A Comparative Analysis of Transformer-Based Topic Modeling Pipelines for Scientific Literature

Farriel Arrianta Akbar Pratama[1], Muhammad Eka Nur Arief[2], Vinna Rahmayanti Setyaning Nastiti[3*]
[1,2,3] *Informatics Engineering, Universitas Muhammadiyah Malang, Indonesia*
*corr-author: vinastiti@umm.ac.id

Abstract - The exponential growth of scientific literature poses a significant challenge for manually identifying thematic trends, thereby necessitating automated analysis methods. This study aims to determine an optimal topic modeling pipeline by conducting a comparative analysis to maximize the coherence of topics extracted from scientific research. Three distinct pipelines were implemented and evaluated on a corpus of 20,972 scientific article abstracts: (i) a custom pipeline combining SBERT, UMAP, and HDBSCAN; (ii) a configuration using RoBERTa, PCA, and k-means; and (iii) the integrated BERTopic model. Performance evaluation, quantitatively benchmarked using the C_v coherence score, revealed that the integrated BERTopic model achieved the highest score of 0.7012. This result significantly surpassed the custom SBERT-based pipeline and the RoBERTa-based pipeline, which scored 0.6079 and 0.4756, respectively. The findings demonstrate that an integrated, purpose-built model like BERTopic is superior for generating highly coherent and interpretable thematic structures from scientific text. This research provides empirical guidance for researchers and shows how integrated models offer a more robust solution for large-scale literature analysis compared to modular pipeline designs.

Keywords: BERTopic; Natural Language Processing; topic modeling; transformer models.

I. INTRODUCTION

Research plays a crucial role in advancing human knowledge, it enables discoveries that drive innovation and development. However, the exponential growth of research publications has made it increasingly difficult to identify emerging themes and trends, creating challenges in extracting insights and prioritizing areas of focus [1-3]. This has created a strong demand for quantitative and predictive approaches, where natural language processing (NLP) and text mining provide essential tools for automated knowledge discovery and thematic analysis [4]. Topic modeling, in particular, has emerged as a powerful unsupervised method for uncovering latent structures in specialized collections ranging from industrial accident reports [5] to broader corpora [6,7], with recent advancements highlighting its growing impact in analyzing trending social media narratives, particularly within the X (formerly Twitter) ecosystem [8].

The landscape of topic modeling has evolved considerably over time. Early methods such as latent semantic analysis (LSA) and latent dirichlet allocation (LDA) primarily relied on bag-of-words representations to capture themes in text [9]. Comparative studies on unstructured text, ranging from customer service transcripts [10] to academic journal corpora [11], have highlighted the coherence strengths of probabilistic models. However, these approaches ignore contextual meaning and often fail to capture nuanced semantic relationships, prompting a shift toward neural topic models that leverage deep learning architectures [12]. The advent of deep learning and short-text modeling techniques [13], especially transformer-based models that include BERT and RoBERTa [14,15], marked a paradigm shift by introducing contextual embeddings that represent words based on surrounding text, enabling richer semantic relevance and coherence even in specific domains such as educational feedback [16].

Building on these advances, integrated models such as BERTopic combine transformer embeddings with clustering algorithms and class-based TF-IDF (c-TF-IDF), achieving superior performance in diverse applications [17]. Studies have shown that BERTopic and NMF outperform LDA and LSI in urgent MOOC forum posts [18] and are effective in bibliometric analyses of research trends [19]. However, scientific literature poses unique challenges: structural words such as “introduction” and “conclusion” often dominate topic extraction and obscure thematic content [20]. This issue has led to specialized tools such as BERTeley, while [21] provided valuable benchmarks with tuned RoBERTa models. However, their work focused on a single architectural pattern, and a broader comparison is necessary to identify best practices for scientific corpora. This study addresses this gap by extending their findings through a comprehensive comparative analysis of distinct pipeline philosophies.

The key novelty of this research lies in its direct, controlled comparison of an integrated framework (BERTopic) against both a modular, custom-built pipeline (SBERT-UMAP-HDBSCAN) and a replicated benchmark (RoBERTa-PCA-k-means). By strictly controlling the architectural variables, this study isolates the performance impact of integrated versus modular design philosophies. Specifically, it investigates transformer-based embeddings (SBERT, RoBERTa), dimensionality reduction methods (UMAP, PCA), and clustering algorithms (HDBSCAN, k-means), drawing on recent comparative frameworks that validate these components for short-text analysis [22,23], to determine an optimal configuration for maximizing topic coherence (C_v) in scientific literature. We also briefly consider the potential application of large language models (LLMs), whose remarkable pre-training and prompt-based capabilities [22,24], suggest a new paradigm in topic discovery.

In summary, this research contributes by benchmarking advanced pipelines for scientific article topic modeling, establishing practical insights into their strengths and limitations, and investigating the potential of LLMs for enhancing topic discovery. To ensure comparability, the same Kaggle dataset of 20,972 English-language scientific articles used in [21], spanning computer science, physics, mathematics, statistics, quantitative biology, and quantitative finance [25], is employed. The findings are expected to inform best practices in applying NLP techniques to scientific corpora. The remainder of this paper is structured into five sections. Section II reviews related work, Section III describes the methodology and experimental setup, Section IV presents results and discussion, and Section V concludes with key insights and future directions.

II. METHOD

This research aims to model topics from a collection of scientific research paper abstracts by comparing several advanced topic modeling pipelines. The primary objective is to achieve high coherence scores (C_v) and analyze the resulting topics. The methodology (Fig. 1) encompasses dataset preparation, pre-processing, implementation of three pipelines, and coherence evaluation.

A. Dataset Description

This study utilized a publicly available dataset of research articles previously employed in [21], sourced from Kaggle [25]. The dataset comprises 20,972 scientific articles, primarily from disciplines such as computer science, physics, and mathematics, along with statistics, quantitative biology, and quantitative finance. Each article entry contains a TITLE filed and an ABSTRACT field. For the purpose of this research, these two fields were concatenated to form a single text column for each document, providing a richer textual basis for topic discovery. Potential missing values in TITLE or ABSTRACT were handled by imputing empty strings to ensure completeness before concatenation. It should be noted that the dataset is primarily composed of articles from computer science, physics, and mathematics, which may introduce a domain bias. This concentration could influence the nature and diversity of the extracted topics, and the model's performance may vary on corpora with different disciplinary distributions.

B. Data Pre-Processing

Effective topic modeling relies heavily on clean, processed text. The pre-processing pipeline applied in this study aims to standardize the text and remove noise, thereby improving the quality of subsequent topic extraction. The pre-processing steps, primarily utilizing the natural language toolkit (NLTK) library, include lowercase conversion, removal of hyperlinks and emails, removal of special characters and numbers, tokenization, removal of stopwords, and lemmatization. The output of this stage is a list of processed tokens for each document. Additionally, a processed_text_for_bert column was created by joining these tokens back into a string, which serves as input for sentence-embedding models. While boilerplate academic phrases such as ‘this paper proposes’ were not explicitly removed via a custom list, the stopword removal step, combined with the c-TF-IDF mechanism used later in the pipelines, effectively down-weights such common, non-descriptive terms, minimizing their impact on topic formation.

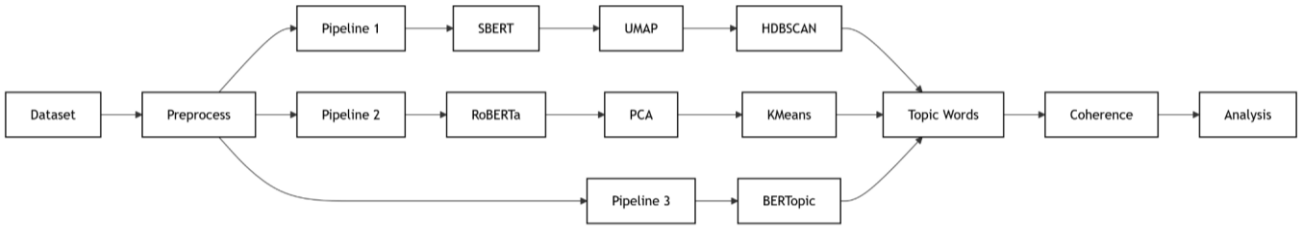

Fig. 1 Research flow diagram

C. Topic Modelling Pipelines

To investigate the thematic structure of the scientific articles, three distinct topic modeling pipelines were implemented and evaluated. Each pipeline was designed with a unique combination of techniques for document embedding, dimensionality reduction, and clustering, allowing for a comparative analysis of their effectiveness.

1) Pipeline 1 (SBERT – UMAP – HDBSCAN): The first pipeline, inspired by common structures in modern topic modeling such as BERTopic, began by generating document embeddings using Sentence-BERT (SBERT). Specifically, the all-MiniLM-L6-v2 model was loaded via the sentence-transformers library, chosen for its excellent balance of performance and efficiency. This model encodes texts into 384-dimensional dense vectors, leveraging SBERT's proven capability to build expressive embeddings, which are essential for maintaining interpretability in short-text environments [23,26]. Subsequently, uniform manifold approximation and projection (UMAP) was applied to reduce the dimensionality of these SBERT embeddings. The choice of n_components=5 is a common practice for UMAP when used as an intermediate step for clustering, as it is low enough to mitigate the curse of dimensionality while retaining sufficient structure for density-based algorithms [22,23]. The final stage employed hierarchical density-based spatial clustering of applications with noise (HDBSCAN), which recent studies highlight as advantageous for its ability to find clusters of varying shapes without pre-specifying their number [22]. The hyperparameters for UMAP and HDBSCAN in this custom pipeline were deliberately chosen to align with the default and commonly recommended settings for BERTopic to ensure a fair comparison.

2) Pipeline 2 (RoBERTa – PCA – k-means): The second pipeline explored an alternative combination. Document embeddings were first generated using RoBERTa (robustly pptimized BERT pre-training approach) [21], specifically the stsb-roberta-base-v2 model, which was selected to directly replicate the architecture used in the benchmark study [21]. RoBERTa is recognized for its robust performance stemming from optimized pre-training procedures [14,15]. Dimensionality was then reduced using principal component analysis (PCA), a linear technique [21,22]. PCA was set to produce 50 components, which captured 71.84% of the total explained variance, providing a reasonable trade-off between dimensionality reduction and information preservation. For the clustering phase, k-means was applied, partitioning data into a pre-defined number of clusters (k) based on Euclidean distance, aligning with recent approaches for clustering transformer-based embeddings [27].

3) Pipeline 3 (BERTopic): The third pipeline utilized BERTopic, an advanced technique that leverages transformer embeddings and class-based TF-IDF (c-TF-IDF). Implemented with the BERTopic library, this pipeline used the all-mpnet-base-v2 sentence-transformer model for generating document embeddings, which is the default and officially recommended model for BERTopic due to its high-quality sentence embeddings. A core parameter, min_topic_size, was set to 40. BERTopic internally handles dimensionality reduction (typically UMAP) and clustering (typically HDBSCAN) before applying its c-TF-IDF mechanism to extract the final topic representations [17].

D. Computational Setup and Evaluation

All experiments were conducted on a cloud-based system (Google Colab) equipped with an NVIDIA Tesla T4 GPU, 12.7GB of RAM, and a Linux operating system. The pipelines were implemented in Python 3.12.12 using core libraries including scikit-learn, pandas, nltk, sentence-transformers, and bertopic.

In terms of computational cost, Pipeline 1 and Pipeline 3 exhibited similar runtimes due to their reliance on UMAP and HDBSCAN. Pipeline 2 was marginally faster due to the efficiency of PCA and k-means, but this came at a significant cost to coherence. All pipelines were most computationally intensive during the embedding generation phase, requiring GPU acceleration for timely processing. To ensure transparency and reproducibility, the complete source code, random seeds, and hyperparameter configurations used in this comparative study are publicly available at [28].

Performance was quantitatively evaluated using the C_v coherence score, a metric that measures the semantic similarity of the top words within a topic. The parameter choices described previously were selected based on common practices in the literature that balance performance and interpretability. An exhaustive hyperparameter sweep was beyond the scope of this comparative study.

III. RESULT AND DISCUSSION

A. Data Preparation and Pre-Processing

The study commenced with the loading of a dataset comprising 20,972 scientific articles. For each article, the TITLE and ABSTRACT fields were concatenated to create a unified text column, serving as the primary input for analysis (Table I).

Subsequently, a pre-processing pipeline was applied. This involved converting text to lowercase, removing hyperlinks, email addresses, special characters, and numerical digits using regular expressions. The text was then tokenized into individual words using NLTK's word_tokenize. Common English stopwords and single-character words were removed. Finally, each remaining token was lemmatized using NLTK's WordNetLemmatizer. The outcome of this stage was a list of processed_tokens for each document, and a processed_text_for_bert column where these tokens were joined back into a string (Table II).

No documents yielded empty text after pre-processing. A Gensim dictionary was successfully created from all_processed_tokens, resulting in a vocabulary size of 81,394 unique terms.

B. Topic Modeling Pipeline Outcomes

1) Pipeline 1 (SBERT – UMAP – HDBSCAN): The first pipeline, combining SBERT, UMAP, and HDBSCAN, began by generating 384-dimensional embeddings for the 20,972 documents using the all-MiniLM-L6-v2 SBERT model, resulting in an embedding matrix of shape (20972, 384). These embeddings were subsequently processed using UMAP, configured with parameters n_neighbors=15, n_components=5, min_dist=0.0, and metric='cosine', which reduced their dimensionality to five, yielding a matrix of shape (20972, 5). Finally, HDBSCAN clustering, with min_cluster_size=15 and a 'euclidean' metric, was applied to the reduced embeddings, identifying 175 distinct clusters (topics). Notably, 8,863 documents (42.26% of the dataset), were classified as noise points (label -1) by HDBSCAN. After extracting the top 10 words for each of the 175 topics using a c-TF-IDF approach, this pipeline achieved a C_v coherence score of 0.6079. As shown in Fig. 2, the UMAP projection reveals distinct clusters, though with significant noise dispersion. This scattering visually represents the 42.26% of unclassified documents and is a direct result of HDBSCAN's strict density thresholds, which prioritize cluster purity over total coverage. By filtering out ambiguous data points that lack strong semantic neighbors, the algorithm ensures that the remaining clusters represent highly coherent topics, albeit at the cost of excluding a substantial portion of the corpus.

TABLE I
COMBINED DATASET SAMPLE

No.	TITLE
0	Reconstructing Subject-Specific Effect Maps
1	Rotation Invariance Neural Network
	ABSTRACT
0	Predictive models allow subject-specific inf...
1	Rotation invariance and translation invarian...
	TEXT
0	Reconstructing Subject-Specific Effect Maps ...
1	Rotation Invariance Neural Network Rotation ...

TABLE II
DATASET SAMPLE AT EACH TEXT PROCESSING STAGE

Text Processing Stage	Sample
Original Text	Reconstructing Subject-Specific Effect Maps Predictive models allow subject-specific...
Processed Tokens	[reconstructing, subjectspecific, effect, map, predictive, model, allow, subjectspecific...]
Processed Text for BERT	reconstructing subjectspecific effect map predictive model allow subjectspecific...

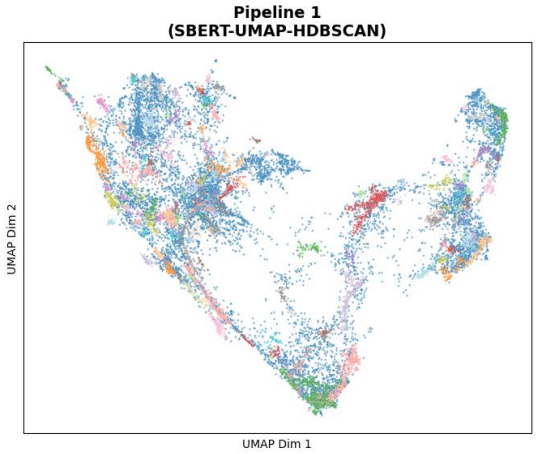

Fig. 2 UMAP projection of pipeline 1

2) Pipeline 2 (RoBERTa – PCA – k-means): The second pipeline, utilizing RoBERTa, PCA, and k-means, employed the stsb-roberta-base-v2 RoBERTa model to produce 768-dimensional embeddings for the 20,972 documents, forming an embedding matrix of shape (20972, 768). PCA was then applied to reduce the dimensionality to 50 components, which successfully captured 71.84% of the total explained variance. The resulting matrix had a shape of (20972, 50). K-means clustering, with n_clusters=6, was subsequently performed, partitioning the documents into six distinct groups. The distribution of documents across clusters was as follows: Cluster 0 (3,661), Cluster 1 (4,135), Cluster 2 (3,846), Cluster 3 (3,345), Cluster 4 (1,659), and Cluster 5 (4,326). The top 10 words for each topic were extracted via a c-TF-IDF method, and this pipeline yielded a C_v coherence score of 0.4756. Visualization of embeddings (Fig. 3) reveals a dense, globular structure with poor cluster separation, which dominates the embedding space and lacks the clear separation seen in the non-linear projections. This compression stems from a methodological mismatch. The complex, non-linear semantic relationships captured by RoBERTa cannot be effectively unfolded by the linear dimensionality reduction of PCA. Consequently, the visible color partitions do not follow natural data boundaries but are artifacts of the k-means algorithm forcing the continuous cloud into arbitrary spherical segments. This explains the lower coherence score and the model's inability to distinguish subtle thematic differences.

3) *Pipeline* 3 (BERTopic): The third pipeline featured BERTopic, configured with the all-mpnet-base-v2 embedding model, a min_topic_size of 40, and a unigram CountVectorizer incorporating NLTK English stopwords for its internal c-TF-IDF process. Upon fitting the model to the preprocessed documents, BERTopic identified 69 distinct topics, excluding the conventional outlier topic (-1). The top 10 words representing each topic were extracted using BERTopic's inherent c-TF-IDF mechanism. This pipeline achieved a C_v coherence score of 0.701. As demonstrated in Fig. 4, the UMAP projection displays the most refined structural definition, characterized by tight and well-separated density islands.

The extracted topics form the most refined and well-separated density islands among all configurations. This structural clarity highlights the efficacy of the integrated framework, where the specific alignment of embeddings and the c-TF-IDF mechanism successfully consolidates thematic content into tight groups. Unlike the scattered result in Pipeline 1, the sharper clustering here indicates a superior disentanglement of closely related scientific sub-fields, allowing for high semantic independence with minimal noise interference.

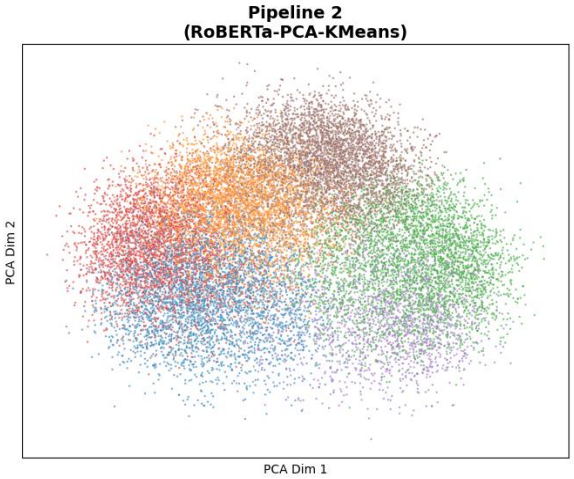

Fig. 3 PCA projection of pipeline 2

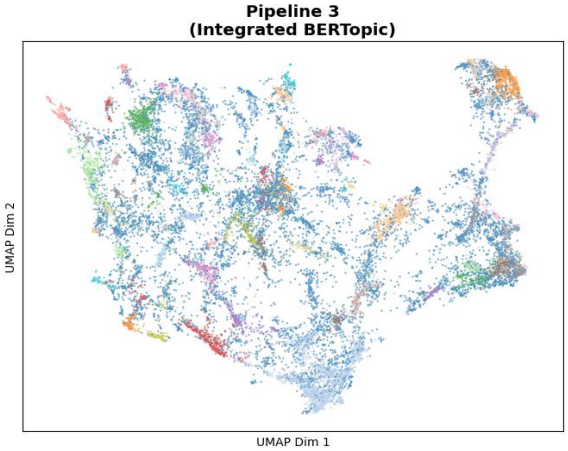

Fig. 4 UMAP projection of pipeline 3 (BERTopic)

C. Comparative Summary of Topics and Coherence

To qualitatively assess the semantic quality of the generated models, the top keywords extracted by each pipeline were aligned across five major scientific disciplines represented in the corpus. This side-by-side comparison reveals significant differences in granularity. As observed in the extracted keywords, Pipeline 2 (RoBERTa) tends to collapse distinct sub-fields, such as geophysics and physics, into broad, generic categories due to the rigid nature of k-means clustering. In contrast, Pipeline 1 and Pipeline 3 preserve specific disciplinary nuances, successfully distinguishing between "distributed computing" and "robotics" within the computer science domain. These representative topics are presented in Table III.

Following the qualitative inspection, the models were quantitatively benchmarked using the C_v coherence metric to provide an objective measure of performance. The evaluation highlights a clear correlation between the structural separation observed in the visualizations and the coherence score. Pipeline 2, which struggled to separate clusters visually, produced the lowest coherence score and a limited number of topics. Conversely, the integrated BERTopic framework achieved the highest validation score, confirming that its density-based approach yields more semantically meaningful topics. The final performance ranking and topic counts are summarized in Table IV.

The results clearly identify Pipeline 3 (BERTopic) as the superior methodology, achieving the highest C_v coherence score of 0.7012. This score was obtained from an optimized configuration that incorporated an (1,3) n-gram range, representing the most superior performance across all experiments.

Pipeline 1, which implemented a custom workflow combining SBERT, UMAP, and HDBSCAN, also demonstrated robust performance. It yielded a C_v score of 0.6079, significantly surpassing its initial performance target and establishing it as a highly coherent model. Notably, this pipeline generated the largest number of distinct topics (175).

In contrast, Pipeline 2, which was designed to replicate the benchmark study's architecture (k-means-RoBERTa-PCA), achieved a C_v score of 0.4756. This outcome falls below the benchmark target of 0.5554 reported in the reference study and represents the lowest performance among the three tested pipelines. This configuration was constrained to produce exactly six topics, in line with the benchmark methodology.

In summary, the empirical findings reveal a clear performance ranking, with the integrated BERTopic framework outperforming both the custom SBERT-based pipeline and the replicated RoBERTa-based pipeline. A detailed methodological interpretation of these differences, including the interplay between embedding models, dimensionality reduction techniques, and clustering algorithms, is presented in the following Discussion section.

TABLE III
COMPARISON OF TOP EXTRACTED TOPICS ACROSS PIPELINES

Category	Pipeline 1 (SBERT)	Pipeline 2 (RoBERTa)	Pipeline 3 (BERTopic)
Geophysics	seismic, earthquake, tectonic..	(Merged into Physics)	seismic, earthquake, fault..
Computer Science	distributed, sgd, gradient..	network, data, system..	robot, policy, reinforcement..
Physics	eternal, inflation, model..	system, model, state..	galaxy, star, mass..
Medicine	epidemic, disease, infection..	(Merged into General)	estimator, distribution, regression..
General	-	model, method, algorithm..	-

TABLE IV
COHERENCE SCORES BETWEEN EACH PIPELINES

Pipeline	Achieved C_v	Number of Topics
1 (HDBSCAN-SBert-UMAP)	0.606	175
2 (KMeans-RoBERTa-PCA)	0.475	6
3 (BERTopic)	0.701	69

D. Performance of Individual Pipelines

The results offer valuable insights into the performance of different topic modeling pipelines applied to a corpus of scientific research abstracts. This section discusses these findings in detail, compares them whith stated targets and existing literature, and highligts their implications.

1) Pipeline 1 (HDBSCAN-SBERT-UMAP): This pipeline achieved a C_v coherence score of 0.6079. The use of SBERT embeddings, known for capturing nuanced semantic meaning, followed by UMAP's ability to preserve data structure in lower dimensions [23], and

HDBSCAN's strength in identifying clusters of varied densities [22], contributed to this strong performance. However, this pipeline generated a high number of topics (175) and identified a substantial portion of documents (42.26%) as noise. This indicates that many documents did not meet the density requirements under the specified parameters. While this prevents the formation of noisy, incoherent topics, it also reduces the number of classifiable documents, presenting a trade-off between coverage and topic specificity.

2) Pipeline 2 (RoBERTa–PCA–k-means): This pipeline achieved a C_v coherence of 0.4756. Although RoBERTa embeddings are recognized for their robustness [14], and PCA is a standard dimensionality reduction technique, this configuration yielded the lowest performance among the tested models. Analytically, the underperformance can be attributed to a methodological mismatch. PCA's linear projection is less effective at preserving the complex, non-linear semantic relationships captured by RoBERTa embeddings compared to non-linear methods like UMAP [23]. Furthermore, k-means imposes rigid, centroid-based clustering and forces documents into a fixed number of spherical clusters, which is not well suited to the irregular thematic groups in textual data, often leading to poor topic separateness and independence compared to density-based approaches [22]. As the result, the extracted topics tend to be overlay broad, leading to lower C_v scores.

3) Pipeline 3 (BERTopic): BERTopic achieved the highest C_v coherence score of 0.7012, while generating 69 distinct topics. This outstanding performance highlights the efficacy of its integrated approach. BERTopic's superior coherence can be directly attributed to its integrated architecture, particularly its class-based TF-IDF (c-TF-IDF) mechanism. Unlike traditional TF-IDF, c-TF-IDF considers all documents within a cluster as a single composite document, which enables the identification of words that are truly descriptive of the topic as a whole. This sophisticated handling of topic word generation, combined with its seamless use of UMAP and HDBSCAN, creates a synergistic effect that produces more distinct and semantically tight topics. These findings align with recent literature showcasing the ability of integrated transformer-based approaches to produce highly coherent topics. Such performance has been observed across diverse corpora, ranging from complex neural document modeling to scientific articles [7,17,26,29].

E. Comparative Analysis and Implications

The comparative evaluation reveals a clear performance hierarchy, with BERTopic emerging as the most effective method. These findings highlight a paradigm shift from ad-hoc combinations of modular components toward integrated, transformer-based frameworks.

To analytically validate these performance disparities, the embedding structures across the three pipelines were visualized. These visualizations provide immediate insight into the methodological mismatch observed in Pipeline 2. As shown in Fig. 3, the combination of RoBERTa embeddings with linear PCA dimensionality reduction results in a dense, globular manifold in which data points are not naturally separated. Because PCA is a linear technique, it cannot effectively capture the complex, non-linear semantic relationships inherent in scientific abstracts. Consequently, the k-means algorithm slices this continuous blob into arbitrary spherical partitions, explaining the low C_v score (0.4756).

Pipeline 1 (Fig. 2) and Pipeline 3 (Fig. 4) utilize UMAP, which successfully disentangle the high-dimensional data into distinct density islands. Fig. 4 (BERTopic) exhibits the most refined structure , tighter clusters and reduced peripheral noise compare to Fig. 2. This observation that the integrated c-TF-IDF mechanism effectively consolidates thematic content within these density islands, whereas Pipeline 2's geometric assumptions fail to capture the natural topology of the data.

While this study utilizes a corpus of scientific articles, the identified performance hierarchy, favoring integrated, non-linear, and density-based approaches, provides a generalizable insight applicable to other domains with complex, high-dimensional text data. The superior ability of the BERTopic framework suggests its potential transferability to fields such as legal text analysis or financial reports.

The practical implications of these findings are significant for researchers in bibliometrics and digital library science. The demonstrated superiority of BERTopic provides a clear recommendation for practitioners building tools for automated literature reviews and research trend analysis. By generating more coherent topics, such tools can more accurately map the intellectual structure of a field and help researchers navigate the growing body of scientific literature.

It is important to note, however, that C_v coherence, while a standard automated metric, does not fully capture topic interpretability. Future validation should incorporate human-in-the-loop evaluation methods, such as intruder tests or expert review, to better assess the practical quality of the generated topics. In summary, the results converge with supporting literature, demonstrating that integrated frameworks like BERTopic are more effective in producing coherent and interpretable topics from scientific text.

IV. CONCLUSION

This study conducted a comparative benchmark of three topic modeling pipelines to identify the optimal approach for analyzing scientific literature. The findings indicate the superior performance of the integrated BERTopic model, which achieved a C_v coherence score of 0.701. This result highlights that the combination of transformer embeddings, non-linear dimensionality reduction, and density-based clustering is crucial for uncovering meaningful thematic structures. Based on these automated metrics, integrated frameworks offer a more robust solution than custom modular pipelines for large-scale text analysis. Future work should address the limitations of automated evaluation metrics by validating these findings across diverse domains and incorporating human evaluation alongside complementary metrics, such as topic diversity. Integrating large language models (LLMs) also represents a promising direction for further advancement. In particular, future research may explore the use of LLMs for zero-shot topic labeling, thematic summarization, and hybrid refinement of unsupervised clusters to enhance both interpretability and analytical depth.

REFERENCES

[1] A. H. Suyanto, T. Djatna, and S. H. Wijaya, "Mapping and predicting research trends in international journal publications using graph and topic modeling," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 30, no. 2, p. 1201, May 2023, doi: 10.11591/ijeecs.v30.i2.pp1201-1213.

[2] S. Kavvadias, G. Drosatos, and E. Kaldoudi, "Supporting topic modeling and trends analysis in biomedical literature," *J. Biomed. Inform.*, vol. 110, p. 103574, 2020, doi: https://doi.org/10.1016/j.jbi.2020.103574.

[3] T. Silwattananusarn and P. Kulkanjanapiban, "A text mining and topic modeling based bibliometric exploration of information science research," *IAES Int. J. Artif. Intell. IJ-AI*, vol. 11, no. 3, p. 1057, Sept. 2022, doi: 10.11591/ijai.v11.i3.pp1057-1065.

[4] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Inf. Syst.*, vol. 112, p. 102131, 2023, doi: https://doi.org/10.1016/j.is.2022.102131.

[5] S. H. Mohammed and S. Al-augby, "LSA & LDA topic modeling classification: comparison study on e-books," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 1, p. 353, July 2020, doi: 10.11591/ijeecs.v19.i1.pp353-362.

[6] 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," *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 298–304, 2022, doi: https://doi.org/10.1016/j.gltp.2022.03.015.

[7] S.-S. SHIN and H.-C. Yang, "A Study on Leadership Trends from the Perspective of Domestic Researcher's Using BERTopic and LDA," *East Asian J. Bus. Econ. EAJBE*, vol. 11, no. 1, pp. 53–71, Mar. 2023, doi: 10.20498/EAJBE.2023.11.1.53.

[8] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Front. Artif. Intell.*, vol. 3, p. 42, July 2020, doi: 10.3389/frai.2020.00042.

[9] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.

[10] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short Text Topic Modeling Techniques, Applications, and Performance: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1427–1445, Mar. 2022, doi: 10.1109/TKDE.2020.2992485.

[11] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Front. Sociol.*, vol. 7, p. 886498, May 2022, doi: 10.3389/fsoc.2022.886498.

[12] G. Papadia, M. Pacella, M. Perrone, and V. Giliberti, "A Comparison of Different Topic Modeling Methods through a Real Case Study of Italian Customer Care," *Algorithms*, vol. 16, no. 2, p. 94, Feb. 2023, doi: 10.3390/a16020094.

[13] M. Hankar, M. Kasri, and A. Beni-Hssane, "A comprehensive overview of topic modeling: Techniques, applications and challenges," *Neurocomputing*, vol. 628, p. 129638, May 2025, doi: 10.1016/j.neucom.2025.129638.

[14] 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, p. e35606, Apr. 2022, doi: 10.2196/35606.

[15] 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, p. 8184, Oct. 2022, doi: 10.3390/s22218184.

[16] C. Y. Sy, L. L. Maceda, N. M. Flores, and M. B. Abisado, "Unsupervised Machine Learning Approaches in NLP: A Comparative Study of Topic Modeling with BERTopic and LDA".

[17] A. Madrid-García, D. Freites-Núñez, B. Merino-Barbancho, I. Pérez Sancristobal, and L. Rodríguez-Rodríguez, "Mapping two decades of research in rheumatology-specific journals: a topic modeling analysis with BERTopic," *Ther. Adv. Musculoskelet. Dis.*, vol. 16, p. 1759720X241308037, Jan. 2024, doi: 10.1177/1759720X241308037.

[18] N. Khodeir and F. Elghannam, "Efficient topic identification for urgent MOOC Forum posts using BERTopic and traditional topic modeling techniques," *Educ. Inf. Technol.*, vol. 30, no. 5, pp. 5501–5527, Apr. 2025, doi: 10.1007/s10639-024-13003-4.

[19] L. Kun, H. Alli, and K. A. A. A. Rahman, "The Trends of Potential User Research from 2014-2023 Based on Bibliometric and Bertopic," *Rev. Gest. Soc. E Ambient.*, vol. 18, no. 9, p. e06100, May 2024, doi: 10.24857/rgsa.v18n9-068.

[20] E. Chagnon, R. Pandolfi, J. Donatelli, and D. Ushizima, "Benchmarking topic models on scientific articles using BERTeley," *Nat. Lang. Process. J.*, vol. 6, p. 100044, Mar. 2024, doi: 10.1016/j.nlp.2023.100044.

[21] M. C. Wijanto, I. Widiastuti, and H.-S. Yong, "Topic Modeling for Scientific Articles: Exploring Optimal Hyperparameter Tuning in BERT," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 3, pp. 912–919, June 2024, doi: 10.18517/ijaseit.14.3.19347.

[22] D. Hanny and B. Resch, "Clustering-Based Joint Topic-Sentiment Modeling of Social Media Data: A Neural Networks Approach," *Information*, vol. 15, no. 4, p. 200, Apr. 2024, doi: 10.3390/info15040200.

[23] 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, p. 114741, Aug. 2021, doi: 10.1016/j.eswa.2021.114741.

[24] X. Han, "Evolution of research topics in LIS between 1996 and 2019: an analysis based on latent Dirichlet allocation topic model," *Scientometrics*, vol. 125, no. 3, pp. 2561–2595, Dec. 2020, doi: 10.1007/s11192-020-03721-0.

[25] B. Densil, "Topic Modeling for Research Articles." Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/blessondensil294/topic-modeling-for-research-articles/data

[26] X. Wu, T. Nguyen, and A. T. Luu, "A survey on neural topic models: methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 57, no. 2, p. 18, Jan. 2024, doi: 10.1007/s10462-023-10661-7.

[27] K. Datchanamoorthy, A. Mala. G. S, and Padmavathi. B, "TEXT MINING: CLUSTERING USING BERT AND PROBABILISTIC TOPIC MODELING," *Soc. Inform. J.*, vol. 2, no. 2, pp. 1–13, Dec. 2023, doi: 10.58898/sij.v2i2.01-13.

[28] F. A. A. Pratama, M. E. N. Arief, and V. R. S. Nastiti, "Transformer-Based Topic Modeling Pipeline for Scientific Literature." Nov. 2025. [Online]. Available: https://github.com/reddishowo/topic-modelling-project

[29] 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.