International Journal on
Advanced Science
Engineering
Information Technology

Vol.14 (2024) No. 3
ISSN: 2088-5334

Topic Modeling for Scientific Articles: Exploring Optimal Hyperparameter Tuning in BERT

Maresha Caroline Wijanto [a,b,*], Ika Widiastuti [a,c], Hwan-Seung Yong [a]

[a] Computer Science & Engineering, Department of Artificial Intelligence and Software, Ewha Womans University, Seoul, Republic of Korea
[b] Faculty of Information Technology, Maranatha Christian University, Bandung, Indonesia
[c] Department of Information Technology, State Polytechnic of Jember, Jember, Indonesia
Corresponding author: *mareshacw@ewhain.net

Abstract—**Topic modeling has emerged as a successful approach to uncovering topics from textual data. Various topic modeling techniques have been introduced, ranging from traditional algorithms to those based on neural networks. In this research, we explore advanced topic modeling techniques, including BERT-based approaches, to enhance the analysis of scientific articles. We first investigate a widely used Latent Dirichlet Allocation (LDA) model and then explore the capabilities of BERT, to automatically uncover latent topics within scientific papers. The goal of this study is to identify the optimal hyperparameter setting for BERT-based topic modeling of scientific articles. We conduct experiments across several scenarios involving combinations of word embedding, dimension reduction, and clustering methods. The results were analyzed based on the coherence values, average execution time, number of topics generated, visualization through the inter-topic distance map, and the top-N-words of each topic. Our findings suggest that combination of RoBERTa for word embedding, PCA for dimension reduction, and K-Means for clustering yields superior results among the tested scenarios. Further evaluation of BERT-based topic modeling is necessary to validate these findings and explore its applications in various academic and industrial contexts. The implications of these advanced techniques could significantly streamline the process of staying updated with scientific literature, potentially revolutionizing research methodologies across disciplines.**

Keywords— **BERT-based; hyperparameter; scientific articles; topic modeling.**

Manuscript received 1 Aug. 2023; revised 4 Nov. 2023; accepted 23 Apr. 2024. Date of publication 30 Jun. 2024.

I. INTRODUCTION

Topic modeling has proven to be a successful approach in extracting meaningful information from vast text corpora. By analyzing a collection of documents, topic modeling aims to uncover the underlying subjects present in the corpus, without the need for explicit supervision [1]. This technique enables efficient processing of large datasets while preserving the essential statistical relationships required for various task such as classification, novelty detection, summarization, ad-hoc information retrieval, and analyzing historical documents. Moreover, topic modeling can compress extensive corpora into a brief summary by identifying and presenting the most frequently occurring topics as groups of linked terms.

In addition to its utility in processing large volumes of text, topic modeling also provides a comprehensible representation of documents, which finds applications in various Natural Language Processing (NLP) tasks [2]. The primary objective of topic modeling in NLP is to uncover topics, which are collections of words expressed as a mixture of closely related terms. Furthermore, each document is represented as a combination of one or more relevant themes. Topic modeling has proven valuable in understanding the diverse domains of science, particularly within the field of scientific publications. However, this task presents challenges due to the specificity and evolving nature of scientific documents over the past few decades [3].

Identifying the topics of scientific articles is instrumental in enabling researchers to track research trends and identify emerging areas of interest within their field [4], [5], [6], [7]. Moreover, it allows researchers to contextualize their own work within the broader landscape of their discipline and highlight how their work addresses critical questions or contributes to existing knowledge gaps [8], [9]. Building upon the work of others and integrating relevant findings into one's own research is a common practice for researchers [6], [10], [11]. Accurate topic identification in scientific articles facilitates efficient literature reviews, enabling researchers to swiftly locate relevant papers and stay up to date with the latest findings and developments, ultimately saving valuable time.

Topic models can be categorized into four types based on their underlying modeling techniques: algebraic, fuzzy, probabilistic, and neural [2]. The algebraic topic model, such as Latent Semantic Allocation (LSA), was developed in the 1990s [2], represents the corpus as a document term matrix (DTM). Zengul et al. [12]state that LSA and topic modeling are among the most commonly employed methods. LSA is a natural language processing approach that examines associations between text-based terms and documents, assuming that words with similar meanings will occur in similar contexts. On the other hand, Latent Dirichlet Allocation (LDA) is a probabilistic topic model that represents a document as a vector of probabilities [2]. Several studies, including one conducted by [4] that combined LDA and SciBERT, have used LDA to enhance classification quality. This study confirmed that adding topic modeling features can improve the quality of topical text classification in the scientific domain.

An empirical comparative study between LSA and LDA was conducted by [3] to investigate the impact of bi-gram collocation and lemmatization on both models. They found that LDA performs relatively better than LSA for topic numbers less than the optimal number reached (17 topics). Topic coherence was assessed in this study using both C_*v* and *Umass* metrics. The C_*v* performance of LSA decreases rapidly as the number of topics increases, while the C_*v* performance of LDA continues to rise until reaching a peak, after which it progressively declines. This study also discovers that lemmatization benefits C_*v* coherence when the number of topics is fewer than optimal. Additionally, comparisons between LSA and LDA have also been made in terms of divergence, throughput, quality, and response time [13]. According to this study, LDA shows considerably greater accuracy than LSA. However, LSA's computing time is significantly less than LDA's.

Furthermore, a research study [14] reports that LDA is an effective tool for extracting features from text by determining the latent topics present in the collection of documents. However, while LDA is a powerful tool for topic modeling and feature extraction in text data, it faces challenges in topic number optimization. Another study by [15] states that LDA method does not perform well if the number of topics ‘k’ is not adequately chosen. It struggles to identify topic correlations as well as topic evolution. Moreover, LDA exhibits inferior performance in inference compared to NMF (Non-negative Matrix Factorization) [16].

Recently, topic modeling approaches have seen a growing trend toward integrating neural components, leveraging contextualized representations instead of the traditional bag-of-words approach [17]. A prominent example is Bidirectional Encoder Representations from Transformers (BERT) is designed to pretrain deep bidirectional representations from unlabeled text. BERT’s ability to capture contextual semantic significantly improves the depth and accuracy of topic mining, overcoming limitations of traditional LDA which might ignore such context [18]. Through its attention mechanisms, BERT can automatically form topical word cluster similar to those generated by LDA [19]. One study [20] utilized the BERT encoder model to encode sentences from textual documents to obtain positional embeddings of topic word vectors. Additionally, BERT has a smaller and faster version called DistilBERT [21], which is trained to mimic the behavior of the larger BERT model while being more computationally efficient and requiring fewer resources.

Additionally, BERT and LDA have shown successful applications in clustering tasks [22]. This study employed a hybrid model, combining the probabilistic subject assignment vector from the LDA model with the sentence vectors derived from the BERT model. Research by [23] utilized a combination of LDA and BERT, where LDA identified the most frequently discussed topics in the dataset and BERT classified the sentiment present. Furthermore, BERT embeddings have been successfully used to explore the evolution of topics in scientific publications [24]. In this application, LDA was used to create topics. Then the LDA probability value for each word in a topic was multiplied by the averaged tensor similarity using monolingual or multilingual BERT embeddings. Another study [25] employed a hybrid approach, integrating BERT with an incremental community detection algorithm. In this instance, BERT established semantic relations between words in different contexts, while graph mining techniques, supported by simple structural rules, enhanced the resulting topics. Another hybrid model combining BERT and LDA in topic modeling with dimensionality reduction has been thoroughly investigated [26]. Clustering algorithms are computationally complex, and their difficulty increases with the number of features. Hence, dimensionality reduction methods such as PCA, t-SNE, and UMAP are used. This framework demonstrates that clustering with dimensionality reduction can lead to more coherent topics.

A Robustly Optimized BERT Pretraining approach (RoBERTa) is a variant of BERT that modifies the original BERT pretraining procedure to enhance end-task performance [27]. Developed by Facebook AI, RoBERTa utilizes a significantly larger corpus and more training data, leading to improved language representations and enhanced generalization capabilities. Conversely, DistilRoBERTa is a streamlined version of RoBERTa designed to reduce model size and computational resources while maintaining competitive performance.

This study aims to achieve the best hyperparameter tuning for topic modeling of scientific articles. The primary focus is to fine-tune the parameters of the topic modeling algorithm to ensure the most accurate representation of topics within scientific articles. The paper's organization is as follows: Section 2 presents the data and methodology, Section 3 discusses the results and findings, and Section 4 concludes the paper.

II. Materials and Method

A. *Dataset and Evaluation*

This research utilized dataset for Research Articles from Kaggle [28] containing 20,972 rows of data in English. Each row includes a title and abstract from a set of research articles. Additionally, there is a column describing the topic based on the actual information provided. The topics included in this dataset are Computer Science, Physics, Mathematics, Statistics, Quantitative Biology, and Quantitative Finance. A detailed count of data per category is shown in Figure 1.

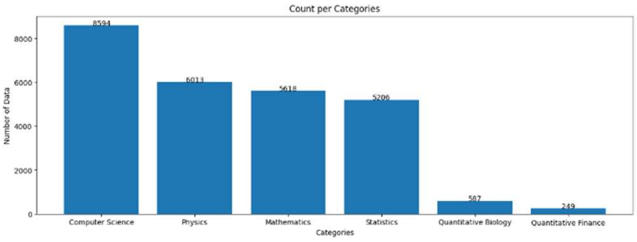

Fig. 1 Amount of data per category

Before applying any topic modeling techniques, the dataset underwent a standard preprocessing pipeline. This involved removing hyperlinks, special characters, numbers, and words of one character. We also converted the text to lowercase, as some of the methods we applied are case-sensitive. Stopwords were removed using a list from the NLTK (Natural Language Toolkit) Library.

Tokenization was performed on the dataset, followed by lemmatization using NLTK's library WordNet Lemmatizer, which ensures only base words are used. The final preprocessing step involves converting texts into TF-IDF (term-frequency inverse-document-frequency) weight for the machine learning model. For the deep learning model, the preprocessing step stopped after lemmatization.

Topic coherence is an evaluation metric used to measure the quality and interpretability of topics generated by topic modeling algorithms. It aims to assess how semantically meaningful and coherent the identified topics are. Higher coherence scores indicate that the topics are more coherent and representative topics. Two commonly used methods for calculating topic coherence are C_v and UMass, widely used in topic modeling research to evaluate and compare different algorithms and configurations.

The C_v coherence measures the coherence of topics by evaluating the pairwise word similarity among the most probable words in each topic. It calculates coherence using a pre-defined word embedding model that captures semantic relationships between words [29]. C_v coherence correlates well with human judgment regarding topic quality and interpretability. A higher C_v score indicates that the topics are more coherent and linguistically meaningful.

The UMass coherence, also known as the u_mass metric, is an alternative method for evaluating topic coherence [30]. Unlike C_v, UMass does not rely on pre-trained word embeddings but uses a document-based approach. This method provides a fast and efficient way to measure topic coherence without requiring external word embeddings. However, it may be less correlated with human judgment than C_v coherence.

B. *Proposed Method*

This section outlines the methodology employed for conducting the topic modeling experiments. These experiments involve the application of both traditional LDA and BERT-based models, including BERTopic, RoBERTa, and DistilRoBERTa. The leading architecture for this experiment is depicted in Figure 2.

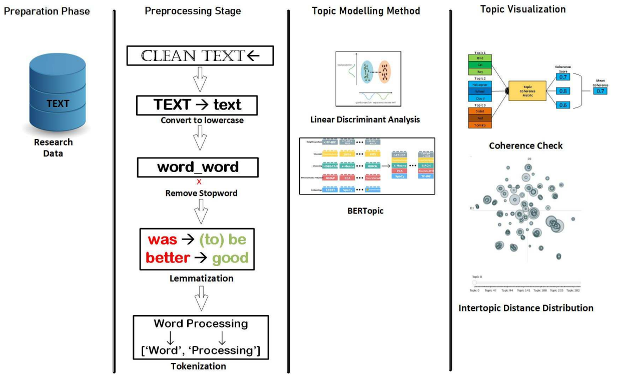

Fig. 2 Main architecture

In this study, we present our methodology for topic modeling, which encompasses data collection, preprocessing, and applying various topic modeling techniques. As mentioned in the previous section, the dataset used in this research was obtained from Kaggle. Before proceeding with topic modeling, we performed a series of preprocessing steps to ensure the consistency and quality of the data.

To investigate the performance of various topic modeling techniques, we employed two distinct approaches: Latent Dirichlet Allocation (LDA) and BERT-based. We implemented LDA using the Gensim library in Python and conducted a systematic grid search to identify the optimal number of topics (K) for achieving coherent and interpretable results. In contrast, we used the BERTopic library and experimented with various settings to optimize its performance. We also used some hyperparameter tuning settings for this BERTopic. We modify the hyperparameter of word embedding, dimension reduction, and clustering methods for the BERTopic.

We employed the inter-topic distance distribution method to visualize the generated topics, which provides an intuitive representation of the relationships between different topics. By visualizing these distance distributions, we can gain insights into the cohesion and separation among the identified topics, thereby enhancing the interpretability of the results.

Furthermore, we employed both the C_v and UMass coherence metrics to assess the quality of the generated topics. These measures provide valuable insights into the semantic coherence of the issues and their ability to represent distinct concepts within the dataset.

Overall, our proposed method integrates data collection, preprocessing, topic modeling using both LDA and BERTopic, visualization through inter-topic distance distribution, and topic coherence evaluation to analyze the topic modeling process comprehensively.

III. Results and Discussion

A. *Experiment Scenario*

The abundance of scientific articles across diverse fields necessitates efficient and interpretable methods for topic modeling. In this research, in addition to implementing the state-of-the-art topic modeling technique, LDA, we also aim to uncover the most effective hyperparameter settings for topic-modeling scientific articles by leveraging the powerful BERTopic model. Our approach's novelty lies in combining various word embedding, dimension reduction, and clustering methods, enabling us to unlock nuanced and coherent topic representations.

We have integrated various word embedding methods within the BERTopic model to leverage the rich semantic information embedded in scientific articles. Our options include Default Sentence-BERT (S-Bert), RoBERTa, DistilRoBERTa, Gensim FastText, and paraphrase-MiniLM-L3-v2. Each word embedding method offers unique strengths in capturing context and meaning from textual data. We systematically explore the impact of these embeddings on topic modeling performance, aiming to identify the optimal choice for effectively distilling knowledge from scientific research.

As scientific articles often comprise large corpora, dimension reduction methods play a pivotal role in enhancing both the scalability and interpretability of topic modeling. We consider two prominent techniques: Uniform Manifold Approximation and Projection (UMAP), and Principal Component Analysis (PCA). UMAP excels in preserving local structure, making it ideal for maintaining nuanced relationships within the data. In contrast, PCA offers an efficient approach to reducing dimensions, simplifying large dataset while retaining significant variance.

Through extensive experimentation, we investigate the impact of these dimension reduction strategies on topic modeling results, seeking the best approach for balancing computational efficiency and topic coherence. To unveil the underlying structures within scientific articles and facilitate topic segmentation, we deploy two clustering methods: K-Means, and HDBScan. K-Means, a classic partitioning algorithm, seeks to divide data into K clusters. In contrast, HDBScan employs density-based clustering to identify clusters of varying shapes and sizes. By applying these clustering techniques to our BERTopic model, we aim to discover the optimal approach for grouping scientific articles into coherent topics that genuinely reflect their inherent thematic content.

Table 1 presents the details of our experiment scenario. Our research involves an extensive experimental setup, systematically exploring the vast hyperparameter space defined by word embedding methods, dimension reduction techniques, and clustering algorithms across 18 scenarios. We assess the quality and coherence of the generated topics for each configuration through rigorous evaluation. The goal is to identify the hyperparameter settings that yield the most interpretable and semantically meaningful topics, thereby equipping researchers with a robust and effective toolkit for exploring the vast landscape of scientific articles.

TABLE I
HYPERPARAMETER TUNING

Word Embedding	Dimension Reduction	Clustering Method
- Default Sentence-BERT - RoBERTa - DistilRoBERTa - Gensim FastText	- UMAP - PCA	- HDBScan - K-Means

By combining state-of-the-art language representations, dimension reduction strategies, and clustering methodologies, our research strives to advance the field of topic modeling for scientific articles. The knowledge gained from this study has the potential to enhance our understanding of and ability to navigate the wealth of scholarly literature significantly. This facilitates knowledge discovery and accelerates scientific progress.

B. Experimental Results

When applying a topic modeling technique, such as Latent Dirichlet Allocation (LDA), setting the number of topics (K) is a crucial hyperparameter that must be determined before model training. For each value of K, the coherence score is calculated to evaluate the coherent and semantic meaningfulness of the generated topics. The score serves as a guide to help researchers select the optimal number of topics. A higher coherence score for a particular value of K indicates that the topics are more coherent and represent distinct concepts within the dataset. We evaluate the coherence score for a total of 12 topics as shown in Figure 3.

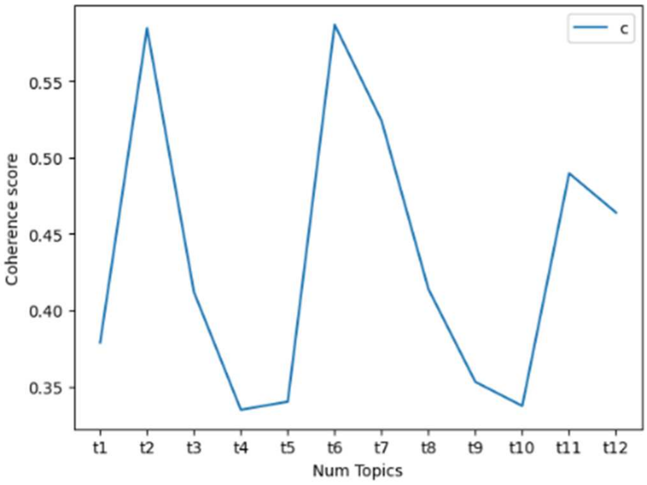

Fig. 3 Topic coherence value in LDA

The experiments indicate that the highest coherence score is achieved with six topics, recording a score of 0.58701, while the second highest is two topics, scoring 0.58471; the difference is a mere 0.0023. We will set six as the default number of topics for subsequent experiments based on these results.

To visualize our LDA model, we employ t-SNE plot, as recommended by Genender-Feltheimer for exploratory data analysis [31]. T-distributed Stochastic Neighborhood Embedding (tSNE) is an unsupervised Machine Learning algorithm introduced by Maaten and Hinton [32] to visualize high-dimensional data in a low-dimensional space [33]. Figure 4 displays the t-SNE plot of our LDA model configured with six topics.

Figure 4 uses six different colors to represent the six generated topic models. The top words and their corresponding subjects for each topic are as follows:

- Topic 1 (Statistics): paper function group case proof complexity problem results in series model.
- Topic 2 (Physics): mass model star formation emission gas galaxy density evolution field.
- Topic 3 (Computer Science): network model paper performance approach information learning method work analysis.

- Topic 4 (Physics): paper group space problem number case graph results in theory function.
- Topic 5 (Mathematics): model energy field phase state temperature theory transition spin quantum.
- Topic 6 (Computer science): model problem method algorithm paper time approach number network analysis.

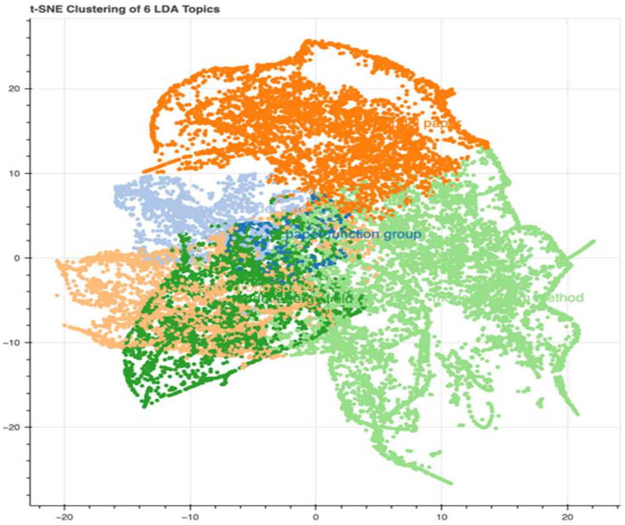

Fig. 4 t-SNE clustering of LDA

We attempt to map the generated topic by category from the initial dataset. In our analysis, these topics fit into four categories, despite the dataset initially having six. This reduction is consistent with the distribution of articles in the original categories, where the other two categories had very few articles. As mentioned in the previous section, we conducted several experiments using Google Collab with GPU settings enabled. Each experiment was run five times across all scenarios, with the number of topics set to six. Table II displays the average execution time results.

TABLE II
AVERAGE EXECUTION TIME

Hyperparameter Tuning			Time-A	Time-T
Clustering	Word Embedding	Dimension Reduction		
HDBScan	S-BERT	UMAP	**177.89**	**197.52**
		PCA	**61.77**	**25.79**
	RoBERTa	UMAP	406.77	395.74
		PCA	377.05	327.53
	DistilRoBERTa	UMAP	4273.88	1837.05
		PCA	4292.33	1784.70
	FastText	UMAP	**35.56**	**27.41**
		PCA	**11.52**	**8.12**
	-	-	173.09	167.93
K-Means	S-BERT	UMAP	**97.96**	**30.67**
		PCA	**47.71**	**12.23**
	RoBERTa	UMAP	404.44	489.12
		PCA	513.99	343.36
	DistilRoBERTa	UMAP	4747.02	2131.22
		PCA	4774.97	2035.53
	FastText	UMAP	**34.14**	**27.59**
		PCA	**11.79**	**3.66**
	-	-	64.37	29.31

We compared the execution times when integrating data from abstracts or titles with various hyperparameter scenarios. Time-A indicates the execution time for abstract data, while Time-T represents the execution time for title data. The results suggest that, generally, the execution time for abstract data is longer than for title data due to the more incredible amount of text in abstracts. The shortest execution times were observed when using FastText followed by S-BERT. It can also be concluded that, in general, UMAP requires a longer processing time than PCA.

The variation in execution time can be attributed to several factors. Gensim FastText is a lightweight and relatively simple word embedding model compared to the transformer-based models like S-BERT, RoBERTa and DistilRoBERTa. Transformer models are deep neural networks with multiple layers and many parameters, making them computationally more intensive during training and inference. In contrast, FastText uses a shallow neural network and character-level n-grams, resulting in a smaller model size and faster execution. Transformer-based models incorporate complex self-attention mechanisms that require more computations. Conversely, FastText employs a straightforward averaging mechanism, which is computationally less demanding, which contributes to its quicker execution time.

This research focuses on topic modeling, and we also analyze the number of topics that are generated from the scenarios. Specifically, we concentrate on scenario using HDBScan as the clustering method. We do not consider K-Means because the number of topics formed is pre-determined at six, as we manually set the number of K cluster. The results can be displayed on Table III. Similarly to the time measurement experiments, these experiments also compare the number of topics generated using abstract data (# Topics-A) and title data (# Topics-T) as the inputs.

TABLE III
NUMBER OF TOPIC

Hyperparameter Tuning			# Topics-A	# Topics-T
Clustering	Word Embedding	Dimension Reduction		
HDBScan	S-BERT	UMAP	290	344
		PCA	30	3
	RoBERTa	UMAP	3	201
		PCA	4	7
	DistilRoBERTa	UMAP	3	3
		PCA	3	4
	FastText	UMAP	171	290
		PCA	3	5
	-	-	296	347

UMAP is renowned for its ability to preserve both local and global data structures, making it particularly effective at capturing complex and nonlinear relationships. It strives to maintain the relative distances between data points, ensuring that similar points are clustered closely in the reduced space. Consequently, UMAP may produce more fine-grained topic representations and reveal more subtle differences. This capability might generate more topics as BERTopic attempts to encapsulate the increased diversity and nuance in the topic space.

On the other hand, PCA is a linear dimension reduction technique that focuses on capturing the most considerable variance within the data. Although it is efficient at reducing dimensionality and can be computationally quicker, PCA may not preserve complex relationships and subtle differences between data points as effectively as UMAP. Therefore, PCA may yield more compact topic representations, which could result in a smaller number of generated topics in BERTopic.

This observation aligns with the results of experiments conducted. Generally, using UMAP as a dimension reduction method results in a highly diverse number of generated topics. In this research, such variability in topic numbers does not aid in effectively identifying scientific articles.

Subsequently, we measured the coherence values for all scenarios, as shown in Table IV. We used C_v and u_mass coherence metrics. These values represent the average outcomes of the five experiments performed and indicate the coherence values for abstract data only.

TABLE IV
EXPERIMENTS COHERENCE VALUE

Hyperparameter Tuning			**C_V**	**U_MASS**
Clustering	**Word Embedding**	**Dimension Reduction**		
HDBScan	S-BERT	UMAP	**0.4857**	-9.1402
		PCA	0.4272	-5.8565
	RoBERTa	UMAP	0.3976	**-4.9033**
		PCA	0.4251	-6.9820
	DistilRoBERTa	UMAP	0.3694	-6.7159
		PCA	0.3949	-6.9514
	FastText	UMAP	0.4419	-9.9408
		PCA	0.4409	-5.9161
	-	-	0.4721	-9.2919
K-Means	S-BERT	UMAP	0.5305	-2.4883
		PCA	**0.5412**	-2.4856
	RoBERTa	UMAP	**0.5531**	-2.0048
		PCA	**0.5554**	**-1.8291**
	DistilRoBERTa	UMAP	0.5003	**-1.8192**
		PCA	0.4950	**-1.7787**
	FastText	UMAP	0.5351	-1.8363
		PCA	0.4782	-2.3326
	-	-	0.5283	-2.4852
Average		UMAP	0.4769	-4.8341
		PCA	0.4695	-4.2885

We achieved the best coherence values for both metrics when using HDBScan as a clustering method and UMAP for dimension reduction. However, PCA generally outperforms UMAP in terms of coherence. This pattern was also observed when using K-Means as a clustering method. Moreover, both clustering methods showed that S-BERT and RoBERTa yield better coherence values.

Overall, we conclude that the best coherence values are obtained using K-Means as a clustering method, RoBERTa as a word embedding technique, and PCA as a dimension reduction method. These findings are consistent with the results of previous experiments and are quite close to the coherence value of the LDA experiments.

Figure 5 displays the inter-topic distance map for our data using RoBERTa-PCA-K-Means hyperparameter settings. The prevalent size of the circles indicates that the volume of data on each generated topic is relatively uniform. Additionally, the absence of overlapping topic circles suggests no overlapping topics. An optimal topic model is characterized by large, non-overlapping bubbles that are evenly distributed across the chart.

From this result, the top words appearing in each topic are as follows:

- Topic 0 (Computer Science): model method algorithm network data machine learning (5,425 data entries)
- Topic 1 (Computer Science): model data network method system approach analysis statistics (4,988 data entries)
- Topic 2 (Statistics): paper group prove space result function boundary (3,097 data entries)
- Topic 3 (Mathematics): problem function result equation shows fully order (2,962 data entries)
- Topic 4 (Physics): state magnetic phase system field energy influence (2,812 data entries)
- Topic 5 (Physics): galaxy star mass model planet observations cosmological (1,688 data entries)

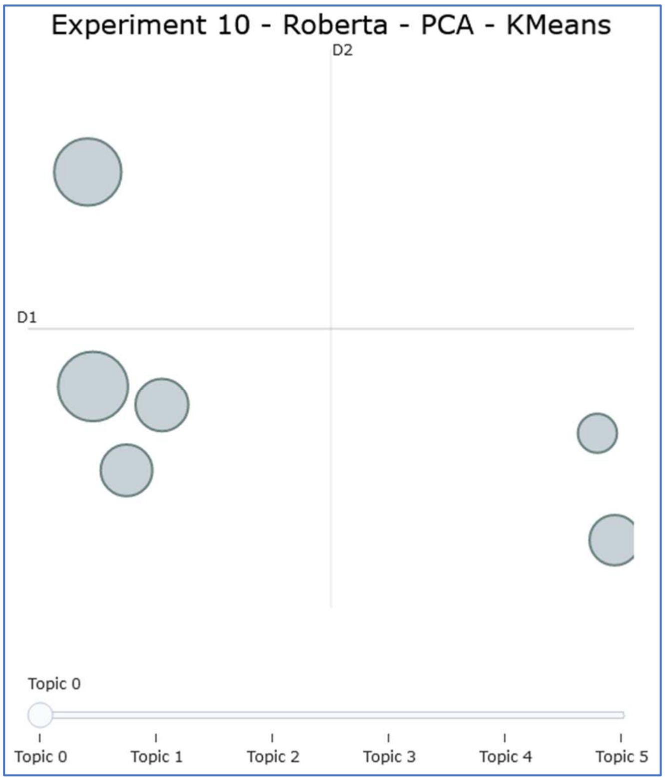

Fig. 5 Inter-topic distance map

Similar to the LDA results, we have also mapped the generated topics to match the original categories from the dataset. Our mapping aligns with the hierarchical relationship results, as shown in Figure 6. From this figure, it is evident that there are only three major topics. Notably, statistics and mathematics are in the same cluster.

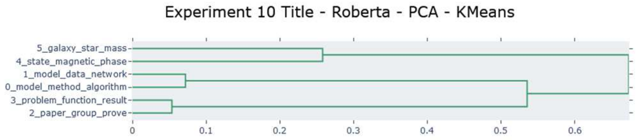

Fig. 6 Hierarchical clustering

We then focused our BERTopic implementation exclusively on computer science data. Figure 7 displays the inter-topic distance map for this subset of data. Similar to the result from the entire dataset, the circles representing each topic are relatively evenly sized and do not overlap. For topics 0 through 5, the respective number of articles in a generated topic are 2211, 1662, 1367, 1261, 1126, and 967.

Upon closer examination, each generated topic corresponds to a specific learning category within Computer Science. Table V displays the top N words for each topic alongside our estimated categories.

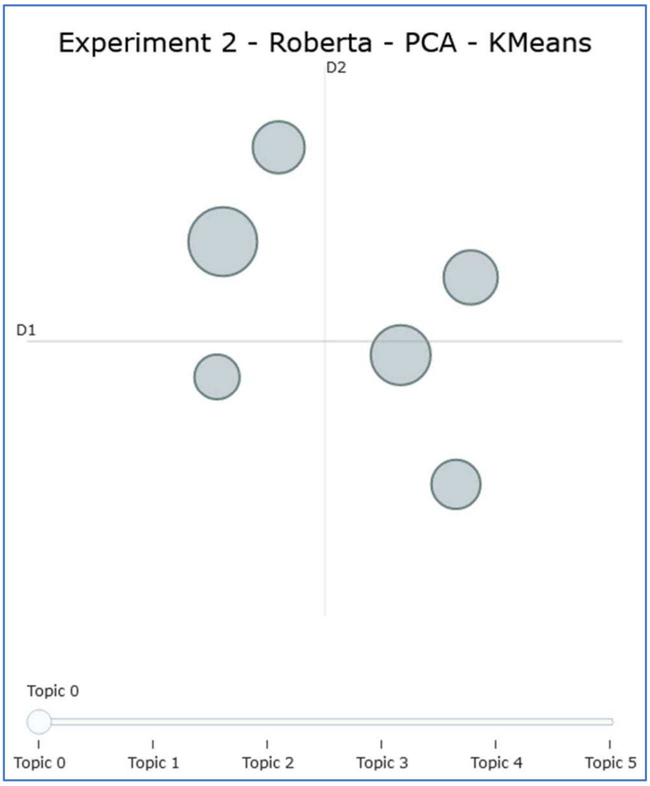

Fig. 7 Inter-topic distance map for Computer Science Articles

TABLE V
TOP-N-WORDS FOR COMPUTER SCIENCE DATA

Topic	Top-N-Words	Estimated Categories
0	neural network method algorithm problem text	Data Science
1	model deep learning data approach method setting	Machine Learning
2	mobile network performance time data traffic internet	Computer Networks
3	algorithm graph bound function number variable	Statistics
4	social network model information research science	Data Analytics
5	graph problem analysis study complex result signal	Hardware

Given that the fields within computer science often overlap, the top-N-word results listed in Table V are logical and coherent. Despite the diversity of existing documents, we can still construct a coherent topic model.

IV. CONCLUSION

This paper proposes a BERT-based topic modeling approach for scientific articles. Our experiments primarily focused on exploring the hyperparameter tuning for these models while also assessing traditional methods. We found that combining RoBERTa for word embedding, PCA for dimension reduction, and K-Means for clustering yielded the best results across all experiments based on the inter-topic distance map, coherence values and execution time. We also conducted a comparison specifically for computer science articles, and the results demonstrated a consistent trend with those from the broader scientific corpus. Thus, BERT-based models show promise as effective methods for topic modeling. Additionally, we discovered that more evaluation metrics are needed for this problem. Unlike traditional LDA methods, evaluation of BERT-based topic modeling results is still limited, so further exploration is needed.

REFERENCES

[1] W. Chen, F. Rabhi, W. Liao, and I. Al-Qudah, “Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study,” *Electronics (Basel)*, vol. 12, no. 12, p. 2605, Jun. 2023, doi: 10.3390/electronics12122605.

[2] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, “Topic modeling algorithms and applications: A survey,” *Information Systems*, vol. 112. Elsevier Ltd, Feb. 01, 2023. doi:10.1016/j.is.2022.102131.

[3] S. Bellaouar, M. M. Bellaouar, and I. E. Ghada, “Topic modeling: Comparison of LSA and LDA on scientific publications,” *ACM International Conference Proceeding Series*, pp. 59–64, Feb. 2021, doi: 10.1145/3456146.3456156.

[4] A. Glazkova, “Identifying Topics of Scientific Articles with BERT-Based Approaches and Topic Modeling,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12705 LNAI, pp. 98–105, 2021, doi: 10.1007/978-3-030-75015-2_10.

[5] R. K. Gupta, R. Agarwalla, B. H. Naik, J. R. Evuri, A. Thapa, and T. D. Singh, “Prediction of research trends using LDA based topic modeling,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 298–304, Jun. 2022, doi: 10.1016/J.GLTP.2022.03.015.

[6] S. Kavvadias, G. Drosatos, and E. Kaldoudi, “Supporting topic modeling and trends analysis in biomedical literature,” *Journal of Biomedical Informatics*, vol. 110. Academic Press Inc., Oct. 01, 2020. doi: 10.1016/j.jbi.2020.103574.

[7] A. H. Suyanto, T. Djatna, and S. H. Wijaya, “Mapping and predicting research trends in international journal publications using graph and topic modeling,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 2, pp. 1201–1213, May 2023, doi:10.11591/ijeecs.v30.i2.pp1201-1213.

[8] E. H. J. Kim, Y. K. Jeong, Y. H. Kim, and M. Song, “Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction,” *J Informetr*, vol. 16, no. 1, p. 101242, Feb. 2022, doi:10.1016/j.joi.2021.101242.

[9] P. Kathiria and H. Arolkar, “Trend analysis and forecasting of publication activities by Indian computer science researchers during the period of 2010–23,” *Expert Syst*, vol. 39, no. 10, p. e13070, Dec. 2022, doi: 10.1111/exsy.13070.

[10] X. Chen *et al.*, “Exploring science-technology linkages: A deep learning-empowered solution,” *Inf Process Manag*, vol. 60, no. 2, p. 103255, Mar. 2023, doi: 10.1016/j.ipm.2022.103255.

[11] T. Silwattananusarn and P. Kulkanjanapiban, “A text mining and topic modeling based bibliometric exploration of information science research,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 3, pp. 1057–1065, Sep. 2022, doi:10.11591/ijai.v11.i3.pp1057-1065.

[12] F. Zengul *et al.*, “A Practical and Empirical Comparison of Three Topic Modeling Methods Using a COVID-19 Corpus: LSA, LDA, and Top2Vec,” Jan. 2023, doi: 10.24251/hicss.2023.116.

[13] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne, “Effective Comparison of LDA with LSA for Topic Modelling,” *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, pp. 1245–1250, May 2020, doi:10.1109/iciccs48265.2020.9120888.

[14] M. Hasan, A. Rahman, M. R. Karim, M. S. I. Khan, and M. J. Islam, “Normalized approach to find optimal number of topics in latent dirichlet allocation (lda),” *Advances in Intelligent Systems and Computing*, vol. 1309, pp. 341–354, 2021, doi: 10.1007/978-981-33-4673-4_27/figures/7.

[15] A. Goyal and I. Kashyap, "Latent Dirichlet Allocation - An approach for topic discovery," *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing*, COM-IT-CON 2022, pp. 97–102, 2022, doi: 10.1109/com-it-con54601.2022.9850912.

[16] N. N. Hidayati, S. Rochimah, and A. B. Rahardjo, "Software Traceability in Agile Development Using Topic Modeling," *Int J Adv Sci Eng Inf Technol*, vol. 12, no. 4, pp. 1410–1420, 2022, doi:10.18517/ijaseit.12.4.15195.

[17] Z. Fang, Y. He, and R. Procter, "BERTTM: Leveraging Contextualized Word Embeddings from Pre-trained Language Models for Neural Topic Modeling," May 2023, doi:10.48550/arXiv.2305.09329.

[18] G. Tang, X. Chen, N. Li, and J. Cui, "Research on the Evolution of Journal Topic Mining Based on the BERT-LDA Model," *SHS Web of Conferences*, vol. 152, p. 03012, 2023, doi: 10.1051/shsconf/202315203012.

[19] M. Talebpour, A. García Seco de Herrera, and S. Jameel, "Topics in Contextualised Attention Embeddings," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13981 LNCS, pp. 221–238, 2023, doi: 10.1007/978-3-031-28238-6_15/figures/2.

[20] H. Gupta and M. Patel, "Method of Text Summarization Using Lsa and Sentence Based Topic Modelling with Bert," *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, pp. 511–517, Mar. 2021, doi:10.1109/ICAIS50930.2021.9395976.

[21] R. Silva Barbon and A. T. Akabane, "Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study," *Sensors*, vol. 22, no. 21, Nov. 2022, doi:10.3390/s22218184.

[22] E. Atagün, B. Hartoka, and A. Albayrak, "Topic Modeling Using LDA and BERT Techniques: Teknofest Example," *Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021*, pp. 660–664, 2021, doi:10.1109/ubmk52708.2021.9558988.

[23] S. E. Uthirapathy and D. Sandanam, "Topic Modelling and Opinion Analysis on Climate Change Twitter Data Using LDA And BERT Model.," *Procedia Comput Sci*, vol. 218, pp. 908–917, Jan. 2023, doi:10.1016/j.procs.2023.01.071.

[24] Q. Xie, X. Zhang, Y. Ding, and M. Song, "Monolingual and multilingual topic analysis using LDA and BERT embeddings," *J Informetr*, vol. 14, no. 3, Aug. 2020, doi: 10.1016/j.joi.2020.101055.

[25] M. Asgari-Chenaghlu, M. R. Feizi-Derakhshi, L. farzinvash, M. A. Balafar, and C. Motamed, "TopicBERT: A cognitive approach for topic detection from multimodal post stream using BERT and memory–graph," *Chaos Solitons Fractals*, vol. 151, p. 111274, Oct. 2021, doi: 10.1016/j.chaos.2021.111274.

[26] L. George and · P Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *International Journal of Information Technology*, vol. 15, no. 4, pp. 2187–2195, 2023, doi:10.1007/s41870-023-01268-w.

[27] Y. Sun, D. Gao, X. Shen, M. Li, J. Nan, and W. Zhang, "Multi-Label Classification in Patient-Doctor Dialogues With the RoBERTa-WWM-ext + CNN (Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach With Whole Word Masking Extended Combining a Convolutional Neural Network) Model: Named Entity Study," *JMIR Med Inform*, vol. 10, no. 4, Apr. 2022, doi: 10.2196/35606.

[28] B. Densil, "Topic Modeling for Research Articles." Accessed: Feb. 19, 2023. [Online]. Available: https://www.kaggle.com/datasets/blessondensil294/topic-modeling-for-research-articles

[29] D. Bretsko, A. Belyi, and S. Sobolevsky, "Comparative Analysis of Community Detection and Transformer-Based Approaches for Topic Clustering of Scientific Papers," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13956 LNCS, pp. 648–660, 2023, doi: 10.1007/978-3-031-36805-9_42/figures/6.

[30] A. F. Pathan and C. Prakash, "Unsupervised Aspect Extraction Algorithm for opinion mining using topic modeling," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 492–499, Nov. 2021, doi:10.1016/j.gltp.2021.08.005.

[31] C. Flexa, W. Gomes, I. Moreira, R. Alves, and C. Sales, "Polygonal Coordinate System: Visualizing high-dimensional data using geometric DR, and a deterministic version of t-SNE," *Expert Syst Appl*, vol. 175, Aug. 2021, doi: 10.1016/j.eswa.2021.114741.

[32] W. Zhu, Z. Webb, X. Han, K. Mao, W. Sun, and J. Romagnoli, "Generic Process Visualization Using Parametric t-SNE," Elsevier B.V., Jan. 2018, pp. 803–808. doi: 10.1016/j.ifacol.2018.09.262.

[33] T. T. Cai and R. Ma, "Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data," *Journal of Machine Learning Research*, vol. 23, pp. 1–54, May 2022, doi:10.48550/arXiv.2105.07536.