdisciplinary efforts (Jasimuddin, 2012). This makes KM literature to be spread over too many facets, leading to the de facto specialization of KM subfields as a consequence of its current fragmentation as a field (Lambe, 2011; Serenko, 2013; González-Valiente et al., 2019). As a potential backlash, KM theoreticians may become more dogmatic, in an attempt to keep all KM subfields under a cohesive theoretical frame of reference (Jevnaker & Olaisen, 2022). Whereas this fragmentation may be seen as pervasive and unwelcome from a unified theoretical point of view, it also allows new research topics and avenues to emerge. Hence, KM - both as a parent field or as in the form of a continuum of closely related areas - still merits an evaluation of its current state, in its conceptual, methodological, substantive domains.

To do so, we propose an automated analysis of the present literature production on Knowledge Management. We used BERTopic, a new modeling technique that generates deep clusters of interpretable topics, to evaluate the 4,983 most recent papers on KM. Results point to 58 topics organized in four well-established clusters - three of which roughly correspond to conceptual and methodological domains and one large cluster of substantive domain applications. These suggest a settled theoretical core feeding studies in many different areas, as is common in other theories.

2 Literature review

The exponential growth in research in Knowledge Management has long been a recorded fact (Despres & Chauvel, 1999; Ragab & Arisha, 2013), that is currently fueled by the growth of digitalization of society (Hawamdeh, 2022). Its growth comes with both advantages and disadvantages. On the pro side, it has been recognized as a key component in scientific

research in the past decades and as a source of theoretical explanations in various roles (antecedent, mediator, moderator and consequent) to a profusion of theoretical.
This makes KM an important field of study that should be considered as a potential component in studies for many situations in the social sciences.

On the cons side of things, this exponential growth comes with its shortcomings. First and foremost, this growth means the field has become increasingly fragmented (Despres & Chauvel, 1999; Gray & Meister, 2003; Lambe, 2011; Farooq, 2018; Kassaneh et al., 2021). As a natural development, specialization of KM subfields as their own fiefdoms has occurred. This can be attested by the proliferation of specialized reviews of literature that explore KM in specific contexts. Examples of this fragmentations are KM in Digital transformation (de Bem Machado et al., 2022), KM in information systems (Ramy et al., 2018), KM in artificial intelligence (Al Mansoori et al., 2021), KM in social media (Panahi et al., 2021), KM in higher education (Quarchioni et al., 2022), KM in virtual learning (Hantoobi et al., 2021), KM in healthcare (Karamitri et al., 2017), KM in supply chains (Cerchione and Esposito, 2016), KM in disaster management (Oktari et al., 2020), KM in gender (Heisig and Kannan, 2020), KM in business model innovation (Bashir and Farooq, 2019), KM in organizational innovation (Areed et al., 2021), and KM in family business (Su and Daspit, 2021), to name a few.

This fragmentation and proliferation of KM research in such a wide area suggests it is slowly becoming a theoretical and practical silver bullet (Spender, 2004) - though some authors defend that it is not the case (Millar et al., 2016). On top of this, the current Covid-19 pandemic has boosted research in KM as a whole (Hasan et al., 2022) along with the interpolations of KM, Covid-19 and an almost endless list of subjects. This is a two-fold phenomenon - the growth in relevance and research in KM along with a growing complexity in analyzing KM as a research field. This poses additional challenges for researchers in KM to understand what KM is as a consolidated theme and future research avenues thereof.

2.1 Natural Language Processing in Knowledge Management literature

Developing literature reviews is becoming increasingly difficult since the quantity and quality of research is growing exponentially and, as such, evaluating scientific research is becoming ever more dependent on compiling large lists of papers. Consequently, bringing together results and theory from scientific papers is becoming more complex and dependent on consolidating multiple sources, which takes its toll on human cognition. Therefore, it also allows biases and expectations from researchers to creep in (Young, 2009; Almeida & Goulart, 2017).

Whereas protocols to debias syntheses of literature have been put together (systematic reviews of literature, protocols such as PRISMA, contributor roles systems such as CRediT), and software to analyze scientific literature have appeared, a frontier that still is largely untouched is the analysis of content. Analyzing the content of a theoretical field spread over hundreds or thousands of separate but interconnected pieces is, in itself, a colossal task. Theories such as bounded rationality show that humans tend to fail in long-term, repeated tasks and that computers tend to fare better in such circumstances.

Analyzing large amounts of text is part of a key growing area of computer science, called Natural Language Processing (NLP) (Khurana et al., 2022). This field has been long accepted as a common framework for analysis in several scientific areas, and it is considered one of the key trends in techniques for analysis in social sciences, particularly in management (Kang et al., 2020). As a subfield of management, KM has also seen the growth in research having NLP as

one possible method for analysis (Basyal et al., 2020; Amarsson et al., 2021). For instance, there have been efforts in integrating NLP and KM in several instances such as healthcare (Basyal et al., 2020), construction engineering (Kim & Chi, 2019) and BIM (Wang et al., 2022), shop floor management (Müller et al., 2021), crowdsourcing (Sabou et al., 2012), as well as software engineering (Gilson & Weyns, 2019).

NLP as a technique has evolved in a way that nowadays has become more useful for analyzing texts such as in Knowledge Management. It has slowly grown from the simpler symbolic approach (rule-based processes), into statistical modeling. That means that initially its goal was to start from pre-made rules to find patterns. Nowadays, modern NLP has largely abandoned these approaches in favor of neural network-based solutions (Goldberg, 2016). Among these, large language models (models pre-trained for thousands of hours and feeding on billions of parameters) have emerged as trustworthy, production-wise options (Tamkin et al., 2021). In addition, the transformer architecture has particularly been accepted as a central starting point to efficiency and accuracy in NLP both in academia and industry by introducing the concept of attention layers (Galassi et al., 2020; Gillioz et al., 2020).

Within these large language models using transformer architecture, BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019) has gained considerable attention from the NLP research community - attested by its more than 45,000 citations. While previous models work by guessing the next word in a string of words using a probabilistic calculation (from left-to-right or right-to-left orientations), the transformer architecture takes the whole string as input and masks the target guessed word during the training phase. As such, BERT proposes a model in which a guessed word is evaluated from its connections in both directions, and its probability is guessed in a dynamic mechanism in which the guessed word is masked from the algorithm during its training phase from these left-leaning and right-learning connections (the attention layer).

Whereas generic models as BERT have their strengths, specifically trained models have emerged. For instance, an area that has evolved in the last decade is text classification. Among the many flavors of BERT, a fine-tuned version thereof, whose goal is to feed on a large corpus of text and classify and cluster topics within the corpus is BERTopic (Grootendorst, 2022). Simply put, one extracts large amounts of text of a specific significance (such as top scientific production in Knowledge Management, which is our goal) and feeds it to BERTopic, which allows us to analyze this collection of texts under a perspective of consolidated topics clustered by similarity.

3. Materials and methods

As discussed before, human-based, manual reviews of literature may carry over biases and expectations from researchers, especially in terms of content interpretation and classification. By developing a complementary automated analysis, researchers can gain a deeper understanding about the inner composition of the underlying topics in the KM research field. As such, in this paper, we used a four-fold strategy - see Figure 1. First (*in yellow in the figure*), we used traditional techniques for article selection. In this phase, we have chosen a specific scientific database as a starting point (Web of Science). We follow de Lima Araújo *et al.* (2021) in choosing the Web of Science database instead of a composite sampling comprising several databases, since it covers virtually all scientific production (Martín-Martín et al., 2018).

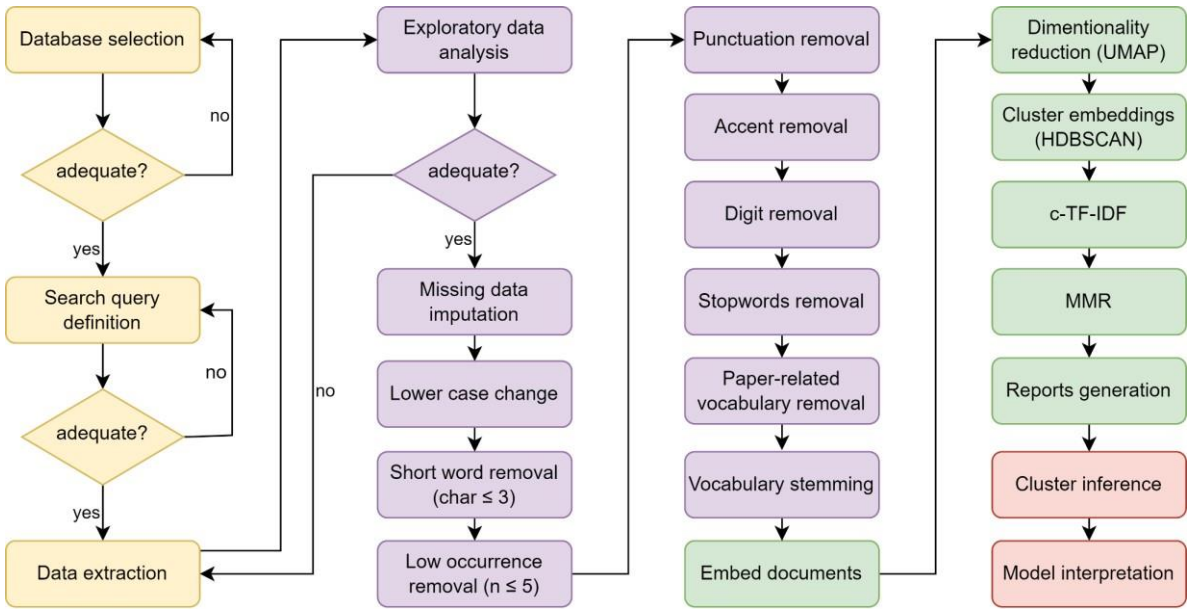

Figure 1. *Methodological procedures*

Next, we have chosen "*knowledge management*" as a search query since our goal was to extract all possible theoretical perspectives, contexts and applications. As a result, 13,373 articles were found and we decided to cap the analysis of the 5,000 most recently published papers (excluding proceeding papers, white papers, etc.). This subset was deemed adequate since it covers mainly the years 2019-2022, complementing several previously published reviews of literature - such as (Costa & Monteiro, 2016; Mariano & Awazu, 2016; Di Vaio et al., 2021). This subset also allows us to analyze the current scenario of KM research, especially considering the changes fostered by the current Covid-19 crisis and its impact on research.

The following phase was concentrated on corpus pre-processing (*in purple*). After collection, text data is unstructured data, i.e. it needs to be pre-processed and cleaned. Dealing with text is a cumbersome task and several wrangling procedures were executed to ensure the corpus was adequate for further analyses. Initially an exploratory data analysis was carried out to ensure the previous phase was properly executed. Then, some missing data imputation was performed on the 'year' column of the data frame (not all papers display this information and 705 papers did not have this information). In addition, several common natural language processing data cleaning tasks were performed: changing the corpus to lowercase; short word removal (≤ 3 characters); low occurrence words removal (n ≤ 5); punctuation removal; accent removal; digit removal; and stop words removal. Since we are dealing with scientific papers, some highly frequent paper-specific words (*paper, results, findings*, etc.) were excluded, otherwise they could form clusters due to their abnormal frequency - see Appendix 1 for more details. Finally, a common corpus simplification technique (stemming) was employed.

Turning now to the main part of the analysis (*in green*), we have used a new topic modeling technique based on natural language processing procedures - BERTopic (Grootendorst, 2022). This technique "leverages transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions" (Grootendorst, 2022:1). In lay terms, this algorithm considers that documents containing the same topic are semantically similar. (Grootendorst, 2022). This algorithm works in three phases - 1) extracting document embeddings using a transformer-based architecture

(BERT, for instance); 2) clustering documents using UMAP (*Uniform Manifold Approximation and Projection*) to reduce embedding dimensionality and HDBSCAN (*Hierarchical Densitybased Spatial Clustering of Applications with Noise*) to cluster reduced embeddings and generate clusters of related documents by semantical analysis; and 3) creating topic representations using c-TF-IDF (*Class Term frequency-Inverse Document Frequency*), and ensuring coherence by using MMR (*Maximal Marginal Relevance*).

The last part of the methodological procedures (*in pink*) were generating reports and interpreting results based on previously published literature. The detailed procedures, code, data and results may be obtained from the authors.

4. Results and discussion

Despite NLP being used in KM research to extract information (Xia et al., 2020; Wang et al., 2022), there is little to no research on NLP used as an automatization tool for literature review on scientific papers. Given that research on KM spans a wide area, this approach presents a new avenue for research of use cases of NLP in surveying literature review for KM. As we mentioned before, we started out with our analysis of 4,983 paper abstracts, and 58 topics have been generated.

As shown in Figure 2, we can observe the terms selected for the first 20 topics of the cTF-IDF scores for each topic representation. Considering these topics, we can get an idea of what is being addressed in the documents (the full list of topics and other technical details may be obtained from the authors). For each topic, one finds a list of words in decreasing order of importance with the numbers in the x-axis displaying proportion of each given word within each topic (the ‘c’ in c-TF-IDF stands for *class* TF-IDF since the topic is its own corpus in terms of evaluation) – see Figures 2 and 3 (the remaining clusters can be found in Appendix 2).

Since most of the sample comprises late 2019 onwards, the corpus reflects the changes imposed to the world during the first years of the Covid-19 pandemic. This becomes clear after analyzing the content of the first two clusters. The first cluster (Cluster 0) deals with how Knowledge Management both affected as was affected by the current pandemic. This points to both sound research relating KM to the issues stemming from the pandemic (Wang & Wu, 2021) but also opportunistic research capitalizing on the novelty of Covid-19 within the KM field.

The same could be observed in the second cluster (Cluster 1), in concentrated educationrelated research. The same behavior was present (both good and opportunistic papers). These 2 topics had great prominence in the Covid-19 pandemic period, and the papers collected for this research are concentrated in the last 2 years that coincide with the Covid-19 pandemic period, this can characterize the volume of publications of articles with these subjects.

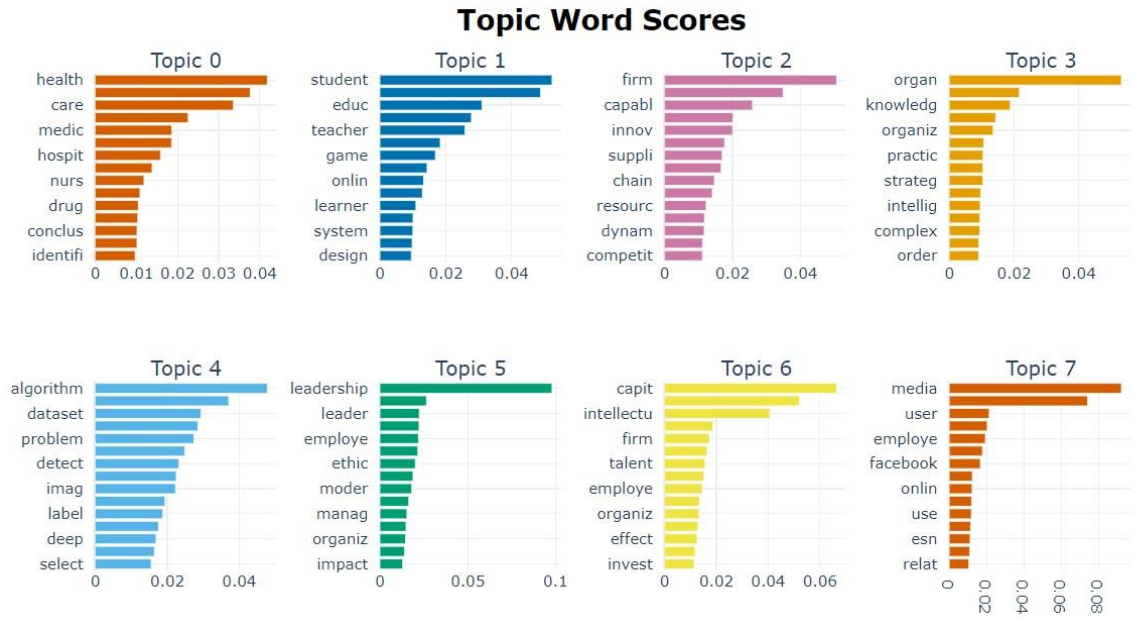

***Figure 2**. Topic Word Scores*

The order of the topics reflects their importance (proportion) within the sample of all papers. One can see that these topics are a mixture of conceptual, methodological, substantive domains. Topics 0, 1, 7 and 13 (the later in Figure 4) for example, are clear instances of substantive domain, while topics 2, 4, 5, 6, 10, 11 (the later in Figure 4) etc. are instances of conceptual domain. Some examples of methodological domain such as topics 4 and 9 are also present.

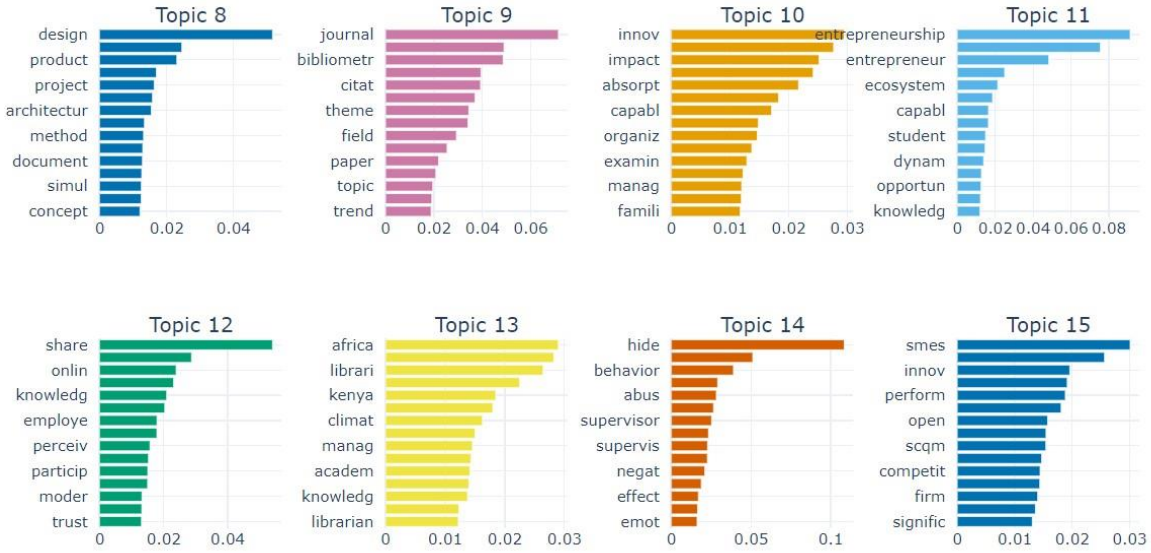

***Figure 3**. Topic Word Scores (continued)*

According to Figure 4, the intertopic distance map generated 4 sharply contrasting clusters. The upper-left quadrant has 7 topics and focuses on the relationship between KM and projects, resources and technology. The upper-right quadrant has 6 topics and concentrates theoretical

aspects related to KM and entrepreneurship, networks, dynamic and absorptive capabilities along with sustainability. The lower-right quadrant has 10 topics and gathers results dealing with KM and firm, performance, employee relationships, teamwork and service.

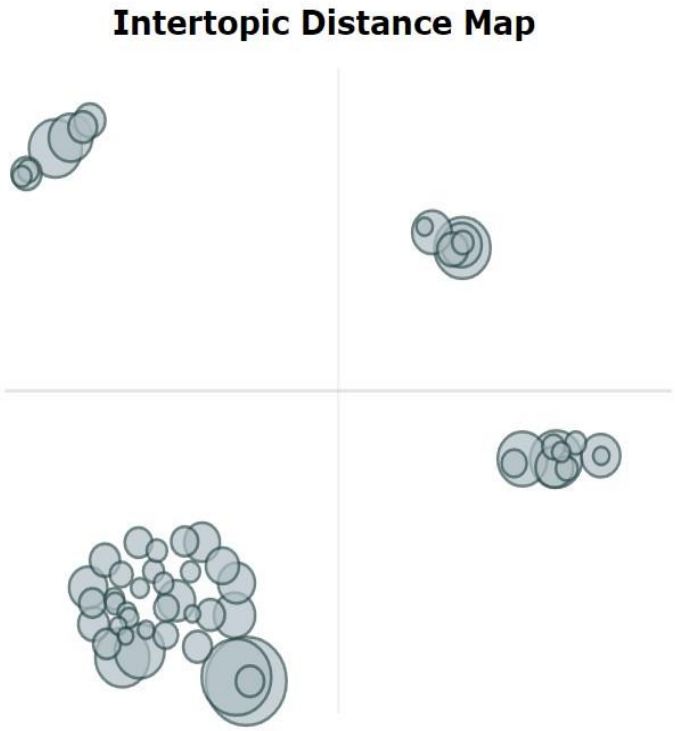

Figure 4. *Intertopic Distance Map*

The lower left quadrant contains the highest concentration of topics totaling 35 topics out of 57 generated. This cluster of topics conflates several diverging cases of KM and methodological and substantive domains.

As shown in Figure 5, we observed that the degree of similarity of the cosine between any two topics is mostly above 0.5, which demonstrates that the topics are related, this is due to the fact that the analyzed papers have the same core, which is knowledge management.

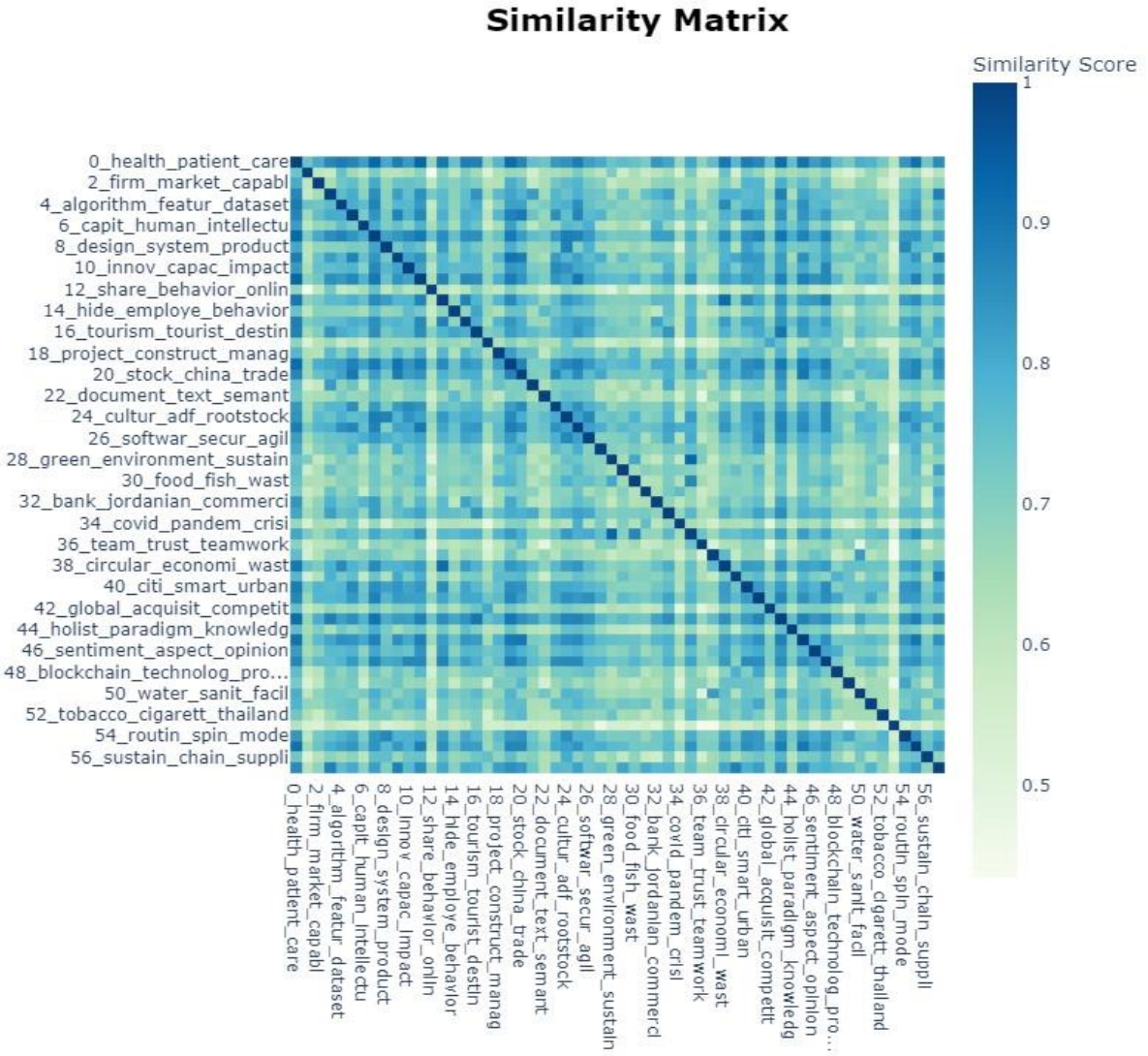

Figure 5*. Similarity Matrix*

We can see in Figure 6 that 10 clusters of topic representations were generated at the hierarchy level, highlighting that previously 58 topics were generated. In this way, hierarchical cluster analysis allows us to merge topics that are similar to each other, which makes it possible to reduce topics. However, we must observe which reductions make sense for our analysis, for example in our study, we have noticed a large concentration in the blue cluster, which has a diversity of subjects in the topics, perhaps for this cluster it is not interesting to reduce topics to a single, but possibly for the black group we could narrow it down to a single topic.

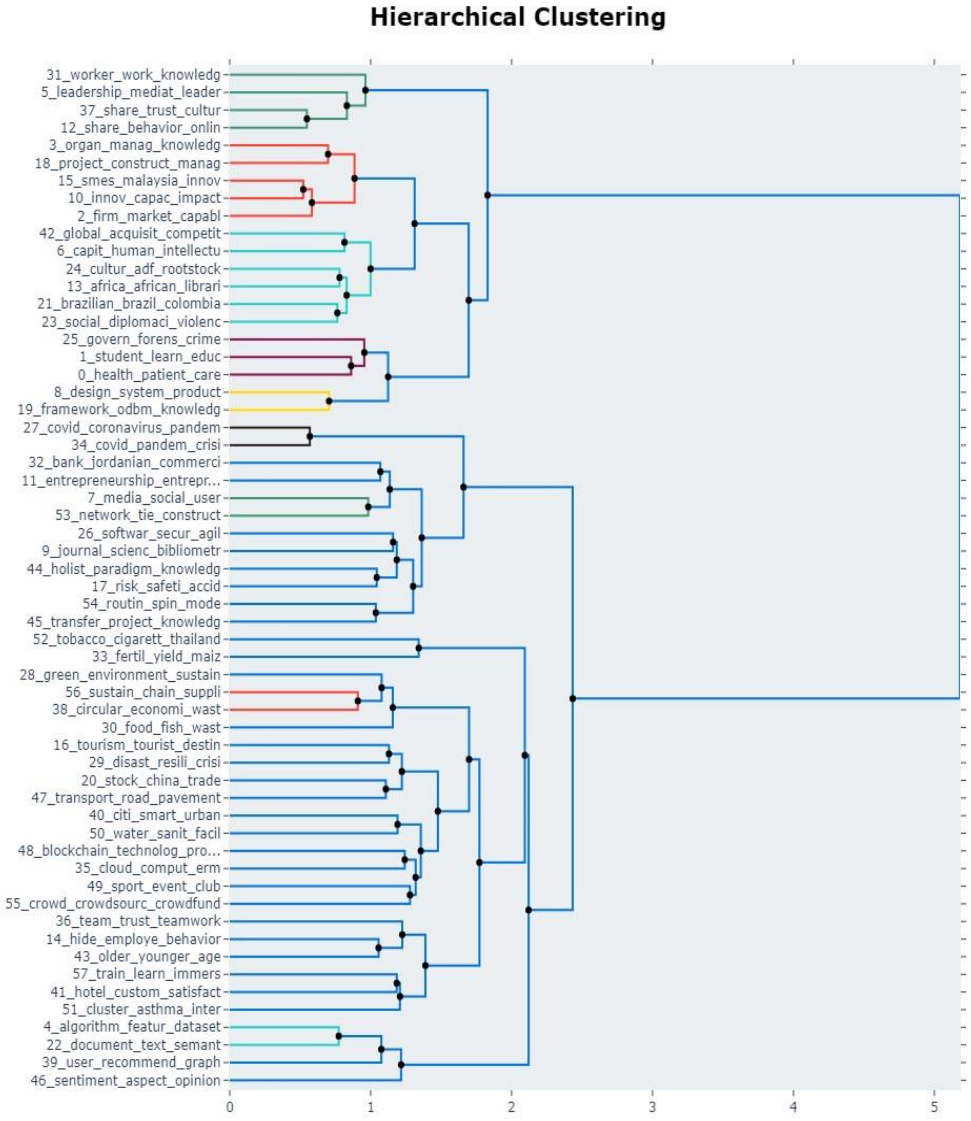

Figure 6*. Hierarchical clustering*

5. FINAL REMARKS

Our results show that it is possible to extract and cluster research topics from a corpus of scientific papers in Knowledge Management using Natural Language Processing. Considering the state of the art in NLP for topic extraction and clustering (and BERTopic as a main method), we believe that while the results obtained are promising, they are still a long way from replacing traditional, human-made literature analyses. On the other hand, considering the advantages of reproducibility, transparency, open source code, automation and low probability of biasing the analyses, we are confident in NLP-based strategies to serve at least as a companion, complementary approach to traditional procedures such as systematic literature reviews.

This paper comes with its share of limitations, though. First, whereas innovative and based on tried and tested previous algorithmic advancements, BERTopic is still a nascent algorithm which still requires caution in interpreting the end results. In addition, the corpus in this paper was established through a simplistic search query ("knowledge management"). This makes the corpus prone to pollution from papers that are not key to understanding KM but rather any paper that contains the expression thereof. In future use cases, combining a traditional, manual filtering of papers with the features of NLP-based, automated topic extraction and clustering would potentially yield better results.

Finally, the results obtained also allow us to infer possible new additions to the state of the art in using NLP to analyze scientific corpora, particularly in the field of social sciences or text-heavy scientific outlets. First, while using BERTopic allows one to extract topics and cluster them, topic content inference is still a manual, cognitive-intense task. Future studies could improve on this by employing NLP-based text summarization techniques to further the research procedures pipeline. Further still, NLP-based techniques for inferring relationships between topics could also advance the use of NLP to analyze scientific texts corpora and bring it closer to human-level analyses.

REFERENCES

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS quarterly*, 107-136.

Al-Emran, M., Mezhuyev, V., Kamaludin, A., & Shaalan, K. (2018). The impact of knowledge management processes on information systems: A systematic review. *International Journal of Information Management*, *43*, 173-187.

Al Mansoori, S., Salloum, S. A., & Shaalan, K. (2021). The impact of artificial intelligence and information technologies on the efficiency of knowledge management at modern organizations: a systematic review. *Recent advances in intelligent systems and smart applications*, 163-182.

Almeida, C. P. B. D., & Goulart, B. N. G. D. (2017). How to avoid bias in systematic reviews of observational studies. *Revista CEFAC*, *19*, 551-555.

Arnarsson, I. Ö., Frost, O., Gustavsson, E., Jirstrand, M., & Malmqvist, J. (2021). Natural language processing methods for knowledge management—Applying document clustering for fast search and grouping of engineering documents. *Concurrent Engineering*, *29*(2), 142-152.

Areed, S., Salloum, S. A., & Shaalan, K. (2021). The role of knowledge management processes for enhancing and supporting innovative organizations: a systematic review. *Recent advances in intelligent systems and smart applications*, 143-161.

Bashir, M., & Farooq, R. (2019). The synergetic effect of knowledge management and business model innovation on firm competence: A systematic review. *International Journal of Innovation Science*.

Basyal, G. P., Rimal, B. P., & Zeng, D. (2020). A systematic review of natural language processing for knowledge management in healthcare. *arXiv preprint arXiv:2007.09134*.

Cerchione, R., & Esposito, E. (2016). A systematic review of supply chain knowledge management research: State of the art and research opportunities. *International Journal of Production Economics*, *182*, 276-292.

Costa, V., & Monteiro, S. (2016). Key knowledge management processes for innovation: a systematic literature review. VINE Journal of Information and Knowledge Management Systems.

de Bem Machado, A., Secinaro, S., Calandra, D., & Lanzalonga, F. (2022). Knowledge management and digital transformation for Industry 4.0: a structured literature review. *Knowledge Management Research & Practice*, *20*(2), 320-338.

de Lima Araújo, H. C., Martins, F. S., Cortese, T. T. P., & Locosselli, G. M. (2021). Artificial intelligence in urban forestry—A systematic review. *Urban Forestry & Urban Greening*, *66*, 127410.

Despres, C., & Chauvel, D. (1999). Knowledge management (s). *Journal of knowledge Management*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Di Vaio, A., Palladino, R., Pezzi, A., & Kalisz, D. E. (2021). The role of digital innovation in knowledge management systems: A systematic literature review. Journal of business research, 123, 220-231.

Elia, G., Margherita, A., & Passiante, G. (2020). Digital entrepreneurship ecosystem: How digital technologies and collective intelligence are reshaping the entrepreneurial process. *Technological Forecasting and Social Change*, *150*, 119791.

Farooq, R. (2018). Developing a conceptual framework of knowledge management. *International Journal of Innovation Science*.

Galassi, A., Lippi, M., & Torroni, P. (2020). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(10), 4291-4308.

Gaviria-Marin, M., Merigó, J. M., & Baier-Fuentes, H. (2019). Knowledge management: A global examination based on bibliometric analysis. *Technological Forecasting and Social Change*, *140*, 194-220.

Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020, September). Overview of the Transformer-based Models for NLP Tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)* (pp. 179-183). IEEE.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420.

González-Valiente, C. L., Santos, M. L., & Arencibia-Jorge, R. (2019). Evolution of the Socio-cognitive Structure of Knowledge Management (1986–2015): An Author Co-citation Analysis. *Journal of Data and Information Science*, *4*(2), 36-55.

Gray, P. H., & Meister, D. B. (2003). Introduction: fragmentation and integration in knowledge management research. *Information Technology & People*.

Grootendorst, Maarten. (2022). BERTopic: Neural topic modeling with a class-based TFIDF procedure. arXiv preprint arXiv:2203.05794.

Grootendorst, Maarten. (2022). BERTopic. <https://maartengr.github.io/BERTopic/index.html>

Gilson, F., & Weyns, D. (2019, March). When natural language processing jumps into collaborative software engineering. In *2019 IEEE International Conference on Software Architecture Companion (ICSA-C)* (pp. 238-241). IEEE.

Hantoobi, S., Wahdan, A., Salloum, S. A., & Shaalan, K. (2021). Integration of knowledge management in a virtual learning environment: A systematic review. *Recent Advances in Technology Acceptance Models and Theories*, 247-272.

Hasan, I., Dhawan, P., Rizvi, S. A. M., & Dhir, S. (2022). Data analytics and knowledge management approach for COVID-19 prediction and control. *International Journal of Information Technology*, 1-18.

Hawamdeh, S. (2022). Foundations of Knowledge Management. In *Understanding, Implementing, and Evaluating Knowledge Management in Business Settings* (pp. 1-13). IGI Global.

Heisig, P., & Kannan, S. (2020). Knowledge management: does gender matter? A systematic review of literature. *Journal of Knowledge Management*, *24*(6), 1315-1342.

Jasimuddin, Sajjad M. *Knowledge management: An interdisciplinary perspective*. Vol. 11. World Scientific Publishing Company, 2012.

Jevnaker, B. H., & Olaisen, J. (2022). A comparative study of knowledge management research studies: making research more relevant and creative. *Knowledge Management Research & Practice*, 1-12.

Kakabadse, N. K., Kakabadse, A., & Kouzmin, A. (2003). Reviewing the knowledge management literature: towards a taxonomy. *Journal of knowledge management*.

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

Karamitri, I., Talias, M. A., & Bellali, T. (2017). Knowledge management practices in healthcare settings: a systematic review. *The International journal of health planning and management*, *32*(1), 4-18.

Kassaneh, T. C., Bolisani, E., & Cegarra-Navarro, J. G. (2021). Knowledge management practices for sustainable supply chain management: A challenge for business education. *Sustainability*, *13*(5), 2956.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 1-32.

Kim, T., & Chi, S. (2019). Accident case retrieval and analyses: Using natural language processing in the construction industry. *Journal of Construction Engineering and Management*, *145*(3), 04019004.

Lambe, P. (2011). The unacknowledged parentage of knowledge management. Journal of knowledge management.

Mariano, S., & Awazu, Y. (2016). Artifacts in knowledge management research: a systematic literature review and future research directions. Journal of Knowledge Management.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of informetrics*, *12*(4), 1160-1177.

Millar, C. C., Lockett, M., & Mahon, J. F. (2016). Guest editorial: knowledge intensive organisations: on the frontiers of knowledge management. *Journal of knowledge management*.

Müller, M., Alexandi, E., & Metternich, J. (2021). Digital shop floor management enhanced by natural language processing. *Procedia CIRP*, *96*, 21-26.

Oktari, R. S., Munadi, K., Idroes, R., & Sofyan, H. (2020). Knowledge management practices in disaster management: Systematic review. *International Journal of Disaster Risk Reduction*, *51*, 101881.

Panahi, S., Ghalavand, H., & Sedghi, S. (2021). How social media facilitates the knowledge management process: a systematic review. *Journal of Information & Knowledge Management*, *20*(04), 2150042.

Quarchioni, S., Paternostro, S., & Trovarelli, F. (2022). Knowledge management in higher education: A literature review and further research avenues. *Knowledge Management Research & Practice*, *20*(2), 304-319.

Ragab, M. A., & Arisha, A. (2013). Knowledge management and measurement: a critical review. *Journal of knowledge management*.

Ramy, A., Floody, J., Ragab, M. A., & Arisha, A. (2018). A scientometric analysis of Knowledge Management Research and Practice literature: 2003–2015. *Knowledge Management Research & Practice*, *16*(1), 66-77.

Sabou, M., Bontcheva, K., & Scharl, A. (2012, September). Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* (pp. 1-8).

Sambamurthy, V., Bharadwaj, A., & Grover, V. (2003). Shaping agility through digital options: Reconceptualizing the role of information technology in contemporary firms. *MIS quarterly*, 237-263.

Serenko, A. (2013). Meta-analysis of scientometric research of knowledge management: discovering the identity of the discipline. *Journal of Knowledge Management*.

Su, E., & Daspit, J. (2021). Knowledge management in family firms: a systematic review, integrated insights and future research opportunities. *Journal of Knowledge Management*.

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Xia, H., An, W., Li, J., & Zhang, Z. J. (2020). Outlier knowledge management for extreme public health events: Understanding public opinions about COVID-19 based on microblog data. Socio-Economic Planning Sciences, 100941.

Wallace, D. P., Van Fleet, C., & Downs, L. J. (2011). The research core of the knowledge management literature. *International Journal of Information Management*, *31*(1), 14-20.

Wang, H., Meng, X., & Zhu, X. (2022). Improving knowledge capture and retrieval in the BIM environment: Combining case-based reasoning and natural language processing. Automation in Construction, 139, 104317.

Wang, W. T., & Wu, S. Y. (2021). Knowledge management based on information technology in response to COVID-19 crisis. *Knowledge management research & practice*, *19*(4), 468-474.

Weber, R. O. (2007). Addressing failure factors in knowledge management. *Electronic Journal of Knowledge Management*, *5*(3), pp334-347.

Young, S. N. (2009). Bias in the research literature and conflict of interest: an issue for publishers, editors, reviewers and authors, and it is not just about the money. *Journal of psychiatry & neuroscience: JPN*, *34*(6), 412.

Appendix 1 - Paper-related terms excluded from the corpus

The following paper-related words were excluded from the corpus. These words are very likely to appear in paper abstracts, yet they are not related to theoretical, practical or technical aspects of the research. The list of words: 'study', 'research', 'data', 'based', 'results', 'model', 'paper', 'findings', 'approach', 'information', 'analysis', 'purpose', 'used', 'relationship', 'also', 'process', 'development', 'methodology' and 'literature'.

Appendix 2 – Remaining clusters


Topic 40
citi
urban
definit
sustain
energi
soft
review
need
Topic 41
hotel
satisfact
servic
replic
star
capit
manag
user
Topic 42
global
competit
learn
continu
shrm
strategi
start
employe
Topic 43
older
age
employe
adult
apprentic
alumni
elder
workforc
Topic 44
holist
knowledg
quantum
innov
theori
abil
dynam
scientif
Topic 45
transfer
knowledg
factor
willing
team
turnov
allianc
enterpris
Topic 46
sentiment
opinion
dataset
task
featur
text
micro
predict
Topic 47
transport
pavement
electr
traffic
obstacl
vehicl
commiss
system
Topic 48
blockchain
protocol
proof
consensus
system
sustain
stake
interest
Topic 49
sport
club
injuri
prevent
nonprofit
perceiv
botswana
recommend
Topic 50
water
facil
access
drink
smart
safe
farmer
septic
Topic 51
cluster
inter
belong
compani
famili
strateg
smes
connect
Topic 52
tobacco
thailand
retail
alcohol
child
price
vet
thai
Topic 53
network
construct
hoard
social
layer
ambidexter
exchang
identifi
Topic 54
routin
mode
dynam
univers
concentr
forget
organiz
interorganiz
Topic 55
crowd
crowdfund
open
innov
challeng
integr
backer
egyptian
Topic 56
sustain
suppli
larg
paradigm
lean
product
circular
impact
Topic 57
train
immers
charact
profession
chines
percept
link
use