Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 5, No. 2, June, 2024

Evaluating the Performance of Topic Modeling Techniques for Bibliometric Analysis Research: An LDA-based Approach

Lan Thi Nguyen [1], Wirapong Chansanam [1*], Nalatpa Hunsapun [1], Vispat Chaichuay [1], Suparp Kanyacome [2], Akkharawoot Takhom [3], Yuttana Jaroenruen [4], Chunqiu Li [5]

[1] Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen 40002, Thailand.

[2] Faculty of Science and Engineering, Kasetsart University, Sakon Nakhon 47000, Thailand.

[3] Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Pathum Thani 12120, Thailand.

[4] Informatics Innovative Center of Excellence, Walailak University, Thai Buri, Nakhon Si Thammarat 80160, Thailand.

[5] School of Government, Beijing Normal University, Beijing 100875, China.

Received 02 February 2024; Revised 08 May 2024; Accepted 11 May 2024; Published 01 June 2024

Abstract

Digital technologies have been used for a vast amount of bibliometric analysis research. Although these technologies have made scientific investigation more accessible and efficient, scholars now face the daunting task of sifting through an overwhelming number of documents. This study aims to identify bibliometric research analysis's primary topics, categories, and latent topics from a global perspective. This study utilized topic modeling techniques to analyze the abstracts of 16,039 eligible papers published between 1977 and 2023 in the Scopus database. Through the use of Latent Dirichlet Allocation (LDA) topic modeling, the study was able to identify four distinct research topics and observe how they have evolved over time. The research topic has shifted its focus from individual concepts and words to relationships between nodes and conceptual, intellectual, and social structures. The study's findings have significant implications for bibliometric analysis-related research, providing valuable insights into trends and patterns in bibliometric analysis content within large digital article archives. The LDA has proven to be an efficient tool for analyzing these trends and patterns quickly. This study's novel approach considers factors for word embedding usage and optimal topic numbers. It focuses on a full understanding of the LDA results and combines statistical analysis, domain knowledge, and temporal exploration to better understand how data structures work.

Keywords: Bibliometric; LDA; Topic Modeling; Topic Trends; Performance Evaluation.

1. Introduction

Clustering topics through bibliometric analysis is a valuable approach for gaining a better understanding of the content and relationships between publications in a specific field [1, 2]. Researchers can use different techniques like co-citation analysis, bibliographic coupling, and co-word analysis to identify co-occurring terms or citation patterns among publications and group them based on similarity. Software tools such as VOSviewer and CiteSpace can also be

* Corresponding author: wirach@kku.ac.th

http://dx.doi.org/10.28991/HIJ-2024-05-02-07

used to visualize and analyze clusters of publications [3]. Li & Lei [4] conducted a bibliometric analysis on topic modeling studies from 2000 to 2017, using data from Web of Science to evaluate bibliometric indices for productive authors, countries, and institutions, as well as investigate thematic changes over time. The study revealed that the number of publications on topic modeling has steadily increased, with a peak in 2015, and that the United States and the University of California are the most productive countries and institutions, respectively. Thematic changes in topic modeling research were the most frequent topics, followed by text mining and natural language processing. This study also detected a trend towards utilizing topic modeling in social media analysis and incorporating external knowledge sources in topic modeling.

Topic modeling is a machine learning technique that can identify common patterns in words and phrases to uncover the primary themes and topics present in a collection of documents [5]. It has numerous applications, including text classification and building recommender systems [6]. However, it is worth mentioning that topic modeling is a form of unsupervised learning that aims to discover underlying topics or themes in a collection of documents without predefined categories. The set of possible topics is unknown a priori and is defined as part of generating the topic models. Topic modeling algorithms group words based on similarities and analyze the patterns in the use of words across multiple documents to identify key topics that best capture the content of the documents [7–10].

Topic modeling is useful for text classification and building recommender systems and has been proven to be effective in identifying significant and relevant topics from extensive text data [6]. It has been widely used in social science research but less commonly in educational research [11–13]. Topic modeling involves using natural language processing techniques, such as Latent Dirichlet Allocation (LDA), to discover recurring topics from a set of texts. LDA is commonly used in research because it provides an objective analysis of the corpus data [14, 15]. Moreover, topic modeling combines both objective data analysis and subjective data labeling processes [16]. It can be useful for large-scale literature research to uncover hidden topics and is more flexible and effective than other methods like document clustering [6].

The bibliometric studies involve various approaches, such as defining data metrics, exploring the field's development from an information and library science perspective, or making cross-disciplinary comparisons. However, most studies overlook the unique bibliographic nature of the field. Thus, this article addresses a comprehensive overview of the bibliographic research by analyzing the subject's referral network to identify patterns and hierarchical semantic relationships. Rather than forcing the literature into predetermined classifications, this approach allows for a more profound understanding of bibliometric research [17–19].

Previous studies have been conducted on topic modeling and bibliometric research. Ayaz et al. [20] explore trends in gamification research, while Robledo & Zuluaga [21] examine topic modeling's evolution. Mifrah & Benlahmar [22] compare LDA and NMF techniques, and Cui et al. [23] propose a recognition method. Motamedi et al. [24] analyze information systems in maternal health, and Almenara [25] focuses on eating disorder literature. Sharma et al. [26] investigate smart cities' trends, Gurcan & Cagiltay [27] analyze bioinformatics research. Cobelli and Blasi [28] examined ATI in healthcare, and Chen & Xie [29] reviewed sentiment analysis. Chen et al. [30] explore semantic computing, Jiang et al. [31] evaluate global hydropower literature, and Linnenluecke et al. [32] outline methods for literature reviews. Lastly, Chen et al. [33] study learning analytics research trends. In addition, Amaro & Bacao [34] underscored the pivotal role of Topic Modeling (TM) in the analysis of digital text data. Their study addresses the intricate challenges associated with evaluating TM algorithms, presenting a meticulous comparative analysis involving five distinct algorithms across varied datasets and metrics. Their findings notably advocate for Top2Vec as the preeminent model, thus contesting the conventional dominance typically attributed to LDA. These studies contribute to understanding trends, methodologies, and future research directions in topic modeling and bibliometric analysis. However, there is a gap in the research on topic modeling and bibliometric analysis because not many hybrid models have been studied. Additionally, there may be a lack of standardized approaches for determining the optimal number of topics and evaluating topic coherence in bibliometric studies. Integrating such techniques could enhance the accuracy and interpretability of topic models in this domain.

A gap in the body of bibliometric research literature about topic modeling methods is the area of hybrid models that combine Latent Dirichlet Allocation (LDA) with word embedding methods. This model type has yet to be studied much. Also, there should be more standardized ways to find the best topic count and check for topic coherence. These are necessary for topic modeling results in bibliometric analyses to be easily understood and used. Integrating such approaches bolsters the efficacy and depth of topic modeling endeavors within this scholarly domain.

This study employs topic modeling and text analysis technology to examine 16,039 papers on bibliometric analysis research from 1977 to 2023 in the Scopus database core collection. Its objective is to identify the primary research topics, categories, and latent topics of bibliometric analysis from a global perspective. The study addresses three main research questions, namely, the main research topics, categories, and latent research topics in bibliometric analysis. To do this, we extracted TF-IDF keywords from the abstracts of the 16,039 papers, developed a keyword corpus, and constructed a time-phased word cloud to analyze word cloud feature evolution. Additionally, we used LDA topic modeling for the first time in bibliometric analysis research to efficiently analyze a large corpus data of text data and obtain essential parameters such as the number of topics through machine learning training. Through our analysis, we discovered the research topics and development trends of bibliometric analysis based on large amounts of text data from 16,039 documents in the Scopus database from 1977 to 2023. This study contributes theoretical and methodological references for future research in bibliometric analysis and underscores the significance of considering the bibliographic nature of the field.

2. Research Methodology

This study aimed to quantitatively analyze bibliometric research articles, with a particular focus on the content of their abstracts. The methodology used was outlined in Figure 1, which depicted the three sub-processes: data retrieval, preprocessing, and topic analysis. The study employed Latent Dirichlet allocation (LDA) topic models, which allowed for the identification of key topics within the texts. However, the study was not without its challenges, including the need to appropriately preprocess text collections, select appropriate model parameters, evaluate model parameters, evaluate model reliability, and internet the resulting topics. Figure 1 provided an excellent visual representation of the study's methodology, making it easy for readers to understand the steps involved in conducting the analysis.

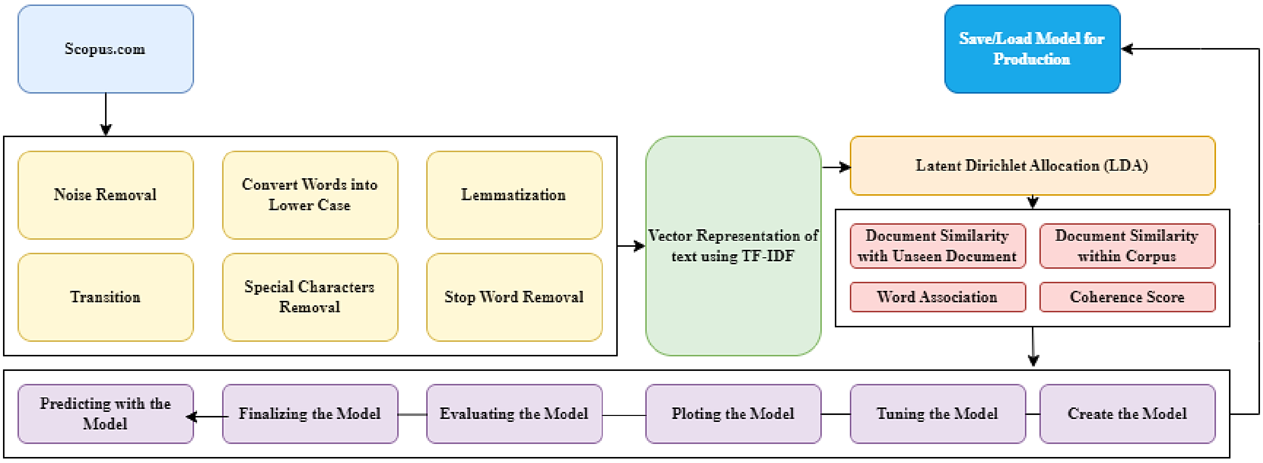

Scopus.com
Noise Removal
Convert Words into Lower Case
Lemmatization
Transition
Special Characters Removal
Stop Word Removal
Vector Representation of text using TF-IDF
Latent Dirichlet Allocation (LDA)
Document Similarity with Unseen Document
Document Similarity within Corpus
Word Association
Coherence Score
Save/Load Model for Production
Predicting with the Model
Finalizing the Model
Evaluating the Model
Ploting the Model
Tuning the Model
Create the Model

Figure 1. Topic modeling pipeline with LDA

2.1. Data Retrieval and Pre-Processing

This section provides a detailed account of the data retrieval and pre-processing process for this study on bibliometric analysis. The researchers retrieved high-quality literature from the Scopus Core Collection between 1977 to 2023, using a well-crafted search string (for example, TS = "bibliometrix", OR TS = "bibliometric" OR TS = "bibliometrics") to ensure that only relevant papers were collected. Scopus was deemed suitable for the study as it covers a wide range of fields and is a widely used database for bibliometric analysis. The inclusion and exclusion criteria were then used to select the 16,039 relevant papers, which were analyzed based on the title, abstract, year of publication, and journal-title.

To prepare the data for topic model mining, the researchers employed several pre-processing steps, including splitting the text into tokens, removing punctuation, numbers, and unnecessary words, and converting words to their base form [35]. Python programming language was used, and the text pre-processing was conducted in PyCaret, a reliable tool that is open-source and readily available [36]. A series of pre-processing steps were conducted to extract representative keywords from the research articles. These steps involved removing commonly used but insignificant words, using the bi-gram algorithm to select common collocation phrases, and applying the TF-IDF algorithm to extract essential keywords from the abstract [37, 38]. Afterwards, the researchers created a keyword corpus with 87,065 words and 16,039 documents by removing keywords with low weights.

2.2. Topic Modeling

Topic modeling is a powerful tool that can help researchers uncover hidden structures in collections of documents, allowing them to make data-driven decisions and gain insights into complex topics [39]. However, selecting the right model can be challenging, as different models have different strengths and weaknesses [40]. For example, LDA is known for its ability to learn descriptive topics, while LSA is better at creating a sematic representation of documents in a corpus [41].

Despite its potential benefits, topic modeling can be difficult to understand and interpret [42]. To ensure that the results are reliable and meaningful, researchers use metrics, such as perplexity and coherence to evaluate the modeling results. Perplexity measures the likelihood value of the model, while coherence is calculated using the Normalized

Pointwise Mutual Information ($NPMI$) formula [43]. The NPMI formula assigns high co-occurrence probability to word pairs, resulting in highly understandable and interpretable modeling results.

$$NPMI(w_i, w_j) = \frac{log\frac{p(w_i,w_j)+\varepsilon}{p(w_i).p(w_j)}}{-log\left(p(w_i,w_j+\varepsilon)\right)} \quad (1)$$

where; $p(w)$ represents the probability that the word w exists in the provided document, and $p(w_i, w_j)$ represents the probability that two words w_i, w_j appear together in the same context.

The LDA Model module in Gensim [44] was used for topic modeling, using the $NPMI$ coherence indicator. $NPMI$ measures the association between word pairs, thereby providing an effective means to assess the model's quality. Following this, we were able to analyze the probability of word co-occurrences within a given document, and subsequently assess the statistical significance of the results [42, 43].

2.3. Topic Modeling Evaluation

This study uses Latent Dirichlet Allocation (LDA) on Scopus documents to evaluate topic modeling techniques. Various distinct research topics will be identified through LDA, tracing their evolution. It emphasizes shifting focus from individual concepts to network relationships, uncovering intellectual and social structures within bibliometric analysis. Factors considered include word embedding utilization, optimal topic numbers, statistical analysis, domain expertise, temporal exploration, and enhancing comprehension of LDA results and data structures. This approach offers valuable insights into the dynamics of research topics and their evolution over time, contributing to advancing topic modeling techniques in bibliometric analysis.

3. Results

The methodology involved retrieving and preprocessing literature from the Scopus Core Collection, selecting 16,039 relevant papers based on specified criteria. Latent Dirichlet Allocation (LDA) topic models were then applied to identify key topics. Challenges included text preprocessing, model parameter selection, reliability evaluation, and topic interpretation, considering the bibliographic nature of the field. A visual representation of the methodology, including the LDA topic modeling pipeline, was provided. This approach enabled the study to analyze the dataset efficiently, identifying key topics and trends in bibliometric analysis research. The results were displayed in text analysis, topic modeling, and clustering evaluation.

3.1. Text Analysis

A word cloud is a visual representation of a text that highlights important words in a given text. The technique has been widely used in various fields [12, 35, 45]. In this study, word clouds were generated using the Word Cloud library in Python to represent the main research content and each topic. The library enables users to customize the appearance of the word cloud, such as font size, colour, and layout, to provide a clear and concise summary of the text [46]. Figure 2 displays a word cloud generated in this study.

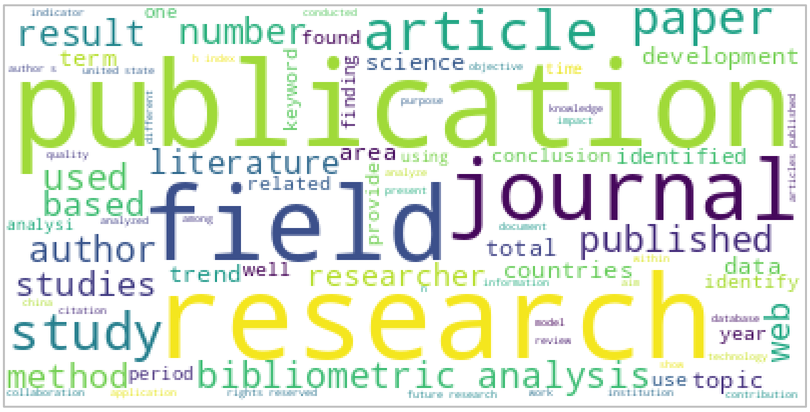
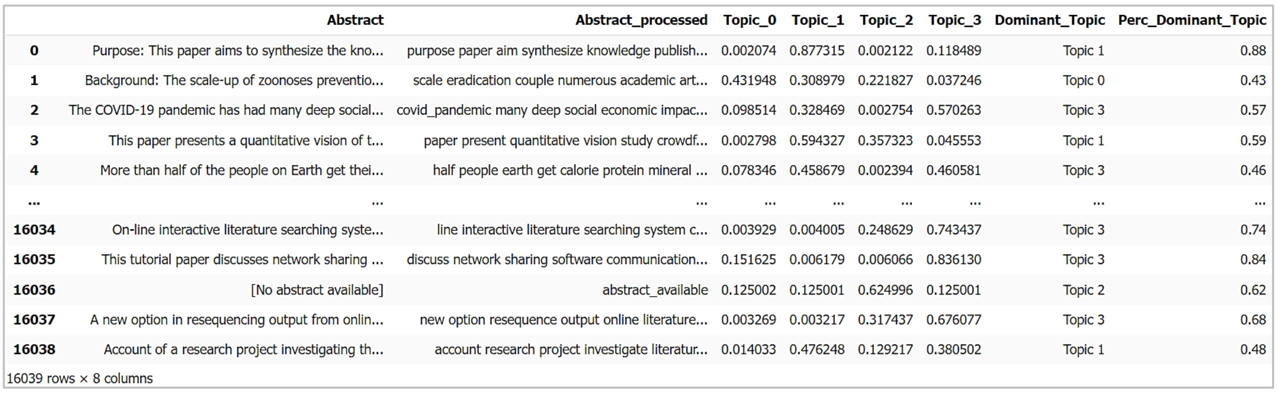


Figure 2. Most common words (1977–2023)

LDA topic modeling methods are useful for constructing document's embedding vector with the number of topics determining the dimension. Each document is represented by a vector of topic probabilities, which can be utilized for document classification. After completing the training dataset, a group label to each document was assigned to each document in the Dominant Topic column, and additional columns can be added to the result. However, observing the behaviors of each group and giving them meaningful names can be challenging and time-consuming. One indirect method of observing the frequency of words used in each group (known as word distribution) was conducted to address this challenge. This allows for a better understanding of the results of document clustering and gives each group a name that accurately describes its meaning. Figure 3 illustrates the results of this observation.



	Abstract	Abstract_processed	Topic_0	Topic_1	Topic_2	Topic_3	Dominant_Topic	Perc_Dominant_Topic
0	Purpose: This paper aims to synthesize the kno...	purpose paper aim synthesize knowledge publish...	0.002074	0.877315	0.002122	0.118489	Topic 1	0.88
1	Background: The scale-up of zoonoses preventio...	scale eradication couple numerous academic art...	0.431948	0.308979	0.221827	0.037246	Topic 0	0.43
2	The COVID-19 pandemic has had many deep social...	covid_pandemic many deep social economic impac...	0.098514	0.328469	0.002754	0.570263	Topic 3	0.57
3	This paper presents a quantitative vision of t...	paper present quantitative vision study crowdf...	0.002798	0.594327	0.357323	0.045553	Topic 1	0.59
4	More than half of the people on Earth get thei...	half people earth get calorie protein mineral ...	0.078346	0.458679	0.002394	0.460581	Topic 3	0.46
...	...	...	...	...	...	...	...	...
16034	On-line interactive literature searching syste...	line interactive literature searching system c...	0.003929	0.004005	0.248629	0.743437	Topic 3	0.74
16035	This tutorial paper discusses network sharing ...	discuss network sharing software communication...	0.151625	0.006179	0.006066	0.836130	Topic 3	0.84
16036	[No abstract available]	abstract_available	0.125002	0.125001	0.624996	0.125001	Topic 2	0.62
16037	A new option in resequencing output from onlin...	new option resequence output online literature...	0.003269	0.003217	0.317437	0.676077	Topic 3	0.68
16038	Account of a research project investigating th...	account research project investigate literatur...	0.014033	0.476248	0.129217	0.380502	Topic 1	0.48

16039 rows × 8 columns

Figure 3. A vector of topic probability

The distribution of topics in the research paper dataset was analyzed to identify patterns and trends in the data. The results indicate that one topic is significantly more prevalent than others, while the remaining topics are outliers. Figure 4 visualizes the topic distribution, with topic number two being the most frequent. However, further investigation is required to determine the underlying reasons. Factors such as the research focus or the sample size could contribute to the results, and it is necessary to consider these variables before drawing any definitive conclusions.

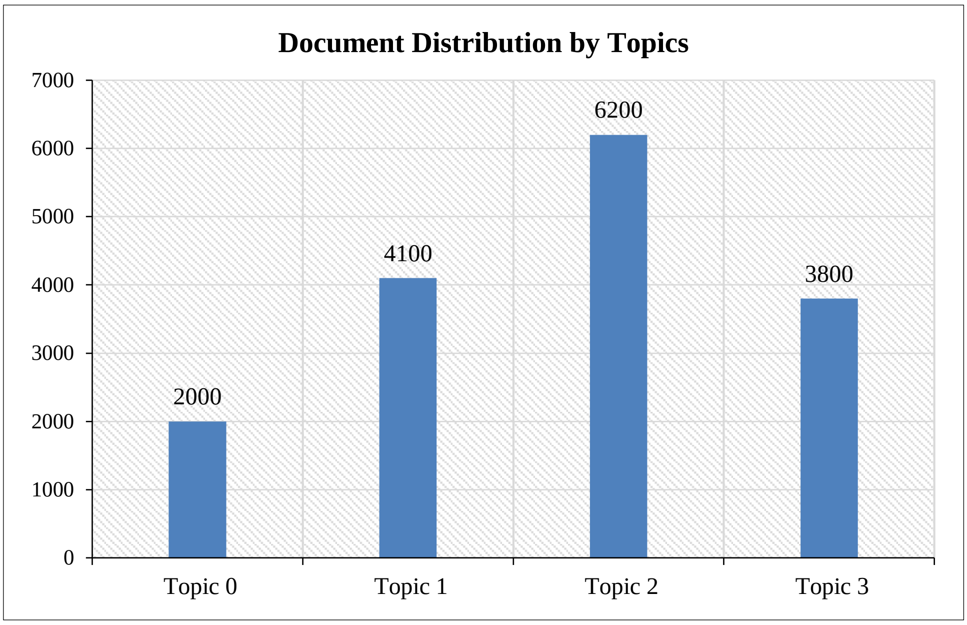


Figure 4. The visualization of the topic distribution

3.2. Topic Modeling

3.2.1. Determination of Optimal Parameters

One of the most widely used algorithms for topic modeling is the Latent Dirichlet Allocation (LDA) model, which has been shown to produce promising results in numerous studies [47]. It's important to note that the LDA model is an unsupervised machine learning technique.

When using the LDA model module of Gensim library in Python, it's crucial to set the topic number K, as well as the apriori values of topic distribution α, and topic word distribution η beforehand. These parameters can significantly impact the effectiveness of your topic mining.

Evaluating topic model algorithms is a challenging task, mainly due to the complexity and volume of text data involved. Human evaluation can be time-consuming and subject to bias, which is why theoretical evaluation models are necessary. Although these models cannot match the accuracy of human-in-the-loop model evaluation [42], they provide a useful framework for assessing the quality of topic models.

One widely used method for evaluating a topic model is Coherence, which measures the degree of semantic similarity between the high-scoring words in a topic. This evaluation matric involves calculating pairwise scores for the top n frequently occurring words in each topic. These scores are then aggregated to determine the final coherence score [48], as shown in Equation 2.

$$Coherence = \sum_{i<j} score(w_i, w_j) \tag{2}$$

Several coherence measures exist in the literature to evaluate topic models. For instance, the UCI (or CV) measure was proposed by Newman et al. [49] as an automatic coherence measure to rate topic understandability. This measure compares word pairs and treats words as facts. Other researchers have also proposed measuring coherence based on word statistics [41, 50, 51].

Metrics, such as perplexity and coherence should be established to assess the effectiveness of topic mining. In addition to evaluating the effectiveness of the model, it is crucial to understand its interpretability for promoting its final application. Coherence, which focuses on the model's interpretability, is an essential metric for evaluating topic mining. Higher coherence indicates better interpretability of the model. Therefore, coherence is often selected as the primary metric for evaluating topic mining. Cao et al. [52] suggested that measuring the average cosine distance between every pair of topics can provide insights into how stable the topic structure is. This can help to identify any inconsistencies or overlaps in the topics, which can further improve the model's interpretability and effectiveness.

To identify the optimal number of topics for our study, we employed the CoherenceModel class in the genism package, which provides fitness scores for various topic numbers. In addition, we used the C_v and C_umass algorithms to compute coherence scores and determine the number of topics with the highest score. A higher C-v value indicates a better model fit. Subsequently, we generated a line graph using Python's Matplotlib package to visualize the results and observed that the coherence score peaked at four topics (as depicted in Figure 5). Therefore, we categorized the collected articles into four topics.

In addition, we experimented with 2-6 topics with a step size of 1 to determine the optimal number of topics, and used the coherence parameter to simplify the model. The results showed that the coherence parameter reached its maximum value of -0.0641 when the number of topics was four (Figure 5). Hence, we chose four topics with α = 0.143 and η = 0.143, which were automatically set.

Subsequently, the model was further fine-tuned by adjusting the auto-set value within a narrow range and testing the coherence parameter to achieve the best results. The coherence score, which measures model performance, reached its maximum of 0.4514 when α = 0.15 and η = 0.18, surpassing the previous score of 0.4360. The results were reviewed and confirmed by experts who found that the four topics were clear and interpretable. Therefore, we determined the optimal parameters and obtained the optimal model results for the four topics, as shown in Figure 5.

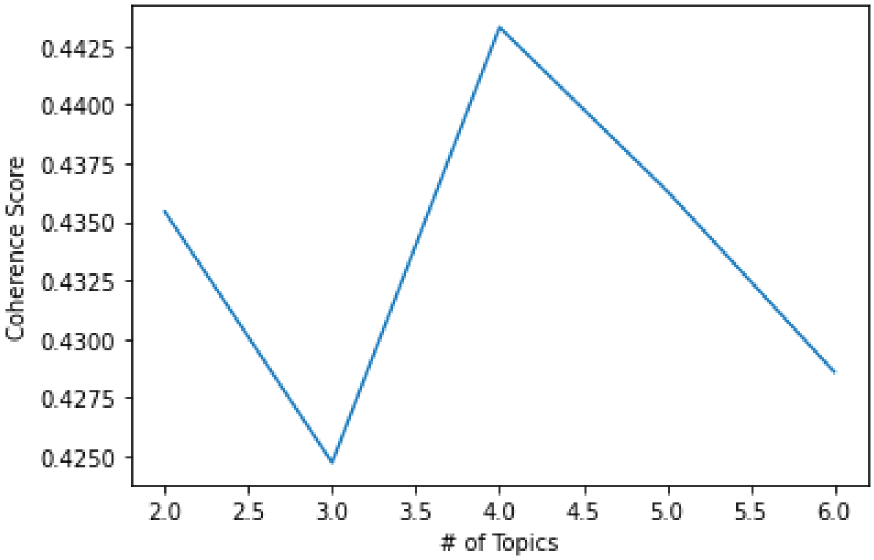


Figure 5. Topics number and coherence

3.2.2. Topic Naming and Topic Details

In a traditional topic modeling, topics are typically named based on the probability of topic words, with high-probability words being used for naming. However, this method can result in low discrimination in topic naming. Thus, Sievert & Kenneth [53] proposed combining keywords with their relationship to the topic for naming purposes. Once the model is determined, topic naming is done manually, which is crucial for ensuring that the results are interpretable. To accomplish his, the correlation formula is used, which is as follows:

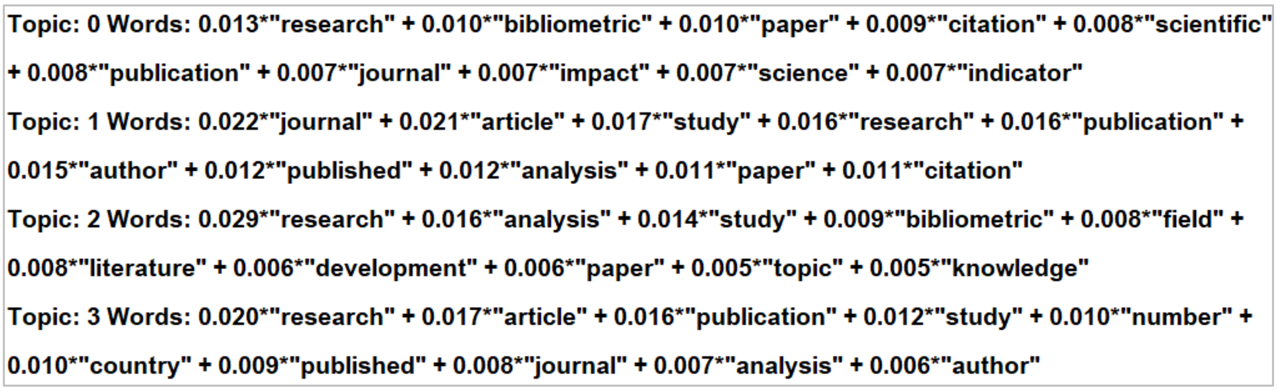
$$\gamma(w,t|\lambda) = \lambda \log[p(w|t)] + (1-\lambda)\, log\, \left[\frac{p(W|t)}{p(w)}\right] \quad (3)$$

The value of λ to determine the weight given to a topic word *w* in relation to its boost under topic *t* was adjusted. After tuning the value, we found that setting λ to 0.6 yielded the best results. The outcome of this process is presented in Table 1, which provides a clear overview of the resulting topics.

To name each topic, we employed three methods: (a) analyzing the top 10 words that are most representative of the topic based on the highest term-topic probability (βk) and frequency in the abstract; (b) creating a word cloud for each topic using the top 50 words, where the size of a term corresponds to its term-topic probability, to identify the most representative terms within each topic; and (c) examining the top 20 articles (θd) with the highest proportion of words to better understand the narratives within each topic and decide on the topic names. The resulting topics and their names are presented in Table 1.

The Latent Dirichlet Allocation (LDA) method is commonly used to divide a collection of bibliometric analysis research into four distinct topics. The process involves several key steps, including:

- Determining the optimal number of topics based on the coherence score;
- Allocating the determined number of topics to the bibliometric analyses;
- Using a group of words to represent the characteristics of each topic;
- Using the resulting representation as a reference for the title of the bibliometric analysis when entering the topic;
- Using numbers 0, 1, 2, and 3 as markers for entry into the topic, based on the probability value shown in Figure 6.



```
Topic: 0 Words: 0.013*"research" + 0.010*"bibliometric" + 0.010*"paper" + 0.009*"citation" + 0.008*"scientific"
+ 0.008*"publication" + 0.007*"journal" + 0.007*"impact" + 0.007*"science" + 0.007*"indicator"
Topic: 1 Words: 0.022*"journal" + 0.021*"article" + 0.017*"study" + 0.016*"research" + 0.016*"publication" +
0.015*"author" + 0.012*"published" + 0.012*"analysis" + 0.011*"paper" + 0.011*"citation"
Topic: 2 Words: 0.029*"research" + 0.016*"analysis" + 0.014*"study" + 0.009*"bibliometric" + 0.008*"field" +
0.008*"literature" + 0.006*"development" + 0.006*"paper" + 0.005*"topic" + 0.005*"knowledge"
Topic: 3 Words: 0.020*"research" + 0.017*"article" + 0.016*"publication" + 0.012*"study" + 0.010*"number" +
0.010*"country" + 0.009*"published" + 0.008*"journal" + 0.007*"analysis" + 0.006*"author"
```

Figure 6. **A list of unique word that represent a formed topic**

To determine the themes that represent each formative topic, this process identifies the most dominant set of words for each topic. A theme would be determined from the list of the most dominant words for each topic, which would be used to represent the topic's name in several words or sentences (Table 1).

Table 1. **Bibliometric analysis topics and category**

Category	Topic No.	Topic name	Representative bigrams
Conceptual Structure	0	Set of publications concepts	Bibliometric analysis, number publication, result total research hotspot, research trend, future research, country institution, number citation, high number, science database
	1	Set of publications words	Bibliometric analysis, number publication, Scopus database, result show, International collaboration, research trend, study aim, science database, research topic, recent year
Intellectual and Social Structure	2	Relate to others in the research field	Bibliometric analysis, web science, research field, research area, research topic, network analysis, bibliometric study, study aim, systematic review, bibliometric method
	3	Relationships between nodes	Article publish, number citation, web science, number publication, highly cite, paper publish, number article, use bibliometric

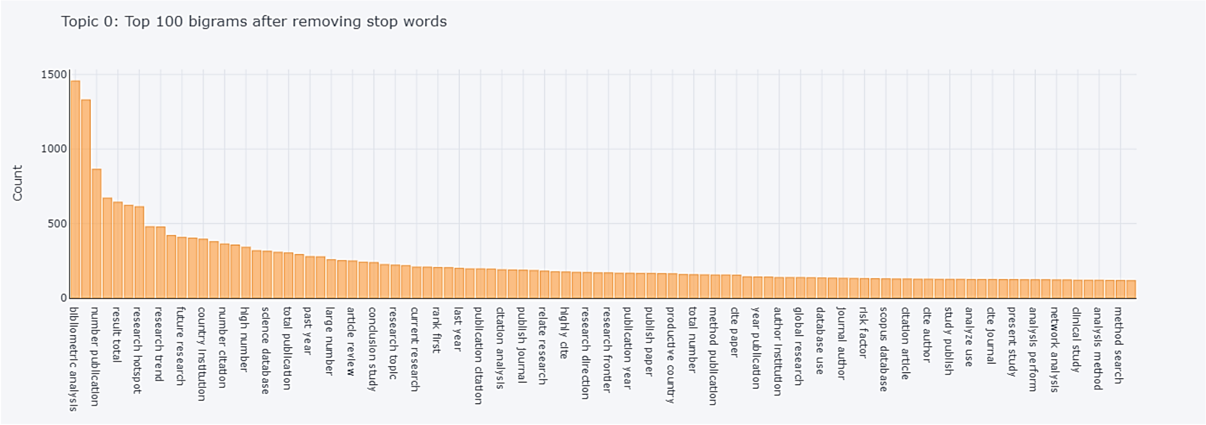


Figure 7. Topic 0 – Top 100 bigrams

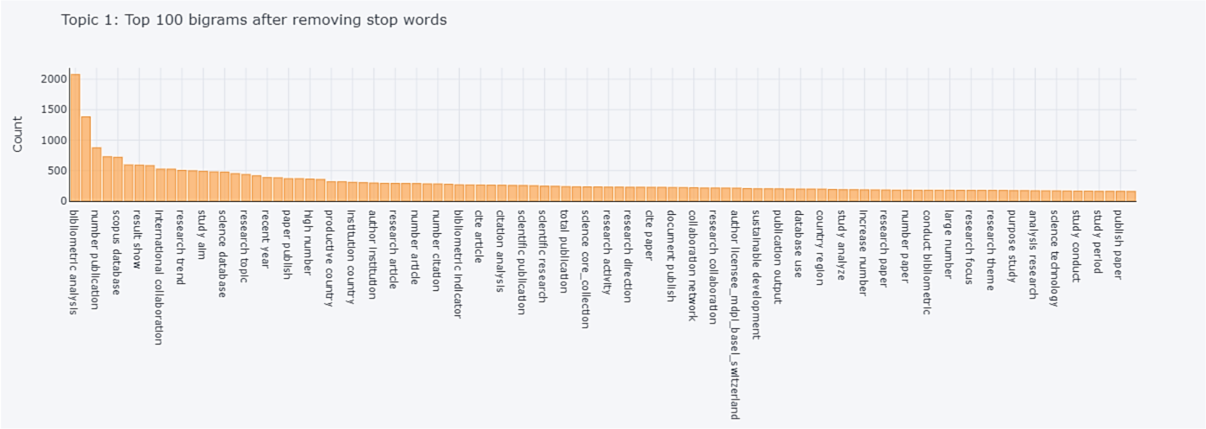


Figure 8. Topic 1 – Top 100 bigrams

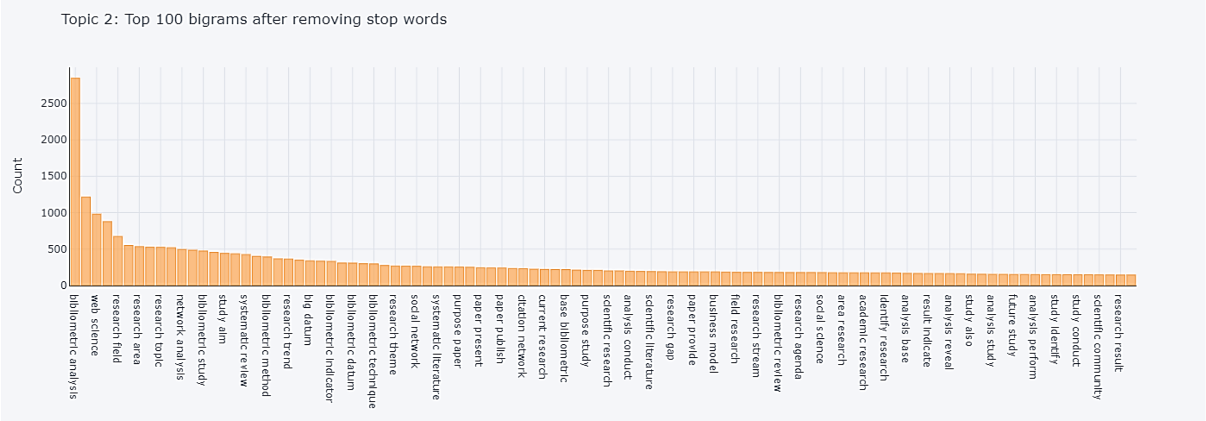


Figure 9. Topic 2 – Top 100 bigrams

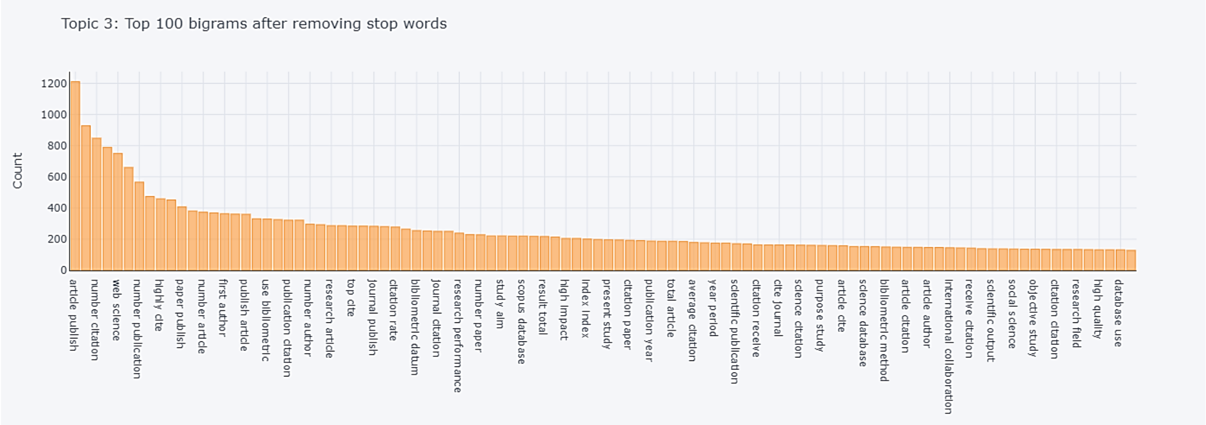


Figure 10. Topic 3 – Top 100 bigrams

Figures 7 to 10 offered a detailed insight into four topics obtained from the LDA algorithm. Each topic is visualized with the top 100 frequently used words. This information is summarized in Table 1, which includes the topic names, top 10 words, and a representative article for each topic. Regarding this, topic 0 (T0) was named "Set of publication concepts." The articles in T0 were conceptual structures representing the interrelationships among concepts in a set of publications. A conceptual framework serves as the foundational structure upon which research is built, systematically aiding in the organization and elucidation of relationships between disparate ideas. Researchers developed it as inherently subjective and inductive, drawing upon existing literature to inform its construction. As a tool for conceptualizing complex phenomena, it provides a roadmap for inquiry, guiding the formulation of hypotheses and the interpretation of findings. While it is not subject to empirical proof, its utility lies in offering a coherent framework for understanding and investigating phenomena within a given field of study. Topic 1 (T1) was labeled "Set of publication words." The articles in T1 examined a conceptual structure representing the relationships between the co-words in those publications. By reading the articles and figuring out the main ideas and variables, a conceptual framework could be created to show how these ideas and variables are likely to be related. This conceptual framework can also provide a useful tool for guiding further research and provide a framework for analyzing and interpreting the data collected from the publications.

Also, it serves as a robust foundation for guiding subsequent research endeavors and a valuable instrument for comprehensively analyzing and interpreting the data gleaned from diverse publications. Its structured approach aids in systematically exploring intricate relationships, uncovering underlying patterns, and facilitating insightful interpretations of the amassed data. Topic 2 (T2) was labeled as "Relate to Others in the Research Field." Articles in T2 examined how an author's work can significantly impact the scientific community, as publishing research outcomes helps researchers gain visibility and acknowledgment. The impact of an author's work transcends mere publication, encapsulating multifaceted dimensions such as research quality, the seminal nature of findings, and their substantive contributions to the field. These factors collectively delineate the significance of an author's scholarly footprint within the scientific community. However, it is imperative to meticulously navigate issues about authorship delineation and acknowledge individual contributions to mitigate potential biases and ensure scholarly integrity. Topic 3 (T3) has been labeled "Relationships between nodes." This research topic delves into the intricate dynamics shaping scientific collaboration networks, particularly exploring the interplay among authors, institutions, and countries. These collaborative networks are subject to various influences, including geographical proximity, linguistic factors, disciplinary domains, and institutional associations. Understanding these multifaceted relationships offers insights into the intricate fabric of global scientific collaboration and its implications for knowledge dissemination and innovation.

Knowledge structure refers to organizing concepts and their relationships in an expert's knowledge structure. An expert's knowledge structure has a rich clustering of concepts, in which each concept is related to many other concepts, and the relationships between concepts are clearly understood. Concepts are arranged hierarchically using umbrella concepts to relate them more closely. In general, drawing a big picture of scientific knowledge has always been desirable. Science mapping is a technique that aims to visually represent the relationships and connections within the scientific knowledge system, including its structure and dynamics [54, 55], and allows the investigation of scientific knowledge from a statistical point of view. Science mapping uses mainly the "structures of knowledge", which includes a conceptual structure which depicts the relationships between concepts or words in a group of publications; and an intellectual and social structure reveals how authors or institutions interact with others in research and the connections between nodes representing references.

3.2.3. Topic Visualization

An interactive visual diagram is a highly effective way to present the results of a topic model. In this regard, the pyLDAvis package [53] was utilized to create an interactive diagram that displays the topics and their most representative words (Figure 11). The size of each bubble in the diagram represents the relevance of the topic in the corpus, and topics that are closer together are more similar to each other. One of the key advantages of the pyLDAvis visualization method is that users can adjust the relevance of words in a topic using a slider [53, 56]. Additionally, the multi-dimensional zoomed model view provides information about the meaning, popularity, and relationships of each topic, making it easy to interpret and analyze the results. Furthermore, the topics are well-differentiated, and the popularity is well-balanced, indicating that the model is robust and accurate. The top 30 relevant words of each topic are displayed in a histogram with saliency and overall term frequency, providing a clear and concise summary of each topic. Finally, the whole model is available on the World Wide Web, allowing readers to use and explore the topic model through an interactive interface. Overall, the pyLDAvis package provides an innovative and effective way to visualize and interpret topic models, enabling researchers to gain deeper insights into the structure and content of their data.

This tool offers a clear and intuitive visualization of the relationships and strengths of each topic by displaying the words that form each topic, using a circle and a horizontal bar chart. The circle on the left panel shows a global view of the model, allowing users to easily comprehend the relationships between topics and their relative strengths. Meanwhile, the horizontal bar chart on the right panel presents the terms that make up each topic, providing users with a detailed understanding of the topics themselves.

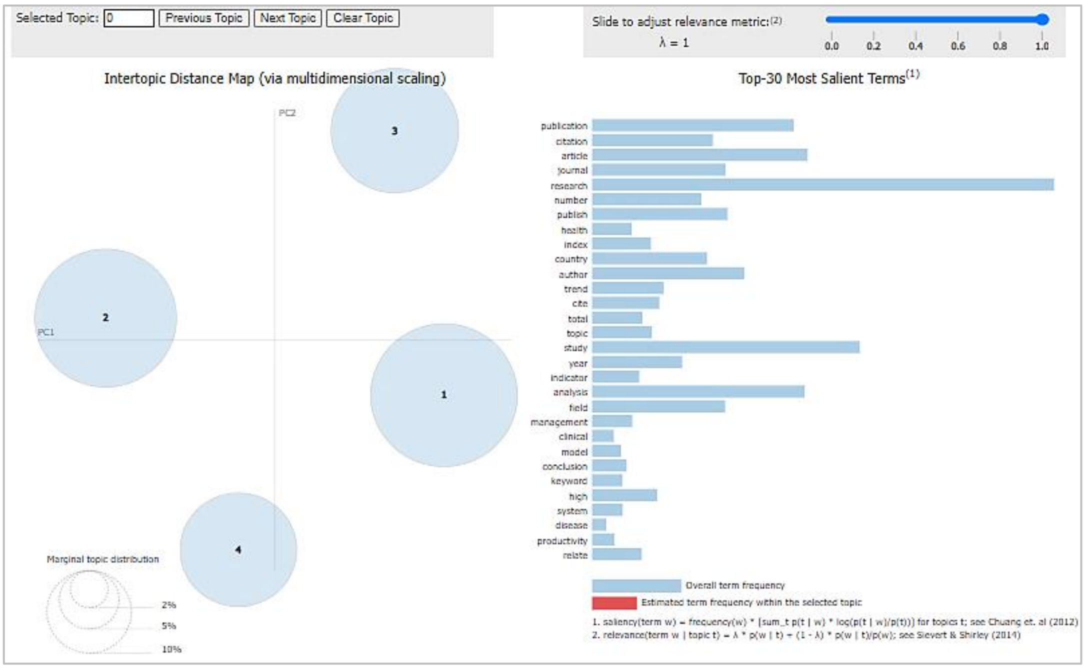

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms(1)
publication
citation
article
journal
research
number
publish
health
index
country
author
trend
cite
total
topic
study
year
indicator
analysis
field
management
clinical
model
conclusion
keyword
high
system
disease
productivity
relate
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 11. The Interactive visualization of LDA model [*]

It is evident that the four identified topics are distinct and belong to different research areas (Figure 11). Upon clicking on each topic circle, the tool generates a bar graph that displays the top 30 most relevant terms for that particular topic. This feature allows users to get a quick and concise summary of the topic's relevance through its most significant keywords. By conducting a lexical analysis of these keywords, it becomes possible to categorize the four topics, as outlined in Table 1. The categories encompassed by these topics are highly pertinent to current research topics, as evidenced by the keywords associated with each theme.

There are two ways to plot the document classification, one of which is t-SNE (t-distributed Stochastic Neighbor Embedding). In this method, each group is represented as a probability distribution, which is essentially a normal distribution. The Euclidean distance is used to measure the distance between groups. This technique is used for 3D projection to visualize similarities between multidimensional vectors and plot clusters of similar documents. The results of applying t-SNE to the case study data are displayed in Figure 12. Additionally, we have made the entire model available on the World Wide Web, providing readers with an interactive interface to explore the document classification. This allows users to gain a deeper understanding of the classification process and to interact with the data in a more dynamic way.

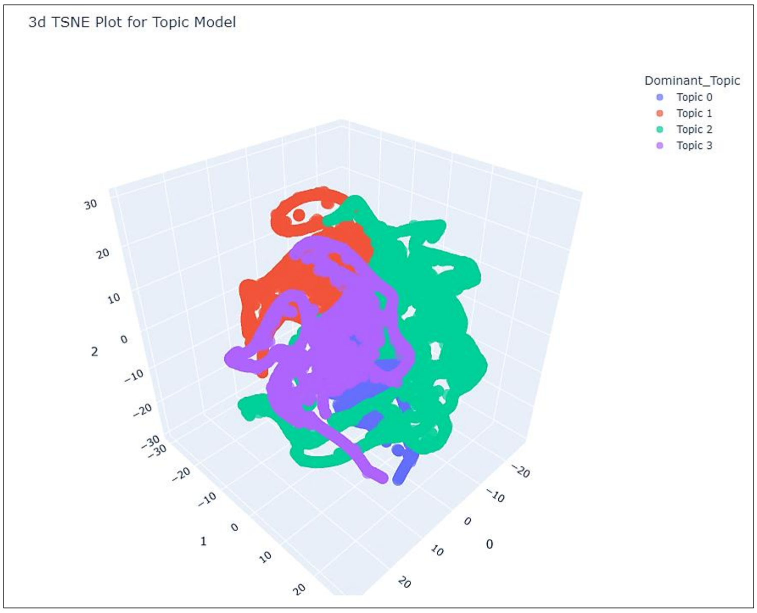


Figure 12. The visualization of the t -distributed Stochastic Neighbor Embedding (TSNE) [†]

[*] https://ischool.kku.ac.th/bibliometric/lda_bibliometric.html

[†] https://ischool.kku.ac.th/bibliometric/TSNE.html

Further experimentation is required to derive a concrete interpretation and justification regarding the phenomena illustrated in Figure 12. To sum up, the t-SNE plotting shows that larger documents could have more variety in nature, and some statistically related words from one category could also be relevant to other categories to some extent.

3.3. Overall Clustering Evaluation

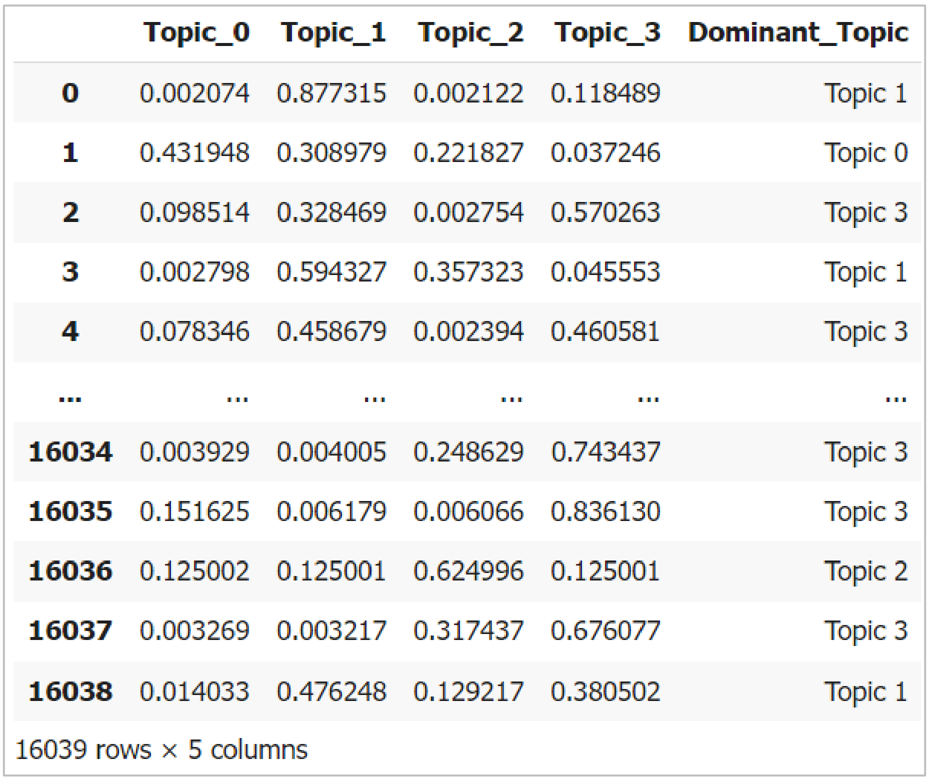
The NLP evaluation model is a powerful visualization tool that facilitates the understanding of topic models by providing additional insights. One of its features is the automatic analysis of each document's sentiment polarity allowing for the identification of positive and negative feelings associated with the topics. Another insightful feature is the word cloud visualization, which displays the frequency of terms used in the topics identified by the LDA model. The size of the text indicates the term's frequency, with the larger text indicating the high probability terms, highlighting their importance.

The word cloud visualization technique is commonly used to represent the results of topic modeling due to its simplicity and clarity. It allows for the comparison of topic-term matrices obtained from different models, providing a clear and straightforward way to evaluate the effectiveness of the LDA topic modeling approach. This technique has been widely adopted in the literature and has been shown to be effective in improving the interpretability of topic models [57, 58].

3.3.1. Document Classification

Document classification is an essential task that involves automatically assigning an input document or predefined group or class based on certain criteria. It can be viewed as a classification problem, where the goal is to convert an input document into an embedding vector. The process of embedding a document begins with transforming it into a corresponding vector.

A traditional approach to embedding a document is called TF·IDF. Following this, TF is a vector of term frequencies (how many times each word appears in this document), and IDF is a vector of inverted document frequencies (how many documents each word appears in). The main drawback of TF·IDF is that each embedding vector can be quite large because its size is equal to the vocabulary size. A more modern approach is the use of topic modeling by utilizing the topic distribution of each document as an embedding vector. Then, these embedding vectors are used to classify the documents. Currently, any classifier can be trained based on these embedding vectors (Figure 13).



	Topic_0	Topic_1	Topic_2	Topic_3	Dominant_Topic
0	0.002074	0.877315	0.002122	0.118489	Topic 1
1	0.431948	0.308979	0.221827	0.037246	Topic 0
2	0.098514	0.328469	0.002754	0.570263	Topic 3
3	0.002798	0.594327	0.357323	0.045553	Topic 1
4	0.078346	0.458679	0.002394	0.460581	Topic 3
...	...	...	...	...	...
16034	0.003929	0.004005	0.248629	0.743437	Topic 3
16035	0.151625	0.006179	0.006066	0.836130	Topic 3
16036	0.125002	0.125001	0.624996	0.125001	Topic 2
16037	0.003269	0.003217	0.317437	0.676077	Topic 3
16038	0.014033	0.476248	0.129217	0.380502	Topic 1

16039 rows × 5 columns

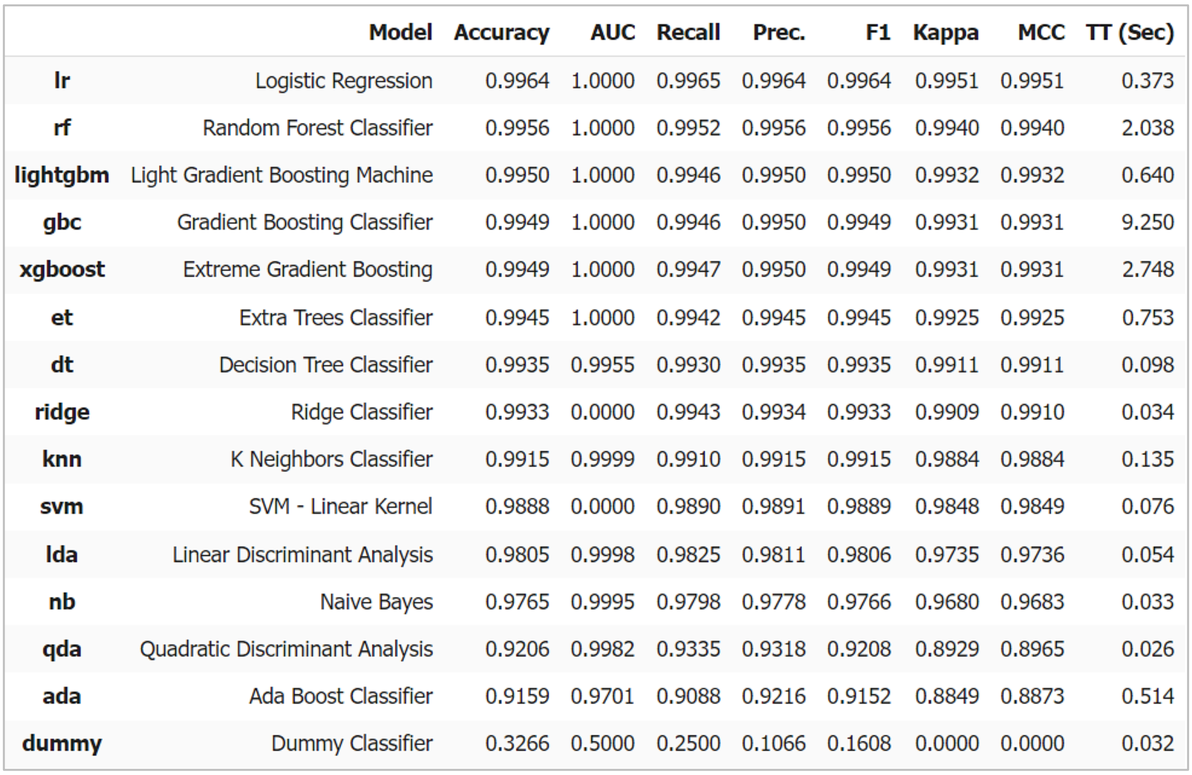
Figure 13. Document Classification

After defining the document classification task and selecting the appropriate classifier, we can proceed to train the model using the provided training data. The first argument of the training command specifies the classifier type, and the model will be trained and tested on the data provided, using default hyper parameters to obtain a general idea of the classifier's effectiveness. For this task, we chose to build a random forest classifier, which is often the first choice for any supervised machine learning task. This function trains all of the models in the model library with the default hyper parameters and uses cross-validation to measure performance metrics. The trained model object class is then returned for further use.

For classification tasks, several evaluation metrics are commonly used, including:

- Accuracy: the percentage of correct predictions over the total number of predictions made;
- Precision: accuracy of predicting an item as a particular class;
- Recall: accuracy of recognizing a member of a particular class;
- F1: geometric mean of precision and recall, providing a single metric to balance both measures;
- AUC (Area under Curve): a metric that measures the model's ability to distinguish between positive and negative instances;
- Cohen's Kappa;
- MCC (Matthew's correlation coefficient);
- TT (training time): the time taken to train the model, which is an essential metric to consider when comparing different models and selecting the most efficient one for a given application.

An extreme gradient boosting (XGBoost) classifier was created based on the dataset and its resulting accuracies were compared with those of the random-forest classifier. XGBoost is known for its robustness in classification tasks, high accuracy, and F1-score. Although XGBoost may outperform random forest in some cases, the choice of which algorithm to use ultimately depends on the problem specifics and available data. To evaluate and compare the performance of different models on the same dataset using appropriate metrics, we recommend starting with comparing all models for performance. This can be achieved by using the PyCaret setup which trains and scores all models in the library with cross-validation, evaluating metrics such as accuracy and precision. The output provides the average scores across folds and training times, giving a performance overview of all models as shown in Figure 14.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.9964	1.0000	0.9965	0.9964	0.9964	0.9951	0.9951	0.373
rf	Random Forest Classifier	0.9956	1.0000	0.9952	0.9956	0.9956	0.9940	0.9940	2.038
lightgbm	Light Gradient Boosting Machine	0.9950	1.0000	0.9946	0.9950	0.9950	0.9932	0.9932	0.640
gbc	Gradient Boosting Classifier	0.9949	1.0000	0.9946	0.9950	0.9949	0.9931	0.9931	9.250
xgboost	Extreme Gradient Boosting	0.9949	1.0000	0.9947	0.9950	0.9949	0.9931	0.9931	2.748
et	Extra Trees Classifier	0.9945	1.0000	0.9942	0.9945	0.9945	0.9925	0.9925	0.753
dt	Decision Tree Classifier	0.9935	0.9955	0.9930	0.9935	0.9935	0.9911	0.9911	0.098
ridge	Ridge Classifier	0.9933	0.0000	0.9943	0.9934	0.9933	0.9909	0.9910	0.034
knn	K Neighbors Classifier	0.9915	0.9999	0.9910	0.9915	0.9915	0.9884	0.9884	0.135
svm	SVM - Linear Kernel	0.9888	0.0000	0.9890	0.9891	0.9889	0.9848	0.9849	0.076
lda	Linear Discriminant Analysis	0.9805	0.9998	0.9825	0.9811	0.9806	0.9735	0.9736	0.054
nb	Naive Bayes	0.9765	0.9995	0.9798	0.9778	0.9766	0.9680	0.9683	0.033
qda	Quadratic Discriminant Analysis	0.9206	0.9982	0.9335	0.9318	0.9208	0.8929	0.8965	0.026
ada	Ada Boost Classifier	0.9159	0.9701	0.9088	0.9216	0.9152	0.8849	0.8873	0.514
dummy	Dummy Classifier	0.3266	0.5000	0.2500	0.1066	0.1608	0.0000	0.0000	0.032

Figure 14. A performance overview of all models

We conducted a thorough comparison of over 15 models using PyCaret and generated a comprehensive table that highlights the best performing models based on N-Fold cross-validation. Our findings, as illustrated in Figure 14 of the table, reveal that the models with the highest performance metrics sorted by "Accuracy" are as follows: Logistic Regression, Random Forest Classifier, Light Gradient Boosting Machine, Gradient Boosting Classifier, Extreme Gradient Boosting, and Extra Trees Classifier. However, our analysis also shows that when sorted by AUC, the aforementioned models are the top performers, with an impressive average ten-fold cross-validated AUC of 1.0000.

Ensemble Models and Other Variations

To enhance the accuracy of the current model, one promising approach is to transform it into an ensemble model. This involves integrating weak classifiers that have been trained on the same dataset. During prediction, the output of the weak classifiers is used to cast a vote for the final output class. The PyCaret tool can assist in building an ensemble model from a base classifier in a straightforward manner. To implement this method, the first step is to specify the base classifier as an argument. Next, the integration method can be specified in the options method, with two options available: Bagging (all weak classifiers cast a vote) and Boosting (weak classifiers hierarchically separate classes). The option n_estimators is used to set the number of weak classifiers in the ensemble model.

Hyper-Parameter Tuning

After evaluating the performance of the different models, we select the best-performing one based on the evaluation metric of our choice. However, it is important to note that the default hyperparameters of each model are used in this selection process. To achieve even better performance, the hyperparameters need to be fine-tuned. To tune the hyperparameters, we can use the command "classification.tune_model", where we specify the classification model in the first argument. This command helps improve the general performance of the model. Additionally, if we want to improve a specific metric, such as F1, we can specify it in the "optimize" option.

3.3.2. Model Evaluation and Interpretation

Finally, the tuned model with the command classification evaluation model (rf_model) was evaluated.

Figure 15 shows the AUC: ROC curve, which is a way to measure the performance of classification models at different thresholds. AUC indicates how well the model can distinguish between classes, with higher values indicating better performance. However, translating these metrics into business value requires further analysis.

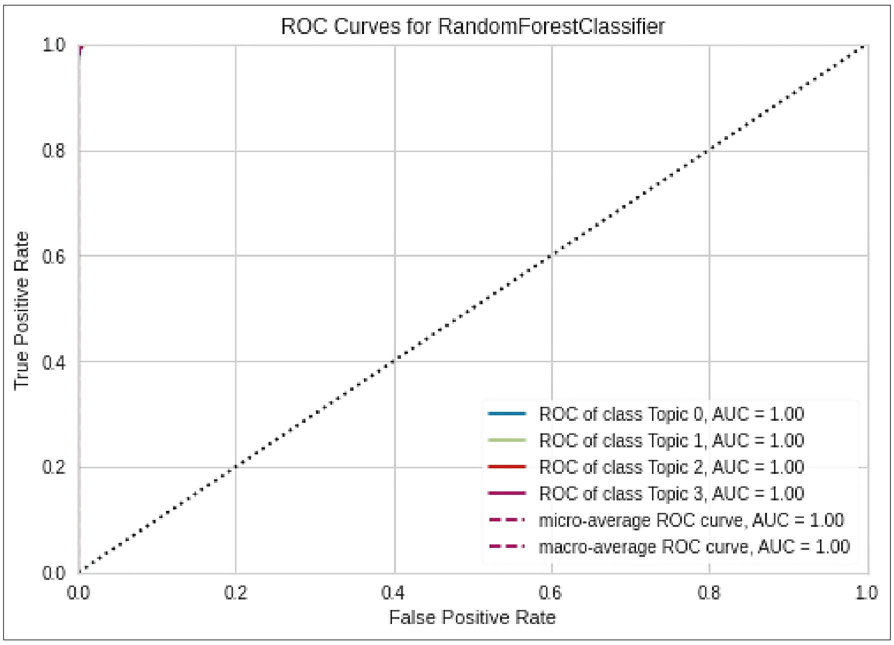


Figure 15. AUC plot of the best model

The confusion matrix is a simple yet effective way to evaluate model performance. It compares predicted labels with actual labels and divides them into four quadrants. The total sum of all quadrants is equal to the number of documents leads in the test set (405 + 2 + 782 + 2 + 1 + 729 + 2 + 3 + 3 + 2 + 475 = 2,406).

Various confusion matrices of randomly selected classifiers were examined in this section. The probability of true and false classification for each language model's classifiers, with reasonable processing time was visualized in Figure 16. Some classifications have a prediction probability of 0.0000, indicating a small number of false positives.

The bigram features worked better for prediction than unigram and trigram attributes were found in the previous section. This confirms the previous finding that there are overlapping structural relationships in the bibliometric analysis research in the dataset, despite having opposite sentiment polarities.

It is initialized with a fitted model and generates a class prediction error chart on the draw. The support for each class in the fitted classification model is shown as a stack of bars on the Class Prediction Error Chart. Each bar is segmented to show the distribution of predicted classes for each class.

The RandomForestClassifier is good at predicting Topic 0 based on the clustering features, but it often gets Topic 2 and Topic 3 mixed up and labels Topic 2 as Topic 1 (Figure 17).

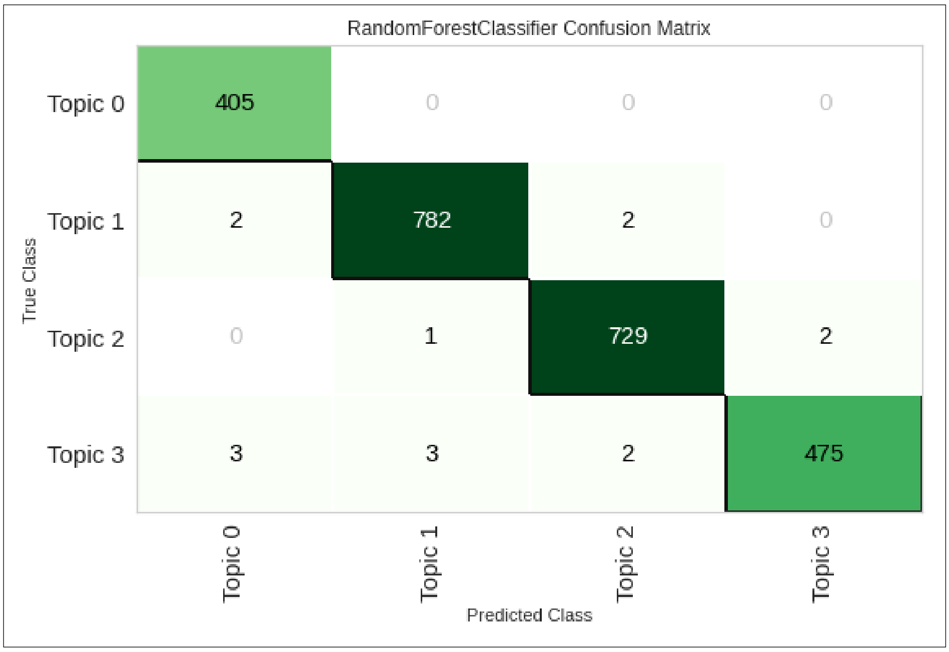


Figure 16. Confusion Matrix

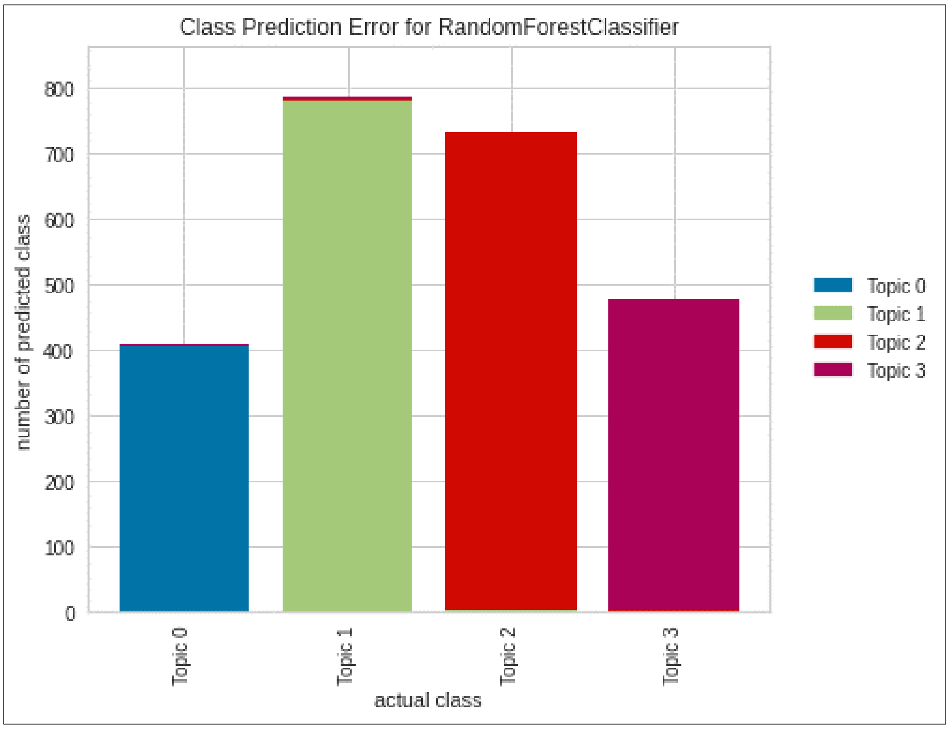


Figure 17. **Class Prediction error for Random Forest Classifier**

The confusion matrix and calculate precision was examined, recalled, and F1 scores done manually. If the model is binary, the impact of each feature on the prediction and the correlation between the two most impactful features using SHAP values could be also analyzed. A variable importance plot and probability values for each topic are represented by dots. The redder the dot, the higher the probability value. The probability values are sorted and colored with a gradient.

The position of each dot, which is a SHAP value, represents a correlation between the probability value and the output class. The more the dot goes to the right, the stronger the correlation between them and a correct prediction. The more it goes to the left, the stronger the correlation between it and an incorrect prediction. Besides, the lower the probability value, the bluer the dot is. The more separable the dots of different colors are, the better.

The result shows the probability distribution of topics for each document with $\theta_{m,k}$. values. The $\theta_{m,k}$. values for the first 10 documents and k=1…4 was displayed in Figure 18. We have prepared the data, trained and selected the best model based on AUC, and analyzed performance using various plots such as AUC-ROC, Confusion Matrix, Class Prediction error, and Topic probability.

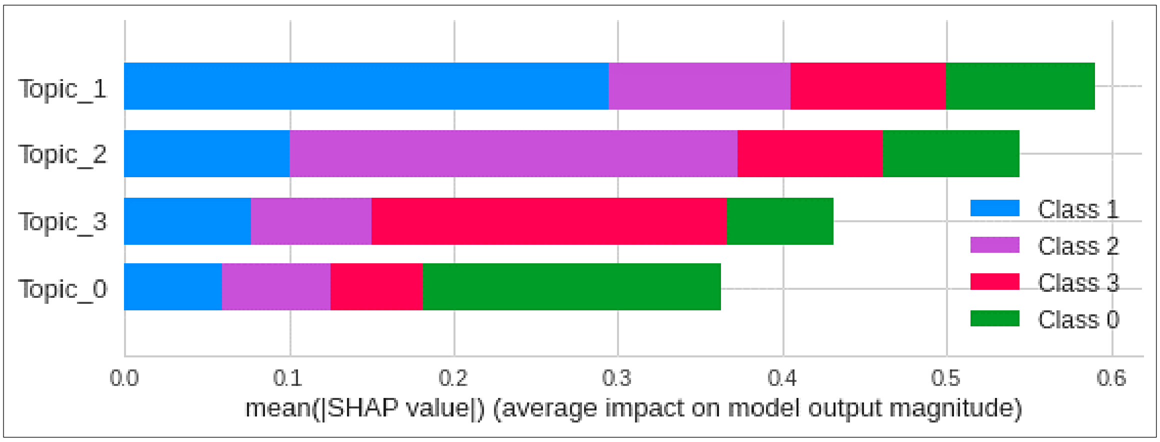


Figure 18. The probability of transcription from 1 to 10 is analyzed.

4. Results and Discussion

In this study, we analyzed bibliometric data across four different aspects: topic focus, characteristics, category, and key modes. To perform this analysis, we utilized Python libraries to carry out topic modeling and assessed the performance of each model using metrics such as perplexity and coherence. Additionally, we also used Python libraries to effectively visualize and present our results. By leveraging these libraries, we were able to create clear and informative visualizations that highlighted the key insights and trends from our analysis.

In this study, we examined the global research trends and context of bibliometric analysis using text analysis and topic modeling techniques on a sample of 16,379 papers from the Scopus database. Through our analysis, we identified four distinct topics and their development trends, revealing a shift in the focus of topic development trends. Specifically, the "Set of publication concepts" topic was found to be the most popular, accounting for 34% of tokens, while the "Relate to others in the research field" topic was less popular, accounting for only 10.4% of tokens. Building on the work of Aria & Cuccurullo [59], we classified two types of bibliometric analysis and knowledge structure synthesis: conceptual structure and intellectual and social structure. This classification is consistent with their findings and provides a useful framework for understanding bibliometric analysis. The study's insights provide a structured framework for understanding the evolving landscape of bibliometric analysis content, facilitating the tracking of research themes, identifying trends, and assessing methodological improvements. Furthermore, this study makes theoretical and methodological contributions by applying LDA topic modeling technology to bibliometric analysis and identifying research topics and development trends for the first time through large-scale text analysis of 16,039 documents from the Scopus database, this research contributes to a deeper understanding of bibliometric research patterns, offering valuable insights for researchers and practitioners. By obtaining core parameters, such as the number of topics, through machine learning training, our study provides a valuable reference for future research on bibliometric analysis.

In recent years, topic modeling has emerged as a powerful technique for discovering hidden topics within large collections of texts. With the increase in online activities of academics, businesses, and the general public, there has been a surge in interest in this field. This article explores the major topic modeling algorithms and their applications, with a specific focus on the Python programming language libraries and tools that can assist researchers in implementing their ideas. By using topic modeling, individuals can update and develop their interests in different aspects, which could lead to new research questions and ideas.

Bibliometrics, a quantitative methodology that uses statistical methods to analyze published works, has become increasingly important in recent years due to the vast amounts of available digital data [1]. One of the essential tools used in bibliometrics is topic modeling, which enables researchers to identify hidden patterns and relationships among texts [4]. Topic modeling allows researchers to discover new research areas and trends, and to learn about how certain publications or authors have affected their fields [3]. Additionally, topic modeling helps overcome some of the limitations of traditional bibliometric analysis, such as the over-reliance on citation counts as a measure of importance [3]. This study contributes to the field by highlighting the importance of analyzing knowledge structure and conceptual relationships within scientific publications and suggests that LDA is a powerful tool for efficient and accurate bibliometric analysis. The findings of this study are valuable for researchers in information science, data mining, and bibliometrics.

The result of this study contributes significantly to the field of bibliometric analysis research by evaluating the performance of topic modeling techniques and offering valuable insights into their advancements and implications. By comparing our findings with previous studies, we can discern the evolution and impact of our research. Our study introduces a novel approach by integrating word embedding with Latent Dirichlet Allocation (LDA) for enhanced topic extraction performance. Unlike previous studies that primarily relied on traditional topic modeling methods [20–31, 33, 34], our approach showcases a more sophisticated and potentially more effective method.

By identifying key topics within bibliometric research articles, our study sheds light on primary research topics, categories, and latent themes. Comparing these findings with previous studies enables researchers to track the evolution of research themes over time and assess consistency or changes in dominant topics within bibliometric analysis. In addition, our method, which includes getting data, cleaning it up, and using LDA models to analyze topics, gives a structured framework for doing bibliometric analysis. Researchers can identify improvements, challenges, and best practices by evaluating the methodology of previous studies compared to ours.

Even though there are some problems with the generalizability and granularity of the datasets, like the fact that abstracts were used instead of full texts, our study shows how important it is to use a variety of datasets and thorough text analysis methods to get reliable and applicable results. We suggest future research directions, such as exploring the full texts of authoritative articles in multiple languages. Comparing these recommendations with previous studies can reveal emerging trends and gaps in bibliometric analysis, guiding future research endeavors and methodological advancements.

5. Conclusion

In summary, the utilization of word embedding in conjunction with LDA has been demonstrated to enhance topic extraction performance, although its appropriateness varies depending on the unique characteristics of each research inquiry and dataset. The decision to employ word embedding should be made judiciously, considering a multitude of factors. Determining the ideal number of topics is a crucial aspect of LDA modeling. Typically, coherence scores offer a valuable metric for evaluating the quality of topic models across various topic quantities. Higher coherence scores signify more coherent topics, aiding researchers in identifying the most suitable number of topics for their specific investigation. Researchers are encouraged to experiment with different topic counts within a predefined range (e.g., 2-6) and compare coherence scores to arrive at an informed decision. To gain meaningful insights from LDA topic modeling results, a comprehensive approach encompassing statistical analysis, domain expertise, and critical thinking is essential. Beyond merely interpreting topics and their associated keywords, researchers should delve into the distribution of topics across documents, track the evolution of topics over time, and explore the interrelationships between topics. This multifaceted approach ensures a richer and more nuanced understanding of the underlying data structure. This study has identified two main limitations. Firstly, the findings of the study may not be generalizable due to the limited dataset used. However, incorporating additional sources such as the Web of Science core repository could increase the general applicability of the results. Secondly, the analysis was limited to abstracts rather than full texts, which may have limited the granularity of the findings. Therefore, future research should explore the full text of authoritative articles in multiple languages to obtain more comprehensive results.

6. Declarations

6.1. Author Contributions

Conceptualization, A.T., L.N., and W.C.; methodology, Y.J., A.T., L.N., and W.C.; software, Y.J., A.T., L.N., and W.C.; validation, S.K., V.C., L.N., and W.C.; formal analysis, S.K., V.C., L.N., and W.C.; investigation, V.C., L.N., and W.C.; resources, Y.J., L.N., and W.C.; data curation, Y.J., L.N., and W.C.; writing—original draft preparation, C.L., L.N., and W.C.; writing—review and editing, C.L., N.H., L.N., and W.C.; visualization, L.N. and W.C.; supervision, L.N. and W.C.; project administration, L.N. and W.C.; funding acquisition, L.N. and W.C. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available in the article.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

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

[2] Mejia, C., Wu, M., Zhang, Y., & Kajikawa, Y. (2021). Exploring topics in bibliometric research through citation networks and semantic analysis. Frontiers in Research Metrics and Analytics, 6, 742311. doi:10.3389/frma.2021.742311.

[3] Ninkov, A., Frank, J. R., & Maggio, L. A. (2022). Bibliometrics: Methods for studying academic publishing. Perspectives on medical education, 11(3), 173-176. doi:10.1007/s40037-021-00695-4.

[4] Li, X., & Lei, L. (2021). A bibliometric analysis of topic modelling studies (2000–2017). Journal of Information Science, 47(2), 161-175. doi:10.1177/0165551519877049.

[5] Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. Transportation Research Part C: Emerging Technologies, 87, 105-122. doi:10.1016/j.trc.2017.12.018.

[6] Nielsen, M. W., & Börjeson, L. (2019). Gender diversity in the management field: Does it matter for research outcomes?. Research Policy, 48(7), 1617-1632. doi:10.1016/j.respol.2019.03.006.

[7] Gohari, P., Wu, B., Hawkins, C., Hale, M., & Topcu, U. (2021). Differential privacy on the unit simplex via the dirichlet mechanism. IEEE Transactions on Information Forensics and Security, 16, 2326-2340. doi:10.1109/TIFS.2021.3052356.

[8] Jiang, H., Qiang, M., & Lin, P. (2016). Finding academic concerns of the Three Gorges Project based on a topic modeling approach. Ecological indicators, 60, 693-701. doi:10.1016/j.ecolind.2015.08.007.

[9] Li, Y., Jiang, D., Lian, R., Wu, X., Tan, C., Xu, Y., & Su, Z. (2021). Heterogeneous latent topic discovery for semantic text mining. IEEE Transactions on Knowledge and Data Engineering, 35(1), 533-544. doi:10.1109/TKDE.2021.3077025.

[10] Zhou, X., Liang, W., Luo, Z., & Pan, Y. (2021). Periodic-aware intelligent prediction model for information diffusion in social networks. IEEE Transactions on Network Science and Engineering, 8(2), 894-904. doi:10.1109/TNSE.2021.3064952.

[11] Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. Policy Studies Journal, 49(1), 300-324. doi:10.1111/psj.12343.

[12] Kwok, S. W. H., Vadde, S. K., & Wang, G. (2021). Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: machine learning analysis. Journal of medical Internet research, 23(5), e26953. doi:10.2196/26953.

[13] Wu, Q., Hare, A., Wang, S., Tu, Y., Liu, Z., Brinton, C. G., & Li, Y. (2021). Bats: A spectral biclustering approach to single document topic modeling and segmentation. ACM Transactions on Intelligent Systems and Technology (TIST), 12(5), 1-29. doi:10.1145/3468268.

[14] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84. doi:10.1145/2133806.2133826.

[15] Yin, B., & Yuan, C. H. (2022). Detecting latent topics and trends in blended learning using LDA topic modeling. Education and Information Technologies, 27, 12689–12712. doi:10.1007/s10639-022-11118-0.

[16] Hwang, S., & Cho, E. (2021). Exploring Latent Topics and Research Trends in Mathematics Teachers' Knowledge Using Topic Modeling: A Systematic Review. Mathematics, 9(22), 2956. doi:10.3390/math9222956.

[17] Schoepflin, U., & Glänzel, W. (2001). Two decades of" Scientometrics". An interdisciplinary field represented by its leading journal. Scientometrics, 50(2), 301-312. doi:10.1023/a:1010577824449.

[18] Jonkers, K., & Derrick, G. E. (2012). The bibliometric bandwagon: Characteristics of bibliometric articles outside the field literature. Journal of the American Society for Information Science and Technology, 63(4), 829-836. doi:10.1002/asi.22620.

[19] Milojević, S., & Leydesdorff, L. (2013). Information metrics (iMetrics): A research specialty with a socio-cognitive identity?. Scientometrics, 95, 141-157. doi:10.1007/s11192-012-0861-z.

[20] Ayaz, A., Ozyurt, O., Al-Rahmi, W. M., Salloum, S., Shutaleva, A., Alblehai, F., & Habes, M. (2023). Exploring Gamification Research Trends Using Topic Modeling. IEEE Access, 11, 119676-119692. doi:10.1109/ACCESS.2023.3326444.

[21] Robledo, S., & Zuluaga, M. (2022). Topic modeling: Perspectives from a literature review. IEEE Access, 11, 4066-4078. doi:10.1109/ACCESS.2022.3232939.

[22] Mifrah, S., & Benlahmar, E. H. (2020). Topic modeling coherence: A comparative study between LDA and NMF models using COVID'19 corpus. International Journal of Advanced Trends in Computer Science and Engineering, 5756-5761. doi:10.30534/ijatcse/2020/231942020.

[23] Cui, W., Jinling, L., Zhang, T., & Zhang, S. (2023). A Recognition Method of Measuring Literature Topic Evolution Paths Based on K-means-NMF. Knowledge Organization, 50(4), 257-271. doi:10.5771/0943-7444-2023-4-257.

[24] Motamedi, N., Ghazimirsaeid, J., Sheikhshoaei, F., Mansourzadeh, M. J., & Dehdarirad, H. (2023). Bibliometric Analysis and Topic Modeling of Information Systems in Maternal Health Publications. International Journal of Information Science and Management, 21(2), 85-101. doi:10.22034/ijism.2023.1977814.0.

[25] Almenara, C. A. (2022). 40 years of research on eating disorders in domain-specific journals: Bibliometrics, network analysis, and topic modeling. PloS one, 17(12), e0278981. doi:10.1371/journal.pone.0278981.

[26] Sharma, C., Batra, I., Sharma, S., Malik, A., Hosen, A. S., & Ra, I. H. (2022). Predicting trends and research patterns of smart cities: A semi-automatic review using latent dirichlet allocation (LDA). IEEE Access, 10, 121080-121095. doi:10.1109/ACCESS.2022.3214310.

[27] Gurcan, F., & Cagiltay, N. E. (2022). Exploratory analysis of topic interests and their evolution in bioinformatics research using semantic text mining and probabilistic topic modeling. IEEE Access, 10, 31480-31493. doi:10.1109/ACCESS.2022.3160795.

[28] Cobelli, N., & Blasi, S. (2024). Combining topic modeling and bibliometric analysis to understand the evolution of technological innovation adoption in the healthcare industry. European Journal of Innovation Management, 27(9), 127-149. doi:10.1108/EJIM-06-2023-0497.

[29] Chen, X., & Xie, H. (2020). A structural topic modeling-based bibliometric study of sentiment analysis literature. Cognitive Computation, 12, 1097-1129. doi:10.1007/s12559-020-09745-1.

[30] Chen, X., Xie, H., Cheng, G., & Li, Z. (2022a). A decade of sentic computing: topic modeling and bibliometric analysis. Cognitive computation, 14(1), 24-47. doi:10.1007/s12559-021-09861-6.

[31] Jiang, H., Qiang, M., & Lin, P. (2016). A topic modeling based bibliometric exploration of hydropower research. Renewable and Sustainable Energy Reviews, 57, 226-237. doi:10.1016/j.rser.2015.12.194.

[32] Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. Australian Journal of Management, 45(2), 175-194. doi:10.1177/0312896219877678.

[33] Chen, X., Zou, D., & Xie, H. (2022). A decade of learning analytics: Structural topic modeling based bibliometric analysis. Education and Information Technologies, 27(8), 10517-10561. doi:10.1007/s10639-022-11046-z.

[34] Amaro, A., & Bacao, F. (2024). Topic Modeling: A Consistent Framework for Comparative Studies. Emerging Science Journal, 8(1), 125-139. doi:10.28991/ESJ-2024-08-01-09.

[35] Cho, S. B., Shin, S., & Kang, D. S. (2018). A study on the research trends on open innovation using topic modeling. Informatization policy, 25(3), 52-74.

[36] Ali, M. (2020). PyCaret: An open source, low-code machine learning library in Python. PyCaret Version, 2.

[37] Bettina, G., & Kurt, H. (2011). Topic models: An R package for fitting topic models. Journal of Statistical Software, 40(13), 1-30. doi:10.18637/jss.v040.i13.

[38] Chowdhury, C. R., & Bhuyan, P. (2010). Information retrieval using fuzzy c-means clustering and modified vector space model. 3rd International Conference on Computer Science and Information Technology, 1, 696-700. doi:10.1109/ICCSIT.2010.5564542.

[39] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

[40] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211. doi:10.1007/s11042-018-6894-4.

[41] Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, 952-961.

[42] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. Advances in Neural Information Processing Systems, 22.

[43] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining, 399-408. doi:10.1145/2684822.2685324.

[44] Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010, 46-50.

[45] Chen, X., Zou, D., & Xie, H. (2020). Fifty years of British Journal of Educational Technology: A topic modeling based bibliometric perspective. British Journal of Educational Technology, 51(3), 692-708. doi:10.1111/bjet.12907.

[46] Ozansoy Çadırcı, T., & Sağkaya Güngör, A. (2021). 26 years left behind: a historical and predictive analysis of electronic business research. Electronic Commerce Research, 21, 223-243. doi.org:10.1007/s10660-021-09459-y.

[47] Zhu, B., Zheng, X., Liu, H., Li, J., & Wang, P. (2020). Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. Chaos, Solitons & Fractals, 140, 110123. doi:10.1016/j.chaos.2020.110123.

[48] Bovens, L., & Hartmann, S. (2003). Solving the riddle of coherence. Mind, 112(448), 601-633. doi:10.1093/mind/112.448.601

[49] Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, 100-108.

[50] Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 530-539.

[51] Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. Proceedings of the 2011 conference on empirical methods in natural language processing, 262-272.

[52] Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing, 72(7-9), 1775-1781. doi:10.1016/j.neucom.2008.06.011.

[53] Sievert, C., & Shirley, K. (2014, June). pyLDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces, 63-70.

[54] Small, H. (1997). Update on science mapping: Creating large document spaces. Scientometrics, 38, 275-293. doi:10.1007/BF02457414.

[55] Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. Annual review of information science and technology, 37(1), 179-255. doi:10.1002/aris.1440370106.

[56] Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. Proceedings of the international working conference on advanced visual interfaces, 74-77.

[57] Zhao, W., Chen, J.J., Perkins, R. et al. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics 16 (Suppl 13), S8 (2015). doi:10.1186/1471-2105-16-S13-S8.

[58] Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures, 12(2-3), 93-118. doi:10.1080/19312458.2018.1430754.

[59] Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959-975. doi:10.1016/j.joi.2017.08.007.