IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 November 2022, accepted 23 December 2022, date of publication 28 December 2022,
date of current version 13 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232939

TOPICAL REVIEW

Topic Modeling: Perspectives From a Literature Review

ANDRÉS M. GRISALES A.[1], **SEBASTIAN ROBLEDO**[1], **AND MARTHA ZULUAGA**[2]
[1]Faculty of Administrative, Economic and Accounting Sciences, Universidad Católica Luis Amigó, Medellín 050004, Colombia
[2]Universidad Nacional Abierta y a Distancia (UNAD), Dosquebradas 661007, Colombia

Corresponding author: Martha Zuluaga (martha.zuluaga@unad.edu.co)

This work was supported in part by the Internal Research Call of Universidad Catóilica Luis Amigó.

ABSTRACT Topic modeling is a Natural Language Processing technique that has gained popularity over the last ten years, with applications in multiple fields of knowledge. However, there is insufficient empirical evidence to show how this field of study has developed over the years, as well as the main models that have been applied in different contexts. The objective of this paper is to analyze the evolution of the topic modeling technique, the main areas in which it has been applied, and the models that are recommended for specific types of data. The methodology applied is based on bibliometric analysis. First, we searched the Web of Science and the Scopus databases. We then used scientometric techniques and a Tree of Science methodology, which allowed us to analyze the search results from the perspectives of classics, structure, and trends. The results show that the USA and China are among the most productive countries in this field and the applications have been mainly in the identification of sub-topics in short texts, such as social networks and blogs. The main conclusion of this work is that topic modeling is a versatile technique that can complement systematic literature reviews and that has been well-received in different academic and research contexts. The results of this study will help researchers and academics to recognize the importance of these techniques for reviewing large volumes of unstructured information, such as research articles, and in general, for systematic literature reviews.

INDEX TERMS Literature review, machine learning, natural language processing, scientometrics, topic modeling.

I. INTRODUCTION

Currently, most information is available in unstructured forms such as video, audio, images, and text formats. However, accessing large volumes of data, particularly in text format, has become a significant challenge in computer science, communication theories, and linguistics, making this dataset manageable and understandable [1]. Examples include computer science techniques for document classification, clustering, named entity extraction, Topic Modeling (TM) [2], general management of unstructured data, and extraction of information from an extensive collection of documents [3]. TM is a statistical technique used to identify underlying themes in a set of documents that facilitate their representation from the occurrence of words that compose them [4]. The basic assumption is that each document is a random mix of topics and words [5]. This technique is based on several methods and strategies to identify these topics.

One TM strategy is Latent Semantic Indexing (LSI), which is based on the definition of a Document Term Matrix (TDM) that relates to the number of times a term appears in a document. Next, singular value decomposition was applied to this matrix to reduce the dimensions and establish the most relevant terms that define thematic issues [6]. Another strategy is Probabilistic Latent Semantic Analysis (pLSA), which considers topics as a probabilistic distribution of words, solving the main problem of the LSI strategy, where a document may not contain a search term, but its synonym [7]. Blei et al. [8] proposed a method for resolving the problem of having documents in the training set without probabilities, leading to the overfitting of the models when working with large volumes of data. This method is called

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

Latent Dirichlet Allocation (LDA) and is based on a Bayesian version of the pLSA method, where each document has a multinomial distribution over the topics, and each topic has a multinomial distribution over the words. This method is one of the most widely reported in the literature [9].

The present work arises from the need for one characterization of scientific production around TM and to identify the leading applications and evolutions of this technique as a complement to other studies that have been developed for similar purposes. Barde and Bainwad [4] discuss various methods and tools. Kherwa and Bansal [5] reviewed 300 papers on TM and concluded that LDA was the most popular technique. Finally, Lie et al. [10] presented an application of TM in short texts named GPU-DMM.

This study aimed to identify the scientific production, evolution, and subtopics of TM through a scientometric analysis. A similar study was presented by Hou et al. [11] but only used the Web of Science (WoS) database, and focused on a specific field of knowledge which is information sciences. This study analyzed the scientific production of TM by Scopus and WoS. Therefore, our research question is what are the main contributions of TM using scientometric techniques?

Using the results of this search in the two databases, scientific mapping was carried out using the analysis of citations and descriptive analysis of the annual scientific production of both journals and researchers. Subsequently, the Tree of Science (ToS) methodology was applied to review the contributions of this topic over time. Finally, a cluster analysis was performed to study articles in different subareas.

This paper contributes to the identification of the main advances in TM and the most efficient strategies for the analysis of literature and unstructured information. It also contributes to the presentation of efficient strategies for systematic literature reviews.

The rest of the article is organized as follows: the methodological part, where the process of selecting papers is explained; followed by the results showing the documents that are part of the root, trunk, and leave; and finally, the conclusions.

II. METODOLOGY

A. SEARCH STRATEGY

The search was performed using WoS and Scopus, because these databases collect the most significant number of research records with the highest impact worldwide [12]. For example, WoS has more than 90 million records and Scopus has approximately 60 million records [13]. Therefore, this study is in accordance with the new trend of merging the most important databases, which is sufficient for scientometric analysis. The parameters used to perform the searches are listed in Table 1.

The results from Scopus and WoS were merged using the bibliometrix R package for the main information, and tosr for references [14]. The final dataset contains 1697 registers,

TABLE 1. Parameters used to perform the search.

Database	Web of Science	Scopus
Time of search	2000 - 2021	1960 - 2021
Search date	November 13, 2021	
Type of document	Paper, book, chapter, conference proceedings	
Search field	Title	
Search words	"Topic modeling"	
Results	500	1679
Total results (WoS + Scopus)	1697	

18 unique in WoS and not in Sopus. This result indicates that Scopus has almost all papers on TM. Surprisingly, conference papers comprised 53.2% (903) and articles comprised 40.1% (680). This is relevant because TM academic literature is positioned more in conference proceedings than in papers and only 2.06% (35) were reviews. All the variables from Scopus and WoS were included in this study; thus, the data analysis was sufficiently rigorous to understand the main contributions of TM. Code and data are available in the GitHub repository.[1]

B. SCIENTIFIC MAPPING

Belmonte et al. [15] described scientific mapping as a scientometric technique that allows the analysis of academic literature using bibliometric indicators of authors and sources. For example, scientometric analysis is widely used to identify disruptive innovations [16] and university performance [17]. This study focuses on four aspects: scientific production, country, journal, and author analysis. This perspective shows readers the complete scope of a research topic, starting from a general overview and ending with a detailed analysis of the collaboration networks. To take advantage of citation analysis, this study applied the method proposed by Marin-Hurtado [18]. This novel method creates a collaboration network using references. Therefore, it is more accurate to identify the network structures behind the scientometric data. All these procedures were developed in the statistical package RStudio (version 4.1.2), with its complementary package Bibliometrix [14], and visualizations were made using the ggraph R package (version 2.0.6) [19] and Gephi (version 0.9.2) [20].

C. TREE OF SCIENCE

Thematic development is based on a citation analysis of references found in research journals. With these references, a large network is constructed in which each node represents a paper, and the edges that join them are the citations (references). The Tree of Science (ToS) algorithm cleans the network removing papers with one citation (in-degree) and zero references to other papers (out-degree). Also, ToS

[1] https://github.com/coreofscience/topic_modeling_review

extracts the most connected subnetwork (giant component) to remove small islands in the research topic (a more detailed explanation is in [21]. With the clean citation network, Blondel et al. [22] are ae applied to identify subcommunities of densely connected nodes (clusters). The clusters with the highest cohesion indicator (in this case, four) are selected. The SAP algorithm was applied [21] to identify works located in the root (hegemonic), trunk (structural), and leaves (perspective) [23].

The Corporation Core of Science has recently deployed two platforms to create ToS using WoS [24] and Scopus [25]. However, this study used the tosr package and a new code to preprocess the data. The ToS is a well-known and applied methodology for identifying the main contributions of various research topics. This tool has been applied to areas such as entrepreneurship [26], management [27], education [28], and marketing [29]. Eggers et al. [30] provided a detailed description of the diffusion process.

III. RESULTS

A. SCIENTIFICA MAPPING RESULTS

This section presents a descriptive analysis of TM using bibliometric techniques. We analyzed five aspects: scientific production, country production, journal production, and author production.

1) SCIENTIFIC PRODUCTION

Fig. 1 shows the evolution of article production on this subject. As can be seen in this graph, research in this field has grown significantly over the last 11 years, with a significant number of articles produced in the Scopus database. However, between 2009 and 2020, the growth rate in the number of publications in WoS was 50.5% per year, while in Scopus, it was 27.2%.

The scientific production of a certain topic per year is used to understand changes in a research field, and citations received by a paper reflect its relative importance in the academic community. In addition, a production comparison between the Scopus and WoS databases is important because it allows us to recognize the limitations and benefits of selecting both databases. Thus, this study analyzed publications in TM between 2004 and 2021 and the total citations received each year to understand the impact of the topic (see Fig. 1). In addition, the total unique production of the two datasets was used to identify the similarities and differences between TM production in Scopus and WoS. Finally, we divided the evolution of production into three stages: initial growth, rapid development, and stability. These stages allow us to understand the different moments of TM over time [31].

Initial growth stage (2004–2013): The total number of publications during this period was 227 (10.54%). WoS and Scopus had 30 and 197 publications, respectively. This difference is because WoS started publishing TM papers in 2009, and Scopus in 2004. Citations received during this stage represent 35.00% (9105) of the total citations; citations received have a laggard effect because they are generated after the papers have been published. During this stage, the metric steadily increased every year. The most cited study was Wang and Blei [32], who proposed an algorithm to recommend papers using TM.

Rapid development stage (2014–2019): The total number of publications and citations increased sharply every year. This stage represents 55.48% (1195) of the total publications and 58.08% (15112) of the total citations. The average growth percentage of publications was 23.08% and the total number of citations peaked in 2016. The paper most cited in 2016 was by Lie et al. [10], and the most productive source was Lecture Notes in Computer Science, with 62 chapters on TM.

Stability stage (2020–2021): We selected only two years in this stage because the total production stay level during this period; the total number of publications was 33.98% (732), and the growth percentage was -4.60%. The total number of citations received decreased because of the lagged effect of this variable.

2) COUNTRY ANALYSIS

Country analysis is becoming a common scientometric technique to identify the most productive places in the world in a specific topic [33]. The country's productivity reflects the investments of governments in science to increase industry innovation [34]. Therefore, it is important to understand the dynamic of scientific production, quality, and impact of countries' research. This study shows the production (number of papers), quality (according to Scimago metrics) and impact (citations received) of a country's research. Also, a collaboration network is created to understand the communities generated through the interactions among researchers.

There are 76 countries researching TM, and the top ten are shown in Table 2. This list is organized according to the percentage of production of each country concerning the total number of records obtained in this search. These 10 countries produce 70.25% of the total papers, but only the first two produce 40% (USA and China). This behavior could be explained by the dynamics of states such as Silicon Valley, which has tech companies such as Metabob[2] specializing in TM. It is important to clarify that we used all affiliations of each author but removed duplicate affiliations in each paper. These results could differ from those of software such as Bibliometrix, which uses only the first author.

The citation column represents the sum of citations in the WoS and Scopus per country. Similar to the production outcomes, the USA and China cited 53.5% of the total citations. It is worth noting that although Switzerland, Finland, and Denmark do not appear on the list because of their low production labels, these countries have high citation indices of 182, 168, and 144, respectively. In contrast, Japan performs better in scientific production, but not in impact.

[2]http://metabob.com/

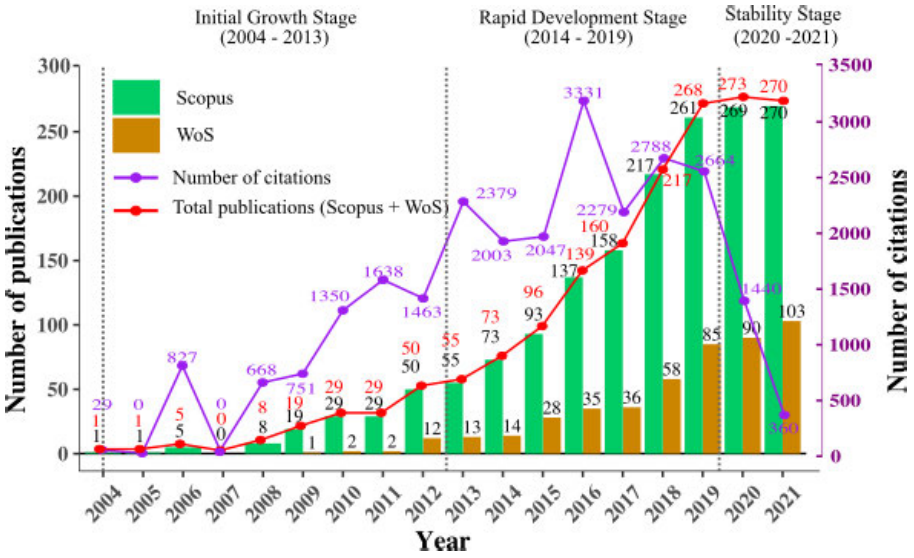

FIGURE 1. Annual scientific production.

TABLE 2. Ten most productive countries in TM research.

Countries	Production	Citations	Q1	Q2	Q3	Q4
USA	445 (25.11%)	8426 (43.34%)	125	25	11	9
China	263 (14.84%)	2610 (13.43%)	99	19	14	6
India	139 (7.84%)	411 (2.11%)	11	5	13	34
Korea	102 (5.75%)	556 (2.86%)	39	10	15	7
United Kingdom	70 (3.95%)	1041 (5.35%)	25	4	4	0
Germany	58 (3.27%)	418 (2.15)	7	5	3	1
Canada	55 (3.10%)	718 (3.69%)	23	0	2	2
Australia	45 (2.54%)	480 (2.47%)	22	2	2	0
Japan	36 (2.03%)	234 (1.20%)	3	1	2	3
Italy	32 (1.81%)	134 (0.69%)	11	2	1	0

Another important variable in the analysis was the quality of production. Table 2 presents the four quartiles of the Scimago dataset. The USA and China have outstanding outcomes: 75.45% of USA production in quartiles is at the top level (Q1), and 69.66% for China's production. Interestingly, India, Japan, and Italy are among the top ten most productive countries, but their quality dropped dramatically in Q1 levels (3, 0, and 4 papers, respectively). Studies without quartile positions in the Scimago dataset were also excluded.

The performance of the USA and China can be explained by the projects together. The five strongest relationships were between the USA and China (75 papers), Australia and China (30 papers), Hong Kong and China (24 papers), the USA and Korea (23 papers), and the USA and the United Kingdom (16 papers). The most recent studies in the USA and China compared eight neural methods for topic modeling in social science [35] and proposed a random walk method for TM [36]. The results show the difficulties to find the optimal number of clusters between TM methods with an experiment using newspapers. Fig. 2 presents six subgroups of countries, and the cluster-by-size inset figure shows a similar size among the five largest clusters. The nodes-and-links through time inner figure shows the interaction between new countries and new relationships over time. According to this figure, new relationships have been increasing since 2016, consolidating TM country collaboration into a strong scientific community.

3) JOURNAL ANALYSIS

IEEE Access has the highest number of publications for each journal. Another outstanding journal is Information Processing and Management, which specializes in computer, decision, engineering, and social sciences. This information is presented in Table 3, which shows that the most important journals regarding factor impact were Knowledge-based systems (1.59), BMC Bioinformatics (1.57), and information sciences (1.52).

Fig. 3 shows the citation analysis using references from the Scopus and WoS searches. The citation network shows different topics of a group of papers. Each node is a journal and the links are references among the journals. This figure presents the three largest communities because the entire network has 20.834 nodes and 37.358 links. The tipping point after cluster four is defined as shown in the inset of figure [18].

TABLE 3. Most productive scientific journals.

Journal	WoS	Scopus	Impact factor	h - index	Quantile
IEEE Access	15	15	0.59	127	Q1
Expert systems with applications	11	14	1.37	207	Q1
Information processing and management	-	13	1.06	101	Q1
Journal of medical internet research	13	12	1.45	142	Q1
IEEE transactions on knowledge and data engineering	11	12	1.36	174	Q1
Sustainability	-	12	0.61	85	Q1
Plos one	6	10	0.99	332	Q1
BMC bioinformatics	9	9	1.57	208	Q1
Knowledge-based systems	8	9	1.59	121	Q1
Information sciences	7	7	1.52	184	Q1

The first community (red) represents new advances in TM technology. For example, the journal most connected to this cluster was the Journal of Machine Learning Research. The last study on TM proposed a new algorithm using a class-specified topic model (CSTM) [37]. Additionally, in this cluster, the ICML (International Conference on Machine Learning and Applications) published papers on TM. An example is an algorithm proposal using semantic-assisted Non-negative Matrix Factorization-based topic modeling (SeNMFk) [38]. SeNMFk was tested with newspapers from BBC, the data has five topics well-identified, and the SeNMFk calculates accurately the right number of clusters.

The second cluster (green) represents the applications of political topics such as debates and climate change. For example, Greene and Cross [39] studied a political agenda using TM, and Grimmer and Stewart [40] presented the advantages and disadvantages of using this method in political texts. In addition, Lesnikowski et al. [41] demonstrated the potential applications of TM in governance literature. The third cluster represents the application of TM in management. This cluster includes journals such as the Journal of Marketing, the Journal of Business Research, and Marketing Science. All of these journals are at the top of marketing and management.

For example, Mustak et al. [42] reviewed Artificial Intelligence in marketing using TM technology. Hyun et al. [43] used TM to analyze the effects of spoilers in online reviews. Fig. 3 presents a clear group of journals related to specific applications of TM. More importantly, TM has a wide range of applications in politics and management.

4) AUTHOR ANALYSIS

Table 4 lists the most productive authors of WoS and Scopus. The author, h-index, number of records, and the most notable publications concerning the number of citations received in the database are shown in each case.

TABLE 4. Most productive authors.

WoS Authors	h - index	Total WoS	Highlighted publication in WoS
Xie, Hao-Ran	6	6	Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education [44].
Chen X.L.	9	5	Fifty years of British Journal of Educational Technology: A topic modeling based bibliometric perspective [45].
Li, Ximing	18	5	Filtering out the noise in short text topic modeling [46].
Park, Haesun	21	5	UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization [47]
Tong, Weida	55	5	Mining FDA drug labels using an unsupervised learning technique - topic modeling [48].
Scopus Authors	**h - index**	**Total Scopus**	**Highlighted publication in Scopus**
Li, Changchun	7	19	Short text topic modeling by exploring original documents [49].
Wang, Cheng	22	19	Multimodal Representation Learning for Recommendation in the Internet of Things [50].
Li, Ximing	11	18	Supervised topic models for multi-label classification [51].
Zhang, Yin	18	17	Supervised coupled dictionary learning with group structures for multi-modal retrieval [52]
Wang, Hanqi	4	13	Identifying Objective and Subjective Words via Topic Modeling [53].

Generally, these studies applied topic modeling to a literature review and analyzed the techniques used in this task. It should also be noted that most authors listed in this table were Chinese.

We generated an Academic Social Network (ASN) with data from the search and their references [54] (see Fig. 3), where each node in the ASN is an author and a link is created when they publish a paper together. The final ASN has 47417 nodes and 318478 links and because of its size, we split the ASN into clusters [22] and show the three biggest clusters with the 10 most connected authors of each one (highest degree).

The first academic community (red) was related to advances in TM. Professor Ximing Li has worked on TM since 2015. In his first paper, he and his colleagues identified some limitations of LDA in identifying topics and proposed a new algorithm called group latent Dirichlet allocation (GLDA) [55]. GLDA shows high performance in evaluating extremely sparse short texts on social media. Dr. Li worked with Professor Yang Wang on topics related to topic extraction [56], and Professor Xinhua Wang on topics related to improving the methods for selecting words in topics [57]. Dr. Li and Dr. Wang worked at the College of

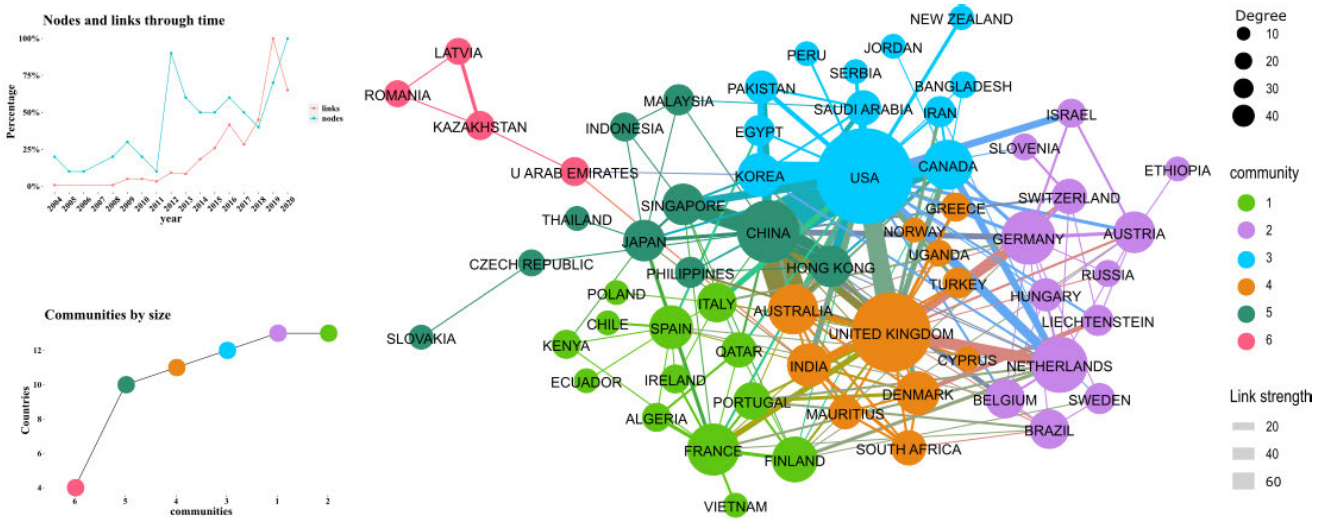

FIGURE 2. Country collaboration map.

Computer Science and Technology, Jilin University (China). Professors Yuefeng Li and Jinglan Zhang worked on several papers. For example, they reviewed the main TM techniques used in customer reviews [58]. Both professors worked at Queensland University of Technology (Australia).

The second community (green) represents the application of TM in different areas. For example, Kim and Lee studied blockchain technology using TM [59]. Both professors were from the School of Management Engineering in the Republic of Korea. Lee, Kim, and other colleagues applied TM to understand the perceptions of smartwatch use [60]. In addition, both the researchers were from the Republic of Korea (Sungkyunkwan University). Another application is perceived trust in educational texts by professors Jaehyun Park, Minyeong Kim, and other co-authors [61]. The two professors were from Incheon National University in Korea. Another study by Korean researchers analyzed diabetes with TM [62]; professors Lee Junghye, Youngji Kim, and Seungmi Park are from different universities but are relatively close to each other.

The last community (blue) shows the contributions of researchers from the United States. Professor David Blein from Columbia University in New York was the most active author of this cluster. He wrote a classic book in TM with professors Andrew Y. Ng (Stanford University) and Michel I. Jordan (University of California) [8]. Dr. Blein also worked with Professor Hanna M. Wallach of Microsoft Research [63] and Professor David Mimno of Princeton University [64]. Finally, Professor Padhraic Smyth from the University of California and Andrew K. McCallum from the University of Massachusetts Amherst worked together [65], [66]. These three clusters show the influence of geographic proximity on academic team development [30].

B. TREE OF SCIENCE (ToS)

The network analysis allowed us to identify the most relevant documents. Records with the highest indicators were selected for review and organized using the metaphor of the tree of science: classic (roots), structural (trunk), and recent (leaves) [21]. The clustering algorithm proposed by Blondel et al. [22] was used to establish the subareas or common areas of research, thus identifying the four main groups that could be observed in the leaves.

1) ROOT (CLASSICS)

Based on the results of the SAP algorithm, Deerwester et al. [6] identified the first study on roots. The authors have described an automatic indexing and retrieval method. This approach takes advantage of the higher-order latent structure by associating terms with documents (''semantic structure'') to improve the detection of related papers from words found in the query. These elements were the starting points for defining the LSI strategy described above. A few years later, Hoffman [67] proposed a variation in LSI that improves synonymy and polysemous word problems.

Latent Dirichlet Allocation (LDA) is a well-known TM technique. This technique is a hierarchical Bayesian model, in which each item in a corpus is modeled as a finite mixture over a set of latent topics. Each topic is modeled from an infinite combination of a set of latent probabilities of words that explicitly represent documents [8]. Subsequently, an essential contribution to the development and application of this methodology was presented in [68], who proposed an algorithm based on the Monte Carlo method with Markov chains to perform inferences in this model. The work presented by Chang and Blei [69] is also important, in which a strategy called the relational thematic model (RTM) is proposed. This binary random variable models the links between documents based on their content, and allows the prediction of word sequences between them.

2) TRUNK (STRUCTURAL)

Among the group of structural articles, we highlight the study by Wang and Blei [32]. They combined collaborative

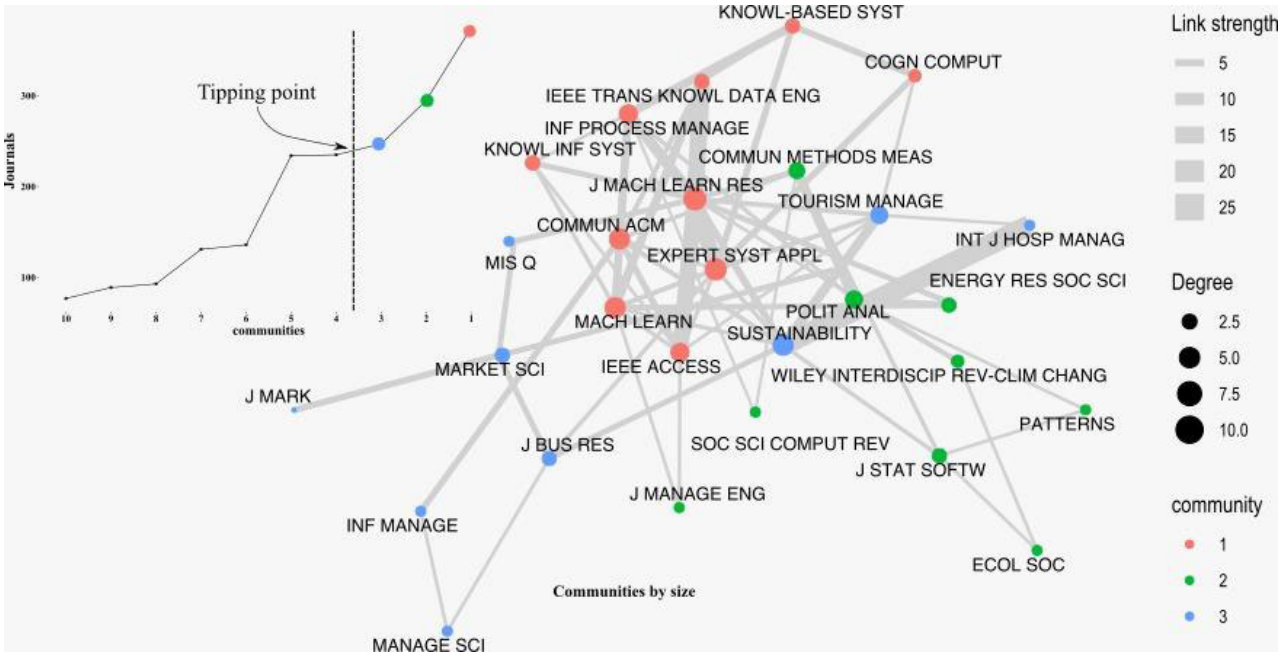

FIGURE 3. Journal citation analysis of TM.

filtering methods with probabilistic topic modeling to find latent structures that are easily interpretable and to generate recommendations on existing and recently published articles. In addition, Cheng et al. [70] showed that LDA and pLSA techniques fail to identify topics in short texts because of the scarcity of coexisting word patterns in these texts. To address this, the authors provide a new method of modeling topics in brief text called bithermal topic modeling (BTM). This model learns topics by directly modeling the generation of word co-occurrence patterns (i.e., bitherms) in a corpus, thereby enabling efficient inference with rich corpus-level information. The experiments developed in this study with short collections of texts show that BTM can explore more extensive and cohesive topics and significantly outperform the results achieved with other techniques [71].

Finally, it is worth highlighting that a series of studies has considered the presence of topics over time. These studies identified and evaluated expert opinions on COVID - 19 [72] during the outbreak period compared with other periods [73].

3) LEAVES (PERSPECTIVES)

The papers that were grouped into categories of perspectives are presented below. The methodology applied made it possible to group these papers into four groups or clusters (Fig. 5).

a: CLUSTER 1: SOCIAL MEDIA

The first perspective is characterized by a set of applications that identify patterns in the interactions of social network users regarding personal and social behavior; therefore, applications of TM with social media data. Alfred et al. [74] proposed a technique called the multi-objective genetic algorithm (MOGA) based on text clustering for topic extraction. This method was applied to perform supervised classification analysis of Twitter interactions. Another work that stands out in this group is that developed by Li et al. [75], in which topic modeling was used to identify situational interactions concerning the COVID - 19 pandemic and how users use social media to acquire and exchange various types of information.

Other examples of TM applications on Twitter include an analysis of the Brexit debate, an investigation of how China influenced the perceptions of Hong Kong's protests, and a study of the controversial Gillette campaign. del Gobbo et al. [76] presented a three-and-a-half-year study on the famous Brexit debate in England. They analyzed 33 million tweets and identified 20 topics. Zhang et al. [77] analyzed 14,412 tweets posted by 13 organizations and identified six strategies: conflict, violence, and calling for a stable order. Xu and Xiong [78] investigated 100,000 tweets from Gillette's campaign on toxic masculinity, and suggested that influencers play an important role in influencing users' perceptions.

Other TM studies on other social media networking sites include those by Jiang et al. [79] and Törnberg and Törnberg [80]. Jiang et al. [79] investigated the spread of information regarding the human papillomavirus vaccine information in China. They studied the information propagation and information acquisition process, and identified the impact of social media on health knowledge. Törnberg and Törnberg [80] studied the patterns of representation of the words Muslim and Islam in 105 million words from Internet forums between 2000 and 2013. This work is one of the most

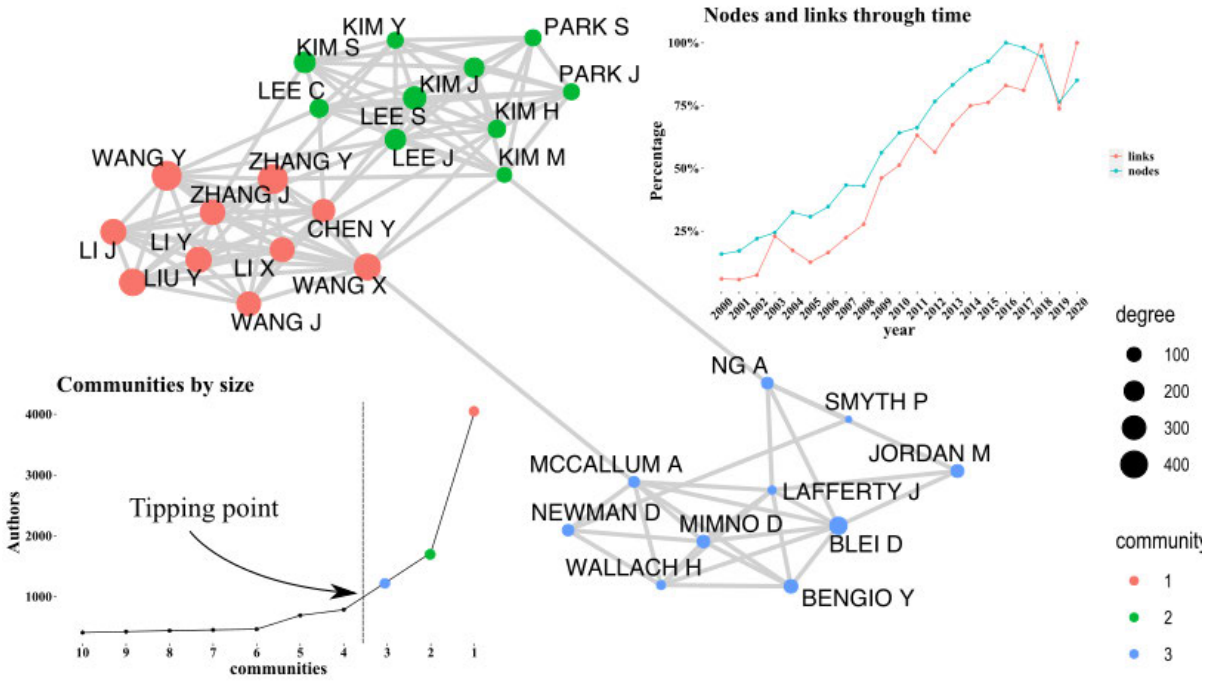

FIGURE 4. Academic social network for TM.

prominent studies on the discursive power of social media in the modern society.

b: CLUSTER 2: TOPIC MODELING FOR SHORT TEXTS

This cluster presents different algorithms in short texts to identify topics. Mai et al. [81] proposed an alternative algorithm called TSSE-MMM that focuses on improving the coherence and interpretability of topics by subdividing the issues and applying semantic enhancement and word embedding to alleviate the problem of low data availability. In addition, Steuber et al. [82] presented an algorithm called A-LDA (Archetypal LDA), which identifies topics without supervision and with co-occurrence evaluation through archetypal analysis. The FastText-based Sentence-LDA (FSL) is another algorithm for short texts [81]. The FSL is divided into two steps: the first trains a word-embedding model with replacement, and the second is a latent model that integrates words in sentences. Lin et al. [84] studied a new method for identifying topics in short sentences by using neural networks and Archimedean copulas. Another new method proposed by He et al. [85] assigns one target from a Dirichlet Multinomial Mistureprodess and discovers several angles to identify the topic (Targeter Aspects Oriented Topic Modeling, TATM).

Similarly, Li et al. [10] considered the word embedding method and proposed a GPU-DMM model based on the Dirichlet multinomial mixture model. The same strategy was used by Qiang et al. [7], where the method was defined as an embedding-based topic model (ETM). In addition, a regularized Markov random field model that provides correlated words with a better chance of being placed on the same topic is presented. A more sophisticated complement to these techniques is based on Recurrent Neural Networks (RNN) to learn these relationships and filter high-frequency words [86].

c: CLUSTER 3: SENTIMENT ANALYSIS

Topic modeling has essential applications in sentiment analysis of information generally posted on blogs and social networks, such as Twitter. Therefore, the third cluster comprised publications reporting results mainly related to the study of information published on social networks, mainly about COVID - 19. For example, Singh et al. [87] analyzed the myths and prevalence of low-quality information circulated on a large scale through social networks about new news about the spread of the disease.

Topic modeling was also applied by Stokes et al. [88] to identify the evolution of COVID-19 discussion topics in an online public forum as of March 2020, and by Ordun et al. [89] to analyze circulating information regarding the spread of cases, healthcare workers, and personal protective equipment in the USA. Similarly, after 2020, studies appeared in which TM was used to analyze the large amounts of information that made it possible to document the experience of patients who have suffered from COVID-19 [90]. In turn, in Ma et al. [91], TM was used to

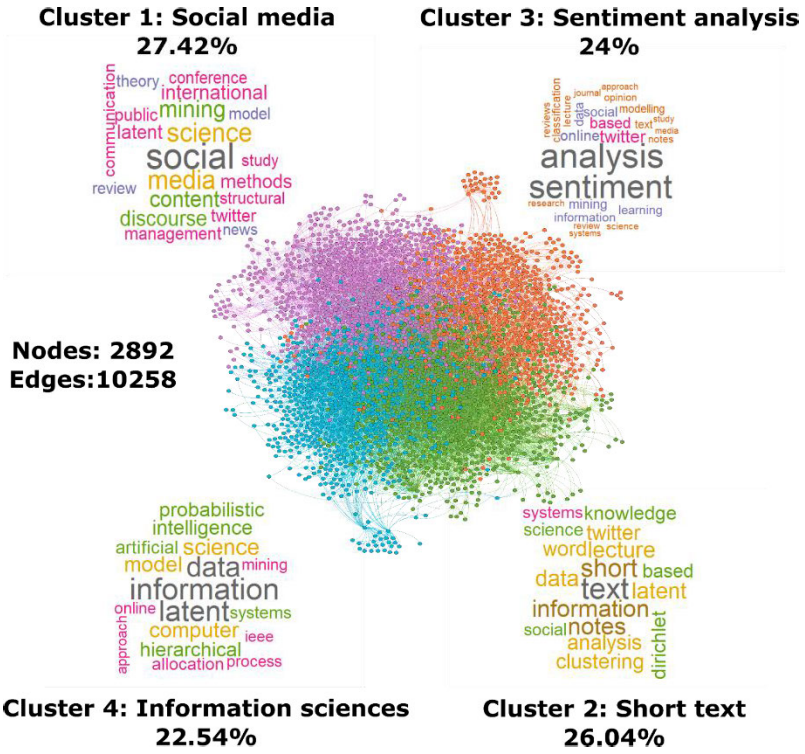

FIGURE 5. Academic social network for TM.

model issues regarding the perception of the effectiveness of vaccines and the indifference of various groups of people toward this protective measure. Wang et al. [92] identified the reactions of people to Twitter regarding the use of masks and vaccines.

We also found studies in which TM was applied to sentiment analysis of information published on social networks and opinion sites on topics such as customer satisfaction in tourism and hotel services [93], [94], [95], [96], [97], consumer behavior [98], [99], [100], [101], and nursing staff experience [102], [103].

d: CLUSTER 4: INFORMATION SCIENCES

This subtopic represents papers that analyze trends in information science. For example, Han [104] investigated the evolution of library and information science from 1996 to 2019, similar to that reported by Miyata et al. [105], who analyzed the thematic transition from 2000 to 2002 to 2015–2017. Kurata et al. [106] analyzed 1648 full-text articles related to information science and published them in the five most prominent journals. Baghmohammad et al. [107] studied the thematic trends in Iran. Ianina et al. [108] implemented additive regularization of thematic models (ARTM) to build a model that met multiple objectives, thereby reducing the cycle of time-query navigation and refinement. All of these studies used the LDA technique to identify the underlying themes in the corpus of documents.

IV. CONCLUSION

This study had three objectives: to map scientific production concerning topical modeling, to identify the most prominent authors and journals, and to identify the main applications and emerging trends in this line of research. We reached these objectives based on 1697 records from the WoS and Scopus databases from 2000 in WoS and from 1960 in Scopus to November 13, 2021. The findings reported here shed light on the evolution and different applications of TM.

This study demonstrates the notorious growth of TM between 2013 and 2019 (Fig 1). It was also observed that the USA and China were countries with high quality and a significant number of contributions (Table 2). The most popular journals in TM are IEEE Access and Expert

Systems with Applications (see Table 3). According to Fig. 3, there are three main subgroups in the academic social network: China, Korea, and the USA. The most prominent authors are Professors Hao-Ran Xie, Chen X.L., Ximing Li, and Changchun Li. The high quality of the journals in which these studies were published was also highlighted.

Furthermore, four clusters were identified in which the main applications of topic modeling were related. The first cluster comprised studies in which different techniques were applied and defined for TM in short text. This group is also related to Cluster two, where TM is applied from a specific perspective: personal and social behavior on social media sites. Cluster three stands out for being formed by works that characterize scientific production in information science. Finally, Cluster four identifies studies related to

sentiment analysis, one of the best-known applications of topic modeling in different contexts.

This study highlights the application of a methodology for the compilation and organization of scientific records. The methodology is based on graph theory and social networks, by grouping authors and documents in order of the number of citations among them. This is a strategy that complements the many existing strategies for systematic literature review. In addition to offering consistent and easy-to-interpret results; also, it has been successfully applied in important literature reviews [54], [109], [110].

A limitation of this study is the multidisciplinary nature observed in applying these techniques in literature review, finding records in medicine, social sciences, and computer science, where new methodologies or combinations of existing methods have been proposed to improve their performance. This leads to the fact that the performance evaluation of these models should be made, not only from theoretical metrics (perplexity, coherence, etc.) but also from the perspective of a subject matter expert.

For future research, we suggest approaching TM from three perspectives: new advances, applications in the definition of underlying topics in specific areas, and latent challenges in computer and basic sciences.

ACKNOWLEDGMENT

The authors would like to thank the Universidad Católica Luis Amigó for supporting the research process, the Universidad de Caldas for arranging training courses in machine learning and deep learning within the framework of the Doctorate in Science, and the Universidad Nacional Abierta y a Distancia (UNAD) for supporting this wonderful cause. It was also based on advances in doctoral studies by the doctorate in the sciences of the Universidad de Caldas.

REFERENCES

[1] L. Rossetto, R. Gasser, S. Heller, M. A. Parian, and H. Schuldt, "Retrieval of structured and unstructured data with vitrivr," in *Proc. ACM Workshop Lifelog Search Challenge*, Ottawa ON, Canada, Jun. 2019, pp. 27–31, doi: 10.1145/3326460.3329160.

[2] P. Lisena, I. Harrando, O. Kandakji, and R. Troncy, "TOMODAPI: A topic modeling API to train, use and compare topic models," in *Proc. 2nd Workshop NLP Open Source Softw. (NLP-OSS)*, 2020, pp. 132–140, doi: 10.18653/v1/2020.nlposs-1.19.

[3] T. Luostarinen and O. Kohonen, "Using topic models in content-based news recommender systems," in *Proc. 19th Nordic Conf. Comput. Linguistics (NODALIDA)*, 2013, pp. 239–251. [Online]. Available: https://www.aclweb.org/anthology/W13-5622.pdf

[4] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Jun. 2017, pp. 745–750, doi: 10.1109/ICCONS.2017.8250563.

[5] P. Kherwa and P. Bansal, "Topic modeling: A comprehensive review," *ICST Trans. Scalable Inf. Syst.*, vol. 7, no. 24, Jul. 2018, Art. no. 159623, doi: 10.4108/eai.13-7-2018.159623.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[7] J. Qiang, P. Chen, T. Wang, and X. Wu, "Topic modeling over short texts by incorporating word embeddings," in *Advances in Knowledge Discovery and Data Mining*. Cham, Switzerland: Springer, 2017, pp. 363–374, doi: 10.1007/978-3-319-57529-2_29.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com

[9] X. Li and L. Lei, "A bibliometric analysis of topic modelling studies (2000–2017)," *J. Inf. Sci.*, vol. 47, no. 2, pp. 161–175, Apr. 2021, doi: 10.1177/0165551519877049.

[10] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Pisa, Italy, Jul. 2016, pp. 165–174, doi: 10.1145/2911451.2911499.

[11] J. Hou, X. Yang, and C. Chen, "Emerging trends and new developments in information science: A document co-citation analysis (2009–2016)," *Scientometrics*, vol. 115, no. 2, pp. 869–892, May 2018, doi: 10.1007/s11192-018-2695-9.

[12] T. D. Marín-Velásquez and D. D. J. Arrojas-Tocuyo, "Revistas científicas de América latina y el caribe en SciELO, scopus y web of science en el área de ingeniería y tecnología: Su relación con variables socioeconómicas," *Revista Española de Documentación Científica*, vol. 44, no. 3, p. e301, Jul. 2021, doi: 10.3989/redc.2021.3.1812.

[13] J. A. Moral-Muñoz, E. Herrera-Viedma, A. Santisteban-Espejo, and M. J. Cobo, "Software tools for conducting bibliometric analysis in science: An up-to-date review," *El Profesional de la Información*, vol. 29, no. 1, pp. 1–20, Jan. 2020, doi: 10.3145/epi.2020.ene.03.

[14] M. Aria and C. Cuccurullo, "Bibliometrix : An R-tool for comprehensive science mapping analysis," *J. Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.

[15] J. L. Belmonte, A. Segura-Robles, A.-J. Moreno-Guerrero, and M. E. Parra-González, "Machine learning and big data in the impact literature. A bibliometric review with scientific mapping in web of science," *Symmetry*, vol. 12, no. 4, p. 495, Mar. 2020, doi: 10.3390/sym12040495.

[16] L. Leydesdorff, A. Tekles, and L. Bornmann, "A proposal to revise the disruption index," *Profesional De La Información*, vol. 30, pp. 1–6, Feb. 2021, doi: 10.3145/epi.2021.ene.21.

[17] H. A. Al-Jamimi, G. M. BinMakhashen, and L. Bornmann, "Use of bibliometrics for research evaluation in emerging markets economies: A review and discussion of bibliometric indicators," *Scientometrics*, vol. 127, no. 10, pp. 5879–5930, Oct. 2022, doi: 10.1007/s11192-022-04490-8.

[18] V. A. Hurtado-Marín, J. D. Agudelo-Giraldo, S. Robledo, and E. Restrepo-Parra, "Analysis of dynamic networks based on the ising model for the case of study of co-authorship of scientific articles," *Sci. Rep.*, vol. 11, no. 1, p. 5721, Mar. 2021, doi: 10.1038/s41598-021-85041-8.

[19] B. Si, Y. Liang, J. Zhao, Y. Zhang, X. Liao, H. Jin, H. Liu, and L. Gu, "GGraph: An efficient structure-aware approach for iterative graph processing," *IEEE Trans. Big Data*, vol. 8, no. 5, pp. 1182–1194, Oct. 2022, doi: 10.1109/TBDATA.2020.3019641.

[20] M. Bastian, S. Heymann, and M. Jacomy. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewPaper/154

[21] D. S. Valencia-Hernandez, S. Robledo, R. Pinilla, N. D. Duque-Méndez, and G. Olivar-Tost, "SAP algorithm for citation analysis: An improvement to tree of science," *Ingeniería e Investigación*, vol. 40, no. 1, pp. 45–49, Jan. 2020, doi: 10.15446/ing.investig.v40n1.77718.

[22] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008, doi: 10.1088/1742-5468/2008/10/P10008.

[23] J. Zhang and Y. Luo, "Degree centrality, betweenness centrality, and closeness centrality in social network," in *Proc. 2nd Int. Conf. Model., Simul. Appl. Math. (MSAM)*, vol. 132, 2017, pp. 300–303, doi: 10.2991/msam-17.2017.68.

[24] M. Zuluaga, S. Robledo, O. Arbelaez-Echeverri, G. A. Osorio-Zuluaga, and N. Duque-Méndez, "Tree of science–ToS: A web-based tool for scientific literature recommendation. Search less, research more!" *Issues Sci. Technol. Librarianship*, vol. 100, no. 100, pp. 1–10, Aug. 2022, doi: 10.29173/istl2696.

[25] S. Robledo, M. Zuluaga, L.-A. Valencia-Hernandez, O. A.-E. Arbelaez-Echeverri, P. Duque, and J.-D. Alzate-Cardona, "Tree of science with scopus: A shiny application," *Issues Sci. Technol. Librarianship*, vol. 100, no. 100, pp. 1–7, Aug. 2022, doi: 10.29173/istl2698.

[26] S. Robledo, A. M. G. Aguirre, M. Hughes, and F. Eggers, "'Hasta la vista, baby'—Will machine learning terminate human literature reviews in entrepreneurship?" *J. Small Bus. Manag.*, pp. 1–30, Aug. 2021, doi: 10.1080/00472778.2021.1955125.

[27] J. A. Vivares, W. Sarache, and J. E. Hurtado, "A maturity assessment model for manufacturing systems," *J. Manuf. Technol. Manag.*, vol. 29, no. 5, pp. 746–767, May 2018, doi: 10.1108/JMTM-07-2017-0142.

[28] P. Duque and L.-S. Cervantes-Cervantes, "Responsabilidad social universitaria: Una revisión sistemática y análisis bibliométrico," *Estudios Gerenciales*, vol. 35, pp. 451–464, Dec. 2019, doi: 10.18046/j.estger.2019.153.3389.

[29] P. Duque-Hurtado, V. Samboni-Rodriguez, M. Castro-Garcia, L. A. Montoya-Restrepo, and I. A. Montoya-Restrepo, "Neuromarketing: Its current status and research perspectives," *Estudios Gerenciales*, vol. 36, pp. 525–539, Nov. 2020, doi: 10.18046/j.estger.2020.157.3890.

[30] F. Eggers, H. Risselada, T. Niemand, and S. Robledo, "Referral campaigns for software startups: The impact of network characteristics on product adoption," *J. Bus. Res.*, vol. 145, pp. 309–324, Jun. 2022, doi: 10.1016/j.jbusres.2022.03.007.

[31] L. Sun, L. Wu, and P. Qi, "Global characteristics and trends of research on industrial structure and carbon emissions: A bibliometric analysis," *Environ. Sci. Pollut. Res.*, vol. 27, no. 36, pp. 44892–44905, Dec. 2020, doi: 10.1007/s11356-020-10915-9.

[32] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Diego, CA, USA, 2011, pp. 448–456, doi: 10.1145/2020408.2020480.

[33] X. Chen, "Does cross-field influence regional and field-specific distributions of highly cited researchers?" *Scientometrics*, pp. 1–16, Nov. 2022, doi: 10.1007/s11192-022-04584-3.

[34] C. Zanardello, "Market forces in Italian academia today (and yesterday)," *Scientometrics*, Nov. 2022. [Online]. Available: https://link.springer.com/journal/11192/online-first?page=2, doi: 10.1007/s11192-022-04579-0.

[35] Q. Fu, Y. Zhuang, J. Gu, Y. Zhu, and X. Guo, "Agreeing to disagree: Choosing among eight topic-modeling methods," *Big Data Res.*, vol. 23, Feb. 2021, Art. no. 100173, doi: 10.1016/j.bdr.2020.100173.

[36] J. Tang, R. Jin, and J. Zhang, "A topic modeling approach and its integration into the random walk framework for academic search," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 1055–1060, doi: 10.1109/ICDM.2008.71.

[37] F. Wang, J. L. Zhang, Y. Li, K. Deng, and J. S. Liu, "Bayesian text classification and summarization via a class-specified topic model," *J. Mach. Learn. Res.*, vol. 22, no. 89, pp. 1–48, 2021. [Online]. Available: https://www.jmlr.org/papers/volume22/18-332/18-332.pdf

[38] R. Vangara, E. Skau, G. Chennupati, H. Djidjev, T. Tierney, J. P. Smith, M. Bhattarai, V. G. Stanev, and B. S. Alexandrov, "Semantic nonnegative matrix factorization with automatic model determination for topic modeling," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2020, pp. 328–335, doi: 10.1109/ICMLA51294.2020.00060.

[39] D. Greene and J. P. Cross, "Exploring the political agenda of the European parliament using a dynamic topic modeling approach," *Political Anal.*, vol. 25, no. 1, pp. 77–94, Jan. 2017, doi: 10.1017/pan.2016.7.

[40] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Anal.*, vol. 21, no. 3, pp. 267–297, 2013, doi: 10.1093/pan/mps028.

[41] A. Lesnikowski, E. Belfer, E. Rodman, J. Smith, R. Biesbroek, J. D. Wilkerson, J. D. Ford, and L. Berrang-Ford, "Frontiers in data analytics for adaptation research: Topic modeling," *WIREs Climate Change*, vol. 10, no. 3, p. e576, May 2019, doi: 10.1002/wcc.576.

[42] M. Mustak, J. Salminen, L. Plé, and J. Wirtz, "Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda," *J. Bus. Res.*, vol. 124, pp. 389–404, Jan. 2021, doi: 10.1016/j.jbusres.2020.10.044.

[43] J. H. Ryoo, X. Wang, and S. Lu, "Do spoilers really spoil? Using topic modeling to measure the effect of spoiler reviews on box office revenue," *J. Marketing*, vol. 85, no. 2, pp. 70–88, Mar. 2021, doi: 10.1177/0022242920937703.

[44] X. Chen, D. Zou, G. Cheng, and H. Xie, "Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of computers & education," *Comput. Educ.*, vol. 151, Jul. 2020, Art. no. 103855, doi: 10.1016/j.compedu.2020.103855.

[45] X. Chen, D. Zou, and H. Xie, "Fifty years of British journal of educational technology : A topic modeling based bibliometric perspective," *Brit. J. Educ. Technol.*, vol. 51, no. 3, pp. 692–708, May 2020, doi: 10.1111/bjet.12907.

[46] X. Li, Y. Wang, A. Zhang, C. Li, J. Chi, and J. Ouyang, "Filtering out the noise in short text topic modeling," *Inf. Sci.*, vol. 456, pp. 83–96, Aug. 2018, doi: 10.1016/j.ins.2018.04.071.

[47] C. Jaegul, L. Changhyun, C. K. Reddy, and P. Haesun, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 1992–2001, Dec. 2013, doi: 10.1109/TVCG.2013.212.

[48] H. Bisgin, Z. Liu, H. Fang, X. Xu, and W. Tong, "Mining FDA drug labels using an unsupervised learning technique–topic modeling," *BMC Bioinf.*, vol. 12, no. 10, pp. 1–8, Oct. 2011, doi: 10.1186/1471-2105-12-S10-S11.

[49] X. Li, C. Li, J. Chi, and J. Ouyang, "Short text topic modeling by exploring original documents," *Knowl. Inf. Syst.*, vol. 56, no. 2, pp. 443–462, 2018, doi: 10.1007/s10115-017-1099-0.

[50] Z. Huang, X. Xu, J. Ni, H. Zhu, and C. Wang, "Multimodal representation learning for recommendation in Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10675–10685, Dec. 2019, doi: 10.1109/JIOT.2019.2940709.

[51] X. Li, J. Ouyang, and X. Zhou, "Supervised topic models for multi-label classification," *Neurocomputing*, vol. 149, pp. 811–819, Feb. 2015, doi: 10.1016/j.neucom.2014.07.053.

[52] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu. (Jun. 2013). *Supervised Coupled Dictionary Learning With Group Structures for Multi-Modal Retrieval*. Accessed: Dec. 7, 2021. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/viewPaper/6323

[53] H. Wang, F. Wu, W. Lu, Y. Yang, X. Li, X. Li, and Y. Zhuang, "Identifying objective and subjective words via topic modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 718–730, Mar. 2018, doi: 10.1109/TNNLS.2016.2626379.

[54] D. Durán-Aranguren, S. Robledo, E. Gomez-Restrepo, J. A. Valencia, and N. Tarazona, "Scientometric overview of coffee by-products and their applications," *Molecules*, vol. 26, no. 24, p. 7605, Dec. 2021, doi: 10.3390/molecules26247605.

[55] X. Li, J. Ouyang, Y. Lu, X. Zhou, and T. Tian, "Group topic model: Organizing topics into groups," *Inf. Retr. J.*, vol. 18, no. 1, pp. 1–25, Feb. 2015, doi: 10.1007/s10791-014-9244-9.

[56] X. Li, Y. Wang, J. Ouyang, and M. Wang, "Topic extraction from extremely short texts with variational manifold regularization," *Mach. Learn.*, vol. 110, no. 5, pp. 1029–1066, May 2021, doi: 10.1007/s10994-021-05962-3.

[57] J. Chi, J. Ouyang, C. Li, X. Dong, X. Li, and X. Wang, "Topic representation: Finding more representative words in topic models," *Pattern Recognit. Lett.*, vol. 123, pp. 53–60, May 2019, doi: 10.1016/j.patrec.2019.01.018.

[58] L. D. C. S. Subhashini, Y. Li, J. Zhang, A. S. Atukorale, and Y. Wu, "Mining and classifying customer reviews: A survey," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 6343–6389, Dec. 2021, doi: 10.1007/s10462-021-09955-5.

[59] S. Kim, H. Park, and J. Lee, "Word2vec-based latent semantic analysis (W2 V-LSA) for topic modeling: A study on blockchain technology trend analysis," *Exp. Syst. Appl.*, vol. 152, Aug. 2020, Art. no. 113401, doi: 10.1016/j.eswa.2020.113401.

[60] T. Ha, B. Beijnon, S. Kim, S. Lee, and J. H. Kim, "Examining user perceptions of smartwatch through dynamic topic modeling," *Telematics Informat.*, vol. 34, no. 7, pp. 1262–1273, Nov. 2017, doi: 10.1016/j.tele.2017.05.011.

[61] Y. Im, J. Park, M. Kim, and K. Park, "Comparative study on perceived trust of topic modeling based on affective level of educational text," *Appl. Sci.*, vol. 9, no. 21, p. 4565, Oct. 2019, doi: 10.3390/app9214565.

[62] J. Lee, Y. Kim, E. Kwak, and S. Park, "A study on research trends for gestational diabetes mellitus and breastfeeding: Focusing on text network analysis and topic modeling," *J. Korean Academic Soc. Nursing Educ.*, vol. 27, no. 2, pp. 175–185, May 2021, doi: 10.5977/jkasne.2021.27.2.175.

[63] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, Pennsylvania, PA, USA, 2006, pp. 113–120, doi: 10.1145/1143844.1143859.

[64] D. Mimno, M. Hoffman, and D. Blei, "Sparse stochastic inference for latent Dirichlet allocation," 2012, *arXiv:1206.6425*.

[65] D. M. Blei and P. Smyth, "Science and data science," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 33, pp. 8689–8692, Aug. 2017, doi: 10.1073/pnas.1702076114.

[66] W. Li, D. Blei, and A. McCallum, "Nonparametric Bayes pachinko allocation," 2012, *arXiv:1206.5270*.

[67] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1999, pp. 50–57, doi: 10.1145/312624.312649.

[68] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004, doi: 10.1073/pnas.0307752101.

[69] J. Chang and D. M. Blei, "Relational topic models for document networks," *J. Mach. Learn. Res.*, vol. 5, pp. 81–88, Apr. 2009. [Online]. Available: https://proceedings.mlr.press/v5/chang09a.html

[70] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic modeling over short texts," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2928–2941, Dec. 2014, doi: 10.1109/TKDE.2014.2313872.

[71] P. Bicalho, M. Pita, G. Pedrosa, A. Lacerda, and G. L. Pappa, "A general framework to expand short text for topic modeling," *Inf. Sci.*, vol. 393, pp. 66–81, Jul. 2017, doi: 10.1016/j.ins.2017.02.007.

[72] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Temporal expert finding through generalized time topic modeling," *Knowl.-Based Syst.*, vol. 23, no. 6, pp. 615–625, Aug. 2010, doi: 10.1016/j.knosys.2010.04.008.

[73] A. R. Alharbi, M. Hijji, and A. Aljaedi, "Enhancing topic clustering for Arabic security news based on K-means and topic modelling," *IET Netw.*, vol. 10, no. 6, pp. 278–294, Nov. 2021, doi: 10.1049/ntw2.12017.

[74] R. Alfred, L. Y. Jie, J. H. Obit, Y. Lim, H. Haviluddin, and A. Azman, "Social media mining: A genetic based multiobjective clustering approach to topic modelling," *IAENG Int. J. Comput. Sci.*, vol. 48, no. 1, pp. 32–42, 2021. [Online]. Available: http://www.iaeng.org/IJCS/issues_v48/issue_1/IJCS_48_1_04.pdf

[75] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T. L. Gao, W. Duan, K. K. F. Tsoi, and F. Y. Wang, "Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 2, pp. 556–562, Apr. 2020, doi: 10.1109/TCSS.2020.2980007.

[76] E. Del Gobbo, S. Fontanella, A. Sarra, and L. Fontanella, "Emerging topics in Brexit debate on Twitter around the deadlines," *Social Indicators Res.*, vol. 156, nos. 2–3, pp. 669–688, Aug. 2021, doi: 10.1007/s11205-020-02442-4.

[77] M. M. Zhang, X. Wang, and Y. Hu, "Strategic framing matters but varies: A structural topic modeling approach to analyzing China's foreign propaganda about the 2019 Hong Kong protests on Twitter," *Social Sci. Comput. Rev.*, vol. 41, pp. 265–285, Sep. 2021, doi: 10.1177/08944393211042575.

[78] S. Xu and Y. Xiong, "Setting socially mediated engagement parameters: A topic modeling and text analytic approach to examining polarized discourses on Gillette's campaign," *Public Relations Rev.*, vol. 46, no. 5, Dec. 2020, Art. no. 101959, doi: 10.1016/j.pubrev.2020.101959.

[79] S. Jiang, P. Wang, P. L. Liu, A. Ngien, and X. Wu, "Social media communication about HPV vaccine in China: A study using topic modeling and survey," *Health Commun.*, pp. 1–12, Sep. 2021, doi: 10.1080/10410236.2021.1983338.

[80] A. Törnberg and P. Törnberg, "Muslims in social media discourse: Combining topic modeling and critical discourse analysis," *Discourse, Context Media*, vol. 13, pp. 132–142, Sep. 2016, doi: 10.1016/j.dcm.2016.04.003.

[81] C. Mai, X. Qiu, K. Luo, M. Chen, B. Zhao, and Y. Huang, "TSSE-DMM: Topic modeling for short texts based on topic subdivision and semantic enhancement," in *Advances in Knowledge Discovery and Data Mining*. Cham, Switzerland: Springer, 2021, pp. 640–651, doi: 10.1007/978-3-030-75765-6_51.

[82] F. Steuber, M. Schoenfeld, and G. D. Rodosek, "Topic modeling of short texts using anchor words," in *Proc. 10th Int. Conf. Web Intell., Mining Semantics*, Jun. 2020, pp. 210–219, doi: 10.1145/3405962.3405968.

[83] F. Zhang, W. Gao, Y. Fang, and B. Zhang, "Enhancing short text topic modeling with FastText embeddings," in *Proc. Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE)*, Jun. 2020, pp. 255–259, doi: 10.1109/ICBAIE49996.2020.00060.

[84] L. Lin, H. Jiang, and Y. Rao, "Copula guided neural topic modelling for short texts," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1773–1776, doi: 10.1145/3397271.3401245.

[85] J. He, L. Li, Y. Wang, and X. Wu, "Targeted aspects oriented topic modeling for short texts," *Int. J. Speech Technol.*, vol. 50, no. 8, pp. 2384–2399, Aug. 2020, doi: 10.1007/s10489-020-01672-w.

[86] H.-Y. Lu, L.-Y. Xie, N. Kang, C.-J. Wang, and J.-Y. Xie, "Don't forget the quantifiable relationship between words: Using recurrent neural network for short text topic discovery," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, pp. 1192–1198. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10670

[87] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, "A first look at COVID-19 information and misinformation sharing on Twitter," 2020, *arXiv:2003.13907*.

[88] D. C. Stokes, A. Andy, S. C. Guntuku, L. H. Ungar, and R. M. Merchant, "Public priorities and concerns regarding COVID-19 in an online discussion forum: Longitudinal topic modeling," *J. Gen. Internal Med.*, vol. 35, no. 7, pp. 2244–2247, Jul. 2020, doi: 10.1007/s11606-020-05889-w.

[89] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of COVID-19 tweets using topic modeling, UMAP, and DiGraphs," 2020, *arXiv:2005.03082*.

[90] S. Chekijian, H. Li, and S. Fodeh, "Emergency care and the patient experience: Using sentiment analysis and topic modeling to understand the impact of the COVID-19 pandemic," *Health Technol.*, vol. 11, pp. 1073–1082, Aug. 2021, doi: 10.1007/s12553-021-00585-z.

[91] P. Ma, Q. Zeng-Treitler, and S. J. Nelson, "Use of two topic modeling methods to investigate COVID vaccine hesitancy," in *Proc. 14th Int. Conf. (ICT), Soc. Hum. Beings, ICT, 18th Int. Conf. Web Based Communities Social Media, (WBC), 13th Int. Conf. e-Health, EH-Held 15th Multi-Conf. Comput. Sci. Inf. Syst. (MCCSIS)*, 2021, pp. 221–226. [Online]. Available: https://www.ict-conf.org/wp-content/uploads/2021/07/04_202106C030_Ma.pdf

[92] Y. Wang, M. Shi, and J. Zhang, "What public health campaigns can learn from people's Twitter reactions on mask-wearing and COVID-19 vaccines: A topic modeling approach," *Cogent Social Sci.*, vol. 7, no. 1, Jan. 2021, Art. no. 1959728, doi: 10.1080/23311886.2021.1959728.

[93] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation," *Tourism Manag.*, vol. 59, pp. 467–483, Apr. 2017, doi: 10.1016/j.tourman.2016.09.009.

[94] H. Q. Vu, G. Li, R. Law, and Y. Zhang, "Exploring tourist dining preferences based on restaurant reviews," *J. Travel Res.*, vol. 58, no. 1, pp. 149–167, Jan. 2019, doi: 10.1177/0047287517744672.

[95] S. Song, H. Kawamura, J. Uchida, and H. Saito, "Determining tourist satisfaction from travel reviews," *Inf. Technol. Tourism*, vol. 21, no. 3, pp. 337–367, Sep. 2019, doi: 10.1007/s40558-019-00144-3.

[96] B. Kim, S. Kim, and C. Y. Heo, "Analysis of satisfiers and dissatisfiers in online hotel reviews on social media," *Int. J. Contemp. Hospitality Manag.*, vol. 28, no. 9, pp. 1915–1936, Sep. 2016, doi: 10.1108/IJCHM-04-2015-0177.

[97] N. Hu, T. Zhang, B. Gao, and I. Bose, "What do hotel customers complain about? Text analysis using structural topic model," *Tourism Manag.*, vol. 72, pp. 417–426, Jun. 2019, doi: 10.1016/j.tourman.2019.01.002.

[98] J. Wang and X. Yu, "The driving path of customer sustainable consumption behaviors in the context of the sharing economy—Based on the interaction effect of customer signal, service provider signal, and platform signal," *Sustainability*, vol. 13, no. 7, p. 3826, Mar. 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/7/3826

[99] P. Brzustewicz and A. Singh, "Sustainable consumption in consumer behavior in the time of COVID-19: Topic modeling on Twitter data using LDA," *Energies*, vol. 14, no. 18, p. 5787, Sep. 2021, doi: 10.3390/en14185787.

[100] E. Park, J. Kwon, and S.-B. Kim, "Green marketing strategies on online platforms: A mixed approach of experiment design and topic modeling," *Sustainability*, vol. 13, no. 8, p. 4494, Apr. 2021, doi: 10.3390/su13084494.

[101] K. Celuch, "Customers experience of purchasing event tickets: Mining online reviews based on topic modeling and sentiment analysis," *Int. J. Event Festival Manag.*, vol. 39, pp. 36–50, Oct. 2020, doi: 10.1108/IJEFM-06-2020-0034.

[102] J. Kang, S. Kim, and S. Roh, "A topic modeling analysis for online news article comments on nurses workplace bullying," *J. Korean Acad. Nursing*, vol. 49, no. 6, pp. 736–747, Dec. 2019, doi: 10.4040/jkan.2019.49.6.736.

[103] S. Yun and J. Kang, "Factors affecting workplace bullying in Korean hospital nurses," *Korean J. Adult Nursing*, vol. 26, no. 5, p. 553, 2014, doi: 10.7475/kjan.2014.26.5.553.

[104] X. Han, "Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model," *Scientometrics*, vol. 125, no. 3, pp. 2561–2595, Dec. 2020, doi: 10.1007/s11192-020-03721-0.
[105] Y. Miyata, E. Ishita, F. Yang, M. Yamamoto, A. Iwase, and K. Kurata, "Knowledge structure transition in library and information science: Topic modeling and visualization," *Scientometrics*, vol. 125, no. 1, pp. 665–687, Oct. 2020, doi: 10.1007/s11192-020-03657-5.
[106] K. Kurata, Y. Miyata, E. Ishita, M. Yamamoto, F. Yang, and A. Iwase, "Analyzing library and information science full-text articles using a topic modeling approach," *Proc. Assoc. Inf. Sci. Technol.*, vol. 55, no. 1, pp. 847–848, Jan. 2018, doi: 10.1002/pra2.2018.14505501143.
[107] M. Baghmohammad, A. Mansouri, and M. Cheashmehsohrabi, "Identification of topic development process of knowledge and information science field based on the topic modeling (LDA)," *Iranian J. Inf. Process. Manag.*, vol. 36, no. 2, pp. 297–328, 2020. [Online]. Available: https://jipm.irandoc.ac.ir/article-1-4480-en.pdf
[108] A. Ianina, L. Golitsyn, and K. Vorontsov, "Multi-objective topic modeling for exploratory search in tech news," in *Artificial Intelligence and Natural Language* (Communications in Computer and Information Science). Cham, Switzerland: Springer, 2018, pp. 181–193, doi: 10.1007/978-3-319-71746-3_16.
[109] D. Landínez-Martínez, C. Quintero-López, and V. D. Gil-Vera, "Working memory training in children with attention deficit hyperactivity disorder: A systematic review," *Revista de Psicología Clínica con Niños y Adolescentes*, vol. 9, pp. 1–11, Sep. 2022, doi: 10.21134/rpcna.2022.09.3.7.
[110] P. Duque, O. E. Meza, D. Giraldo, and K. Barreto, "Economía social y economía solidaria: Un análisis bibliométrico y revisión de literatura," *REVESCO. Revista de Estudios Cooperativos*, vol. 138, Jun. 2021, Art. no. e75566, doi: 10.5209/reve.75566.

ANDRÉS M. GRISALES A. was born in Colombia, in 1983. He received the master's degree in applied mathematics. He is currently pursuing the Ph.D. degree in mathematical sciences with the Universidad de Caldas.

Since 2015, he has been working at Universidad Católica Luis Amigó in the area of basic sciences. He is also a Mathematician at the Universidad Nacional de Colombia, Manizales, Caldas. He is working in natural processing techniques and specifically in topic modeling. His research interests include applied statistics, multivariate analysis, and structural equation modeling.

SEBASTIAN ROBLEDO received the Ph.D. degree in engineering from the Universidad Nacional de Colombia. He is currently an Assistant Professor in marketing at Universidad Católica Luis Amigó, Colombia. He is also the Director of Core of Science, a spin-off development of his doctoral research. He conducted a postdoctoral study in scientometrics at the Centro de Bioinformática y Biología Computacional (BIOS). His research interests include entrepreneurial marketing and networking as drivers of word-of-mouth marketing.

MARTHA ZULUAGA worked as a Visiting Scholar at the West Coast Metabolomics Centre, UC Davis. She also worked as a Research Advisor on bioprospecting projects in cosmetics at the Center of Bioinformatics and Computational Biology of Colombia (BIOS). In her postdoctoral fellowship, she worked on the metabolomics of cocoa post-harvest processes at Agrosavia. She is currently a Researcher with more than ten years of experience in metabolomics and chemometrics based on mass spectrometry. She is also working as a Research Leader of the Western Zone with Universidad Nacional Abierta y a Distancia, Colombia.