Domain-Agnostic Neural Topic Modeling with Contextual Token-Level Semantic Graph Representation

Anonymous EMNLP submission

Abstract

PLM-powered neural topic models depend on general-domain pre-training, leading to degraded topic interpretability on specialized corpora. The root cause is geometric: domain-specific terms unseen during pre-training collapse into an indistinguishable region of the embedding space, and neither domain-specific re-training, word-level graph enrichment, nor parameter-efficient fine-tuning can restructure this space without inheriting the capacity ceiling of the underlying encoder. Our key insight is that a learnable graph layer operating on token-level PLM embeddings can acquire corpus-specific semantic structure that the frozen encoder lacks, because token-level graphs preserve document-local context that word-level representations discard and joint optimization with the topic objective reshapes embedding geometry directly from target-domain evidence. We instantiate this insight as DARTOPIC, a domain-agnostic framework that constructs token-level semantic graphs from frozen PLM embeddings and jointly trains a GNN encoder with variational topic inference. Across three benchmarks spanning general, biomedical, and legal domains, DARTOPIC consistently outperforms strong baselines in topic coherence and document clustering without any encoder fine-tuning, while demonstrating robustness to PLM choice and favorable runtime efficiency over fine-tuning-based alternatives.

1 Introduction

Pre-trained language models (PLMs) have fundamentally transformed neural topic modeling. By replacing sparse bag-of-words inputs with dense contextual document embeddings, recent models such as CombinedTM (Bianchi et al., 2021a), BERTopic (Grootendorst, 2022), and FASTopic (Wu et al., 2024b) achieve state-of-the-art topic coherence, surpassing classical probabilistic approaches (Blei et al., 2003) and earlier VAE-based neural topic models (Miao et al., 2016; Srivastava and Sutton, 2017) by a wide margin. However, this success rests on an implicit assumption: the semantic geometry of the PLM embedding space must adequately cover the vocabulary and relational structures of the target corpus. When this assumption holds, as in general-domain benchmarks whose distributions closely mirror PLM pre-training data, the resulting topics are coherent and interpretable. The critical question is what happens when it does not.

Model	Top-related words
FASTopic	develop, immunize, repeat, elicit, pseudomona, experiment, result, mechanic, correlate, occur, feature, convert
DARTopic	energy, cerevisiae, oxidative, peroxide, nitrate, hydrogen, carbon, kinetics, succinate, respiration, reduction, galactose

Table 1: Comparison of topics generated by FASTopic (Wu et al., 2024b) and DARTOPIC on BioASQ dataset. We report low-coherence words in RED font.

In specialized domains such as biomedicine, law the assumption breaks down. Domain-specific terms that appear rarely or never in general pre-training corpora are mapped to a narrow, poorly differentiated region of the embedding space, a pathology known as representation degeneration (Gao et al., 2019; Godey et al., 2024). For topic modeling, the consequence is direct: when the inference network receives document representations in which domain-critical terms are geometrically indistinguishable, it conflates semantically distinct concepts into incoherent topics. As illustrated in Table 1, even FASTopic (Wu et al., 2024b), a recent state-of-the-art model, produces topics on biomedical text that mix domain-relevant terms with generic, low-coherence words such as *repeat*, *mechanic*, and *feature*, confirming that fixed PLM representations fail under domain shift.

Existing remedies each address only a partial aspect of this problem. Domain-specific PLMs such as BioBERT (Lee et al., 2020) can reshape the embedding geometry but require large-scale pre-

training that is unavailable for many specialized fields. Graph-based neural topic models (Xu et al., 2023; Adhya and Sanyal, 2024; Liu et al., 2025) enrich representations with structural information, yet they operate on word-level graphs with fixed, context-independent node features and therefore cannot capture the contextual variation of the same term across documents. Parameter-efficient fine-tuning via prefix tuning (Li and Liang, 2021; Akash and Chang, 2024) adapts PLM representations at low cost, but if the frozen encoder cannot geometrically distinguish domain-specific terms in the first place, a lightweight prefix cannot overcome this representational ceiling. None of these approaches resolves the fundamental coupling between topic quality and PLM pre-training coverage.

Our key insight is that this coupling can be broken by interposing a learnable, corpus-specific graph layer between frozen token embeddings and topic inference. Rather than treating PLM outputs as fixed final representations or attempting to adapt the encoder itself, we propose to use them as initial node features for a dynamically constructed token-level semantic graph, which a graph neural network (GNN) then refines jointly with the topic objective. Token-level construction captures contextual semantics that word-level graphs miss, because each token node carries document-specific neighborhood structure. Joint GNN-topic optimization discovers corpus-specific relational patterns that the frozen PLM never observed during pre-training. And because the GNN operates over the full token embedding space, it is not bounded by the narrow capacity of a prefix adapter.

Based on this insight, we propose DARTopic (**D**omain-**A**gnostic neu**R**al **Topic** modeling), a lightweight framework that constructs token-level semantic graphs from frozen PLM embeddings and jointly trains a GCN encoder with a variational topic inference module. As shown in Table 1, DARTopic generates topics consistently organized around coherent biomedical concepts such as *oxidative stress*, *respiration*, and *carbon metabolism*, achieved without any domain-specific pre-training or fine-tuning.

Our contributions are as follows:

- We show that the performance ceiling of PLM-powered topic models under domain shift arises from fixed embedding geometry, and propose DARTopic, which breaks this coupling by interposing a corpus-learned token-level semantic graph between frozen PLM embeddings and variational topic inference.
- We reveal that token-level semantic graphs whose edges reflect embedding similarity capture corpus-specific distributional structure that neither word-level graphs nor PLM representations encode, rendering topic quality largely invariant to PLM choice and scale.
- We demonstrate that because the graph layer learns relational structure directly from target-domain evidence, it compensates for distributional gaps in any frozen encoder, enabling a single lightweight architecture to match or exceed domain-specific PLMs across general, biomedical, and legal benchmarks in both topic coherence and document clustering.

2 Related Works

2.1 PLM-Powered Neural Topic Models

Early neural topic models reformulated topic inference within VAE frameworks (Miao et al., 2016; Srivastava and Sutton, 2017), replacing the intractable posteriors of LDA (Blei et al., 2003) with amortized inference networks, though their reliance on bag-of-words representations limited topic coherence. Subsequent work incorporated static word embeddings such as GloVe (Pennington et al., 2014) to encourage semantically similar words to share topic assignments, partially mitigating this limitation.

The introduction of contextual PLMs marked a more substantial shift: CombinedTM (Bianchi et al., 2021a) combines dense Sentence-BERT (Reimers and Gurevych, 2019) embeddings with BoW features, while BERTopic (Grootendorst, 2022) adopts a clustering-based approach leveraging contextual embeddings to group semantically coherent documents. More recent models such as FASTopic (Wu et al., 2024b) and NeuroMax (Pham et al., 2024) further advance this paradigm through optimal transport-based alignment and mutual information maximization, respectively, achieving strong performance on general-domain benchmarks.

Despite their effectiveness, PLM-powered models are fundamentally constrained by their pre-training domain coverage. When applied to specialized domains such as biomedical or legal text, the divergence between pre-training and target distributions leads to degraded topic coherence. Domain-

specific PLMs (Lee et al., 2020) can partially address this gap but require costly large-scale pre-training that is not always feasible. PVTM (Akash and Chang, 2024) mitigates domain shift via prefix tuning (Li and Liang, 2021) of the PLM encoder, yet its quality remains tightly coupled to the underlying PLM capacity. In particular, lightweight models with limited prior knowledge struggle to produce coherent topics under significant domain shift, regardless of the tuning strategy applied.

2.2 Graph-Based Neural Topic Models

GNNs have demonstrated strong representational capacity across NLP tasks (Wu et al., 2023a), and several lines of work have explored their application to topic modeling. Early graph-based models focused on alleviating data sparsity in short-text settings, where limited token co-occurrence information impedes reliable topic inference (Cheng et al., 2014; Zhu et al., 2018; Wang et al., 2021a), enriching structural signals via corpus-level word co-occurrence graphs. Zhou et al. (2020) extended this by explicitly modeling document-level relational structures at the corpus level.

More recent work integrates GNNs with pre-trained embeddings: GINopic (Adhya and Sanyal, 2024) employs graph isomorphism network (GIN) over word graphs to capture higher-order relational structures, CGTM (Liu et al., 2025) fuses PLM-based contextual document embeddings with graph-structured word representations, and Meta-CETM (Xu et al., 2023) uses a variational graph auto-encoder over dependency-parsed word graphs for low-resource settings. However, all of these approaches rely on fixed pre-trained embeddings, propagating domain-mismatched semantics through the graph regardless of architecture.

Taken together, PLM and graph-based topic models share a fundamental vulnerability: their representations are anchored to fixed pre-training knowledge that may not reflect the semantic structures of the target domain. DARTOPIC addresses this gap by constructing token-level semantic graphs whose node representations are initialized from a frozen PLM and jointly refined with the topic inference objective, learning corpus-adaptive representations directly from target data without reliance on domain-specific pre-training.

3 Proposed Methodology

3.1 Problem Statement

Notations. Let $\mathcal{D} = \{d_1, d_2, ..., d_N\}$ denote a corpus consisting of N documents. Each document d_i is represented as a sequence of tokens $d_i = (t_{i1}, t_{i2}, ..., t_{iL_i})$, where L_i denotes the number of tokens in the document and each token t_{ij} corresponds to a word from a vocabulary set $\mathcal{V}$ with size $V = |\mathcal{V}|$. In topic modeling, documents are commonly represented using a bag-of-words (BoW) representation. Let $\mathbf{x}_i \in \mathbb{R}^V$ denote the BoW vector for document d_i, where $\mathbf{x}_{iv}$ indicates the frequency of word v in the document. A topic model assumes the existence of K latent topics. Each document is associated with a document-topic distribution $\boldsymbol{\theta}_i \in \Delta^K$, where Δ^K denotes a K-dimensional probability simplex. Each topic k is characterized by a topic-word distribution $\boldsymbol{\beta}_k \in \Delta^V$, and the topic-word matrix is denoted as $\boldsymbol{\beta} \in \mathbb{R}^{K \times V}$ where β_{kv} represents the probability of word v under topic k. The log-likelihood of document d_i under the topic model is given by (Blei et al., 2003):

$$\log p(d_i \mid \theta_i, \beta) = \sum_{v=1}^{V} x_{iv} \log \left(\sum_{k=1}^{K} \theta_{ik} \beta_{kv} \right) \quad (1)$$

Topic Modeling with Document Representations. Recent neural topic models often incorporate pre-trained language encoders to obtain semantic document representations (Bianchi et al., 2021b,a; Pham et al., 2024; Wu et al., 2024b; Akash and Chang, 2024; Liu et al., 2025). Let $f(\cdot)$ denote a pre-trained encoder that maps a document d_i into a dense vector representation $\mathbf{z}_i \in \mathbb{R}^H$, where H denotes the embedding dimension, i.e., $\mathbf{z}_i = f(d_i)$. Such encoders are typically trained on large-scale corpora and provide contextualized semantic representations of documents. In neural topic modeling frameworks, the document representation $\mathbf{z}_i$ is used as input to an inference network that estimates the document-topic distribution. Formally, the document-topic distribution is obtained as $\boldsymbol{\theta}_i = g(\mathbf{z}_i)$, where $g(\cdot)$ denotes a inference network. The topic model then learns topic-word distributions $\boldsymbol{\beta}$ by maximizing the log-likelihood of documents given the inferred topic mixtures:

$$\max_{\beta, g} \sum_{i=1}^{N} \log p(d_i \mid \theta_i, \beta), \quad \text{where } \theta_i = g(f(d_i)). \quad (2)$$

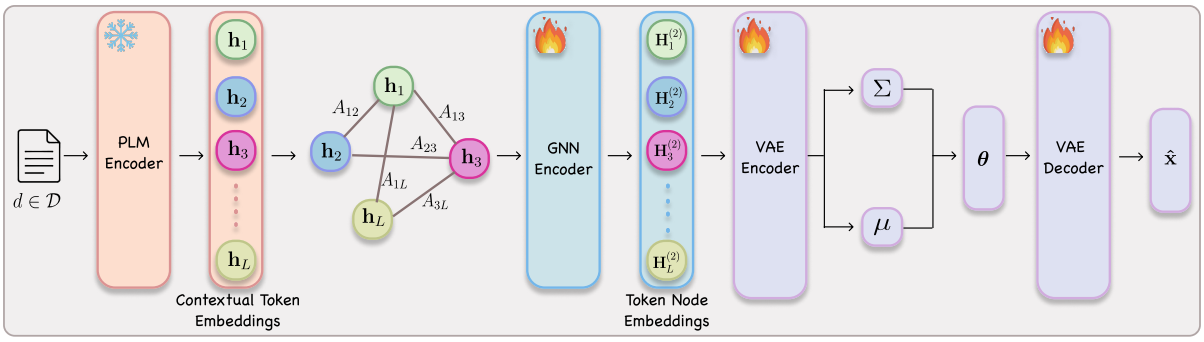

Figure 1: The overall architecture of our proposed DARTOPIC framework.

Since contextualized document representations obtained from a pre-trained encoder often capture rich semantic information, they can significantly contribute to the discovery of coherent topics. Such contextual information can help the topic model better identify semantically related words and improve the overall quality of the learned topic-word distributions. Therefore, our main goal is to effectively leverage contextual information derived from the pre-trained encoder while adapting it to heterogeneous domain corpora.

3.2 Token-Level Semantic Graph Construction

To effectively represent documents from heterogeneous domains, we construct a token-level graph representation by leveraging token embeddings obtained from a pre-trained language model (PLM). Inspired by (Angelov and Inkpen, 2024; Donabauer and Kruschwitz, 2025), we represent each document as a graph where nodes correspond to tokens and edges capture relationships between tokens within the same document.

The motivation for adopting a token-level graph representation is closely related to the domain-agnostic setting considered in this work. When documents originate from diverse domains, contextual information available in the training corpus may be insufficient to learn high-quality word representations without extensive fine-tuning. In such scenarios, relying solely on corpus-level statistics can lead to suboptimal semantic representations.

To alleviate this issue, we explicitly leverage the knowledge embedded in a PLM by constructing graphs based on token embeddings extracted from the PLM. This design allows the model to incorporate rich semantic knowledge acquired during large-scale pretraining while maintaining robustness under low-resource or cross-domain conditions. Unlike the token-level graph construction strategy (Donabauer and Kruschwitz, 2025), which connects tokens based on local co-occurrence windows (i.e., an n-hop or sliding-window graph), we construct a semantic graph where edges are determined by the similarity between token embeddings. The difference in graph construction reflects the different inductive biases required by the target task. Local window-based graphs emphasize positional proximity and syntactic relations, which are beneficial for text classification tasks (Wang et al., 2021b, 2024). In contrast, topic modeling aims to uncover latent semantic structures among words across a document. Therefore, a graph that directly captures semantic similarity provides a more appropriate inductive bias for topic modeling task. A detailed empirical analysis comparing different graph construction strategies is presented in Appendix D.

To obtain token embeddings, we employ the lightweight small-scale pre-trained model[1], which offers a good trade-off between computational efficiency and representation quality, making it suitable for low-resource settings. The tokenizer used in our framework is defined as `WordPiece` tokenizer (Song et al., 2021). Given a document containing L tokens, we denote the token embedding matrix as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L] \in \mathbb{R}^{L \times d}$, where $\mathbf{h}_i \in \mathbb{R}^d$ represents the contextual embedding of the i-th token obtained from the PLM. We construct an undirected and weighted token-level semantic graph $G = (V, E)$, where each node $v_i \in V$ corresponds to a token embedding $\mathbf{h}_i$. The edges are defined based on cosine similarity between token embeddings. Specifically, the adjacency matrix $\mathbf{A} \in \mathbb{R}^{L \times L}$ is defined as follows:

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

$$A_{ij} = \begin{cases} \cos(\mathbf{h}_i, \mathbf{h}_j), & \text{if } \cos(\mathbf{h}_i, \mathbf{h}_j) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $\cos(\mathbf{h}_i, \mathbf{h}_j)$ denotes the cosine similarity between token embeddings and τ is a predefined similarity threshold controlling graph sparsity. The constructed token-level semantic graph is subsequently processed by a graph neural network to learn contextualized node representations, which are then used for topic inference.

3.3 Graph Representation Learning

We employ a two-layer graph convolutional network (GCN) (Kipf and Welling, 2017) to propagate semantic information over the token graph and obtain node embeddings that encode both token semantics and document-specific structural relations.

Let a document graph be denoted as $G = (V, E)$ with adjacency matrix $\mathbf{A} \in \mathbb{R}^{L \times L}$ and node feature matrix $\mathbf{X} \in \mathbb{R}^{L \times d}$, where L is the number of tokens in the document and d is the dimensionality of the PLM-based token embeddings. Following the standard GCN formulation (Kipf and Welling, 2017), we apply symmetric normalization with self-loops: $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\tilde{\mathbf{D}}^{-\frac{1}{2}}$. The first layer computes hidden representations as $\mathbf{H}^{(1)} = \mathrm{ReLU}(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}^{(1)} + \mathbf{b}^{(1)})$, and the second layer produces the final node representations as

$$\mathbf{H}^{(2)} = \mathrm{ReLU}(\hat{\mathbf{A}}\mathbf{H}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)}), \quad (4)$$

where $\mathbf{W}^{(\cdot)}$ and $\mathbf{b}^{(\cdot)}$ are learnable parameters of each layer. The resulting matrix $\mathbf{H}^{(2)} \in \mathbb{R}^{L \times d'}$ contains the contextualized embeddings of all token nodes. Graph convolution enables each token representation to be enriched by semantically related neighbors in the document graph, which is particularly desirable for topic modeling where semantically coherent groups of tokens should be aggregated into document-level latent representations. To derive a fixed-length document representation, we apply mean pooling over all token nodes:

$$\mathbf{g} = \frac{1}{L} \sum_{i=1}^{L} \mathbf{H}_i^{(2)}, \quad (5)$$

where $\mathbf{H}_i^{(2)}$ denotes the contextualized embedding of the i-th token node.

3.4 DARTOPIC Framework

In this subsection, we present the overall DARTOPIC framework. The overall architecture is illustrated in Figure 1, our framework is designed to remain lightweight while effectively leveraging PLM-derived token semantics for neural topic modeling on various domains. To this end, DARTOPIC consists of a graph encoder for token-level document representation learning and a compact VAE-based topic inference module.

Lightweight Architecture. A key design principle of DARTOPIC is to keep the overall architecture lightweight. Unlike prior contextualized neural topic models that use dense contextual representations together with sparse lexical features such as TF-IDF or bag-of-words (BoW) vectors (Bianchi et al., 2021b; Adhya and Sanyal, 2024), we do not employ any joint input structure with sparse document embeddings. Such hybrid architectures often require additional projection layers or high-dimensional transformations, which increase the number of learnable parameters. Instead, we use only the graph-level document representation $\mathbf{g} \in \mathbb{R}^{d'}$ obtained by mean pooling over all token nodes, as defined in Eq.5. By directly using $\mathbf{g}$ as the input of the topic model, DARTOPIC avoids the need to concatenate additional sparse document features, reducing computational cost and minimizing the number of trainable parameters.

Topic Inference. Following (Bianchi et al., 2021b), we employ a variational autoencoder (VAE) to infer the latent topic structure of the documents. The encoder maps $\mathbf{g}$ to the parameters of a Gaussian posterior distribution over the latent variable $\mathbf{z} \in \mathbb{R}^K$, where K denotes the number of topics. Concretely, the encoder produces the mean vector $\boldsymbol{\mu}$ and log-variance vector $\log \boldsymbol{\sigma}^2$ as $\boldsymbol{\mu} = f_\mu(\mathbf{g})$ and $\log \boldsymbol{\sigma}^2 = f_\sigma(\mathbf{g})$, where $f_\mu(\cdot)$ and $f_\sigma(\cdot)$ denote neural transformations of VAE. To enable gradient-based optimization, we adopt the reparameterization trick (Kingma and Welling, 2013) and sample the latent variable as $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ denotes element-wise multiplication. The sampled latent variable is then transformed into the document-topic distribution $\boldsymbol{\theta} \in \mathbb{R}^K$ through a softmax function, i.e., $\boldsymbol{\theta} = \mathrm{softmax}(\mathbf{z})$.

Overall Training Objective. The GNN encoder and the VAE-based neural topic model are trained jointly in an end-to-end unsupervised manner. Let $\mathbf{x}_{\text{bow}} \in \mathbb{R}^V$ denote the target document word-count

vector over a vocabulary set of size V, and let $\hat{\mathbf{x}} \in \mathbb{R}^V$ denote the reconstructed word distribution generated from the inferred topic mixture $\boldsymbol{\theta}$ through the decoder. The reconstruction loss is defined as the negative log-likelihood of the observed document under the reconstructed word distribution:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^{N} \left[-(\mathbf{x}_i)^\top \log \left(\text{softmax}(\boldsymbol{\beta}\boldsymbol{\theta}_{\boldsymbol{i}}) \right) \right]. \tag{6}$$

To regularize the latent topic space, we impose a standard Gaussian prior on $\mathbf{z}$, i.e., $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The Kullback–Leibler (KL) divergence between the approximate posterior $q(\mathbf{z} \mid \mathbf{g})$ and the prior is defined as follows:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\big(q(\mathbf{z} \mid \mathbf{g}) \,\|\, p(\mathbf{z})\big). \tag{7}$$

The overall training objective of DARTopic is given by $\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$. By optimizing this objective, the GCN-based graph encoder and the VAE-based topic model are trained simultaneously. Importantly, the token-level graph representation learning module is not optimized with a separate graph-specific supervision signal; instead, it is directly guided by the neural topic modeling objective. This allows the learned token and document representations to be specialized for latent topic inference while preserving the lightweight nature of the overall framework.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on three benchmarks with heterogeneous domain texts: 20NewsGroup (20NG) (*General*), BioASQ (*Biomedical*) and BillSum (Bills) (Kornilova and Eidelman, 2019) (*Legal*). The statistics of the processed datasets are shown in Table 7 in Appendix A.

Baselines. We compare our framework with the following neural topic models: (1) **FASTopic** (Wu et al., 2024b), the recent state-of-the-art neural topic model using embedding regularization with optimal transport, (2) **BERTopic** (Grootendorst, 2022), a standard clustering-based topic model, (3) **ProdLDA** (Srivastava and Sutton, 2017), a first VAE-based standard neural topic model, (4) **ZeroShotTM** (Bianchi et al., 2021b) which replaces input data with PLM-based document representation, (5) **CombinedTM** (Bianchi et al., 2021a) which integrates BoW embeddings with PLM representations, (6) **PVTM** (Akash and Chang, 2024), a VAE-based neural topic model with parameter-efficient fine-tuning method, (7) **GINopic** (Adhya and Sanyal, 2024), a graph-based neural topic model with graph isomorphism network, (8) **CGTM** (Liu et al., 2025), a neural topic model utilizing both Graph and PLMs, (9) **NeuroMax** (Pham et al., 2024), a lightweight neural topic model using PLM only in training time. The implementation details of baseline models are shown in Appendix B.

Evaluation. We use automatic evaluation metrics such as NPMI (Lau et al., 2014) (topic coherence) and TU (Nan et al., 2019) (topic diversity). For topic coherence evaluation on domain-specific texts, we follow (Han et al., 2023) and use the NPMI-In score. To evaluate coherence and diversity as a whole, we also use Topic Quality (TQ) (Dieng et al., 2020). Furthermore, to evaluate the quality of document-topic distributions, we conduct document clustering task and report Purity and normalized mutual information (NMI) (Manning et al., 2008). NMI measures the agreement between predicted topic assignments and ground-truth labels, while Purity assesses how consistently each cluster contains documents from a single class. Please refer to Appendix C for the detailed.

4.2 Main Results

Topic-Word Distribution Quality. We report the mean value over 5 random runs with 50 topics. Table 2 reports the topic-word distribution quality results on three datasets with different domains. Overall, DARTopic achieves consistently strong performance across all datasets, demonstrating its robustness to both domain variation and document length. In particular, the average document lengths of the three corpora are substantially different, i.e., 48.02 for 20NG, 7.44 for BioASQ, and 76.28 for Bills, yet DARTopic maintains superior or highly competitive results in all cases. This indicates that the proposed framework can effectively capture topic structure regardless of whether the target corpus consists of short or long documents. PVTM shows strong overall performance by leveraging parameter-efficient fine-tuning (PEFT) such as prefix tuning (Li and Liang, 2021) to adapt pre-trained language models to domain-specific corpora. NeuroMax, in contrast, achieves perfect or near-perfect topic diversity through its embedding

Dataset	Domain	Metric	ProdLDA	FASTopic	BERTopic	ZeroShotTM	CombinedTM	NeuroMax	PVTM	GINopic	CGTM	DARTopic (ours)
20NG	*General*	NPMI	0.1581	0.2624	0.2300	0.2604	0.1670	0.1833	0.2707	0.2172	0.2520	**0.2736**
		TU	0.7613	0.8813	0.4547	0.9227	0.7853	**1.0000**	0.9133	0.4747	0.5227	0.9360
		TQ	0.1203	0.2313	0.1046	0.2403	0.1311	0.1833	0.2473	0.1031	0.1317	**0.2561**
BioASQ	*Biomedical*	NPMI	0.0142	0.0789	0.1653	0.1448	0.0120	0.0094	0.1596	0.0213	0.1377	**0.1641**
		TU	0.8213	0.4093	0.7320	0.8640	0.8080	**1.0000**	0.9730	0.8013	0.6947	0.9760
		TQ	0.0117	0.0323	0.1210	0.1251	0.0097	0.0094	0.1553	0.0171	0.0957	**0.1602**
Bills	*Legal*	NPMI	0.0387	0.2291	0.1795	0.2525	0.0425	0.0506	0.2609	0.0371	0.2318	**0.2685**
		TU	0.7853	0.9720	0.4787	0.9000	0.7813	**1.0000**	0.9133	0.8080	0.6013	0.9240
		TQ	0.0304	0.2226	0.0859	0.2272	0.0332	0.0506	0.2383	0.0300	0.1394	**0.2481**

Table 2: Main results on topic-word distribution quality across three domains: *general* (20NG), *biomedical* (BioASQ) and *legal* (Bills). The best performance in each row is shown in **bold** and the second-best is underlined.

clustering regularization (ECR) (Wu et al., 2023b), but this comes at the cost of substantially weaker topic coherence. Compared with these methods, DARTOPIC provides a more balanced and reliable trade-off between coherence and diversity. Specifically, DARTOPIC achieves the best NPMI and TQ scores on all three datasets, while also maintaining highly competitive TU scores, demonstrating that it can generate topics that are not only diverse but also semantically coherent. These results suggest that DARTOPIC is not merely specialized for a particular domain or corpus condition, but instead offers a generally effective topic modeling framework across heterogeneous datasets. Its consistent gains over strong baselines further highlight the advantage of graph-based document representation learning for improving topic quality in neural topic modeling.

Doc-Topic Distribution Quality. We evaluate document-topic distributions via a downstream clustering task. BERTopic cannot infer document-topic distributions (Wu et al., 2024a) and is therefore excluded; the Bills dataset lacks document labels and is also omitted. Following Adhya and Sanyal (2024), each document is assigned to its highest-probability topic. Table 4 reports the results. DARTOPIC shows strong performance on both 20NG and BioASQ, with a clearer advantage on BioASQ, where domain-specific terms are poorly separated in the frozen PLM space and the graph layer's corpus-learned structure provides greater discriminative benefit.

	NPMI	TU	TQ	Purity	NMI
`MiniLM`	0.1641	0.9760	0.1602	0.5235	0.3833
`Roberta-large`	0.1660	0.9680	0.1606	0.4738	0.3492
`Qwen-0.6B`	0.1644	0.9653	0.1587	0.4829	0.3555
`BioBERT`	0.1690	0.9693	0.1638	0.5061	0.3696

Table 3: Comparison of DARTOPIC performance on BioASQ dataset with different pre-trained models.

	20NG		BioASQ	
	Purity	NMI	Purity	NMI
ProdLDA	0.1186	0.1652	0.2032	0.1030
FASTopic	0.4743	**0.4089**	0.4339	0.3650
ZeroshotTM	0.4555	0.3657	0.3217	0.3312
CombinedTM	0.1963	0.2262	0.2466	0.1566
NeuroMax	0.1889	0.2530	0.2561	0.2802
PVTM	**0.4927**	0.3917	0.5174	0.3824
GINopic	0.2122	0.2523	0.2440	0.1619
CGTM	0.4562	0.3900	0.4809	0.3626
DARTopic (ours)	0.4905	0.3968	**0.5235**	**0.3833**

Table 4: Performance comparison on document clustering task using document-topic distribution vector. The best-performing method is highlighted in **bold** and second-best performance is underlined.

5 Analysis

5.1 Robustness Analysis

We analyze the robustness of neural topic models under different PLM choices, including general-purpose PLMs and a domain-specific PLM, `BioBERT`. Tables 3 and 5 report the results on the BioASQ dataset.

Table 3 shows the performance of DARTOPIC with various PLMs, ranging from lightweight models such as `MiniLM` to larger or domain-specific encoders such as `RoBERTa-large`, `Qwen-0.6B`, and `BioBERT`. The results demonstrate that DARTOPIC achieves consistently strong performance across different PLM choices, with only limited variation. This indicates that DARTOPIC is robust to both the scale and type of the underlying encoder, making it practical even when smaller and more efficient PLMs are preferred.

Table 5 further compares recent PLM-based neural topic models using `MiniLM` and `BioBERT`. The results show that existing methods exhibit different levels of sensitivity to PLM choice. For models that combine PLM-based document embeddings with additional representations such as BoW or word embeddings, including CombinedTM, CGTM, and

Methods	PLMs	NPMI	TU	TQ	Purity	NMI
FASTopic	MiniLM	0.0789	0.4093	0.0323	0.4339	**0.3650**
FASTopic	BioBERT	**0.1146**	**0.9920**	**0.1137**	**0.4439**	0.3375
ZeroshotTM	MiniLM	0.1448	0.8640	0.1251	0.3217	0.3312
ZeroshotTM	BioBERT	**0.1628**	**0.9680**	**0.1576**	**0.4967**	**0.3652**
CombinedTM	MiniLM	0.0120	**0.8080**	0.0097	**0.2466**	**0.1566**
CombinedTM	BioBERT	**0.0130**	0.7813	**0.0102**	0.1108	0.1157
BERTopic	MiniLM	**0.1653**	**0.7320**	**0.1210**	–	–
BERTopic	BioBERT	0.1601	0.7013	0.1123	–	–
NeuroMax	MiniLM	**0.0094**	**1.0000**	**0.0094**	**0.2561**	**0.2802**
NeuroMax	BioBERT	0.0073	**1.0000**	0.0073	0.2490	0.1994
PVTM	MiniLM	0.1596	**0.9730**	0.1553	**0.5174**	**0.3824**
PVTM	BioBERT	**0.1716**	0.9707	**0.1666**	0.4908	0.3651
CGTM	MiniLM	**0.1377**	**0.6947**	**0.0957**	**0.4809**	**0.3626**
CGTM	BioBERT	0.1366	0.6040	0.0825	0.4632	0.3437

Table 5: PLM dependency of PLM-powered neural topic models performance on BioASQ dataset. The better result is highlighted in **bold** for each metric.

NeuroMax, MiniLM often outperforms BioBERT. In contrast, PLM-driven models such as FASTopic and ZeroShotTM benefit substantially from replacing MiniLM with BioBERT.

Overall, these results suggest that many baseline methods are highly dependent on the selected PLM, whereas DARTOPIC maintains stable and competitive performance regardless of the encoder. This demonstrates that the proposed graph-based framework can preserve topic quality even with relatively small or generic PLMs, offering a more robust solution for practical neural topic modeling.

	20NG		BioASQ		Bills	
	Training	Inference	Training	Inference	Training	Inference
PVTM	13.55s	12.79s	2.55s	2.22s	9.82s	9.56s
DARTopic	7.67s	7.65s	2.59s	2.44s	8.67s	8.52s

Table 6: Runtime performance comparison with PVTM.

5.2 Runtime Performance

We compare PVTM and DARTOPIC in terms of per-epoch training time and full-corpus inference time, where all experiments were conducted on a single NVIDIA H100 GPU. In Table 6, the results show that DARTOPIC is consistently faster than PVTM on 20NG and Bills, while maintaining comparable efficiency on BioASQ. Notably, the advantage of DARTOPIC becomes more pronounced on datasets with longer documents, such as 20NG and Bills. This trend suggests that DARTOPIC is particularly effective for long-document processing, where computational bottlenecks tend to increase with document length. While PVTM relies on Prefix Tuning to mitigate domain mismatch, such adaptation introduces additional trainable parameters and computational overhead. In contrast, DARTOPIC uses a frozen pretrained encoder, enabling a simpler and more efficient architecture. As a result, DARTOPIC reduces the runtime burden associated with increasing document length, demonstrating that efficient domain-agnostic topic modeling can be achieved without fine-tuning. Notably, as shown in Table 2, DARTOPIC achieves this efficiency while also outperforming PVTM in topic quality across all three datasets.

5.3 Summary of Key Findings

Based on our experiments and analysis, we highlight three key findings:

Decoupling Topic Quality from PLM Capacity DARTOPIC achieves the best NPMI and TQ across all three domains using only a frozen PLM, demonstrating that corpus-learned token-level graph representations can replace domain-specific pre-training.

Robustness to PLM Choice. Unlike existing models that are sensitive to encoder selection, DARTOPIC maintains stable performance across PLMs of varying scale and domain specificity.

Efficiency without Fine-Tuning. Compared to PVTM which relies on prefix tuning, DARTOPIC achieves superior topic quality while being faster, confirming that graph representation learning is a more effective alternative to fine-tuning for cross-domain topic modeling.

6 Conclusion

We introduced DARTOPIC, which decouples topic quality from PLM pre-training coverage by interposing a corpus-learned token-level semantic graph between frozen embeddings and variational topic inference. Experiments across general, biomedical, and legal domains confirm consistent improvements in topic coherence and document clustering, with robustness to PLM choice and lower runtime than fine-tuning alternatives. Our results suggest a broader principle: when a frozen encoder lacks domain-specific geometric structure, a jointly optimized graph layer can reconstruct it from target-corpus evidence alone.

Limitations

Our proposed DARTOPIC framework shows strong effectiveness across heterogeneous domains, consistently improving topic quality and document

representations with an efficient small-scale PLM-powered design. This work has several limitations that should be addressed in future research. Although DARTOPIC shows consistent gains across *general*, *biomedical*, and *legal* benchmarks, our experiments are limited to English corpora. Because the framework relies on token-level semantic graph construction rather than domain-specific resources, it is potentially applicable to broader settings, but multilingual validation remains for future work. In addition, our evaluation mainly uses standard automatic topic coherence metrics. While these protocols are widely adopted in neural topic modeling, human evaluation could provide complementary insight into domain-specific topic interpretability.

References

Suman Adhya and Debarshi Kumar Sanyal. 2024. GINopic: Topic modeling with graph isomorphism network. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6171–6183, Mexico City, Mexico. Association for Computational Linguistics.

Pritom Saha Akash and Kevin Chen-Chuan Chang. 2024. Enhancing short-text topic modeling with LLM-driven context expansion and prefix-tuned VAEs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15635–15646, Miami, Florida, USA. Association for Computational Linguistics.

Dimo Angelov and Diana Inkpen. 2024. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, Miami, Florida, USA. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Gregor Donabauer and Udo Kruschwitz. 2025. Token-level graphs for short text classification. In *European Conference on Information Retrieval*, pages 427–436. Springer.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Nathan Godey, Éric Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, St. Julian's, Malta. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Sungwon Han, Mingi Shin, Sungkyu Park, Changwook Jung, and Meeyoung Cha. 2023. Unified neural topic model via contrastive learning and term weighting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1802–1817, Dubrovnik, Croatia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Anastassia Kornilova and Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality.

In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Jiyuan Liu, Jiaxing Yan, Chunjiang Zhu, Xingyu Liu, Li Qing, and Yanghui Rao. 2025. Neural topic modeling via contextual and graph information fusion. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13247–13263, Suzhou, China. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Duy-Tung Pham, Thien Trang Nguyen Vu, Tung Nguyen, Linh Van Ngo, Duc Anh Nguyen, and Thien Huu Nguyen. 2024. NeuroMax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7758–7772, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Kunze Wang, Yihao Ding, and Soyeon Caren Han. 2024. Graph neural networks for text classification: A survey. *Artificial intelligence review*, 57(8):190.

Yiming Wang, Ximing Li, Xiaotang Zhou, and Jihong Ouyang. 2021a. Extracting topics with simultaneous word co-occurrence and semantic correlation graphs: Neural topic modeling for short texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 18–27.

Ziyun Wang, Xuan Liu, Peiji Yang, Shixing Liu, and Zhisheng Wang. 2021b. Cross-lingual text classification with heterogeneous graph neural network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 612–620, Online. Association for Computational Linguistics.

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. 2023a. Graph neural networks for natural language processing: A survey. *Foundations and Trends in Machine Learning*, 16(2):119–328.

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023b. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*, pages 37335–37357. PMLR.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

Xiaobao Wu, Thong Nguyen, Delvin Zhang, William Yang Wang, and Anh Tuan Luu. 2024b. Fastopic: A fast, adaptive, stable, and transferable topic modeling paradigm. In *Advances in Neural Information Processing Systems*, volume 37.

Xiaobao Wu, Fengjun Pan, and Anh Tuan Luu. 2024c. Towards the TopMost: A topic modeling system toolkit. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Bangkok, Thailand. Association for Computational Linguistics.

Yishi Xu, Jianqiao Sun, Yudi Su, Xinyang Liu, Zhibin Duan, Bo Chen, and Mingyuan Zhou. 2023. Context-guided embedding adaptation for effective topic modeling in low-resource regimes. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Deyu Zhou, Xuemeng Hu, and Rui Wang. 2020. Neural topic modeling by incorporating document relationship graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3790–3796, Online. Association for Computational Linguistics.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4663–4672, Brussels, Belgium. Association for Computational Linguistics.

A Benchmark Datasets

In this section, we provide detailed description of the benchmark datasets: 20NewsGroup[2], BioASQ[3] and Bills[4]. The statistics of the preprocessed datasets are presented in Table 7.

B Implementation Details and Hyperparameter Settings

In this section, we describe the training environment, our proposed DARTOPIC framework and baseline models details. All experiments, including those for DARTOPIC and all baseline methods, were conducted under the same training environment for fair comparison. Specifically, all models were implemented in PyTorch (version 2.10.0) and trained on a single NVIDIA H100 GPU. Following a unified setup, we used `all-MiniLM-L6-v2`[5] as the PLM for all methods, and adopted Adam (Kingma and Ba, 2014) as the optimizer with a learning rate of 1×10^{-3} for 200 epochs. For **FASTopic**, we set $\mathrm{DT}_\alpha = 3.0$, $\mathrm{TW}_\alpha = 2.0$, and $\theta_{\text{temp}} = 1.0$. For **NeuroMax**, we followed the hyperparameter settings provided in the official implementation. In particular, we used $\epsilon = 1 \times 10^{-16}$, OT_max_iter $= 5000$, stopThr $= 0.5 \times 10^{-2}$, num_groups $= 10$, $\beta_{\text{temp}} = 0.2$, weight_loss_ECR $= 250.0$, weight_loss_GR $= 250.0$, $\alpha_{\text{GR}} = 20.0$, $\alpha_{\text{ECR}} = 20.0$, sinkhorn_max_iter $= 1000$, and weight_loss_InfoNCE $= 10.0$. For **GINopic**, we used pre-trained Word2Vec[6] embeddings for graph construction. The edge-weight threshold was set to 0.4 for Bills and 20NG, and 0.05 for BioASQ. For **CGTM**, we set $\gamma_c = 1.0$, $\gamma_g = 0.2$, and GMM weight equal to 20. For graph information, we used $\epsilon_g = 0.004$.

For our proposed DARTOPIC, the edge-weight threshold (τ) was set to 0.2 for BioASQ and 20NG, and 0.3 for Bills.

C Evaluation Metrics

In this section, we explain the details of our used automatic evaluation metrics in subsection 4.1. We implement NPMI and TU metrics based on TopMost (Wu et al., 2024c), a comprehensive toolkit for comparing and optimizing topic modeling in various scenarios[7]. We evaluate topic coherence using Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014). For each topic k, let $W_k = \{w_1, \ldots, w_L\}$ denote the top L words. The NPMI between a pair of words (w_i, w_j) is defined as:

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}. \quad (8)$$

Here, $P(w_i)$ and $P(w_j)$ denote the marginal probabilities of words w_i and w_j, respectively, and $P(w_i, w_j)$ denotes their co-occurrence probability estimated from a reference corpus. The topic-level coherence score is obtained by averaging the NPMI values over all word pairs among the top L words:

$$\mathrm{NPMI}(k) = \frac{2}{L(L-1)} \sum_{1 \le i < j \le L} \mathrm{NPMI}(w_i, w_j). \quad (9)$$

The final NPMI score is computed by averaging across all K topics:

$$\mathrm{NPMI} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{NPMI}(k). \quad (10)$$

A higher NPMI score indicates that the top words within a topic co-occur more frequently in

[2] https://github.com/MIND-Lab/OCTIS/tree/master/preprocessed_datasets/20NewsGroup

[3] https://github.com/AdhyaSuman/GINopic/tree/master/preprocessed_datasets/Bio

[4] https://huggingface.co/datasets/FiscalNote/billsum

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[6] https://huggingface.co/fse/word2vec-google-news-300

[7] https://github.com/BobXWu/TopMost

Dataset	Domain	#Docs	#Vocab	Avg. length	#Labels
20NewsGroup	*General*	16,309	1,612	48.02	20
BioASQ	*Biomedical*	19,448	2,000	7.44	20
Bills	*Legal*	18,945	2,000	76.28	–

Table 7: Statistics of the preprocessed benchmark datasets

the reference corpus, suggesting better semantic coherence. Additionally, we evaluate topic diversity using topic uniqueness (TU) (Nan et al., 2019). For each topic k, the top L words are obtained by ranking the decoder topic-word weights β_k in descending order. Let $\mathrm{cnt}(l, k)$ denote the number of topics whose top-L words contain the l-th top word of topic k. Then, the uniqueness of topic k is defined as:

$$\mathrm{TU}(k) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{\mathrm{cnt}(l, k)}, \quad k = 1, \ldots, K. \tag{11}$$

The final TU score is computed by averaging over all K topics:

$$\mathrm{TU} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{TU}(k). \tag{12}$$

The TU value ranges from $1/K$ to 1, where a higher value indicates that fewer top words are shared across topics, implying greater topic diversity and lower redundancy. The overall topic quality (TQ) (Dieng et al., 2020) is calculated as the product of the coherence and diversity values.

To evaluate the quality of document-topic distributions, we perform a document clustering task based on the inferred document-topic representations. The resulting clustering assignments are then evaluated using Purity and normalized mutual information (NMI) (Manning et al., 2008), which measure how well the predicted clusters align with the ground-truth document labels.

Let $\mathcal{C} = \{C_1, \ldots, C_K\}$ denote the set of predicted clusters and $\mathcal{L} = \{L_1, \ldots, L_J\}$ denote the set of ground-truth classes, where N is the total number of documents. Purity measures the extent to which each cluster contains documents from a single class, and is defined as

$$\mathrm{Purity} = \frac{1}{N} \sum_{k=1}^{K} \max_{j} |C_k \cap L_j|. \tag{13}$$

A higher Purity indicates that each predicted cluster is dominated by documents from one ground-truth class.

NMI measures the agreement between the predicted clustering and the ground-truth labels from an information-theoretic perspective. It is defined as

$$\mathrm{NMI}(\mathcal{C}, \mathcal{L}) = \frac{I(\mathcal{C}; \mathcal{L})}{\sqrt{H(\mathcal{C}) H(\mathcal{L})}}, \tag{14}$$

where $I(\mathcal{C}; \mathcal{L})$ is the mutual information between the predicted clusters and the ground-truth classes, and $H(\mathcal{C})$ and $H(\mathcal{L})$ are their entropies. The mutual information is computed as

$$I(\mathcal{C}; \mathcal{L}) = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{|C_k \cap L_j|}{N} \log \frac{N \cdot |C_k \cap L_j|}{|C_k|\,|L_j|}, \tag{15}$$

and the entropy of the predicted clusters is defined as

$$H(\mathcal{C}) = -\sum_{k=1}^{K} \frac{|C_k|}{N} \log \frac{|C_k|}{N}, \tag{16}$$

with $H(\mathcal{L})$ defined analogously for the ground-truth classes. Higher NMI indicates better consistency between the cluster assignments and the true labels.

D Analysis of Graph Construction Methods

Graph Structure	20NG			BioASQ		
	NPMI	TU	TQ	NPMI	TU	TQ
1-hop	0.2643	0.9307	0.2460	0.1583	0.9600	0.1520
2-hop	0.2667	**0.9400**	0.2507	**0.1687**	0.9600	**0.1620**
3-hop	0.2655	0.9227	0.2449	0.1659	0.9693	0.1608
Semantic (ours)	**0.2736**	0.9360	**0.2561**	0.1641	**0.9760**	0.1602

Table 8: Performance comparison on document clustering task using document-topic distribution vector. The best-performing method is highlighted in **bold**.

We explore two graph construction strategies, the token-level n-hop graph (Donabauer and Kruschwitz, 2025) and the our token-level semantic graph method, within our DARTOPIC framework. In tabel 8, the results show that the n-hop graph is more effective on BioASQ, which consists of

relatively short documents, whereas the semantic graph performs better on 20NG, which contains relatively long documents. This suggests that local structural relations are more useful for short texts, while broader semantic connections become more important for longer documents. Overall, the results highlight that graph construction should be chosen carefully according to the document length characteristics of the target corpus.

E Hyperparameter Analysis

We analyze the effect of the graph construction hyperparameter τ in DARTopic. The parameter τ controls the semantic similarity threshold used to construct the token-level semantic graph. To examine its influence, we evaluate $\tau \in \{0.1, 0.2, 0.3, 0.4\}$ on three datasets: 20NG, BioASQ, and Bills.

As shown in Table 9, the best topic quality is generally achieved when τ is set between 0.2 and 0.3. Specifically, 20NG obtains the highest TQ score when $\tau = 0.2$, while Bills achieves the best TQ score when $\tau = 0.3$. For BioASQ, the performance is also stable in this range, and we select $\tau = 0.2$ for the main experiments. Thus, we use $\tau = 0.2$ for 20NG and BioASQ, and $\tau = 0.3$ for Bills. The results show that both too small and too large values of τ can be less effective. When τ is too small, weak semantic relations may be included, which can introduce noisy edges into the graph. In contrast, when τ is too large, the graph may become overly sparse and discard useful semantic connections. Therefore, moderate values of τ provide a better balance between preserving meaningful semantic relations and filtering out noisy ones.

Interestingly, this tendency differs from n-hop based graph construction methods (Donabauer and Kruschwitz, 2025), whose performance can be highly affected by document length. Although BioASQ consists of relatively short documents, while 20NG and Bills contain longer documents, the optimal range of τ remains similar across all datasets. This indicates that the proposed token-level semantic graph construction is less sensitive to document length and less dependent on dataset-specific hyperparameter tuning. Overall, our analysis suggest that the proposed semantic graph construction method is robust to hyperparameter changes and can effectively adapt to datasets with different document lengths and domains.

F Qualitative Evaluation

In this section, we analyze to further examine the interpretability of the topics generated by different methods. Specifically, we select topics from the BioASQ dataset that contain biomedical-related keywords such as *"acid"*, *"dna"*, *"rna"*, and *"cell"*. For each method, we report two representative topic word lists in Table 10. Words highlighted in blue indicate general or less domain-specific terms that are relatively unrelated to professional biomedical terminology.

As shown in Table 10, several baseline methods generate topics that include many general words, such as "efficiency", "food", "human", "prediction", "experience", and "route". Although these words may frequently appear in the corpus, they are less informative for explaining biomedical topics. This suggests that the baseline methods often mix domain-specific biomedical terms with general-purpose words, which can reduce topic interpretability.

In contrast, DARTopic produces more coherent and domain-specific topic words, including terms such as *"replication"*, *"leukemic"*, *"rna"*, *"avian"*, *"leukemia"*, *"adenovirus"*, *"dna"*, *"ribonucleic acid"*, and *"messenger"*. These words are closely related to biomedical concepts and provide clearer semantic explanations of the learned topics. Therefore, the qualitative results demonstrate that DARTopic can generate more interpretable and specialized topics in domain-specific corpora such as BioASQ.

G Use of AI Assistants

AI assistants were used only for limited writing and language-editing support, as well as preparation of submission materials. All technical content, methodology, implementation, experimental results, and final verification were performed and checked by the authors.

	20NG			BioASQ			Bills		
	NPMI	TU	TQ	NPMI	TU	TQ	NPMI	TU	TQ
$\tau = 0.1$	0.2670	0.9320	0.2488	0.1668	0.9693	0.1617	0.2503	0.9307	0.2329
$\tau = 0.2$	0.2736	0.9360	0.2561	0.1641	0.9760	0.1602	0.2545	0.9160	0.2331
$\tau = 0.3$	0.2689	0.9307	0.2502	0.1737	0.9533	0.1656	0.2685	0.9240	0.2481
$\tau = 0.4$	0.2647	0.9400	0.2488	0.1609	0.9640	0.1551	0.2489	0.9173	0.2283

Table 9: Analysis of graph construction hyperparameter in DARTOPIC.

Methods	**Top-related word examples**
ProdLDA	efficiency citrate acidosis african section diurnal gastrin diabetic tube conformation food guanethidine reagent hydroxy recognition testing acidosis trial dog dark
FASTopic	human cell skin red culture erythrocytes peripheral serum studies normal acid synthesis amino acids enzyme phosphate biosynthesis ribonucleic rna dna
BERTopic	cells human cell cadmium zinc dna electron skin culture synthesis amino intestine intestinal acid albumin trypsin pig pigs small chymotrypsin
ZeroShotTM	plasmid recombination repair dna coli excision rec replication defective division hela nucleic rna poliovirus histone simian simplex sendai herpesvirus newcastle
CombinedTM	mother prediction rna rest virulence insufficiency tracheal sickle dynamic experience range mapping map intracellular head occurrence rhesus angle eye respiration
NeuroMax	extracellular basal probe ontogeny grow subject modify apply react mediate rna person recipient effector border evolution sow kitten rhythm route
PVTM	versus hapten secondary host response cellular suppression class hybrid carrier nucleic hela rna ribonucleic poliovirus replication acid dna messenger double
GINopic	spin cross rec dna head description viral inducible repair hydroxybutyrate polymerase wild dna hela freeze synthase estradiol importance polarity malate
CGTM	human cell vitro peripheral surface antigen response anti normal culture virus murine lymphocyte tumor mice cell friend spleen host lymphocytic
DARTOPIC	replication leukemic rauscher rna avian leukemia murine adenovirus primate dna ribonucleic acid aminoacyl ribosomal deoxyribonucleic nucleic trna synthetase messenger polynucleotide

Table 10: Top-related word examples generated by different baseline methods.