

A comparative topic modeling analysis of geoscience literature using LDA and BERTopic^{*}

Kadriye Karadeniz^{*}, Yasin Ortakçı and İsmail Rakıp Karas

¹ Karabük University, Karabük 78050, Turkey

Abstract

The rapid growth of scientific literature necessitates the automated analysis of research trends and thematic structures. This study applies topic modeling techniques to publications in the field of geosciences found in the Web of Science (WoS) database. The study includes a comparative analysis of Latent Dirichlet Allocation (LDA), a probabilistic approach, and BERTopic, a modern method based on contextual language representations. Using the titles and abstracts of 11,732 scientific publications, both methods were evaluated in terms of topic coherence, topic distribution, and thematic significance. The findings show that the BERTopic method yields more successful results than LDA, particularly in terms of semantic coherence (coherence: 0.757) and topic separation (coherence: 0.623). While LDA generated 11 topics, BERTopic identified 25 meaningful topics. The comparative analyses reveal that the topic generation strategies of BERTopic and LDA differ; this shows that BERTopic offers more detailed topic structures, while LDA creates more general thematic clusters.

Keywords

Topic Modelling, Latent Dirichlet Allocation, BERTopic, Geoscience, Machine Learning, Scientific Literature Analysis.

1. Introduction

The rapid increase in the number of scientific publications is making literature tracking increasingly difficult for researchers. The number of articles published in indexing systems such as Scopus and Web of Science (WoS) is increasing at an average annual growth rate of 4-5% [1]. Especially in interdisciplinary fields, there is a growing need for automated text analysis methods to understand the thematic structure of existing studies and identify research trends. In this context, topic modeling is a widely used approach to uncover hidden themes from large text collections [2]. The field of geoscience is a broad research area encompassing numerous sub-disciplines such as remote sensing, geographic information systems, geology, climate science, and environmental engineering. In recent years, the integration of artificial intelligence and machine learning techniques into this field has led to further diversification of the literature [3]. Therefore, systematically mapping research trends and identifying thematic structures in the field of geoscience is of great importance to both researchers and field experts. Comprehensive bibliometric databases, such as WoS, offer a rich resource for such systematic analyses [1]. Classical topic modeling methods include Latent Semantic Analysis (LSA) [4] and probabilistic LSA [5], but one of the most commonly used methods is Latent Dirichlet Allocation (LDA) [2]. LDA is considered a reference model in bibliometric studies due to its theoretical robustness and interpretability. However, in recent years, deep learning methods based on contextual word representations have made a significant progress in the field of topic modeling. One of these methods, BERTopic, aims to produce more meaningful and consistent topics by combining BERT-

^{*}ADT-ISMDCC 2026: *Advanced Digital Technologies: Intelligent Systems, Modeling, and Data-Driven Control*, September 22–24, 2026, Odesa, Ukraine

¹CST-2026: *Information Control Systems & Technologies*, September 22-24, 2026, Odesa, Ukraine

^{*} Corresponding author: kadriyekaradeniz@karabuk.edu.tr

kadriyekaradeniz@karabuk.edu.tr (K. Karadeniz); yasinortakci@karabuk.edu.tr (Y. Ortakçı);

ismail.karas@karabuk.edu.tr (İ. R. Karas)

 0009-0007-4845-9354 (K. Karadeniz); 0000-0002-0683-2049 (Y. Ortakçı); 0000-0001-5934-3161 (İ. R. Karas)

 © 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

based embeddings with the class-based TF-IDF approach [6]. Recent comparative studies show that BERTopic is superior to classical methods, especially in topic consistency metrics [7]. The main objective of this study is to comparatively examine the traditional topic modeling approach, LDA, and the context-based modern method, BERTopic, on a large-scale collection of publications in the WoS database covering the field of geoscience. Approximately 12,000 scientific articles were analyzed using both methods, and the resulting topic structures were examined to assess their capacity to represent field knowledge. This study addresses three fundamental research focuses. First, the thematic levels and depth of detail at which LDA and BERTopic methods represent the geoscience literature are investigated. Second, the consistency and interpretability of the resulting topics are compared through quantitative metrics and qualitative analyses. Third, within an interdisciplinary context (the intersection of geoscience and artificial intelligence/machine learning), the study analyzes which method more distinctly and meaningfully reveals the sub-themes of the field. In this respect, the study aims to contribute to a better understanding of the thematic structure of the geoscience literature by comparing the types of information offered by different topic modeling approaches in large-scale scientific text collections. The main contributions of the study to the literature can be summarized under four headings. First, a comprehensive topic modeling analysis is presented on a large-scale dataset encompassing 12,701 WoS publications. Second, a traditional probabilistic approach, LDA, and a modern context-based method, BERTopic, are systematically and controllably compared on the same dataset. Third, research trends emerging at the intersection of geoscience, artificial intelligence, and machine learning are analyzed, revealing the thematic structure and topic distributions of this interdisciplinary field. Finally, in light of the findings, concrete methodological inferences are presented regarding which topic modeling approach may be more suitable for different data characteristics and research objectives, providing practical guidance for researchers.

2. Related Works

Topic modeling is an approach aimed at automatically uncovering latent topics in large-scale text collections and is widely used in fields such as information retrieval, text mining, and bibliometrics.

2.1. Classical Topic Modelling Methods

The first studies in this field were based on linear algebra-based methods. LSA aimed to represent documents in a low-dimensional latent space by applying singular value decomposition to the term-document matrix and was one of the pioneers of topic modeling [4]. The problems of LSA producing negative values and interpretability were largely overcome by the Non-negative Matrix Factorization (NMF) approach developed later. NMF offered more intuitive and part-based subject representations using only positive components [8]. Following linear methods, a significant breakthrough in the field of topic modeling occurred with LDA. LDA has become a standard in the field by offering a probabilistic framework that models documents as topic mixtures and topics as word distributions [2]. The need to predetermine the number of topics in LDA has been addressed with hierarchical Dirichlet processes; this non-parametric Bayesian approach has been able to learn the number of topics directly from the data [9]. Dynamic topic models, developed by including the time dimension in the model, have allowed for the analysis of the evolution of topics over time [10]. The independent topic assumption of LDA has been extended with the related topic model, and inter-topic correlations have been modeled using logistic-normal distributions [11]. Similarly, the Pachinko allocation model has offered a more flexible structure by modeling multi-level and directional relationships between topics [12]. Supervised LDA, developed for supervised scenarios, has produced effective results in labeled datasets by optimizing topic inference along with classification or regression objectives [13]. In addition, the author-topic model has made significant contributions, especially in fields such as scientific publication analysis, by including author information in the model [14]. Among the studies focusing on theoretical foundations, the anchor words approach stands out. This method has provided fast, scalable, and theoretically guaranteed

topic modeling algorithms based on the assumption that certain words belong to only one topic [15]. In the literature, methods like SeaNMF, developed for short texts, have also aimed to overcome the sparse data problem; successful results have been obtained in social media posts and title-based datasets [16]. Interactive topic modeling studies, on the other hand, have highlighted human-controlled approaches by guiding topics through user feedback [17].

2.2. Neural and Embedding-Based Approaches

With the rise of deep learning in recent years, neural topic models have gained significant importance in the field of topic modeling. The neural variational document model was one of the first studies to apply the variational autoencoder approach to document modeling [18]. Following this, AVITM/ProdLDA, developed afterwards, offered significant improvements compared to classical LDA by producing sharper and more semantically consistent topics [19]. Embedding-based approaches have created a major transformation in the field of topic modeling in terms of practical use. Top2Vec offered a method that automatically extracts topic vectors using word and document embeddings, eliminating the need to predetermine the number of topics [20]. BERTopic, a continuation of this approach, has become one of the most widely used topic modeling methods today with its combination of transformer-based embeddings, clustering algorithms, and class-based TF-IDF [6].

2.3. Applications of Scientific Literature Analysis

One of the studies on topic modeling applications on scientific literature is a study based on the WoS database and examining research trends in the social sciences. [21] shows that interdisciplinary thematic structures can be revealed by applying topic modeling on bibliometric data. Similarly, in [22] which examines theses and publications of computer engineering and computer science master's programs in Turkey, LDA, Top2Vec and BERTopic methods were compared; it was revealed that the methods have different strengths in terms of topic consistency and interpretability. Studies focusing on the comparative evaluation of topic modeling algorithms also hold an important place in the scientific literature. In particular, in a comprehensive comparison study conducted using BERTelex, classical methods and embedding-based approaches were compared on scientific articles [23]. This study reports that modern methods such as BERTopic produced more consistent and meaningful topics. Similarly, [7] compares studies conducted on Twitter data, LDA, NMF, Top2Vec and BERTopic methods. It shows that embedding-based models yield more successful results in short texts. Topic modeling methods are used not only for literature analysis purposes but also in the context of scientific article recommendation systems and information access. The collaborative topic modeling approach aimed to improve scientific article recommendation performance by combining traditional topic modeling with collaborative filtering methods. This study shows that topic-based representations offer an important source of information in recommendation systems [24]. Similarly, the neural relational topic modeling study managed to capture richer semantic relationships compared to classical methods by using neural network-based topic modeling approaches to model the relationships between scientific articles [25]. Dynamic topic modeling approaches have been used to examine the evolution of topics over time, especially in rapidly evolving research areas such as epidemics. In [26] on the coronavirus literature, the temporal change of research themes was analyzed using dynamic topic modeling, and it was shown that newly emerging topics could be identified at an early stage. Similarly, [27] examined research dynamics in the field of sports management with the help of topic modeling, and the thematic transformation of the field over the years was revealed. Studies focusing on topic tagging and classification problems in short texts are also included in the literature. It has been shown that topic tagging performance can be improved in short texts such as scientific article titles using graph-based models and deep learning approaches [28]. In addition, a systematic review study on Wi-Fi and deep learning integration emphasized that topic modeling techniques are an effective tool in literature review and field mapping processes [29]. Similarly,

interdisciplinary topic modeling studies conducted in the fields of semantic web and ontology-based knowledge representation have presented important findings on the optimization of LDA hyperparameters and the systematic evaluation of model performance on different models [30]. Studies on determining topic relationships in scientific literature draw attention to the importance of considering topic modeling outputs not only as independent themes but also as related structures. These studies show that topic networks and relational analyses offer a deeper analytical perspective compared to classical topic lists [31]. All of these studies demonstrate that topic modeling is a powerful method in scientific literature analysis, identifying research trends, and field mapping processes. However, the literature shows a limited number of studies directly comparing the application of LDA and BERTopic methods on large-scale WoS data. This study aims to fill this gap by systematically evaluating LDA and BERTopic methods on a comprehensive dataset containing more than 12,000 scientific publications and revealing research trends in the field of geoscience.

3. Methodology

3.1. Latent Dirichlet Allocation

LDA, a probabilistic topic modeling method proposed by Blei, Ng, and Jordan that aims to uncover hidden thematic structures in text collections, treats documents as a mixture of topics, and topics as probability distributions defined on words [2]. LDA has gained widespread acceptance in the natural language processing and text mining literature and has become a fundamental reference model for the analysis of large-scale document collections. The process of the generative LDA is described as follows: For each topic $k \in \{1, \dots, K\}$, the word distribution $\phi_k \sim \text{Dirichlet}(\beta)$ is sampled and for each document $d \in \{1, \dots, D\}$ the topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$ is sampled. For each word position $n \in \{1, \dots, N_d\}$ in document d , a subject $z_{d,n} \sim \text{Multinomial}(\theta_d)$ is first assigned, and then the word $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$ associated with this subject is generated. Here: $\alpha \in R_{+^{K \times 1}}$: Dirichlet hyperparameter controlling the concentration of the document-topic distribution, $\beta \in R_{+^{V \times 1}}$: Dirichlet hyperparameter controlling the concentration of the topic-word distribution, $\theta_d \in \mathcal{A}^{K-1}$: Topic distribution of document d , $\phi_k \in \mathcal{A}^{V-1}$: Word distribution of topic k , $z_{d,n}$: Latent topic assignment, $w_{d,n}$: Observed word.

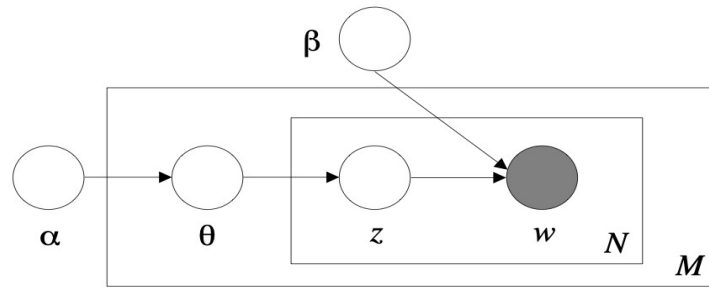


Figure 1: Graphical representation of LDA with plate notation [2].

In the graphical representation of LDA given in Figure 1, the boxes are plates representing repetitions. The outer plate represents documents, and the inner plate represents the repeated selection of topics and words within a document. The joint probability distribution of the model is given in Equation 1:

$$p(\theta, z, w | \alpha, \beta) = \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \beta) \quad (1)$$

Calculating the posterior probability distribution ($p(\theta, z | w, \alpha, \beta)$) of latent variables is not analytically possible. Therefore, two basic inference methods are used in the literature: variational inference [2] and reduced Gibbs sampling [32]. The variational approach proposed by Blei et al. minimizes the Kullback-Leibler divergence by approximating the true posterior probability distribution with a simpler family of distributions. The Markov Chain Monte Carlo method developed by Griffiths and Steyvers [32] is widely preferred in practical applications. The method samples the subject assignments (z) directly by marginalizing the θ ve ϕ parameters. The main advantages of LDA are theoretical robustness, interpretability, and scalability in large datasets. However, it has some important limitations such as the need to predetermine the number of subjects in the model, the loss of contextual information due to the bag of words assumption, and poor performance in short texts. These limitations have paved the way for the development of embedding-based methods such as BERTopic [6] and Top2Vec [20] in recent years.

3.2. BERTopic

BERTopic is a modern topic modeling method that integrates transformer-based language models, density-based clustering algorithms, and class-based TF-IDF (c-TF-IDF) approach [6]. The method aims to overcome the loss of contextual information resulting from the baggage assumption of classical probabilistic models by leveraging the semantic representation capacity of pre-trained deep learning models. BERTopic performs the topic modeling process in three main stages: (1) transformer-based document representation, (2) dimensionality reduction and density-based clustering, and (3) topic characterization with class-based TF-IDF.

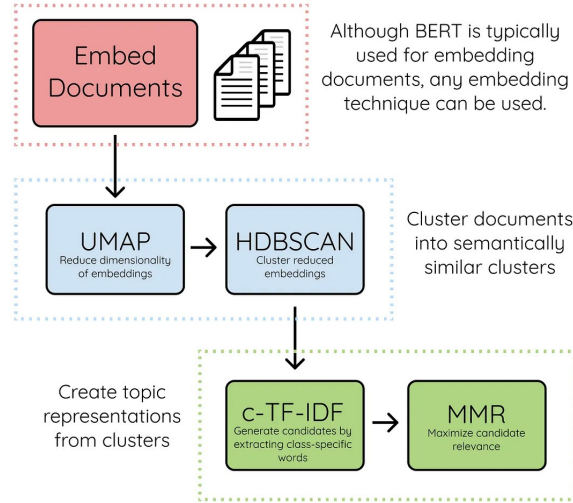


Figure 2: BERTopic methodological flowchart [33].

Figure 2 shows methodological flowchart of BERTopic. It shows the document embedding, dimensionality reduction, clustering and c-TF-IDF stages.

3.2.1. Document Representation

Each document d is represented in a dense vector space using a pre-trained transformer model, as given in Equation 2:

$$e_d = f_{\text{BERT}}(d) \quad (2)$$

Here f_{BERT} is the all-MiniLM-L6-v2 Sentence-BERT model, a model finely tuned to capture semantic similarities at the sentence level, and $e_d \in R^n$ represents the semantic representation of the document.

3.2.2. Dimensionality Reduction and Clustering

High dimensional High-dimensional embedding vectors are projected into low-dimensional space using the UMAP (Uniform Manifold Approximation and Projection) algorithm [34], as shown in Equation 3:

$$z_d = \text{UMAP}(e_d) \quad (3)$$

Then, the documents are assigned to topics by applying the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm, as given in Equation 4:

$$C = \text{HDBSCAN}(z) \quad (4)$$

HDBSCAN has the advantage of automatically determining the number of clusters from the data structure, being able to filter out noisy observations, and detecting clusters with varying densities.

3.2.3. Class-based TF-IDF

Each set is treated as a “class” and topic-word relationships are quantified with c-TF-IDF scores. The equation for this process is given in Equation 5:

$$\text{c-TF-IDF}_{t,c} = \frac{f_{t,c}}{\sum_{t \in c} f_{t,c}} \cdot \log \left(\frac{N}{\sum_c I(t \in c)} \right) \quad (5)$$

Here, $f_{t,c}$ represents the frequency of word t in topic c , N represents the total number of

documents, and I represents the indicator function. This approach enables the identification of topic-specific distinctive terms. BERTopic’s key advantages include its ability to directly model contextual semantic information, automatically determine the number of topics from the data structure, and exhibit high performance in short texts. Comparative studies show that BERTopic performs better than classical methods, especially in topic consistency metrics [7]. However, the method has limitations such as high computational cost, sensitivity to embedding and clustering hyperparameters, and lack of a fully probabilistic framework. In addition, the variability of results depending on random initialization values is a factor that should be considered in terms of the model’s reproducibility. BERTopic is widely applied in current literature in fields such as scientific publication analysis, social media mining, and the identification of interdisciplinary research trends, and is used as a reference model in comparative evaluations of modern topic modeling methods.

4. Experimental Study

4.1. Dataset

The dataset used in this study was created on December 19, 2025, using the Topic field in the WoS database. The literature search was conducted to cover all time periods without any year restrictions. Keywords representing the intersection of geographic information science and

artificial intelligence and machine learning were used in the search process. The WoS query expression used in this context is as follows: *Topic = (“geo ai” OR “artificial intelligence in gis” OR “deep learning for geospatial data” OR “geoscience and remote sensing” OR “geosciences” OR “geospatial artificial intelligence” OR “geospatial machine learning” OR “machine learning in gis” OR “spatial machine learning”)*. A search conducted using the specified keywords resulted in access to a total of 12,701 scientific publications indexed in the WoS database. In this study, the title and abstract information of each publication were included in the dataset for use in text mining and subject modeling analyses. This large-scale dataset allows for a comprehensive examination of research trends based on artificial intelligence and machine learning in the field of geospatial sciences. The basic characteristics of the dataset are presented in Table 1.

Table 1

Dataset statistics.

Feature	Value
Total number of documents	12.701
Year range	1990–2024
Average word count (Title)	12.4
Average word count (Abstract)	187.6
Average word count (Combined)	200.0
Total number of unique words	34.562
Language	English

In order to enable topic modeling methods to produce more meaningful and consistent outputs, some preprocessing steps were systematically applied to the dataset.

4.2. Preprocessing

The text used in the analysis was created by combining the title and abstract texts of each publication. This approach aims to increase the representation of the obtained topics by ensuring that both the title information, which reflects the main idea of the documents, and the abstract text, which presents the content of the study in detail, are evaluated together. In the text preprocessing process, all documents were first converted to lowercase; punctuation marks, numerical expressions, URLs and email addresses, and special characters were removed from the text. This step reduces formal differences between texts and allows the model to focus on semantic patterns. Then, common English stop words (e.g., the, and, of, in, to, for) were removed from the texts using the NLTK library. In addition, common academic expressions (e.g., study, method, result, paper) that could reduce the distinctiveness of the subject modeling results were filtered through a custom stop word list created based on domain knowledge. Tokenization and lemmatization processes were performed using the spaCy library. The texts were parsed at the word level, and words were reduced to their root forms. This process increases conceptual consistency by grouping different conjugations representing the same concept (e.g., classify, classified, classification) under a single term. To remove noise from the model, documents that were too short after preprocessing were excluded from the analysis. In this context, texts with less than 20 words were removed from the dataset, and as a result of the filtering process, a total of 11,732 documents were included in the final analysis. Finally, an n-gram (bigram) model was created using the Gensim library to capture meaningful expressions consisting of two words. In this way, domain-specific terms such as machine learning, remote sensing, and neural network were treated as single tokens. After

completing these preprocessing steps, a cleaned and homogenous text collection suitable for applying subject modeling algorithms was obtained. Statistics for the preprocessed dataset are presented in Table 2.

Table 2

Dataset statistics after preprocessing step.

Feature	Value
Number of documents included in analysis	11.732
Average number of tokens	230.2
Median number of tokens	222.0
Total number of unique tokens	10.000
Top 15 tokens	geoscience, geosciences, high, field, area, time, applications, science, university, geological, spatial, rock, earth, information, learning

In the preprocessing stage, the original 47,468-word dictionary was reduced to 10,000 words by filtering out rare and very common terms. Examining the 15 most frequently used words, terms like “geoscience”, “geosciences”, and “learning” reflect the interdisciplinary nature of the dataset. The texts generated after the preprocessing steps were then fed as input to the LDA and BERTopic algorithms. Word-frequency-based representation was used for LDA, while transformer-based embedded representations were utilized for the BERTopic method. In both methods, the quality of the model outputs was evaluated using topic consistency metrics.

4.3. Experimental Settings

This section discusses the experimental settings of the LDA and BERTopic models; the process for determining hyperparameters and selecting the optimum number of topics is detailed.

Hyperparameters. The key hyperparameters affecting the performance of the LDA model are the number of topics (K), the α parameter which controls the document-topic distribution, and the β parameter which controls the topic-word distribution. In the process of determining the number of topics, coherence scores were calculated on different K values, $K = \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$, and $K = 11$ was selected as the optimum number of topics. The alpha parameter was set to 0.1 to ensure that each document focuses on a limited number of topics, and the beta parameter was set to 0.01 to ensure that each topic contains more distinctive words. To guarantee the convergence of the model, the number of iterations was set to 1000 and the number of passes to 10.

Table 3

Hyperparameter of the LDA model.

Feature	Value
Topic number	11
Alpha	0.1
Beta	0.01

Iterations 1000
Passes 10

Table 3 contains the hyperparameters of the LDA model. The graph in Figure 3 shows the consistency scores according to the K value (K=5: 0.5779, K=7: 0.6132, K=11: 0.6228, K=14: 0.5998, K=17: 0.5817, K=20: 0.5802).

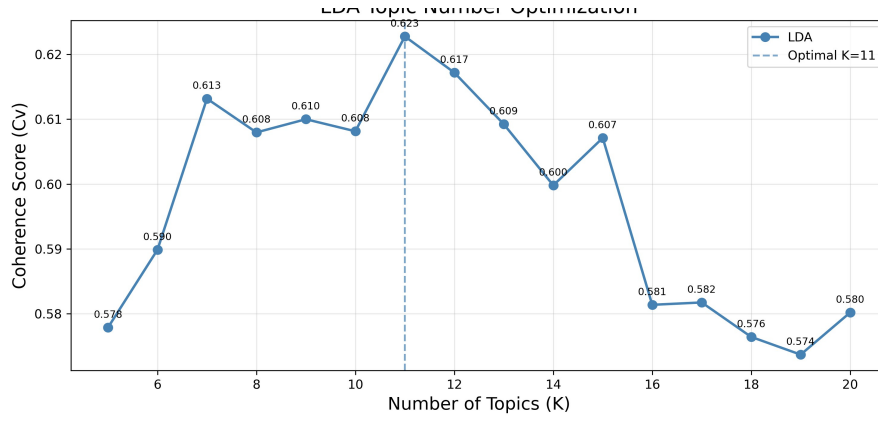


Figure 3: LDA topic number optimization.

Table 4 shows the hyperparameter settings for the components used in the BERTopic model. The UMAP and HDBSCAN hyperparameters were selected via a grid search over $n_neighbors \in \{10, 15, 20, 25, 30\}$ and $min_cluster_size \in \{20, 30, 40, 50, 60\}$ (with $min_samples$ fixed at 8), evaluating every combination with the same coherence metric used for LDA. The combination $n_neighbors = 20$ and $min_cluster_size = 60$ achieved the highest coherence (0.757) among all combinations tested and was adopted for the final model.

Table 4

BERTopic Hyperparameter Settings

Component	Parameter	Value
Embedding	Model	all-MiniLM-L6-v2
	Boyt	384
UMAP	$n_neighbors$	20
	$n_component$	5
	min_dist	0.0
	$metric$	cosine
HDBSCAN	$min_cluster_size$	60
	$min_samples$	8
	$metric$	euclidean
c-TF-IDF	Top N Words	10

BERTopic analyses were performed using the BERTopic 0.16.0 library. Semantic representations of the texts were obtained through sentence-transformer-based embedding models. The high-

dimensional vector space was reduced using the UMAP (umap-learn 0.5.4) method; the resulting representations were clustered using the HDBSCAN (0.8.33) algorithm. All experiments were conducted on an Apple M1 Pro chip (8-core CPU: 6 performance cores and 2 efficiency cores) and 16 GB of unified memory, running a Python 3.10 environment.

5. Results

In this study, a total of 11,732 pre-processed scientific publications obtained from the WoS database, containing title and abstract information, were analyzed. The topics obtained for both methods were evaluated using the number of articles per topic and the topic coherence metric. The LDA model was trained with $K = 11$ topics. As a result of the model, the top 10 keywords with the highest probability were obtained for each topic. Table 5 contains the topic structures, coherence values, document numbers, and dominant keywords obtained with the LDA model, and the topics in the table are limited to the top 10 topics, sorted in descending order according to their coherence values.

Table 5

Topics obtained with LDA and representative keywords.

Topic	Doc. Count	Top Words	Coherence
2	663	forest, height, elevation, surface, glas, cloud, land, laser, satellite, lidar	0.878
4	827	student, geoscience, learning, education, science, service, course, tool, teaching, user	0.732
6	2166	spatial, learning, network, image, algorithm, machine, prediction, classification, deep, datasets	0.716
9	831	rock, age, mantle, zircon, crust, geosciences, crustal, subduction, peking, similar	0.687
5	1000	seismic, flow, earthquake, wave, velocity, fracture, parameter, numerical, stress, inversion	0.670
8	992	satellite, observation, signal, resolution, gravity, earth, measurement, error, space, gps	0.630
0	876	mineral, rock, reservoir, gas, element, material, fluid, isotope, formation, carbonate	0.606
3	2164	geoscience, science, geological, scientific, resource, geosciences, project, technology, geology, earth	0.599
10	548	basin, deposit, fault, zone, ore, south, gold, geological, region, tectonic	0.589
1	1001	water, soil, climate, region, groundwater, temperature, land, environmental, sea, energy	0.553

According to LDA results, the best number of topics calculated for the model was 11, with a topic coherence of 0.6228. A total of 11 topics were obtained as a result of the applied topic modeling analysis. The highest topic coherence was observed in topic 2 with a value of 0.878, while the lowest coherence was obtained in topic 7 with a value of 0.543. All 11,732 documents included

in the analysis were successfully assigned to a topic, indicating that the model completely covers the dataset. The BERTopic model was implemented using transformer-based document embeddings and a density-based clustering approach. The model generated 25 meaningful topics; outlier documents were reassigned to their nearest topic using a class-based TF-IDF strategy, so that no documents were excluded from the analysis. Table 6 includes the topics generated by the BERTopic model, the number of documents for each topic, coherence values, and representative keywords. The topics in the table are limited to the top 10 topics, sorted in descending order according to their coherence values.

Table 6

Topics and document distribution obtained with BERTopic.

Topic	Doc. Count	Top Words	Coherence
23	78	genes, gene, expression, cancer, patients, gene expression, geo, cell, immune, clinical	0.970
19	137	geoscience remote, grss, ieee, symposium, remote sensing, sensing, remote, international geoscience, sensing symposium, igarss	0.913
1	1054	forest, glas, laser, elevation, height, lidar, canopy, altimeter, accuracy, ice	0.908
15	168	bibliometric, publications, journals, articles, collaboration, countries, sciences, papers, citation, citations	0.905
18	224	reservoir, oil, gas, geothermal, reservoirs, production, wells, petroleum, shale, hydrocarbon	0.853
3	900	students, education, teaching, learning, course, student, teachers, women, undergraduate, virtual	0.848
5	863	geoheritage, geological, geotourism, geodiversity, heritage, tourism, cultural, environmental, geoethics, geopark	0.829
7	430	gravity, satellite, gps, orbit, gnss, positioning, grace, satellites, gravitational, accuracy	0.815
17	195	fracture, fractures, rock, stress, flow, permeability, fractured, numerical, fluid, hydraulic	0.810
11	264	geoai, geospatial, intelligence, artificial intelligence, artificial, urban, geospatial artificial, learning, social, intelligence geoai	0.809

The BERTopic model's topic consistency value was calculated as 0.757. The highest topic consistency was observed in topic 23 with a value of 0.970, while the lowest consistency was obtained in topic 21 with a value of 0.375. All 11,732 documents included in the analysis were successfully assigned to a topic, indicating that the model fully covers the dataset. As shown in Figure 4, the coherence performance of two different topic modeling methods was compared. BERTopic showed a coherence score of 0.757, while LDA showed a score of 0.623. BERTopic's 21.6% higher score compared to LDA demonstrates the superiority of transformer-based embeddings and density-based clustering in capturing contextual semantic information. Table 7 shows the comparative results of the LDA and BERTopic models. The results show that the BERTopic method produces higher topic consistency values compared to LDA.

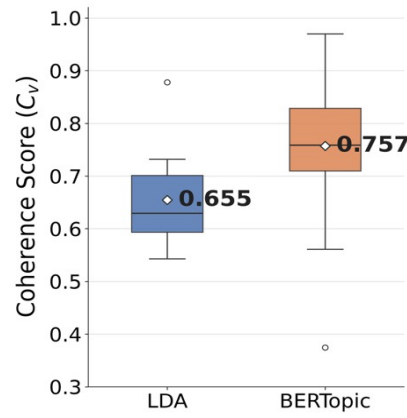


Figure 4: Comparison of topic coherence between LDA and BERTopic models.

Table 7

Comparative performance of LDA and BERTopic.

Model	Topic Number	Coherence
LDA	11	0.6228
BERTopic	25	0.7574

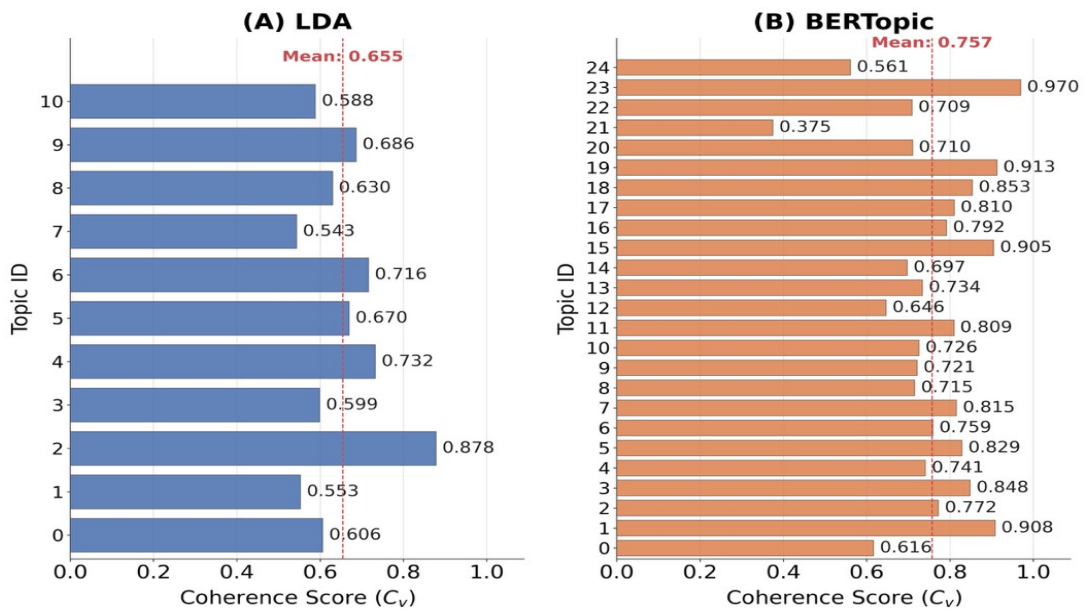


Figure 5: Distribution of coherence scores for topics generated by LDA and BERTopic models.

Figure 5 presents a comparative analysis of the topic coherence scores obtained by the LDA and BERTopic models. The average topic coherence value for the LDA model was calculated as 0.6546, while the average coherence value for the BERTopic model was obtained as 0.7574. Although BERTopic achieved a higher average coherence score and produced more topics than LDA, a qualitative inspection indicates that this does not necessarily reflect a “better”, domain-relevant clustering of the geoscience literature. A small number of the highest-ranked BERTopic topics (accounting for roughly 11% of the corpus) are biomedical, bibliometric, or education-related rather than substantively geoscientific, most likely reflecting noise introduced by the broad WoS search query (Section 4.1) rather than genuine thematic content. In contrast, nearly all LDA topics are directly related to core geoscience sub-fields. This contrast reflects a methodological trade-off:

LDA's document-level mixture averages small, idiosyncratic clusters into broader, more uniformly domain-aligned topics, whereas BERTopic's density-based clustering isolates any sufficiently dense cluster of documents regardless of its relevance to the target domain. The higher coherence and larger topic count obtained with BERTopic should therefore be read as a trade-off between thematic granularity and domain purity, rather than as straightforward evidence of a superior representation of the geoscience literature. This granularity also has a practical benefit for domain experts. Several broad LDA topics correspond to two or more distinct BERTopic topics that separate practically different research communities, such as reservoir engineering from fracture mechanics, or hydrology from renewable-energy research, which LDA's coarser topics merge together. For a researcher searching the literature for a specific subfield, BERTopic's finer resolution can therefore provide more directly navigable, practically distinct clusters than LDA's broader groupings, even though this same granularity is also what produces the off-target topics discussed above.

6. Conclusion

This study presents a comprehensive comparative analysis of LDA and BERTopic methods applied to a large-scale corpus of 11,732 geoscience publications retrieved from the WoS database. While LDA generated 11 topics with an average coherence score of 0.623, BERTopic identified 25 topics with a higher coherence score of 0.757. However, this gap should not be read as straightforward evidence that BERTopic provides a uniformly superior representation of the geoscience literature: part of its additional topics reflect genuinely finer-grained research communities that LDA's broader topics merge together, while another part reflects noise in the underlying WoS corpus. The results nonetheless indicate that BERTopic's integration of BERT embeddings, UMAP dimensionality reduction, and HDBSCAN clustering captures contextual semantic relationships that LDA's bag-of-words representation misses, offering a more granular view of interdisciplinary research at the intersection of geoscience, artificial intelligence, and machine learning. These findings suggest that LDA remains a theoretically robust, interpretable baseline that yields broadly domain-relevant topics with comparatively little manual scrutiny, whereas BERTopic offers higher-resolution topic structures at the cost of requiring closer qualitative inspection. The choice between the two methods should therefore be guided by the specific research objective and dataset characteristics rather than by coherence score alone: LDA's lower computational requirements may still be preferable for scenarios with limited resources, while BERTopic offers a more powerful option for researchers seeking fine-grained thematic structures in rapidly evolving interdisciplinary fields. Future research could explore dynamic topic modeling to analyze the temporal evolution of research themes, integrate citation network analysis with topic structures, and extend these methods to multilingual scientific corpora. This study contributes to the growing literature on automated scientific text analysis and offers practical guidance for selecting topic modeling approaches in bibliometric research.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Grammarly in order to: grammar and spelling check. Further, the author(s) used Claude Code in order to: assist with code writing. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] Bornmann, L. and Mutz, R., "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references", *Journal Of The Association For Information Science And Technology*, 66 (11): 2215–2222 (2015).

- [2] Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent dirichlet allocation", *Journal Of Machine Learning Research*, 3 (Jan): 993–1022 (2003).
- [3] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, F., "Deep learning and process understanding for data-driven Earth system science", *Nature*, 566 (7743): 195–204 (2019).
- [4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., "Indexing by latent semantic analysis", *Journal Of The American Society For Information Science*, 41 (6): 391–407 (1990).
- [5] Hofmann, T., "Probabilistic latent semantic analysis", *ArXiv Preprint ArXiv:1301.6705*, (2013).
- [6] Grootendorst, M., "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", *ArXiv Preprint ArXiv:2203.05794*, (2022).
- [7] Egger, R. and Yu, J., "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts", *Frontiers In Sociology*, 7: 886498 (2022).
- [8] Lee, D. and Seung, H. S., "Algorithms for non-negative matrix factorization", *Advances In Neural Information Processing Systems*, 13: (2000).
- [9] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M., "Hierarchical dirichlet processes", *Journal Of The American Statistical Association*, 101 (476): 1566–1581 (2006).
- [10] Blei, D. M. and Lafferty, J. D., "Dynamic topic models", *Proceedings of the 23rd International Conference on Machine Learning*, 113–120 (2006).
- [11] Blei, D., Lafferty, J., and others, "Correlated topic models", *Advances In Neural Information Processing Systems*, 18: 147 (2006).
- [12] Li, W. and McCallum, A., "Pachinko allocation: DAG-structured mixture models of topic correlations", *Proceedings of the 23rd International Conference on Machine Learning*, 577–584 (2006).
- [13] McAuliffe, J. and Blei, D., "Supervised topic models", *Advances In Neural Information Processing Systems*, 20: (2007).
- [14] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., "The author-topic model for authors and documents", *ArXiv Preprint ArXiv:1207.4169*, (2012).
- [15] Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M., "A practical algorithm for topic modeling with provable guarantees", (2013).
- [16] Shi, T., Kang, K., Choo, J., and Reddy, C. K., "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations", (2018).
- [17] Hu, Y., Boyd-Graber, J., Satinoff, B., and Smith, A., "Interactive topic modeling", *Machine Learning*, 95 (3): 423–469 (2014).
- [18] Miao, Y., Yu, L., and Blunsom, P., "Neural variational inference for text processing", (2016).
- [19] Srivastava, A. and Sutton, C., "Autoencoding variational inference for topic models", *ArXiv Preprint ArXiv:1703.01488*, (2017).

- [20] Angelov, D., "Top2vec: Distributed representations of topics", *ArXiv Preprint ArXiv:2008.09470*, (2020).
- [21] Pandur, M. B., Dobša, J., and Kronegger, L., "Topic modelling in social sciences-case study of web of science", (2020).
- [22] Marwan Tareq Shakir Al Jumaili and Murat Saran, "Identifying research trends in computer engineering and computer science master's programs in Türkiye using topic modeling techniques: LDA, TOP2VEC, AND BERTOPIC ", *Çankaya Üniversitesi*, Ankara, (2025).
- [23] Chagnon, E., Pandolfi, R., Donatelli, J., and Ushizima, D., "Benchmarking topic models on scientific articles using BERTeley", *Natural Language Processing Journal*, 6: 100044 (2024).
- [24] Wang, C. and Blei, D. M., "Collaborative topic modeling for recommending scientific articles", (2011).
- [25] Bai, H., Chen, Z., Lyu, M. R., King, I., and Xu, Z., "Neural relational topic models for scientific article analysis", (2018).
- [26] Guillén-Pacho, I., Badenes-Olmedo, C., and Corcho, O., "Dynamic topic modelling for exploring the scientific literature on coronavirus: an unsupervised labelling technique", *International Journal Of Data Science And Analytics*, 20 (3): 2551–2581 (2025).
- [27] Zelenkov, Y. and Solntsev, I., "Analysis of research dynamics in sport management using topic modelling", *Managing Sport And Leisure*, 30 (4): 527–548 (2025).
- [28] Sumpeno, S., Purwitasari, D., Farandy, B., Navastara, D. A., and Purnomo, M. H., "Graph Model and Deep Learning for Topic Labels in Classifying Short Texts of Scientific Article Titles.", *International Journal Of Intelligent Engineering & Systems*, 15 (2): (2022).
- [29] Atzeni, D., Bacciu, D., Mazzei, D., and Prencipe, G., "A systematic review of Wi-Fi and machine learning integration with topic modeling techniques", *Sensors*, 22 (13): 4925 (2022).
- [30] Stănescu, G. and Oprea, S.-V., "Recent Trends and Insights in Semantic Web and Ontology-Driven Knowledge Representation Across Disciplines Using Topic Modeling", *Electronics*, 14 (7): 1313 (2025).
- [31] Chen, H., Wang, X., Pan, S., and Xiong, F., "Identify topic relations in scientific literature using topic modeling", *IEEE Transactions On Engineering Management*, 68 (5): 1232–1244 (2019).
- [32] Griffiths, T. L. and Steyvers, M., "Finding scientific topics", *Proceedings Of The National Academy Of Sciences*, 101 (suppl_1): 5228–5235 (2004).
- [33] BERTopic: Neural Topic Modeling. <https://bertopic.com/> (2022). Accessed 4 January 2026.
- [34] McInnes, L., Healy, J., and Melville, J., "Umap: Uniform manifold approximation and projection for dimension reduction", *ArXiv Preprint ArXiv:1802.03426*, (2018).