Journal of Information and Knowledge, Vol 60(2), April 2023, p.97-104

ISSN (Print) : 0972-2467
ISSN (Online) : 0976-2477
DOI: 10.17821/srels/2023/v60i2/170967

AI-Based Literature Reviews: A Topic Modeling Approach

Manoj Kumar Verma[1] and Mayank Yuvaraj[2]*

[1]Department of Library and Information Science, Mizoram University, Aizwal - 796004, Mizoram, India;
manojdlis@mzu.edu.in
[2]Central Library, Central University of South Bihar, Gaya – 824236, Bihar, India;
mayank.yuvaraj@gmail.com

Abstract

The purpose of this paper is to highlight the importance of topic modelling in conducting literature reviews using the open-source LDAShiny package in the R environment, with green libraries literature as a case study. To conduct the analysis, a title and abstract dataset were prepared using the Scopus database and imported into the LDAShiny package for further analysis. It was found that the green libraries' literature ranged from 1989-2023, with a sharp increase in research topics since 2003. The study also identified key themes and documents associated with green libraries research, revealing that energy efficiency, waste reduction and recycling, and the use of sustainable materials have been extensively discussed in the literature. However, further research is needed on the implementation of these practices in libraries, as well as the impact of the COVID-19 pandemic on green libraries. The findings will be beneficial to researchers interested in using topic modelling for literature reviews.

Keywords: Green Libraries, Latent Topics, LDA Shiny, Literature Review, Topic Modelling

1. Introduction

A literature review builds upon and refers to existing knowledge and is integral to academic research (Kunisch *et al.*, 2023; Snyder, 2019). An exhaustive literature review typically involves two steps: 1. Identifying a subset of citations, and 2. Manually screening the set of citations (Adam, Wallace and Trikalinos, 2022). However, the manual exploratory literature review is becoming increasingly difficult due to the rapid growth of literature (Asmussen and Moller, 2019). Whenever a field's literature grows faster than the time available for manual reviews, an adequate manual review of the literature cannot be conducted. (Marshall and Wallace, 2019). A variety of factors have made manual literature reviews challenging in recent years. Firstly, the process of searching and gathering relevant papers on any research domain is time-consuming. Furthermore, due to the laborious process of screening entire research literature on a topic, scholars develop rather narrow search terms (Schoot *et al.*, 2021). Secondly, the literature reviews are usually performed manually, with an enormous number of papers that may overwhelm human processing capacity (Wagner, Lukyanenko and Pare, 2022), because of which only a few papers are analyzed. Thirdly, the traditional review method is also compromised by researcher bias in selecting articles for review. In addition, some scholars have criticized the traditional method for its inability to be replicated and proven (Saha, 2021).

To overcome these, the conduct of literature reviews has been transformed by various tools and approaches. The first approach involves mapping the literature by examining relationships through a few papers to discover scholarly articles. Similar papers linked by "citations", "authors", "funders", "keywords", and other metadata can be identified this way. Several cloud-based SaaS platforms enable exploration of these connections, including *Citation Geecko* (https://www.citationgecko.com/), *Connected papers* (https://www.connectedpapers.

**Author for correspondence*

com/), *Inciteful* (https://inciteful.xyz/), *LitMaps* (https://www.litmaps.com/), *Open Knowledge Maps* (https://openknowledgemaps.org/), *JSTOR Text analyzer* (https://www.jstor.org/analyze/). Also, to deal with the volume, machine learning algorithms are employed to screen papers for relevant ones. *AS Review* (https://asreview.nl/) is one such free and open-source tool which uses active learning (a type of machine learning) to train a model that uses limited examples to predict relevance from texts. The AS Review performs automated title-abstract screening and ranks the paper based on the knowledge in the paper. Other similar machine learning systems available for use in systematic reviews are *Rayyan* (https://www.rayyan.ai/), *Colandr* (https://www.colandrcommunity.com/), *Covidence* (https://www.covidence.org/), *EPPI reviewer* (https://eppi.ioe.ac.uk/CMS/Default.aspx?alias=eppi.ioe.ac.uk/cms/er4&), *FASTREAD* (https://github.com/fastread/src), and *SWIFT Review* (https://www.sciome.com/swift-review/).

Although these two approaches are useful to identify relevant papers in any domain that can be reviewed to identify research gaps and to understand key research themes, researchers are faced with a high time cost when they need to read a large number of papers from a set of relevant ones manually. It has been found that a seasoned reviewer can screen about two abstracts per minute on average, but more complex abstracts can take much longer (Wallace *et al.,* 2010). Although, artificial intelligence-based research assistant- Elicit (https://elicit.org/) can be used to perform article summarization tasks yet it cannot identify the latent topics in the literature which are crucial for understanding gaps in research and trends. A simple solution to this can be found in topic modelling which can be automated, making it an ideal tool to conduct an exploratory literature review (Antons *et al.,* 2023; Asmussen and Moller, 2019). According to Kavvadias, Drosatos and Kaldoudi (2020) to help researchers navigate a wide range of publications and get quick overviews of evolving research fields, identification of research topics in published literature has emerged as a powerful tool which can be easily accomplished through topic modelling. The automated topic modelling approach complements traditional approaches to content analysis (Schmiedel, Muller and Brocke, 2019). Topic modelling is the process of condensing the text into topics composed of connected words based on statistical correlation.

With topic modelling, researchers can better understand large volumes of data spanning over extended periods. It has been applied extensively to understand the latent topics in newspapers (Ahmed and Khan, 2022), journals (Ozyurt and Ayaz, 2022), and research articles (Xie, Ning and Sun, 2022; Mostafa, 2022). The present paper seeks to demonstrate the importance of topic modelling using open source LDAShiny package in conducting literature reviews using green libraries literature as a case study to discover the hidden topics in the domain.

2. Topic Modelling

Topic modelling is a text-mining technique often used in machine learning and natural language processing. It is an effective method of analyzing and summarizing large amounts of textual data without human intervention to reveal hidden semantic patterns (latent topics). In the case of a collection of documents, it facilitates identifying hidden topics. Words making sense together are presented with each topic. The topic-associated words can help organize and provide insight into large amounts of unstructured text. The basic idea behind topic modelling is that each document is regarded as a mixture of topics, and each word within the document has a certain probability of belonging to a specific topic.

In the results obtained from topic modelling, two matrices are represented: the word-topic matrix (probability of a certain word belonging to a topic) and the topic-document matrix (probability of a particular topic appearing in a specific document); however, the end-users usually select the top words (words that have the highest probability in a topic) and the most probable texts. In terms of topic modelling algorithms, Latent Dirichlet Allocation (LDA) is the simplest, most well-studied, and most widely accepted method. In addition, LDA can effectively uncover latent topics and co-occurrences between words (Mustak *et al.,* 2021). LDA was used in this study for analyzing the collected papers on green libraries.

While there are other packages in the R environment through which topic modelling can be performed, LDAShiny is the only free statistical software package providing a GUI that allows analysts and researchers to perform LDA-focused scientific literature reviews interactively. LDAShiny is primarily intended for researchers who have little prior knowledge of the research field and would like to explore a large number of documents (such as scientific articles) to identify trends (Hoz-M, Fernandez-Gomez and Mendes, 2021).

Below is a step-by-step guide to installing and using LDAShiny:

- The latest version of the R language (https://cran.r-project.org/bin/windows/base/) and the RStudio platform (https://posit.co/) should be downloaded and installed as a first step.
- Next step is the installation of LDAShiny. Open the RStudio interface and type the following command:
- Install.packages ("LDAShiny")
- To invoke and open LDAShiny programs, enter the following command in the control interface window:
- library(LDAShiny)
- LDAShiny::runLDAShiny()

3. Objectives of the Study

The major objectives of this research study are:

- To use Latent Dirichlet Allocation (LDA) algorithm, a widely used algorithm for topic modelling, to identify latent topics and show its usefulness in conducting a literature review.
- To identify major topics discussed in green libraries literature.
- To understand the growth of research topics.
- To identify key themes in green libraries literature.
- To find out key documents associated with research topics in the green libraries domain.

4. Methodology

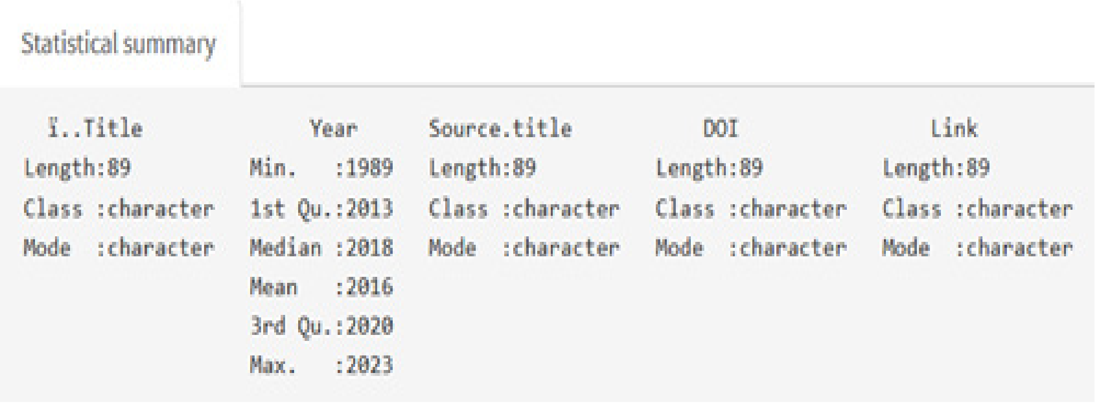
In the following section, we describe the steps we took to collect and analyze data.

4.1 Data Collection

In this study, data are collected for review using Scopus, which has a wider range of academic sources than its counterpart, the Web of Science (Paul *et al.*, 2021) and its coverage of documents is approved for indexing through strict criteria like "ethics and malpractice statement", "minimum of two-year publication history", "ownership", "peer review" (Donthu *et al.*, 2021). The search strategy is based on a single keyword, "green librar*," searched in the title, abstract, and keywords of the articles, following the recommendations of Lim, Yap and Makkar (2021), which recommends using a single keyword for review domains that are sufficiently broad and generic. We used asterisks (*) with the keyword to capture various endings of the terms such as a green library, and green libraries. The initial search using the keyword returned a result of 89 documents which were further used for analysis. We did not use any additional filters like document type, language, or year in the search query. We recorded the document title, year, source title, DOI and abstract of the 89 documents in an Excel file for conducting topic modelling from the database on January 30, 2023.

4.2 Data Processing

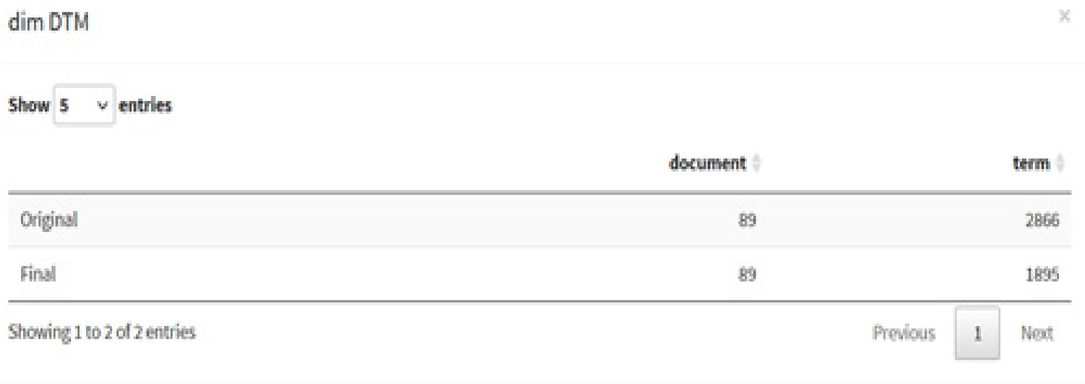
LDAShiny package was used for conducting LDA-based topic modelling. The Excel file (.csv) exported



```
Statistical summary

   I..Title              Year        Source.title          DOI                Link
 Length:89          Min.   :1989   Length:89          Length:89          Length:89
 Class :character   1st Qu.:2013   Class :character   Class :character   Class :character
 Mode  :character   Median :2018   Mode  :character   Mode  :character   Mode  :character
                    Mean   :2016
                    3rd Qu.:2020
                    Max.   :2023
```

Figure 1. Statistical summary of uploaded data.



dim DTM

Show 5 entries

	document	term
Original	89	2866
Final	89	1895

Showing 1 to 2 of 2 entries

Figure 2. Document term matrix dimensions (DTM dim) pre and post-processing.

from the Scopus database was uploaded to the software. A statistical summary of uploaded data is shown in Figure 1.

It can be seen from Figure 1 that the publications on green libraries ranged from 1989 to 2023. The mean and median years of publication are 2016 and 2018 respectively. The length of all five exported metadata is 89. The next step in LDAShiny involves data cleaning where n-gram inclusion is done, stop-words are added, and stemming is done. An N-gram consists of N words. As a result, a 2-gram (or bigram) is a sequence of two words, such as "green libraries", or "library automation", while a 3-gram (or trigram) is a sequence of three words, such as "automation through KOHA", or "sustainable green libraries". According to Hoz-M, Fernandez-Gomez and Mendes (2021), it is more common to analyze words individually or use N-grams, and that was chosen in this study. The stop-words approach in text mining helps reduce computing complexity and improve performance by removing words like "and", "or", and "was", etc. Although there are many possibilities for StopWord lists, we are limited to the words provided by the R StopWord (https://cran.r-project.org/web/packages/stopwords/index.html) as it has been used in previous studies. In stemming, root words are morphologically modified. Stemming is a text pre-processing technique that involves reducing a word to its root form, which can help reduce the dimensionality of the data and improve efficiency. For example, by using this feature the words "library" and "libraries" will be stemmed to librar. Figure 2 shows a snapshot of document term dimensions pre and post-processing.

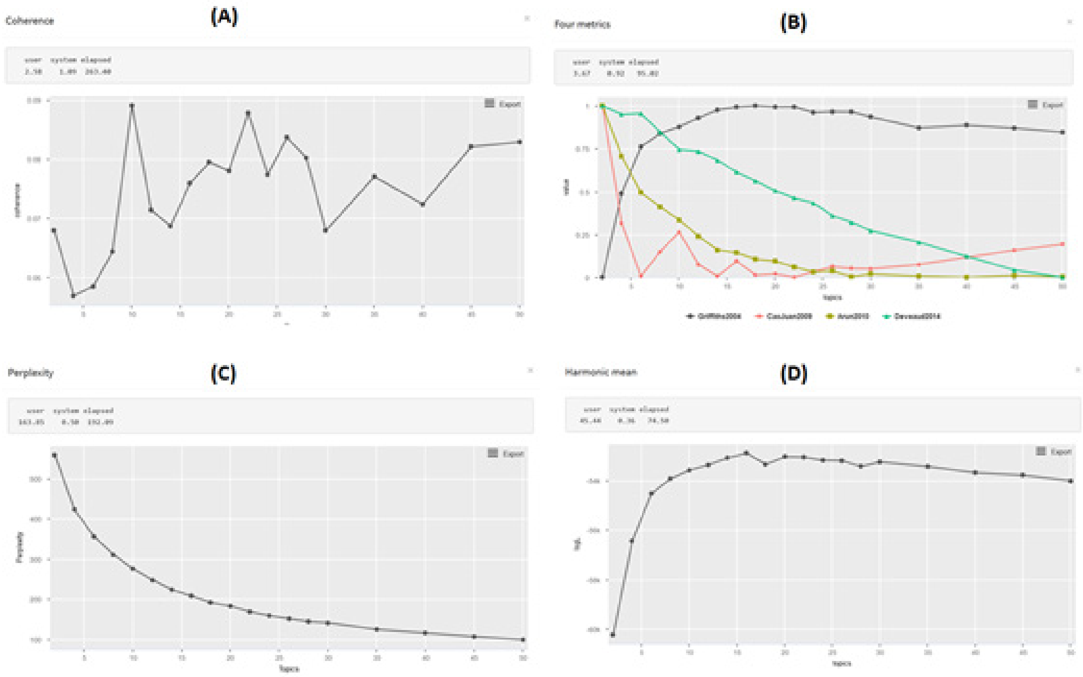
4.3 Number of Topics

Previous studies have stressed that the effectiveness of LDA models depends on the number of topics chosen when categorizing topics. According to Gan and Qi (2021) when the number of topics selected is small, the meaning under each topic will be insufficient; when the number of topics selected is in excess, the data will lead to over-clustering, resulting in redundant topics. To address this issue, there are various methods in LDAShiny such as coherence, four metrics, perplexity and harmonic mean, to find out an optimal number of topics. Based on configuration settings recommended by Hoz-M, Fernandez-Gomez and Mendes (2021), we calculated them which is represented in Figure 3. Among the metrics Griffiths 2004, CaoJuan 2009, Arun 2010, Perplexity and Harmonic mean, the number of suitable topics stands between 45 and 50, while Deveaud 2014 shows 35 topics and Coherence 14. As a result of our analysis, we found that the best value for k, or the number of topics, is between 10 and 12 for our dataset. We have selected 10 topics for the present study as 10 was the optimal coherence score.

5. Results

5.1 Top Terms in Green Libraries Literature Ranked by Term Frequency

Figure 4 shows the top terms in green libraries which are ranked by term frequency. The top term that appears in our dataset is "libraries", with term frequency (TF = 553),



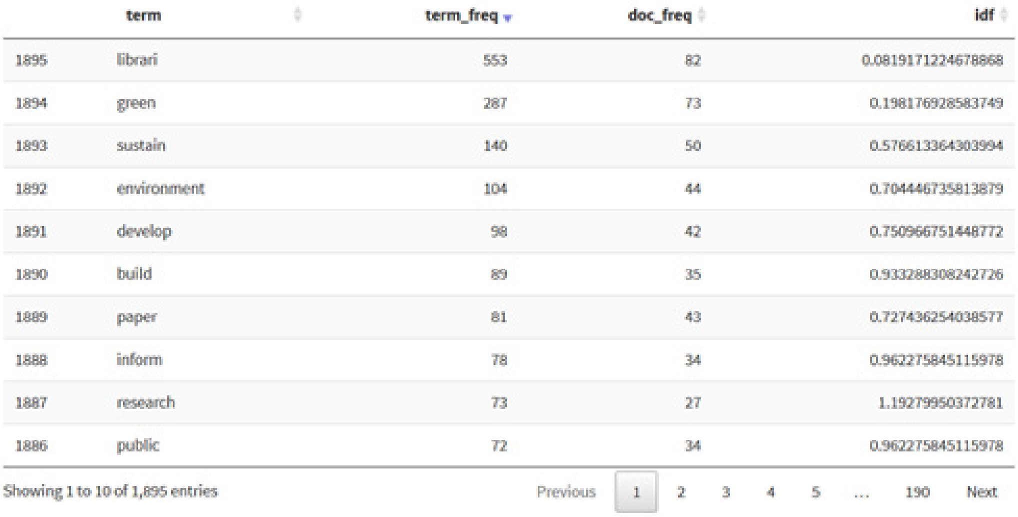
```
   user  system elapsed
   2.58    1.09  263.40
```

```
   user  system elapsed
   3.67    0.92   95.82
```

```
   user  system elapsed
 163.05    0.58  152.09
```

```
   user  system elapsed
  45.44    0.36   74.50
```

Figure 3. . Number of topics. **(A)** Coherence method. **(B)** Comparison of four methods. **(C)** Perplexity. **(D)** Harmonic Mean.



	term	term_freq	doc_freq	idf
1895	librari	553	82	0.0819171224678868
1894	green	287	73	0.198176928583749
1893	sustain	140	50	0.576613364303994
1892	environment	104	44	0.704446735813879
1891	develop	98	42	0.750966751448772
1890	build	89	35	0.933288308242726
1889	paper	81	43	0.727436254038577
1888	inform	78	34	0.962275845115978
1887	research	73	27	1.19279950372781
1886	public	72	34	0.962275845115978

Showing 1 to 10 of 1,895 entries

Figure 4. Top terms in the green libraries dataset.

document frequency (DF = 82) and inverse document frequency (IDF = 0.08) followed by the terms “green” having TF = 287 DF = 73 IDF = 0.19, “sustainable” with TF = 140 DF = 50 IDF = 0.57 and “environment” with TF = 104 DF = 44 IDF = 0.7.

5.2 Topic Trend

The yearly growth of topics in green libraries literature has been represented in Figure 5 through heatmap. There was a growth of topics from 2003 onward. The peak of growth can be seen during the period 2012-2023 for each topic.

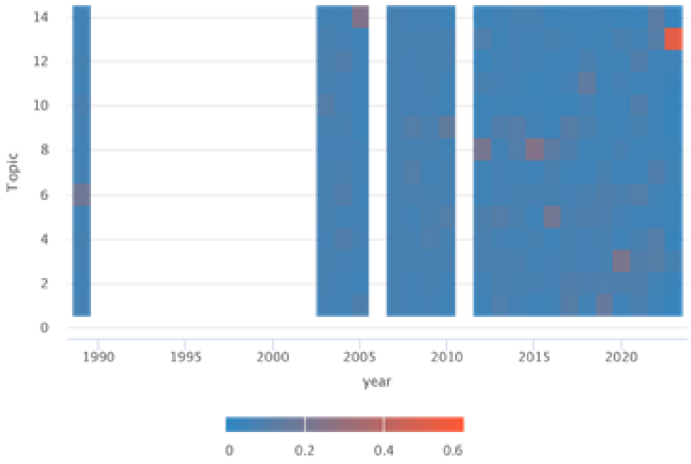


Figure 5. Topic trend in green libraries literature.

5.3 Key Themes in Green Libraries Literature

Table 1 shows the research topics that were obtained from the LDA model. The results are ranked by the prevalence scores. There is a high prevalence of literature on the topics such as green libraries, academic libraries, library programs, and library projects. We could not find any topics related to technology or COVID-19 in the present study from our dataset.

5.4 Key Documents Associated with Research Topics

Through LDAShiny we also identified key documents which were associated with each topic. Table 2 presents

Table 1. Key themes in green libraries literature

Topic	Label_1	Coherence	Prevalence	Top_Terms
t_1	Green libraries	0.052	8.314	Library, green, development, public, environment, research, analysis
t_2	Academic libraries	0.047	8.079	Library, academic, sustainable, strategies, services
t_3	Library program	0.141	7.793	Program, library, social, develop, green, cooper, paper, sustainable, education
t_4	Library project	0.118	7.535	Project, library, activity, green, œgreen, paper, group
t_5	Library building	0.097	7.47	Build, library, design, green, construct, plan, exist, architecture, sustain, environment
t_6	Library research	0.104	7.077	Research, library, studies, publish, management, countries, green
t_7	sustainability	0.044	6.944	Libraries, change, climate, sustainable, environment,
t_8	Library structure	0.079	6.08	Model, smart, structure, earthquake, time

Table 2. Key documents in green libraries research

Document	topic	theta
Green Library and green librarianship- Towards a conceptualization	t_1	0.45473
A perspective on computational research support programs in the library: More than 20 years of data from Stanford University library	t_2	0.36759
Planning Approach with "Better Than Before" Concept: A Case Study of Library Building at SVNIT, Surat, Gujarat, India	t_3	0.3413
More than Just a green building: Developing green strategies at the Chinese University of Hong Kong Library	t_4	0.59044
Library sow the seed of a sustainable society: A comparative analysis of IFLA Green Library Award projects 2016	t_5	0.28296
Operation performance evaluation of green public buildings with AHP-fuzzy synthetic assessment method based on cloud model	t_6	0.5795
The Emergence of Green Library in Kenya: Insights from Academic Library	t_7	0.29348
Environmentally Sustainable Approaches in Academic Library: A Micro-study in Uttar Pradesh	t_8	0.47281

an overview of the documents which can be used as a reference for researchers interested in green libraries to explore the domain.

6. Conclusion

Using topic modelling, this research study presents an efficient method for conducting literature reviews and gaining an overview of latent topics found in the title and abstract datasets. Literature reviews conducted manually suffer from researcher bias, lack replicability and validity, are extremely time-consuming, and are unreliable. An LDA-based method addresses these concerns. As a result of an LDA-based tool such as LDAShiny, researchers can not only understand the key research topics within a document but also identify the key documents associated with each topic, which is an effective alternative to manually screening titles and abstracts to identify relevant papers. In the present case study of green libraries research, we found that research works were scattered between 1989-2023. In summary, the analysis of the literature on green libraries using topic modelling reveals that energy efficiency, waste reduction and recycling, and the use of sustainable materials are important themes in the literature. However, there is a need for more research on the implementation and effectiveness of these practices in libraries, and on the impact of the COVID-19 pandemic on green libraries.

7. References

Adam, G.P., Wallace, B.C. and Trikalinos, T.A. (2022). Semi-automated tools for systematic searches. in: meta-research. methods in molecular biology, edited by Evangelou, E., Veroniki, A.A. New York, NY: Humana; pp. 17-40. https://doi.org/10.1007/978-1-0716-1566-9_2 PMid:34550582

Ahmed, F. and Khan, A. (2022). Topic modeling as a tool to analyze child abuse from the corpus of english newspapers in Pakistan. Social Science Computer Review. OnlineFirst. https://doi.org/10.1177/08944393221132637

Antons, D., Breidbach, C. F., Joshi, A. M. and Salge, T. O. (2023). Computational literature reviews: Method, algorithms, and roadmap. Organizational Research Methods, *25*, 107-138. https://doi.org/10.1177/1094428121991230

Asmussen, C.B. and Moller, C. (2019). Smart literature review: A practical topic modeling approach to exploratory literature review. Journal of Big Data, *6*, 93. https://doi.org/10.1186/s40537-019-0255-7

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N. and Lim, W.M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, *133,* 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Gan, J. and Qi, Y. (2021). Selection of the optimal number of topics for LDA topic model- taking patent policy as an example. Entropy, 23, 1-45. https://doi.org/10.3390/e23101301

Hoz-M, J. De La, Fernandez-Gomez, M. J. and Medes, S. (2021). LDAShiny: An R package for exploratory review of scientific literature based on Bayesian probabilistic model and machine learning tools. Mathematics, *9*. https://doi.org/10.3390/math9141671

Kavvadias, S., Drosatos, G. and Kaldoudi, E. (2020). Supporting topic modeling and trend analysis in biomedical literature. Journal of Biomedical Informatics, *110*, 103574. https://doi.org/10.1016/j.jbi.2020.103574 PMid:32971274

Kunisch, S., Denyer, D., Bartunek, J. M., Menz, M. and Cardinal, L. B. (2023). Review research as scientific inquiry. Organizational Research Methods, *26*, 3-45. https://doi.org/10.1177/10944281221127292

Lim, W.M., Yap, S.F. and Makkar, M. (2021). Home sharing in marketing and tourism at a tipping point: What do we know, how do we know, and where should we be heading? Journal of Business Research, *122,* 534-566, https://doi.org/10.1016/j.jbusres.2020.08.051 PMid:33012896 PMCid:PMC7523531

Marshall, I. J. and Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic Reviews, *8,* 163. https://doi.org/10.1186/s13643-019-1074-9 PMid:31296265 PMCid:PMC6621996

Mostafa, M. (2022). A one-hundred-year structural topic modeling analysis of knowledge structure of international management research. Quality and Quantity. OnlineFirst. https://doi.org/10.1007/s11135-022-01548-w PMid:36249708 PMCid:PMC9549032

Mustak, M., Salminen, J., Ple, L. and Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis and research agenda. Journal of Business Research, *124,* 389-404. https://doi.org/10.1016/j.jbusres.2020.10.044

Ozyurt, O. and Ayaz, A. (2022). Twenty-five years of education and information technologies: Insights from a topic modeling based bibliometric analysis. Education and Information Technologies, *27,* 11025-11054. https://doi.org/10.1007/s10639-022-11071-y PMid:35502161 PMCid:PMC9046010

Paul, J., Lim, W.M. , O'Cass, A., Hao, A.W. and Bresciani, S. (2021). Scientific Procedures and Rationales for Systematic Literature Reviews (SPAR-4-SLR). International Journal of Consumer Studies, *45*, O1-O16, https://doi.org/10.1111/ijcs.12695

Saha, B. (2021). Application of topic modeling for literature review in management research. In: Interdisciplinary research in technology and management, edited by S. Chakrabarti, R. Nath, P. K. Banerji, S. Datta, S. Poddar and M. Gangopadhyaya. London: CRC Press; pp. 249-256.

Schmiedel, T., Muller, O. and Brocke, J.V. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. Organizational Research Methods, *22*, 941-968. https://doi.org/10.1177/1094428118773858

Schoot, R. V., Bruin, J. Schram, R., Zahedi, P., Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoggerwerf, M., Ferdinands, G., Harkema, A., Willemsen, W., Ma, Y., Fang, Q., Hindriks, S., Tummers, L. and Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence, *3*, 125-133. https://doi.org/10.1038/s42256-020-00287-7

Snyder, H. (2019). Literature review as a research methodology: an overview and guidelines, Journal of Business Research, *104,* 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039

Wagner, G., Lukyanenko, R. and Pare, G. (2022). Artificial intelligence and the conduct of literature reviews. Journal of Information Technology, *37*, 209-226. https://doi.org/10.1177/02683962211048201

Wallace, B. C., Small, K., Brodley, C. E. and Trikalinos, T. A. (2010). Active learning for biomedical citation screening. In 16th ACM SIGKDD International Conference on Knowledge discovery and data mining, edited by B. Rao, B. Krishnapuram, A. Tomkins and Q. Yang, Washington DC, USA; pp. 173-182. https://doi.org/10.1145/1835804.1835829 PMid:20565949 PMCid:PMC2903585

Xie, Y., Ning, C. and Sun, L. (2022). The twenty-first century of structural engineering research: A topic modeling approach. Structures, *35,* 577-590. https://doi.org/10.1016/j.istruc.2021.11.018