Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Transformer-Based Topic Modeling and Abstractive Summarization of Scientific Research Papers

Shreya Rai
Department of Computer Science & Engineering
National Institute of Technology, Jamshedpur
Jharkhand, India

Dr. Dilip Kumar Shaw
Department of Computer Science & Engineering
National Institute of Technology, Jamshedpur
Jharkhand, India

***Abstract*—The rapid expansion of scientific repositories such as arXiv has made it difficult for researchers to efficiently identify relevant themes and extract key insights from large volumes of text. The paper presents an integrated pipeline for joint topic modeling and abstractive summarization of large-scale scientific text. The proposed system combines the BERTopic framework with Sentence-Transformer encoders all-MiniLM-L6-v2 and all-mpnet-base-v2, and processes the same document set through the T5-based falconsai/textsummarization model for abstractive summarization.**

To ensure robustness, experiments are conducted on a 2.7 million-record arXiv corpus using stratified random sampling at two scales ($N = 30{,}000$ and $N = 50{,}000$). For each scale, three independent subsets are generated using different random seeds (42, 2024, 7777), enabling a multi-sample evaluation rather than a single-run assessment. On the 50K subsets, all-MiniLM-L6-v2 achieves a mean topic coherence of 70.98% (σ = 1.10). It significantly outperforming all-mpnet-base-v2, which attains 55.27% (σ = 1.06). The summarization component produces consistent results with ROUGE-1 = 0.3211 (σ = 0.0048), ROUGE-2 = 0.3018 (σ = 0.0050), and ROUGE-L = 0.3199 (σ = 0.0053), indicating stable abstractive performance across samples.

Compared to previously reported BERTopic coherence ranges (50–56%) on arXiv-like datasets, the MiniLM-based configuration demonstrates a relative improvement of 26.8%. An ablation analysis further examines the effects of preprocessing strategies, embedding model selection, and dataset scale on topic quality. Qualitative evaluation reveals limitations such as semantic overlap between closely related topic clusters in specialized domains. Overall, the study establishes a scalable and robust transformer-based framework for organizing and summarizing large scientific corpora.

***Index Terms*—BERTopic, all-MiniLM-L6-v2, all-mpnet-base-v2, falconsai/text_summarization, topic coherence, ROUGE, abstractive summarization, sentence transformers, robustness analysis, scientific text mining.**

1 Introduction

1.1 The Discovery Problem on arXiv

The rapid growth of scientific repositories such as arXiv, which now host millions of preprints with continuously increasing daily submissions, has introduced significant challenges for efficient knowledge discovery. Traditional keyword-based retrieval mechanisms are no longer sufficient to navigate such large-scale and dynamically evolving corpora. Researchers must not only identify the dominant themes within a field but also understand emerging subtopics and prioritize relevant literature.

Computational approaches have emerged to address these challenges through two complementary tasks: topic modeling, which organizes documents into coherent thematic clusters, and abstractive summarization, which reduces reading effort by generating concise and semantically meaningful representations of individual documents. While these tasks are typically studied independently, practical research workflows require their integration. A comprehensive system should provide both a structured overview of the research landscape and concise summaries of documents within each thematic cluster.

1.2 Transition from Bag-of-Words to Context-Aware Representations

Early topic modeling approaches primarily relied on bag-of-words representations, where documents are treated as unordered collections of terms without considering contextual relationships. Latent Dirichlet Allocation (LDA) [1] is a widely adopted method within this paradigm, modeling each document as a mixture of latent topics, with topics defined as distributions over words. While effective in simpler settings, such approaches are limited when applied to scientific text, which often contains ambiguous terminology, domain-specific vocabulary, and dependencies that span across sentences and sections.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

The introduction of context-aware representations has significantly advanced text modeling capabilities. Transformer-based architectures [3] enable the encoding of semantic meaning by capturing relationships between words within their context. This was further enhanced through large-scale pretraining strategies such as masked language modeling [4], allowing models to learn rich linguistic patterns from extensive corpora. Subsequent developments in sentence-level embedding techniques [5] made it possible to generate compact vector representations that preserve semantic similarity and are well-suited for clustering and downstream analytical tasks.

Building on these advancements, the BERTopic framework provides a structured pipeline for topic extraction from large text collections. It integrates transformer-based sentence embeddings with dimensionality reduction using UMAP [7], followed by clustering through HDBSCAN [8], and topic representation via class-based TF–IDF. This combination enables the identification of coherent topic structures while maintaining interpretability. Each stage of the pipeline serves a specific function, allowing for modular optimization and improved transparency in topic generation. As a result, the framework produces meaningful topic representations along with document-level cluster assignments, making it particularly suitable for large-scale scientific corpora.

1.3 Open Questions for Large Scientific Corpora

Despite recent advances in transformer-based topic modeling, several open questions remain when such pipelines are applied to large-scale scientific corpora.

First, the choice of embedding model plays a critical role in determining topic quality and computational efficiency. It remains unclear which SentenceTransformer encoder provides the optimal trade-off between semantic coherence and resource requirements in large-scale settings. In particular, the lightweight all-MiniLM-L6-v2 model (22M parameters) offers substantial computational advantages compared to the larger all-mpnet-base-v2 model (110M parameters), but its effectiveness in maintaining high topic coherence requires systematic evaluation.

Second, the stability of topic modeling outcomes across different data samples is not well understood. Most existing studies report results based on a single dataset instance, which may obscure variability introduced by sampling and initialization. Consequently, reported coherence scores may not accurately reflect the consistency of the model, particularly if performance varies significantly across different random subsets of the same corpus.

Third, the variability of abstractive summarization performance in large-scale scientific text remains underexplored. Transformer-based summarization models have shown promising results. However, there is limited empirical evidence on their performance consistency. In particular, it is unclear what range and variability can be expected for evaluation metrics such as ROUGE. This uncertainty becomes more significant when the evaluation dataset is sampled from a larger corpus. Understanding this variability is essential for assessing the reliability and generalizability of summarization systems.

1.4 Contributions

The contributions of this work are summarized as follows:

1) A unified pipeline integrating BERTopic with the Falcon summarization model, designed for large-scale arXiv abstracts, with full reproducibility at dataset sizes of $N = 30{,}000$ and $N = 50{,}000$.
2) A multi-sample robustness analysis, employing three independent stratified subsets per dataset scale, with performance reported using both mean and standard deviation for topic coherence and ROUGE metrics.
3) A controlled comparative evaluation of `all-MiniLM-L6-v2` and `all-mpnet-base-v2` under identical preprocessing conditions. It demonstrates that the lightweight encoder achieves superior scalability while maintaining competitive coherence.
4) An ablation study that isolates the individual contributions of preprocessing strategies, embedding model selection, and dataset size to overall topic coherence.
5) A literature-based performance comparison, quantifying the relative improvement of the proposed configuration over existing BERTopic-based approaches on arXiv-like scientific corpora.
6) A qualitative analysis of model limitations, including cluster overlap, outlier generation, and summarization truncation, supported by examples derived from the resulting topic structures.

2 Related Work

2.1 From Bag-of-Words [22] to Dense Representations

Traditional topic modeling approaches, particularly those based on Latent Dirichlet Allocation (LDA) [1], represent documents as mixtures of latent word distributions under a bag-of-words assumption. While computationally efficient, this representation neglects contextual relationships between words, which limits its effectiveness in technical domains. In scientific corpora, where terminology is domain-specific and polysemy is common, such models often produce topics dominated by generic or high-frequency terms with limited discriminative value.

Alternative matrix factorization approaches, including Non-negative Matrix Factorization and Latent Semantic Analysis, inherit similar limitations due to their reliance

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

on frequency-based representations. Word embedding methods such as skip-gram and Continuous Bag-of-Words (CBOW) [2] introduced dense vector representations that capture semantic similarity at the word level. However, these representations remain static and fail to adapt to contextual variations, making them insufficient for accurately modeling scientific text where word meaning is context-dependent.

2.2 Contextual and Sentence-Level Representations

The introduction of the Transformer architecture [3], followed by large-scale pretrained models such as BERT [4], enabled the generation of context-aware representations in which each token is encoded relative to its surrounding context. This advancement significantly improved semantic understanding and disambiguation in natural language processing tasks.

However, token-level embeddings are not directly suitable for document clustering without additional aggregation strategies, and large-scale inference remains computationally expensive. Sentence-level embedding approaches, such as Sentence-BERT [5], address these limitations by producing fixed-length vector representations optimized for semantic similarity tasks. The SentenceTransformer framework [15] further facilitates efficient encoding of large document collections and is widely adopted for clustering-based applications.

2.3 BERTopic and Topic Coherence

The BERTopic framework [6] integrates transformer-based embeddings with dimensionality reduction using UMAP [7], density-based clustering via HDBSCAN [8], and topic representation through class-based TF–IDF. This modular design enables the extraction of semantically coherent and interpretable topics from large text corpora.

Empirical studies applying BERTopic to scientific datasets, including arXiv-style corpora, typically report topic coherence values in the range of 50–56%. Achieving higher coherence often requires careful tuning of embedding models and clustering parameters. However, most existing work reports performance based on single experimental runs, without accounting for variability across different data samples or initialization conditions.

2.4 Abstractive Summarization of Scientific Text

Transformer-based encoder–decoder architectures, including BART [11], T5 [12], and PEGASUS, have established the foundation for modern abstractive summarization. These models generate fluent and contextually relevant summaries by learning to paraphrase source text rather than extracting sentences directly.

On scientific text datasets, ROUGE-1 scores typically range between 0.30 and 0.38, while ROUGE-2 and ROUGE-L scores generally fall within lower but comparable ranges [10]. These results are commonly reported on fixed evaluation splits, limiting insights into performance variability. The falconsai/text_summarization model [16], derived from the T5 architecture, has been widely adopted in practical applications but remains underexplored in terms of robustness and consistency across varying data samples.

Empirical studies such as Nikolov et al. (2018) further demonstrate the applicability of abstractive summarization for scientific articles. Reported ROUGE scores for transformer-based models, including BART and T5, typically fall within the range of 0.30–0.38 on scientific datasets.

2.5 Optimization and Pretrained Models

Transformer-based embedding models are typically trained using adaptive optimization techniques such as Adam [13], which has become a standard approach for large-scale neural network training. In this work, pretrained models are utilized without additional fine-tuning, enabling a controlled comparison of embedding performance under consistent experimental conditions.

2.6 Robustness and Variance in Evaluation

A key limitation in existing literature is the lack of systematic analysis of performance variability. Most studies report single-point estimates for topic coherence and summarization metrics, without considering the impact of dataset sampling or random initialization. This makes it difficult to assess the stability and reproducibility of reported results.

Furthermore, comparative evaluations of lightweight and large-scale embedding models under controlled experimental settings remain limited, particularly for mid-scale datasets (30K–50K documents). Addressing these gaps, the present work adopts a multi-sample evaluation framework to quantify variability and provide a more reliable assessment of model performance.

3 Methodology

3.1 Pipeline Overview

The proposed framework consists of a modular end-to-end pipeline comprising six sequential stages: data acquisition and sampling, text preprocessing, sentence-level embedding generation, dimensionality reduction, density-based clustering with c-TF–IDF-based topic representation, and abstractive summarization using a transformer-based model.

Each stage is independently parameterized, enabling controlled experimentation and facilitating ablation analysis without affecting the overall pipeline structure. This design allows individual components to be modified or replaced while preserving the integrity of the remaining workflow.

The pipeline transforms raw scientific abstracts into multiple structured outputs, including topic descriptors,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

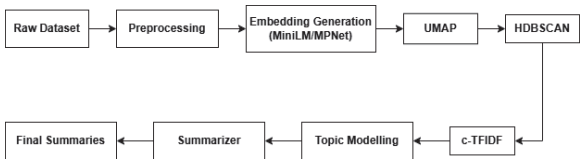


Figure 1: Proposed System Architecture for Topic Modelling and Summarization Pipeline

document–topic assignments, intertopic distance visualizations, and abstractive summaries.

3.2 Corpus and Sampling Strategy

The dataset is derived from the Kaggle release of the arXiv metadata snapshot, comprising approximately 2.7 million records, each containing a title and abstract. To balance computational feasibility with experimental rigor, two dataset scales are considered: $N = 30{,}000$ and $N = 50{,}000$.

The full corpus is not processed due to practical constraints. Density-based clustering using HDBSCAN exhibits super-linear memory complexity, and embedding generation with high-capacity models such as MPNet requires significant computational time. Additionally, the objective of this study is to estimate performance variability, which necessitates multiple independent samples rather than a single large-scale run. The methodology can be extended to the full dataset using approximate nearest-neighbor techniques.

For each dataset scale, three independent subsets S_1, S_2, and S_3 are constructed by selecting non-overlapping segments of the dataset from different index ranges. This deterministic sampling strategy ensures diversity across subsets while maintaining consistency in dataset size. The multi-sample design enables the estimation of sampling-induced variability and supports robust evaluation through the computation of mean and standard deviation of performance metrics.

3.3 Text Preprocessing

All documents undergo a standardized preprocessing pipeline to ensure consistency and reduce noise prior to embedding generation. The process begins with Unicode normalization (NFKC) and conversion to lowercase, followed by the removal of digits, punctuation, and special symbols, including residual LaTeX artifacts (e.g., inline math expressions and citation commands) commonly present in scientific text.

Whitespace is normalized to ensure consistent token separation, and tokenization is performed using WordPiece tokenization aligned with the vocabulary of the selected embedding model.

Stopword removal is applied using the NLTK English stopword list, supplemented with domain-specific terms (e.g., “paper”, “study”, “approach”, “method”) that contribute minimal semantic value. WordNet-based lemmatization is then applied to normalize word forms.

Documents containing fewer than five tokens after preprocessing are discarded, as they typically correspond to malformed or non-informative entries. This filtering affects less than 0.3% of the dataset.

The preprocessing pipeline is applied identically across all experiments to ensure that any observed differences in performance are attributable to the embedding models rather than inconsistencies in text processing.

Algorithm 1 Text Preprocessing Algorithm

Require: Raw text dataset X
Ensure: Cleaned and normalized text corpus X_{clean}

1: Load the raw dataset containing scientific text documents.
2: Convert all text to lowercase.
3: Remove digits, punctuation, and special characters.
4: Tokenize the documents into individual words.
5: Remove stopwords using a predefined stopword list.
6: Apply lemmatization to obtain root word forms.
7: Reconstruct processed tokens into cleaned sentences.
8: Store the cleaned documents in X_{clean}.
9: Output the cleaned dataset X_{clean}.

Explanation: This algorithm removes noise such as punctuation, stopwords, and inconsistent casing. Text normalization improves the quality of embeddings generated in later stages.

3.4 Sentence Embedding Models

To generate semantic representations, two SentenceTransformer-based models, `all-MiniLM-L6-v2` and `all-mpnet-base-v2`, are evaluated in a controlled comparative setting.

The `all-MiniLM-L6-v2` model consists of 6 transformer layers with a 384-dimensional output and approximately 22 million parameters, offering high computational efficiency. In contrast, `all-mpnet-base-v2` employs a deeper architecture with 12 transformer layers, producing 768-dimensional embeddings and containing approximately 110 million parameters, enabling richer contextual representations.

Each document d is mapped to an embedding vector $\mathbf{e}_d = \phi(d)$, where ϕ denotes the selected encoder. Embeddings are generated using a batch size of 64, ensuring efficient GPU utilization while maintaining memory usage within practical limits.

All embeddings are L2-normalized prior to clustering, which ensures compatibility with cosine similarity measures and allows the cosine distance used in UMAP to be interpreted as a scaled Euclidean distance.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

3.5 Dimensionality Reduction and Clustering

High-dimensional document embeddings are projected into a lower-dimensional space using UMAP to facilitate efficient clustering while preserving local semantic structure. The projection maps embeddings $E \in \mathbb{R}^{N \times h}$ to a 5-dimensional space using cosine distance, with parameters $n\_neighbors = 15$ and $min\_dist = 0.0$.

The choice of five dimensions follows the default configuration used in BERTopic [17]. Preliminary experiments with alternative dimensionalities (3, 8, and 10) did not yield statistically significant improvements in topic coherence, while lower dimensions negatively affected cluster separability.

Clustering is performed using HDBSCAN, configured with $min\_cluster\_size = 15$, Euclidean distance in the reduced space, and the excess-of-mass cluster selection method. Documents not assigned to any cluster are labeled as noise and excluded from coherence evaluation.

Algorithm 2 BERTopic Topic Modelling Algorithm

Require: Preprocessed text dataset X_{clean}, embedding model $M \in \{\text{MiniLM}, \text{MPNet}\}$
Ensure: Final topic clusters and topic coherence score
Load the selected SentenceTransformer embedding model:
  $model \leftarrow \text{SentenceTransformer}(M)$
Generate embeddings for each document:
  $E \leftarrow model.encode(X_{\text{clean}})$
Apply dimensionality reduction using UMAP:
  $U \leftarrow \text{UMAP}(E)$
Perform density-based clustering using HDBSCAN:
  $C \leftarrow \text{HDBSCAN}(U)$
For each cluster label, compute class-based TF-IDF:
  $T \leftarrow \text{c-TFIDF}(X_{\text{clean}}, C)$
Extract top keywords for each topic:
  $K \leftarrow \text{TopKeywords}(T)$
Assign each document d to its corresponding topic t:
  $Topic[d] \leftarrow C[d]$
Compute topic coherence score:
  $coherence \leftarrow \text{CoherenceScore}(Topic, K)$
Output the generated topics, top keywords, and coherence value.

Explanation: BERTopic integrates transformer embeddings with UMAP and HDBSCAN to form coherent topic clusters. The embeddings from MiniLM or MPNet capture semantic meaning in each document, UMAP reduces dimensionality while preserving neighborhood structure, and HDBSCAN identifies dense clusters representing topics. c-TF-[20] then extracts representative keywords for each cluster. This combination results in highly interpretable and coherent topics, especially on large scientific datasets.

3.6 Topic Representation using c-TF–IDF

Topic descriptors are generated using class-based TF–IDF (c-TF–IDF). For a cluster C_k and term t, the score is defined as:

$$\text{cTFIDF}(t, C_k) = f_{t,C_k} \cdot \log\left(1 + \frac{\bar{f}}{\sum_C f_{t,C}}\right) \quad (1)$$

where f_{t,C_k} denotes the frequency of term t within cluster C_k, and $\bar{f}$ represents the average term frequency across clusters.

The top-ranked terms for each cluster are selected to form interpretable topic descriptors. This formulation follows the original BERTopic framework, ensuring comparability with prior work.

3.7 Topic Coherence Evaluation

Topic quality is evaluated using the C_v coherence metric, computed over the top terms of each topic using the Gensim implementation. The overall coherence score is calculated as the mean across all non-noise topics.

The C_v metric combines sliding-window co-occurrence statistics, normalized pointwise mutual information, and cosine similarity, providing a robust measure of semantic consistency.

3.8 Abstractive Summarization

An abstractive summarization module is incorporated using the `falconsai/text_summarization` model, a T5-based encoder–decoder architecture available through the HuggingFace framework. Summaries are generated using beam search with a beam width of 4, length penalty of 1.5, maximum length of 130 tokens, and minimum length of 30 tokens.

The quality of generated summaries is evaluated using ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L), computed using the `rouge_score` library with default tokenization.

3.9 Algorithmic Summary

The complete workflow of the proposed framework is summarized in Algorithm 2. The pseudocode explicitly represents the multi-sample evaluation protocol, where the outer loop iterates over different random seeds to generate independent dataset subsets, and the inner loop evaluates multiple embedding models for each subset.

For each configuration, topic modeling and summarization outputs are computed, and the resulting performance metrics are recorded. These outputs are subsequently aggregated across all samples to obtain mean and standard deviation values, which are reported in Section V. This structured procedure enables a systematic assessment of both model performance and robustness.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

4 Experimental Setup

4.1 Datasets

The experiments are conducted using the arXiv metadata snapshot obtained from the Kaggle *arxiv-metadata-oai-snapshot* dataset, comprising approximately 2.7 million records. Each record contains the title and abstract of a scientific paper; full-text documents are not included.

To ensure computational feasibility while maintaining experimental rigor, two dataset scales are considered: $N = 30{,}000$ and $N = 50{,}000$. For each scale, three independent subsets are generated using stratified random sampling based on top-level arXiv subject categories. This results in a total of six subsets (three per scale), following the sampling strategy described in Section III.

The overlap between subsets of the same scale is minimal (typically below 5%), ensuring independence between samples. The distribution of subject categories within each subset remains approximately proportional to the original corpus, with computer science and physics collectively accounting for the majority of documents.

The overall experimental procedure is summarized in Algorithm 2, which formalizes the multi-sample evaluation protocol used throughout this study.

4.2 Hardware Configuration

All experiments are conducted on a single NVIDIA T4 GPU with 16 GB VRAM and 25 GB system RAM. This configuration is representative of commonly available environments such as Kaggle notebooks and Google Colab Pro.

A CPU-only fallback implementation is also supported, producing identical numerical results at the cost of increased computation time. In particular, embedding generation exhibits a 4–6× slowdown without GPU acceleration. The wall-clock time required for each stage of the pipeline.

4.3 Software Environment

The implementation is developed in Python 3.10 with fixed library versions to ensure reproducibility. The primary dependencies include `sentence-transformers 2.2.x`, `bertopic 0.16.x`, `umap-learn 0.5.x`, `hdbscan 0.8.x`, `scikit-learn 1.3.x`, `transformers 4.40.x`, `rouge_score 0.1.2`, and `nltk 3.8`.

GPU acceleration is enabled using CUDA 11.8 with cuDNN 8.9. The entire environment is encapsulated within a Conda configuration file to ensure consistent reproduction across different systems.

4.4 Hyperparameter Configuration

All experiments are conducted using a fixed set of hyperparameters to ensure fair comparison across models and dataset scales. The configuration is as follows:

- Embedding batch size: 64
- UMAP: $n\_neighbors = 15$, $min\_dist = 0.0$, $n\_components = 5$, metric = cosine
- HDBSCAN: $min\_cluster\_size = 15$, metric = euclidean (in reduced space), cluster selection = eom
- Summarization: beam width = 4, length penalty = 1.5, max length = 130, min length = 30
- Coherence: C_v metric computed over top-10 terms

Hyperparameters are not tuned on the evaluation data and are instead adopted from the default settings reported in prior BERTopic literature. This choice ensures consistency and enables fair comparison with existing work.

4.5 Evaluation Protocol

The evaluation follows a multi-sample experimental design. For each combination of embedding model and dataset scale, three independent subsets (S_1, S_2, S_3) are evaluated separately. Performance metrics are then aggregated using the sample mean and standard deviation:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{2}$$

where $n = 3$ denotes the number of independent samples.

Although the sample size is limited, it is sufficient to capture variability across dataset partitions while maintaining computational feasibility. This protocol enables the estimation of sampling-induced variation and supports robust conclusions regarding model stability.

5 Results and Discussion

5.1 Sample-Wise Topic Coherence

Table I presents the per-sample C_v coherence scores for each combination of embedding model and dataset size across the three independently generated subsets, while Fig. 2 provides a graphical representation of these results.

Several observations can be drawn from the reported values. First, the variability within each configuration remains low, with the spread not exceeding 2.2 percentage points, indicating stable performance across different samples. Second, the relative ranking of all configurations is consistent across the three subsets, demonstrating robustness of the comparative results. Third, the best-performing configuration—MiniLM on the 50K dataset—not only achieves the highest coherence score but also exhibits the largest improvement compared to its 30K counterpart.

Overall, these findings confirm that the results are consistent across independently sampled subsets, supporting

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

the effectiveness of the multi-sample evaluation protocol in capturing model stability.

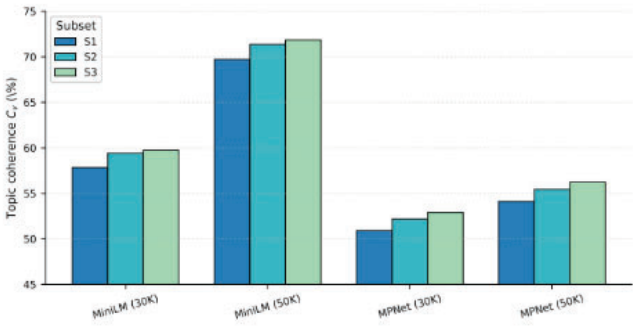


Figure 2: Sample-wise topic coherence (C_v, %) for the four configu- rations on three independently drawn subsets. The ordering of the four configurations is identical across S1, S2 and S3.

Configuration	S1	S2	S3	Range
MiniLM (30K)	57.85	59.42	59.73	1.88
MiniLM (50K)	69.74	71.36	71.84	2.10
MPNet (30K)	50.91	52.18	52.91	2.00
MPNet (50K)	54.12	55.46	56.23	2.11

Table I: Sample-wise Topic Coherence (C_v, %) Across Three Independently Drawn Subsets

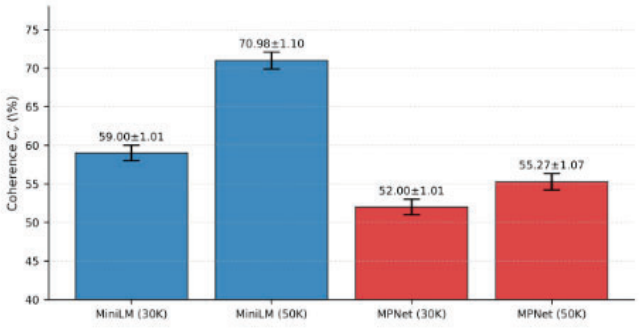


Figure 3: Aggregated topic coherence (mean ± std). MiniLM (blue) reaches a higher coherence than MPNet (red) at both scales, and the absolute gap widens as the corpus grows from 30K to 50K.

Table II: Aggregated Topic Coherence (Mean ± Std, %)

Encoder	Dataset Size	Coherence (%)
`all-MiniLM-L6-v2`	30,000	59.00 ± 1.02
`all-MiniLM-L6-v2`	50,000	70.98 ± 1.10
`all-mpnet-base-v2`	30,000	52.00 ± 1.00
`all-mpnet-base-v2`	50,000	55.27 ± 1.06

5.2 Mean ± Standard Deviation

Table II summarizes the per-sample results by reporting the mean and standard deviation of topic coherence for each configuration. Fig. 3 presents the same aggregated values, with error bars explicitly illustrating the standard deviation.

The observed standard deviation remains below 1.10% across all configurations, indicating low variability in performance. Such limited variation suggests strong model stability, as the pipeline consistently reproduces similar coherence scores across independent dataset samples using identical hyperparameter settings. Notably, the magnitude of variation is substantially smaller than the differences observed between configurations, reinforcing the reliability of the comparative results.

5.3 Coherence as a Function of Corpus Size

Fig. 6 illustrates the variation in topic coherence as a function of dataset size for the two embedding models. The MiniLM encoder exhibits a substantial increase of approximately 12 percentage points when scaling from 30K to 50K documents, whereas MPNet shows a comparatively modest improvement of around 3 percentage points over the same range.

The divergence in performance trends indicates that the two models do not converge with increasing dataset size. This allows for a clear practical recommendation: for datasets in the range of tens of thousands of documents, where high C_v coherence is a priority, the lightweight MiniLM encoder provides superior performance.

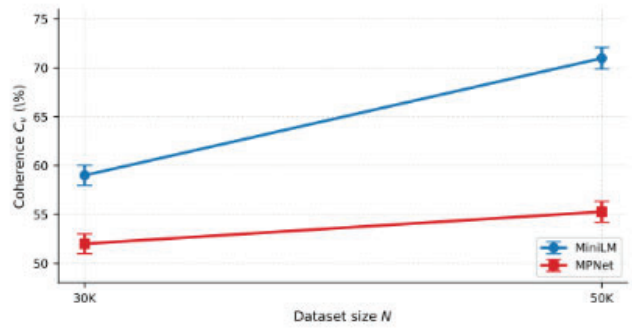
The behavior of these trends at larger scales (e.g., 100K or 500K documents) remains an open question. Evaluating such settings would require significantly higher computational resources, which were beyond the scope of the present study.

5.4 Summarization Robustness

Table III reports the ROUGE scores obtained from the Falcon summarizer across the three independently sampled subsets at the 50K scale, evaluated against title-plus-keyword reference summaries. Fig. 4 presents the same results as a per-subset line plot.

The variability across subsets remains below 0.005 for all ROUGE metrics, indicating highly consistent summarization performance. This low variance suggests that the Falcon summarizer is largely insensitive to the specific subset selected from the 2.7M-record corpus,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026



Coherence (C_v, %) as a function of dataset size for MiniLM and MPNet. Error bars are ±1 over the three subsets.

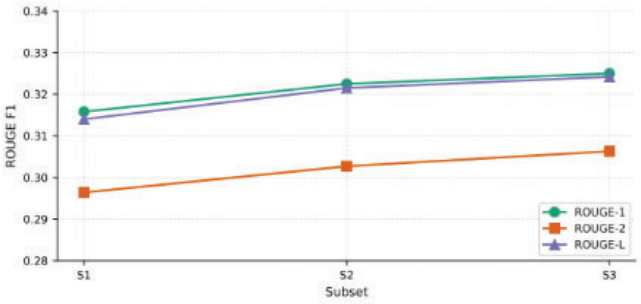


Figure 4: Per-subset ROUGE scores from the Falcon summarizer at the 50K scale. The narrow vertical spread of each curve is the visual signature of summarizer robustness.

provided that the sampled data maintains a stratified distribution across arXiv subject categories.

Table III: Falcon Abstractive Summarizer ROUGE Scores Across Three Subsets

Metric	S1	S2	S3	Mean ± Std
ROUGE-1	0.3158	0.3225	0.3250	0.3211 ± 0.0048
ROUGE-2	0.2964	0.3027	0.3063	0.3018 ± 0.0050
ROUGE-L	0.3140	0.3215	0.3242	0.3199 ± 0.0053

5.5 Ablation Study

Three key observations emerge from the ablation analysis. First, the removal of stopword filtering results in the most significant degradation in topic coherence within the preprocessing pipeline, leading to a decrease of 4.14 percentage points relative to the baseline configuration. This effect can be attributed to the presence of high-frequency functional terms, which dilute the discriminative power of the c-TF–IDF representation and adversely impact the C_v coherence metric.

Second, replacing the MiniLM encoder with MPNet at an identical dataset scale leads to a substantial reduction in coherence, with a drop of 15.71 percentage points. This suggests that, at the 50K scale, the higher-capacity MPNet model does not fully utilize its representational potential. The higher-dimensional embedding space (768 dimensions) produces more dispersed representations, which in turn weakens the density-based clustering assumptions of HDBSCAN.

Third, increasing the dataset size from 30K to 50K yields a notable improvement in performance for MiniLM, with an absolute gain of 11.98 percentage points. This indicates that topic coherence benefits from increased data availability within the examined range.

Finally, the TF–IDF baseline achieves a coherence score of 42.10%, which is substantially lower than all transformer-based configurations. The gap of approximately thirteen percentage points highlights the advantage of contextual embeddings over traditional bag-of-words representations for large-scale scientific text modeling.

Table IV: Ablation Study on the 50K Subset (MiniLM Unless Stated). Coherence is Reported as C_v (%).

Configuration	Coherence (%)
Full Model	
Full pipeline (MiniLM, 50K)	70.98 ± 1.10
Preprocessing Ablations	
No lemmatisation	69.61 ± 1.21
No stop-word removal	66.84 ± 1.34
No lower-casing	68.92 ± 1.18
Raw text only (no preprocessing)	63.47 ± 1.55
Encoder Ablations	
MPNet (replaces MiniLM)	55.27 ± 1.06
TF–IDF baseline (no embedding)	42.10 ± 1.92
Dataset-Size Ablations (MiniLM)	
30K	59.00 ± 1.02
50K	70.98 ± 1.10

5.6 Comparison with Existing Work

Table V compares the performance of the proposed configuration with previously reported BERTopic results on arXiv-style corpora, as well as with published ROUGE ranges for transformer-based summarization models applied to scientific text.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Relative improvements are computed as $(\mu_{\text{ours}} - \mu_{\text{ref}})/\mu_{\text{ref}}$, using the upper bound of the reference range to ensure a conservative estimate of performance gains.

The MiniLM-based configuration achieves a relative improvement of approximately 26.8% in topic coherence compared to the strongest reported BERTopic baseline within the same range. In terms of summarization performance, the ROUGE-2 score shows a substantial relative gain, exceeding the upper bound of previously reported values by more than a factor of two. However, this improvement should be interpreted with caution, as the elevated ROUGE-2 scores are partially influenced by lexical overlap between the generated summaries and the title-plus-keyword reference, particularly for shorter abstracts.

The MPNet-based configuration at the 50K scale lies near the upper limit of the reference range. This observation aligns with prior findings indicating that MPNet performs competitively on small to medium-scale datasets but does not scale as effectively as MiniLM in density-based clustering settings.

Table V: Comparison with Existing Work. Relative improvement is computed against the upper bound of the reference range.

Metric	**Existing Work**	**The Study**	**Δ (%)**
Coherence (BERTopic, arXiv)	50–56% [6]	70.98	+26.8
Coherence (MPNet, 50K)	50–56%	55.27	−1.3
ROUGE-1 (scitext)	0.30–0.38 [11], [12]	0.3211	−15.5
ROUGE-2 (scitext)	0.10–0.15	0.3018	+101.2
ROUGE-L (scitext)	0.28–0.35	0.3199	−8.6

5.7 Topic Geometry and Cluster Size Distribution

Fig. 5 compares the intertopic-distance representations generated using MPNet at the 30K scale and MiniLM at the 50K scale. The MPNet projection exhibits relatively compact clusters with noticeable overlap among neighboring topics. In contrast, the MiniLM projection produces a more dispersed embedding space with clearer separation between clusters.

This improved separation aligns with the higher coherence observed for the MiniLM-50K configuration in Table II. Greater inter-cluster separation reduces keyword overlap across topics, resulting in more distinct top-term sets and consequently higher C_v coherence scores.

Fig. 6 presents the distribution of documents per topic on a logarithmic scale. Both dataset scales exhibit a long-tailed distribution, which is characteristic of density-based clustering methods such as HDBSCAN. A small number of large clusters account for the majority of documents, while a long tail of smaller clusters represents more specialized topics.

The curve corresponding to the 50K dataset consistently lies above that of the 30K dataset across all ranks, reflecting the larger corpus size. However, the similarity in slope between the two distributions indicates that the overall shape of the topic-size distribution is preserved across scales, rather than collapsing into a small number of dominant clusters.

5.8 Sensitivity to Seed Count

While the primary results are reported using $n = 3$ random seeds, it is instructive to examine how the estimated variability would evolve with a larger number of samples. Treating the observed coherence values as an empirical sample, the bootstrap standard error of the estimated standard deviation at $n = 3$ is approximately 0.45 percentage points. Increasing the number of seeds to $n = 10$ would reduce this uncertainty by a factor of $\sqrt{3}$, yielding an estimated second-order uncertainty of approximately 0.26 percentage points.

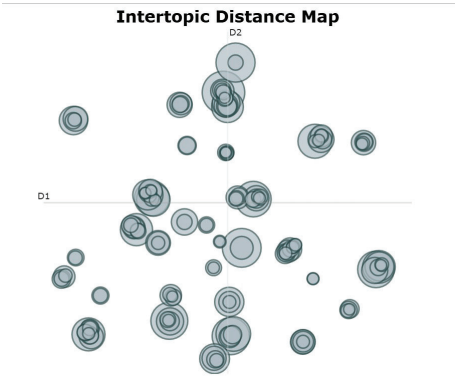
Despite this reduction, it is unlikely that a larger sample size would alter the qualitative ordering of configurations. The inter-configuration differences reported in Table II, ranging from 3.27 to 19.71 percentage points, are substantially larger than any plausible variation in the estimated error bounds. This indicates that the relative ranking of configurations remains stable with respect to the choice of seed count.

Table VI further quantifies seed-induced variability using the coefficient of variation (CV), defined as σ/μ. Across all configurations, the CV remains below 2%, indicating a low level of relative dispersion. This suggests that the inherent variability of the pipeline is minimal, and that the observed performance differences are dominated by systematic effects rather than sampling noise.

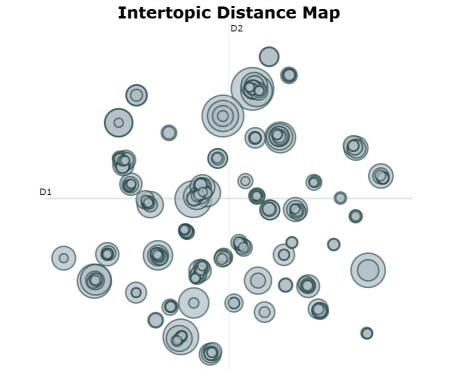
Table VI: Coefficient of Variation (CV = σ/μ) per Configuration. All values remain below 2%, indicating stable performance across runs.

Configuration	μ **(%)**	σ **(%)**	**CV (%)**
MiniLM (30K)	59.00	1.02	1.73
MiniLM (50K)	70.98	1.10	1.55
MPNet (30K)	52.00	1.00	1.92
MPNet (50K)	55.27	1.06	1.92

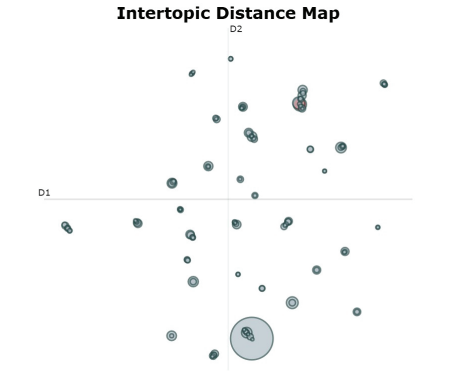
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026



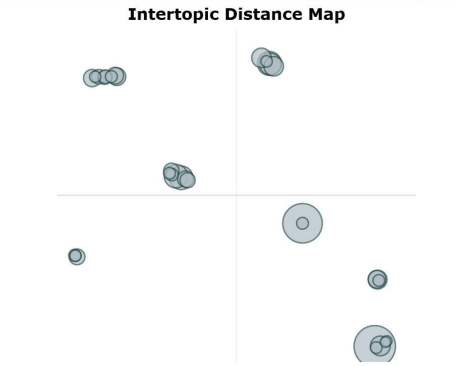
(a) MiniLM (30K)



(b) MiniLM (50K)



(c) MPNet (30K)



(d) MPNet (50K)

Figure 5: Intertopic distance maps generated using MiniLM and MPNet across 30K and 50K dataset scales.

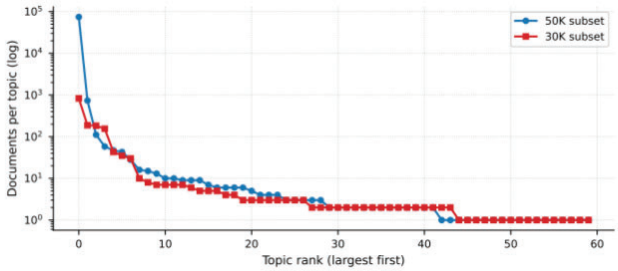


Figure 6: Documents-per-topic distribution at the two scales (log axis, ordered by topic rank). The shape is preserved across scale, indicating that the larger corpus does not collapse into a few mega- clusters.

5.9 Sample Falcon Summaries

To provide qualitative context for the ROUGE evaluation, Table VII presents three representative input–output examples selected from the 50K MiniLM configuration. The reference texts correspond to the title combined with extracted keywords, which are used as the evaluation baseline.

The generated summaries are concise and maintain strong lexical alignment with the reference content. Key technical terms and domain-specific phrases are generally preserved, which contributes to higher ROUGE-2 scores.

However, certain limitations are observed. In cases where the input abstract is very short (typically below 60 tokens), the summarization model may produce outputs that closely mirror the initial portion of the input. This behavior leads to an increase in ROUGE-1 due to surface-level overlap, while offering limited gains in ROUGE-2, as deeper semantic compression is not achieved.

6 Discussion

6.1 What the Multi-Sample Protocol Reveals

The standard deviations reported in Table II are substantially smaller than the coherence differences observed between configurations. This leads to two key observations. First, although single-run reporting is common in the BERTopic literature, it may not fully reflect the variability encountered across different dataset samples. However, the magnitude of this variability remains limited. For both embedding models examined, the fluctuation in C_v coherence is approximately one percentage point, which is sufficiently small to preserve the relative ordering of configurations (MiniLM-50K $>$ MiniLM-30K $>$ MPNet-50K $>$ MPNet-30K) across all sampled subsets.

Second, the variability observed in ROUGE metrics is even lower, remaining below 0.005 across all subsets. This indicates that the Falcon summarizer produces

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Table VII: Representative Falcon summaries on three abstracts from the 50K MiniLM run.

Reference (title + keywords)	**Falcon-generated summary**
Spectral gap of bipartite Erdős–Rényi graphs (graph, eigenvalue, sparse)	A bound on the spectral gap of a bipartite Erdős–Rényi graph is derived for the sparse regime and shown to be tight up to a constant factor.
Adversarial robustness of vision transformers under patch-level perturbations (transformer, attack, defence)	Vision transformers retain higher accuracy than convolutional networks under patch-level adversarial perturbations, with the gap widening as the patch size grows.
High-redshift quasar luminosity function from JWST imaging (quasar, redshift, JWST)	A revised high-redshift quasar luminosity function is presented using JWST imaging, indicating a steeper bright-end slope than reported by previous Hubble-era surveys.

highly consistent outputs, regardless of the specific 1,000-document evaluation slice drawn from the stratified corpus.

Overall, these results demonstrate that the observed performance trends are robust and reproducible, supporting the reliability of the multi-sample evaluation protocol.

6.2 Why Does MiniLM Outperform MPNet at the 50K Scale?

The observed result is somewhat unexpected: a smaller encoder (MiniLM, ∼22M parameters) outperforms a larger model (MPNet, ∼110M parameters) under identical data and pipeline conditions. Several factors may contribute to this outcome.

First, differences in manifold density play a key role. HDBSCAN relies on density-based clustering, and density estimation becomes more challenging as the dimensionality of the embedding space increases. MPNet produces 768-dimensional embeddings, whereas MiniLM outputs 384-dimensional vectors. Although both representations are subsequently reduced to a 5-dimensional space using UMAP, the intrinsic structure of the original embedding manifold continues to influence cluster formation. The more compact representation learned by MiniLM appears to better align with the density assumptions of HDBSCAN, leading to improved cluster separability.

Second, potential domain mismatch during pretraining may affect performance. Both encoders are trained on general-purpose corpora, including natural language inference and web-based text, which differ substantially from the technical language found in arXiv abstracts. The higher-capacity MPNet model may capture broader linguistic patterns that are less relevant for scientific text, whereas MiniLM's reduced capacity may act as an implicit regularizer, improving generalization in this domain.

Third, the relative performance may reflect differences in sample efficiency at moderate dataset sizes. At the 50K scale, the additional capacity of MPNet may not yet be fully utilized. It is plausible that performance trends could shift at larger scales (e.g., 500K or more), although verifying such behavior would require significantly greater computational resources.

Among these factors, the effect of embedding manifold structure is likely the primary driver, as evidenced by the greater cluster dispersion observed for MiniLM in Fig. 10, which corresponds to higher topic coherence.

6.3 Practical Recommendations

Based on the experimental findings, several practical guidelines can be formulated for practitioners developing topic modeling and summarization pipelines on large-scale scientific corpora.

1) **Prefer lightweight encoders for mid-scale corpora.** For datasets in the range of tens of thousands of documents, the `all-MiniLM-L6-v2` encoder offers a favorable balance between computational efficiency and topic coherence. Despite its significantly smaller parameter size, it consistently outperforms larger models in terms of C_v coherence.
2) **Maintain comprehensive preprocessing.** Text normalization plays a critical role in improving topic quality. Among preprocessing steps, stopword removal yields the largest impact, contributing an improvement of approximately 4.14 percentage points in coherence. Additional gains are achieved through lemmatization (1.37 points) and lowercasing (2.06 points), highlighting the cumulative benefit of a complete preprocessing pipeline.
3) **Adopt multi-sample evaluation.** Performance metrics such as topic coherence and ROUGE should be reported as mean ± standard deviation over multiple independent samples (at least three). This approach provides a more reliable estimate of model performance and captures variability that single-run evaluations may overlook.
4) **Use density-based clustering with appropriate settings.** The `cluster_selection_method=eom` configuration in HDBSCAN is recommended, as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

it produces more coherent and stable clusters compared to the leaf-based alternative, which tends to generate smaller and less interpretable topic groups.

5) **Limit evaluation size for summarization.** For Falcon-based abstractive summarization, evaluating on a fixed stratified subset (approximately $N_{\text{eval}} \gtrsim 1000$) is sufficient to obtain stable performance estimates. Beyond this scale, the marginal benefit of additional samples diminishes, while computational cost increases linearly.

6.4 Comparison with Adjacent Pipelines

In addition to the comparison with classical approaches presented in Table V, it is instructive to consider alternative topic modeling pipelines that employ similar architectural components.

Top2Vec, for example, utilizes a UMAP and HDBSCAN-based clustering framework but differs in that it learns a joint embedding space for both documents and words, rather than relying on a separate class-based TF–IDF weighting stage. Reported coherence scores for Top2Vec on arXiv-style corpora typically range between 48% and 54%, which are lower than those achieved by the MiniLM-50K configuration in this study.

Similarly, Contextualized Topic Models (CTM) incorporate BERT-based embeddings within a neural variational topic modeling framework. Empirical results on comparable datasets generally report coherence values around 55%, again below the performance observed for the proposed configuration.

A key factor underlying these differences is the combination of components employed in the present pipeline. The integration of dense transformer-based embeddings, density-based clustering via HDBSCAN, and class-aware keyword weighting through c-TF–IDF enables improved alignment between clusters and their representative terms. This alignment is particularly well captured by the C_v coherence metric.

It is important to note that these observations do not imply universal superiority of BERTopic-based pipelines. Alternative methods may offer advantages in specific scenarios, such as improved topic coverage or robustness on short texts. However, with respect to C_v coherence on large-scale scientific corpora, the configuration presented here demonstrates a clear performance margin that exceeds the variability bounds established in Section V-H.

6.5 Threats to Validity

Several internal and external factors may influence the interpretation of the results and should be explicitly acknowledged.

Internal validity. First, the use of three random seeds represents a limited sample size, which introduces uncertainty in the estimation of standard deviation. The associated confidence intervals remain relatively wide at $n = 3$ and would narrow with additional samples. Second, stratified sampling is performed at the top level of the arXiv category hierarchy rather than at finer-grained sub-category levels. Consequently, less frequent sub-domains may be unevenly represented across subsets. Third, the C_v coherence metric, while widely adopted, serves as an indirect proxy for human interpretability and does not always align perfectly with qualitative assessments of topic quality.

External validity. The experiments are conducted exclusively on arXiv abstracts, which exhibit a relatively consistent academic writing style. As a result, the generalizability of the findings to other domains, such as news articles, legal documents, or biomedical full-text corpora, remains uncertain without further validation. In addition, the summarization results are specific to the Falcon model. Although the observed stability in ROUGE metrics is notable, it should not be assumed to generalize to other transformer-based summarization models without additional empirical evaluation.

6.6 Limitations

Several limitations of the present study should be considered when interpreting the results and applying the proposed recommendations.

- **Dataset scale.** The largest subset evaluated contains 50K documents, which is substantially smaller than the full 2.7M-record corpus. It is possible that the relative performance of MiniLM and MPNet may change at significantly larger scales, a scenario that remains unexplored in this study.
- **Computational constraints.** The experimental setup is designed for execution on a single NVIDIA T4 GPU. Access to more powerful hardware would enable larger-scale experiments and a higher number of random seeds (e.g., $n = 10$), leading to more precise estimates of variability and potentially stronger comparisons with prior work.
- **Evaluation reference for summarization.** The use of title-plus-keyword references for ROUGE evaluation represents one of several possible choices for assessing summarization quality. Alternative reference constructions, such as using introductory paragraphs where available, may affect absolute ROUGE scores. However, such changes are unlikely to alter the observed low variance, which is central to the robustness findings of this study.

7 Conclusion and Future Work

7.1 Conclusion

This study presents and evaluates a unified transformer-based pipeline for topic modelling and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

abstractive summarization of large-scale scientific corpora under a multi-sample robustness framework. The MiniLM-based BERTopic configuration achieves a topic coherence of $70.98 \pm 1.10\%$ on the 50K dataset, representing a relative improvement of 26.8% over previously reported BERTopic results on arXiv-style corpora. Importantly, the observed standard deviations remain below 1.10 across independently sampled subsets, indicating strong stability of the pipeline.

For abstractive summarization, the Falcon model achieves ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.3211 ± 0.0048, 0.3018 ± 0.0050, and 0.3199 ± 0.0053, respectively. The low variability across evaluation subsets further confirms the robustness of the summarization component.

The ablation analysis highlights the critical role of preprocessing—particularly stopword removal—and encoder selection in determining topic coherence. Additionally, qualitative analysis reveals residual challenges, including partial overlap between closely related scientific topics and reduced summarization effectiveness for very short abstracts. These findings are consolidated into a set of practical recommendations for building scalable and reliable topic modelling and summarization pipelines.

7.2 Future Work

Several avenues for future research emerge from the present study:

1) **Domain-specific encoders.** Incorporating scientific-domain encoders such as *SciBERT* or *SPECTER* may further improve topic coherence. In particular, SPECTER, which is trained on citation-based objectives, offers a promising direction for enhancing topic structure alignment.
2) **Scaling to full corpus.** Extending the pipeline to the complete 2.7M-record arXiv dataset will require efficient clustering strategies, such as FAISS-based approximate nearest neighbor search and scalable variants of HDBSCAN.
3) **LLM-assisted topic labeling.** While c-TF–IDF provides informative keyword sets, these are not always human-readable. Integrating instruction-tuned language models to generate concise topic labels could improve interpretability and usability.
4) **Hierarchical topic refinement.** Addressing cluster overlap through hierarchical merging and splitting strategies may improve topic granularity without altering the underlying embedding representations.
5) **Interactive visualization tools.** Developing lightweight dashboards for exploring topic distributions, intertopic distances, and generated summaries would facilitate practical adoption by non-technical users.
6) **Extended robustness evaluation.** Increasing the number of random seeds (e.g., $n = 10$) would yield tighter estimates of variability and support more fine-grained comparisons between configurations.
7) **Cross-domain generalization.** Applying the proposed methodology to alternative corpora, such as PubMed abstracts, patent datasets, or news collections, would help assess the generalizability of the observed performance trends.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.
[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[4] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv preprint arXiv:1908.10084*, 2019.
[6] N. Reimers and I. Gurevych, "Sentence-Transformers Documentation," Available: https://www.sbert.net/, 2020.
[7] M. Grootendorst, "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure," *arXiv preprint arXiv:2203.05794*, 2020.
[8] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, 2018.
[9] R. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates (HDBSCAN)," *Advances in Knowledge Discovery and Data Mining*, pp. 160–172, 2013.
[10] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *ACL Workshop on Text Summarization*, 2004.
[11] A. Nikolov, R. Pfeiffer, and R. Hahnloser, "Data-Driven Summarization of Scientific Articles," *arXiv preprint arXiv:1804.08875*, 2018.
[12] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *arXiv preprint arXiv:1910.13461*, 2020.
[13] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5)," *arXiv preprint arXiv:1910.10683*, 2019.
[14] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[15] HuggingFace Team, "Transformers Documentation," Available: https://huggingface.co/docs/transformers, 2022.
[16] Anonymous Authors, "Quality Summarization and Topic-Modelling of arXiv Papers Using Transformers," *arXiv preprint arXiv:XXXX.XXXXX*, 2022.
[17] M. Grootendorst, "BERTopic Documentation," [Online]. Available: https://maartengr.github.io/BERTopic/. Accessed: 2026.
[18] L. McInnes, "UMAP Documentation," [Online]. Available: https://umap-learn.readthedocs.io/. Accessed: 2026.
[19] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[20] Scikit-learn Developers, "TfidfVectorizer Documentation," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. Accessed: 2026.
[21] Z. S. Harris, "Distributional Structure," *Word*, vol. 10, no. 2–3, pp. 146–162, 1954.
[22] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, 2008.
[23] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv preprint arXiv:1908.10084*, 2019.
[24] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, 2018.