HiCOT: Improving Neural Topic Models via Optimal Transport and Contrastive Learning

Hoang Tran Vuong[1*], Tue Le[1*], Tu Vu[2*], Tung Nguyen[1*], Linh Van Ngo[1†], Sang Dinh[1], Thien Huu Nguyen[3]
[1]Hanoi University of Science and Technology (HUST), Vietnam
[2]Bytedance Inc
[3]University of Oregon, USA

Abstract

Recent advances in neural topic models (NTMs) have improved topic quality but still face challenges: weak document-topic alignment, high inference costs due to large pretrained language models (PLMs), and limited modeling of hierarchical topic structures. To address these issues, we introduce HiCOT (**Hi**erarchical Clustering and **C**ontrastive Learning with **O**ptimal **T**ransport for Neural Topic Modeling), a novel framework that enhances topic coherence and efficiency. HiCOT integrates Optimal Transport to refine document-topic relationships using compact PLM-based embeddings, captures semantic structure of the documents. Additionally, it employs hierarchical clustering combine with contrastive learning to disentangle topic-word and topic-topic relationships, ensuring clearer structure and better coherence. Experimental results on multiple benchmark datasets demonstrate HiCOT's superior effectiveness over existing NTMs in topic coherence, topic performance, representation quality, and computational efficiency.

1 Introduction

Topic modeling has been recognized as a fundamental task in natural language processing (NLP), aims to uncover latent topic structures within a corpus, while simultaneously providing document topic distributions (Hofmann, 1999; Griffiths et al., 2003; Srivastava and Sutton, 2017; Wu et al., 2024a). In recent years, neural topic models (NTMs) (Zhao et al., 2021; Xu et al., 2022; Wu et al., 2024d; Nguyen et al., 2025a,b) have emerged as a promising alternative, leveraging deep neural networks to enhance flexibility, improve the quality of discovered topics, help to overcome limitations about inefficient parameter inference of traditional topic models.

Most NTMs (Dieng et al., 2020; Wu et al., 2023b; Pham et al., 2024b) are built upon the Variational Autoencoders (VAEs) (Kingma and Welling, 2013) framework, where an inference encoder generates document-topic distributions, and a generative decoder reconstructs the original texts. Beyond VAEs, other architectures such as using dual semantic relation reconstruction paradigm (Wu et al., 2024b), or through the development of variations of TF-IDF (Grootendorst, 2022), have been proposed to improve topic coherence and interpretability. In addition to architectural innovations, various advanced techniques have been introduced to enhance topic modeling performance. One widely explored approach involves integrating contextual information, such as word embeddings (Pennington et al., 2014), or sentence embeddings (Reimers and Gurevych, 2019a), to provide more semantically meaningful topic distributions. Another methods based on Optimal Transport also have been employed to model relationships between documents, topics, and words more effectively (Zhao et al., 2022; Wu et al., 2023b; Xu et al., 2023), while some integrating pretrained language models (PLMs) (Devlin, 2018; Reimers and Gurevych, 2019a) to capture complex linguistic structures (Han et al., 2023; Pham et al., 2024b).

Despite these advancements, NTMs still face several challenges. Recent models exhibit limitations in effectively capturing document-topic relationships, lacking information and showing weak alignment between document representations and topic proportions. While some approaches leverage PLMs to enhance topic modeling (Wu et al., 2023a; Han et al., 2023), the process of extracting embeddings from large PLMs significantly increases inference costs, making them less practical for situations requiring low inference times. It is entirely feasible to replace PLM-based embeddings with alternative methods that still achieve good performance, while significantly reducing computational overhead. To

[*]These authors contributed equally to this work.
[†]Corresponding author: linhnv@soict.hust.edu.vn

Findings of the Association for Computational Linguistics: ACL 2025, pages 13894–13920

address the lack of document-topic relationships, we propose an apporach that effectively utilizes compact PLMs with a simple neural network. Our method incorporates with Optimal Transport to refine relationships between documents and topics, enhancing the model's ability to learn more structured and semantic topic distributions through pretrained document embeddings.

Another limitation of NTMs is their tendency to overlook hierarchical topic structure, limiting their ability to model semantic dependencies at different levels of abstraction. To tackle this issue, we introduce a hierarchical clustering framework combined with contrastive learning, explicitly disentangling relationships between topics and words, as well as between topics themselves. Our approach regularizes semantic relations among topic and word embeddings, ensuring clear separation between topic clusters while maintaining meaningful relationships within each cluster. This leads to enhanced topic interpretability, coherence, and representation quality. In this work, we propose HiCOT (**Hi**erarchical Clustering and **C**ontrastive Learning with **O**ptimal **T**ransport for Neural Topic Modeling), a novel framework that effectively addresses these challenges. We summarize the contributions of our study as follows:

- We introduce an efficient topic modeling approach leveraging compact PLMs and Optimal Transport to enhance document-topic alignment, capture semantic relationships between documents while ensuring computational efficiency.
- We develop a novel framework that integrates hierarchical clustering with contrastive learning to strengthen topic-word associations and topic-level relationships, enhancing interpretability and coherent topic representations.
- We conduct comprehensive experiments on multiple benchmark datasets, demonstrating that our approach effectively improves overall topic modeling performance compared to existing NTMs.

2 Background

Consider a collection of Bag-of-Words (BoW) representations, denoted as $\mathbf{X} = \{\mathbf{x}_d\}_{d=1}^{D}$, corresponding to D documents and defined over a vocabulary of V distinct words. Topic modeling aims to uncover K latent topics within $\mathbf{X}$. Each topic k is associated with a topic-word distribution $\beta_k \in \mathbb{R}^{V\times 1}$, forming a topic-word distribution matrix of K desired topics $\beta = (\beta_1, \ldots, \beta_K) \in \mathbb{R}^{V\times K}$. Let L be the word embedding dimension, we define the word embedding matrix as $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_V) \in \mathbb{R}^{V\times L}$ and the topic embedding matrix as $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_K) \in \mathbb{R}^{K\times L}$. Recent advances in neural topic modeling (Wu et al., 2023b; Pham et al., 2024b; Wu et al., 2024c) have shifted away from factorizing β as the product of word embeddings W and topic embeddings T. Instead, β is expressed as:

$$\beta_{ij} = \frac{\exp\left(-\|\mathbf{w}_i - \mathbf{t}_j\|^2/\tau\right)}{\sum_{j'=1}^{K}\exp\left(-\|\mathbf{w}_i - \mathbf{t}_{j'}\|^2/\tau\right)}, \quad (1)$$

where τ is a temperature hyperparameter. Another objective of neural topic models is the inference of topic proportions for each document x_i, represented as $\theta_i \in \mathbb{R}^K$ which characterizes the distribution of topics within the document. In VAE-based topic models, topic proportion θ is inferred through a latent variable z that follows a logistic-normal prior distribution $p(z) = \mathcal{N}(z|\mu_0, \Sigma_0)$. Given a document x_i, its Bag-of-Words (BoW) representation is passed through an inference network to compute the parameters of a Gaussian posterior distribution, where mean and diagonal covariance matrix are given by $\mu = h_\mu(x_d, \gamma)$ and $\Sigma = \mathrm{diag}(h_\Sigma(x_d, \gamma))$, respectively, with γ denoting the parameters of inference network. Using the reparameterization trick (Kingma and Welling, 2013), the latent variable z is subsequently sampled from the posterior distribution $q(z|x_i) = \mathcal{N}(z|\mu, \Sigma)$. Finally, the BoW representation is reconstructed through β and θ, where the generative process follows a multinomial distribution $\hat{\mathbf{x}}_{\mathrm{BoW}} \sim \mathrm{Multi}(\mathrm{softmax}(\beta\theta))$. The loss function for topic modeling is composed of two components: a reconstruction term and a regularization term, as detailed below:

$$\mathcal{L}_{\mathrm{TM}} = \frac{1}{D}\sum_{i=1}^{D}\Big[-(\mathbf{x}_{i\mathrm{BoW}})^\top \log(\mathrm{softmax}(\beta\theta_i)) + \mathrm{KL}(q(z|\mathbf{x}_i)\|p(z))\Big]. \quad (2)$$

Unlike VAE-based models, FASTopic (Wu et al., 2024c) adopts a different inference mechanism. FASTopic utilizes Optimal Transport to directly compute document-topic and topic-word distributions. Specifically, it defines cost matrices $C^{(1)} \in$

$\mathbb{R}^{D\times K}$ and $C^{(2)} \in \mathbb{R}^{K\times V}$ based on Euclidean distances between document embeddings, topic representations, and word embeddings. The resulting optimal transport plans ψ^* and ϕ^* yield document-topic and topic-word distributions $\theta = D\psi^*$ and $\beta = K\phi^*$, with the overall loss:

$$\begin{aligned}\mathcal{L} = &-\frac{1}{D}\sum_{i=1}^{D}(x_{i,\text{BoW}})^\top \log(\beta\theta_i) \\ &+\sum_{i,k} C^{(1)}_{ik}\psi^*_{ik} + \sum_{k,j} C^{(2)}_{kj}\phi^*_{kj}\end{aligned} \tag{3}$$

Then, given a new document $\mathbf{x}'$, FASTopic performs inference by mapping it to an embedding $\mathbf{d}'$ and estimates its topic distribution θ' as follows:

$$\theta'_k = \frac{p_k}{\sum_{k'=1}^{K} p_{k'}}, \quad p_k = \frac{\exp(-\|\mathbf{t}_k - \mathbf{d}'\|^2/\tau)}{\sum_{i=1}^{D}\exp(-\|\mathbf{t}_k - \mathbf{d_i}\|^2/\tau)} \tag{4}$$

where τ is a temperature hyperparameter. Despite its efficiency, FASTopic is limited by inconsistencies between training and testing, and its dependence on the choice of pretrained language model.

3 Proposed Method

In this section, we present a detailed analysis of our proposed, which leverages compact PLMs and Optimal Transport to improve document-topic alignment. Furthermore, we introduce a novel framework that incorporates hierarchical clustering and contrastive learning to reinforce topic-word associations and topic-level relationships.

3.1 Efficient and Consistent Topic Modeling with MLP and Optimal Transport

Existing topic models, such as FASTopic (Wu et al., 2024b), suffer from inconsistencies between training and inference, where inference relies on a pretrained language model, leading to significant computational overhead. To address this, we employ a unified VAE architecture and replace the reliance on large Transformer-based encoders (e.g., BERT (Devlin, 2018), Sentence-BERT (Reimers and Gurevych, 2019b)) with a lightweight alternative, that ensures consistency between training and inference while enabling efficient inference without reducing representation quality.

Specifically, we initialize document embeddings using Doc2Vec (Le and Mikolov, 2014), where $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_D\} \in \mathbb{R}^{D\times M}$ denotes the document embedding matrix, with M is the embedding dimension. Then we use an MLP network to project document embeddings into the topic embedding space: $\phi_{\mathbf{E}}(\mathbf{E})$, with learnable weights $\mathrm{W}_{\phi_{\mathbf{E}}} \in \mathbb{R}^{M\times L}$. This projection is designed to preserve the relationships of document embeddings generated from the pretrained language model, ensuring that topic embeddings can capture highly meaningful semantic structures. To achieve this, we model the relationship between document and topic embeddings through optimal transport, where the transportation cost between a document i and a topic k is defined as: $\mathrm{C}^{(ik)}_{\mathrm{DT}} = ||\phi_{\mathbf{E}}(\mathbf{e}_i) - \mathbf{t}_k||^2$. The objective is then to minimize the weighted transport distance between outputs of the MLP network and topic embeddings, formulated as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{DT}} &= \sum_{i=1}^{D}\sum_{k=1}^{K} \mathrm{C}^{(ik)}_{\mathrm{DT}}\pi^*_{ik} \\ \text{where } \pi^* &= \arg\min_{\pi\in\mathbb{R}^{D\times K}} \langle \mathrm{C}_{\mathrm{DT}}, \pi\rangle - \nu H(\pi) \\ &\text{s.t. } \pi\mathbb{1}_K = \frac{1}{D}\mathbb{1}_D,\ \pi^\top\mathbb{1}_D = \frac{1}{K}\mathbb{1}_K\end{aligned} \tag{5}$$

By optimizing this objective, the distance between projected document embeddings and topic embeddings is minimized. This results in a topic structure where the topic embeddings inherit the good semantic structure of document embeddings from the pretrained model. This alignment enhances the interpretability and coherence of the topic embeddings. Our approach still achieves computational efficiency while maintaining topic modeling performance, making it well-suited for scaling to larger datasets.

3.2 Topic Regularization via Optimal Transport for Semantic Relationships

We propose a novel topic regularization approach leveraging Optimal Transport (Peyré and Cuturi, 2020). The core idea is to align topic distributions with semantic relationships by utilizing a transport plan that integrates both topic proportions and semantic similarities between documents, utilizing the strengths of PLMs. This ensures that topic distributions are preserved while also capturing the underlying meaningful structure of the document collection. Let D be the number of documents, we define $C \in \mathbb{R}^{D\times D}$ as the cost matrix in Euclidean space for topic proportions

$\{\theta_1, \theta_2, \ldots, \theta_D\}$, where $c_{ij} = ||\theta_i - \theta_j||_2^2$. In parallel, we define a matrix P to capture the semantic similarity between documents. The elements of P are computed using cosine similarity between document embeddings:

$$p_{ij} = \frac{\langle \mathbf{e}_i, \mathbf{e}_j \rangle}{\|\mathbf{e}_i\|_2 \cdot \|\mathbf{e}_j\|_2} \tag{6}$$

Since document embeddings $\mathbf{E}$ are initialized using Doc2Vec, they encode high-level semantic information by capturing both contextual and distributional similarities between documents. This allows matrix P to effectively capture the semantic structure of the document corpus. To ensure the alignment between topic-based relationships and semantic structures, we incorporate a regularization term that forces the transport plan Q to approximate the semantic relationships captured in matrix P. This is formulated by introducing an additional KL-divergence constraint into the following optimization problem:

$$\begin{aligned} &\min_{Q \in \mathbb{R}^{D \times D}} \langle Q, C \rangle - \lambda_1 H(Q) + \lambda_2 \mathrm{KL}(Q||P) \\ &\text{s.t. } Q\mathbb{1}_D = Q^\top \mathbb{1}_D = \frac{1}{D}\mathbb{1}_D \end{aligned} \tag{7}$$

where $\lambda_1,\ \lambda_2 > 0$, $\mathbb{1}_D$ is D-dimensional vector with all elements equal to 1, and $H(Q) = -\langle Q, \log Q - 1 \rangle$ represents the Shannon entropy of Q (Cuturi, 2013a). The transport plan Q uncovers the topic relationships by considering topic proportions, effectively capturing information is transferred between topics based on their interrelations. By solving the problem in Equation 7, we can also minimize $\mathrm{KL}(Q||P)$. Adding a KL-divergence constraint forces Q to learn the semantic relationships between documents while maintaining the consistency of topic proportions. The motivation behind using Optimal Transport is its ability to model the transfer of information between topics while preserving the mass (i.e., topic proportions). Similar to the transportation of mass, our approach ensures that topic proportions remain consistent while enabling information transfer between topics. This not only improves the coherence of the topic model but also enhances the understanding of semantic within the document.

3.3 Contrastive Learning for Topic Clustering

In this paper, we introduce an novel regularization method for topic modeling, integrating hierarchical clustering with contrastive learning. By leveraging hierarchical clustering, our method automatically identifies topic clusters, enabling us to impose contrastive losses that enhance both intra-cluster and inter-cluster relationships. Contrastive learning is employed between topic and word embeddings in a same cluster, as well as among topic embeddings between clusters. This solution enhances the semantic relationships between topic and word embeddings, as well as at the topic level, improving both coherence and quality of topic representations. Specifically, we use hierarchical agglomerative clustering (HAC) (Murtagh and Contreras, 2012) to divide topic embeddings into clusters. Assume that after performing hierarchical clustering, we partition K topics into G clusters. For each cluster i, we denote its set of topic embeddings as: $\mathcal{T}_i = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{N_i}\}$, where N_i is the number of topics within cluster i, ensuring that $\sum_{i=1}^{G} N_i = K$.

3.3.1 Contrastive Loss between Topic Embeddings in a Cluster

Once the topics are clustered, we apply contrastive learning to the topic embeddings within the same cluster. To achieve this, we first compute the cosine similarity between each word embedding and all topic embeddings. A word is then assigned to the topic with the highest similarity. Within each cluster, we consider only topics that contain at least one assigned word embedding. Formally, consider topic i in cluster j with a set of word embeddings $Z_{ij} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{M_{ij}}\}$, where M_{ij} denotes the number of word embeddings assigned to topic i in cluster j. The total number of assigned words across all topics and clusters satisfies the constraint: $\sum_{j=1}^{G} \sum_{i=1}^{N_j} M_{ij} = V$, where V is the vocabulary size. We then define the contrastive components based on Triplet Loss (Schroff et al., 2015), as follows:

- Anchor a_{ij}: The average embedding of all words in topic i within cluster j, computed as:

$$a_{ij} = \frac{1}{M_{ij}} \sum_{k=1}^{M_{ij}} \mathbf{w}_k \tag{8}$$

- Positive sample p_{ij}: A randomly selected word from the same topic, i.e., $p_{ij} \sim Z_{ij}$
- Negative sample n_{ij}: A randomly selected

word from another topic within the same cluster, i.e.,

$$n_{ij} \sim Z_j \setminus Z_{ij} \tag{9}$$

where $Z_j = \bigcup_{i=1}^{N_j} Z_{ij}$ is the set of all word embeddings assigned to topics in cluster j.

The objective is to pull words within the same topic closer in the embedding space, while simultaneously pushing words from different topics apart, improving the topic coherence within each cluster. To achieve this, we minimize the following contrastive loss:

$$\mathcal{L}_{\text{CLT}} = \sum_{j=1}^{G} \sum_{i=1}^{N_j} \max \left(d(a_{ij}, p_{ij}) - \frac{1}{k} \sum_{s=1}^{k} d(a_{ij}, n_{ijs}) + m, 0 \right) \tag{10}$$

where k denotes the number of negative samples, $d(x, y)$ represents a distance function, and m is a margin ensuring that the anchor-positive distance is smaller than the anchor-negative distance.

3.3.2 Contrastive Loss between Clusters

Beyond intra-cluster contrastive learning, we extend contrastive framework to the clustering level. We also define the following components based on Triplet Loss (Schroff et al., 2015), as follows:

- Anchor a_i: The mean embedding of all topics within a given cluster i, computed as:

$$a_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{t}_j \tag{11}$$

- Positive sample p_i: A randomly selected topic embedding from the same cluster, i.e., $p_i \sim \mathcal{T}_i$
- Negative sample n_i: A randomly selected topic embedding from a different cluster, i.e.,

$$n_i \sim \mathbf{T} \setminus \mathcal{T}_i \tag{12}$$

To promote tighter topic coherence within each cluster and enhance the separation between clusters, we minimize the following contrastive loss:

$$\mathcal{L}_{\text{CLC}} = \sum_{i=1}^{G} \max \left(d(a_i, p_i) - \frac{1}{k} \sum_{s=1}^{k} d(a_i, n_{is}) + m, 0 \right) \tag{13}$$

Here, $d(x, y)$ denotes the chosen distance metric, k is the number of negative samples, and m serves as a margin parameter. This loss encourages topics within the same cluster to have similar representations, while ensuring that topics from different clusters are distinct in the embedding space. As a result, it improves the separation between clusters while maintaining meaningful relationships within each cluster.

3.4 Overall objective function

Building on the approach of (Wu et al., 2023b), we integrate the Embedding Clustering Regularization (ECR) regularizer to minimize the weighted distance between word and topic embeddings, formulated as:

$$\mathcal{L}_{\text{ECR}} = \sum_{j=1}^{V} \sum_{k=1}^{K} ||\mathbf{w}_j - \mathbf{t}_k||^2 \psi_{jk}^* \tag{14}$$

Let C_{ECR} denote the transport cost between word and topic embeddings measured by Euclidean distance. The optimal transport plan ψ^* is computed through Sinkhorn's algorithm (Cuturi, 2013b):

$$\min_{\psi \in \mathbb{R}^{V \times K}} \langle \psi, \text{C}_{\text{ECR}} \rangle - \epsilon H(\psi) \quad \text{s.t. } \psi \mathbb{1}_K = \frac{1}{V} \mathbb{1}_V,\ \psi^\top \mathbb{1}_V = \frac{1}{K} \mathbb{1}_K \tag{15}$$

We now formalize our training process as a two-stage approach to ensure effective topic modeling and clustering. In **Stage 1**, the model undergoes initial training to capture the semantic structure of the data and refine the learned embeddings before clustering is introduced. During this phase, we update transport plan Q following Equation 7, and then optimize the following objective function:

$$\mathcal{L}_{\text{stage}_1} = \mathcal{L}_{\text{TM}} + \lambda_{\text{DT}} \mathcal{L}_{\text{DT}} + \lambda_{\text{ECR}} \mathcal{L}_{\text{ECR}} \tag{16}$$

After the model has learned meaningful representations, we proceed to **Stage 2**, where hierarchical clustering is applied. Clustering is first initialized at a predefined epoch, and subsequently refined at fixed intervals. During this stage, the model still updates transport plan Q according to the iterative process of Equation 7, and optimizes additional contrastive losses alongside the topic modeling objective from **Stage 1**:

$$\mathcal{L}_{\text{stage}_2} = \lambda_{\text{CLC}}\mathcal{L}_{\text{CLC}} + \lambda_{\text{CLT}}\mathcal{L}_{\text{CLT}} + \mathcal{L}_{\text{TM}} + \lambda_{\text{DT}}\mathcal{L}_{\text{DT}} + \lambda_{\text{ECR}}\mathcal{L}_{\text{ECR}} \quad (17)$$

where λ_{CLC}, λ_{CLT}, λ_{DT}, λ_{ECR} are weight hyperparameters. This two-stage approach ensures that clustering is performed only after the model has sufficiently learned the meaningful representations of the data, resulting in more coherent and well-separated topic clusters. The full algorithm are described in Appendix A.

4 Experiments

4.1 Settings

Datasets. Our experiments employ five prominent datasets, covering a diverse range of domains. We utilize three widely recognized benchmarks in topic modeling: **20 News Groups (20NG)** (Lang, 1995), a standard dataset for topic modeling; **AGNews** (Zhang et al., 2015) which comprises news articles from over 2000 sources; and IMDB (Maas et al., 2011), a collection of movie reviews. Additionally, we examine two short-text datasets: **SearchSnippets** (Phan et al., 2008), which contains over 12,000 web search snippets across 8 domains; and **GoogleNews** (Yin and Wang, 2016), featuring more than 10000 news article titles categorized into 152 topics. This dataset collection enables a comprehensive analysis across different text types and domains. The preprocessing steps and statistics of all datasets are detailed in Appendix G.2.

Evaluation metrics. We evaluate our model using the framework proposed by (Wu et al., 2023b), focusing on both topic quality and document-topic distributions. Topic quality is assessed using coherence and diversity metrics. Coherence is quantified by $\text{C}_\text{V}15$, which measures the semantic consistency of the top 15 words within each topic and has been shown to strongly correlate with human interpretability (Röder et al., 2015). Coherence scores are computed using a modified version of the Wikipedia corpus[1] as a reference source. For topic diversity, we use $\text{TD}15$, which computes the proportion of unique words among the top 15 topic words. To evaluate document-topic distributions, we employ Normalized Mutual Information (NMI) and Purity (Manning et al., 2008), following the approach in (Wang et al., 2022), where each document is assigned to its most probable topic.

[1] https://github.com/dice-group/Palmetto/

Baseline models. We consider several advanced topic modeling frameworks, including **ETM** (Dieng et al., 2020), which incorporates word embeddings; **NTM + CL** (Nguyen and Luu, 2021), which leverages contrastive learning to model relationships between similar and dissimilar documents; **ECRTM** (Wu et al., 2023b), which improves topic coherence via clustering regularization; **FASTopic** (Wu et al., 2024b), which models document-word-topic relationships using Optimal Transport; and **NeuroMax** (Pham et al., 2024b; Nguyen et al., 2025d), which refines topic distributions through pretrained embeddings and mutual information maximization.

4.2 Topic and Doc-Topic Distribution Quality

We conduct experiments to assess both topic quality and the effectiveness of document-topic distributions across five benchmark datasets: 20NG, AGNews, IMDB, SearchSnippets, and GoogleNews. Tables 1 and 2 present the evaluation results for models trained with 50 and 100 topics, respectively. Our approach consistently surpasses baseline models, improving overall topic quality. Furthermore, it significantly enhances the quality of document-topic distributions, as reflected in higher Purity and NMI scores, which indicate better clustering performance and greater distinguishability of topic groups. In addition to improving document-topic distributions, HiCOT also strengthens topic coherence across most datasets, as demonstrated by its superior C_V score compared to other models.

4.3 Contrastive Learning Strategies

We evaluated the impact of contrastive learning strategies within our topic modeling framework, focusing on sampling strategies and contrastive learning method. **For sampling strategies**, we examine the selection process for positive and negative samples relative to the anchor. Beyond the default of random selection for both positive and negative samples ("Random / Random"), we evaluated two alternative approaches:

- Hard negatives: Selecting negative samples that exhibit the smallest embedding space distance to the anchor.
- Hard positives: Selecting samples at an intermediate distance, intended to increase learning difficulty while maintaining relevance.

We performed these experiments on 20NG and GoogleNews datasets, setting number of topics

50 Topics	20NG				AGNews				IMDB			
	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD
ETM ‡	0.375	0.347	0.319	0.704	0.364	0.679	0.224	0.819	0.346	0.660	0.038	0.557
NTM + CL	0.437	0.582	0.491	0.802	0.440	0.322	0.100	0.441	0.396	0.657	0.044	0.617
ECRTM ‡	0.431	0.560	0.524	0.964	**0.466**	0.802	0.367	0.961	0.393	0.694	0.058	**0.974**
FASTopic	0.427	0.583	0.528	**0.980**	0.379	0.831	0.352	0.960	0.371	0.683	0.055	0.969
NeuroMax ‡	0.435	0.623	0.570	0.912	0.385	0.804	0.410	0.952	0.402	0.709	0.061	0.936
HiCOT	**0.451**	**0.626**	**0.583**	0.852	0.446	**0.857**	**0.412**	**0.992**	**0.404**	**0.737**	**0.082**	0.837
100 Topics	**20NG**				**AGNews**				**IMDB**			
	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD
ETM ‡	0.369	0.394	0.339	0.573	0.371	0.674	0.204	0.773	0.341	0.648	0.037	0.371
NTM + CL	0.420	0.626	0.490	0.624	0.415	0.280	0.050	0.277	0.382	0.705	0.044	0.492
ECRTM ‡	0.405	0.555	0.494	0.904	0.416	0.812	**0.428**	**0.981**	0.373	0.694	0.049	**0.887**
FASTopic	0.400	0.622	0.522	0.861	0.385	0.833	0.330	0.912	0.369	0.680	0.048	0.886
NeuroMax ‡	0.412	0.602	0.516	**0.913**	0.406	0.828	0.389	0.957	0.381	0.706	0.059	0.870
HiCOT	**0.424**	**0.652**	**0.568**	0.741	**0.435**	**0.862**	0.388	0.960	**0.388**	**0.739**	**0.071**	0.733

Table 1: Evaluation results on standard datasets, assessed using C_V, TD, Purity and NMI under $K = 50$ topics and $K = 100$ topics. The highest-performing results are marked in bold, while the second-best values are underlined. Results reported in (Pham et al., 2024b).

$K = 50$ in Table 3. The results indicate that no single strategy consistently outperforms others across all metrics. "Random / Random" achieving highest scores in both cases for topic coherence (C_V). However, incorporating "Hard" sampling introduces trade-offs: "Hard Negatives" enhance TD on GoogleNews but slightly reduce C_V, "Hard Positives" improve Purity and NMI on 20NG ("Hard / Random") yet show different impacts on GoogleNews. These findings suggest that optimal contrastive sampling configuration is not fixed but depends on dataset characteristics. While "Random / Random" offers a reliable default, the results strongly motivate future research into developing more sampling techniques tailored to specific dataset properties.

For contrastive learning methods, alongside Triplet Loss employed in our main experiments, we evaluated performance using two contrastive losses: InfoNCE (Oord et al., 2018) and Circle Loss (Sun et al., 2020). To ensure a fair comparison, these experiments utilized same sampling strategy as applied with Triplet Loss. The results, presented in Table 4, demonstrate that all three loss functions achieve comparable performance across both datasets and all metrics. While minor fluctuations exist (e.g., slight advantages Purity and NMI for Circle Loss on AGNews, Triplet Loss on SearchSnippets), the differences are consistently small. These results highlight the robustness of our contrastive learning framework, suggesting that its effectiveness primarily from contrastive paradigm rather than the choice of loss function.

4.4 Robustness of HiCOT across embeddings

While Doc2Vec captures less semantic structure than SentenceTransformer, using a lightweight model like Doc2Vec instead of advanced pretrained language models (PLMs) demonstrates HiCOT's superior performance without relying heavily on extensive pretrained knowledge. As shown in Tables 1 and 2, HiCOT with Doc2Vec surpasses PLM-based models such as FASTopic and NeuroMax. This highlights HiCOT's robustness, which relies primarily on its architecture rather than the representational power of the pretrained language model. To further investigate this, we evaluated HiCOT with Doc2Vec and SentenceTransformer (`all-MiniLM-L6-v2` (Reimers, 2019)) embeddings on 20NG dataset with $K = 50$ and

	$K = 50$								$K = 100$							
	SearchSnippets				**GoogleNews**				**SearchSnippets**				**GoogleNews**			
	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD
ETM ‡	0.397	0.688	0.389	0.594	0.402	0.366	0.560	0.916	0.389	0.691	0.365	0.448	0.398	0.554	0.713	0.677
NTM + CL	0.403	0.215	0.030	0.532	0.433	0.041	0.005	0.301	0.406	0.217	0.020	0.394	0.432	0.039	0.005	0.367
ECRTM ‡	0.450	0.711	0.419	0.998	0.441	0.396	0.615	0.987	0.432	0.789	0.443	**0.966**	0.418	0.342	0.491	**0.991**
FASTopic	0.356	0.793	**0.497**	0.519	0.401	0.252	0.570	0.235	0.350	0.801	0.466	0.463	0.366	0.237	0.459	0.100
NeuroMax	0.427	0.743	0.427	0.920	0.409	0.359	0.590	**1.000**	0.439	0.854	0.472	0.960	0.427	0.664	0.834	0.915
HiCOT	**0.460**	**0.818**	0.478	**1.000**	**0.454**	**0.465**	**0.657**	0.920	**0.449**	**0.857**	**0.480**	0.940	**0.470**	**0.763**	**0.864**	0.802

Table 2: Evaluation results on short datasets, assessed using C_V, TD, Purity and NMI under $K = 50$ topics and $K = 100$ topics. The highest-performing results are marked in bold, while the second-best values are underlined. Results reported in (Nguyen et al., 2025a).

20NG

Positive	Negative	C_V	Purity	NMI	TD
Random	Random	0.451	0.626	0.583	0.852
Hard	Random	0.434	0.641	0.584	0.856
Random	Hard	0.438	0.620	0.575	0.901
Hard	Hard	0.445	0.638	0.582	0.843

GoogleNews

Positive	Negative	C_V	Purity	NMI	TD
Random	Random	0.454	0.465	0.657	0.920
Hard	Random	0.446	0.458	0.663	0.987
Random	Hard	0.439	0.456	0.661	0.995
Hard	Hard	0.443	0.462	0.667	0.989

Table 3: Performance of contrastive sampling strategies on 20NG and GoogleNews datasets under $K = 50$ topics.

Dataset	CL Method	C_V	Purity	NMI	TD
AGNews	InfoNCE	0.442	0.855	0.412	1.000
	Circle Loss	0.442	0.858	0.416	0.997
	Triplet Loss	0.446	0.857	0.412	0.992
SearchSnippets	InfoNCE	0.464	0.814	0.476	1.000
	Circle Loss	0.464	0.809	0.476	1.000
	Triplet Loss	0.460	0.818	0.478	1.000

Table 4: Comparison of contrastive loss functions on AGNews and SearchSnippets datasets with 50 topics.

20NG – 50 topics

Method	C_V	Purity	NMI	TD
HiCOT + Doc2Vec	0.451	0.626	0.583	0.852
HiCOT + SBERT	0.438	0.649	0.585	0.868

20NG – 100 topics

Method	C_V	Purity	NMI	TD
HiCOT + Doc2Vec	0.424	0.652	0.568	0.741
HiCOT + SBERT	0.416	0.632	0.556	0.779

Table 5: HiCOT performance on 20NG with two document embeddings at K=50 and K=100 topics.

$K = 100$ topics. Table 5 show improvements on several evaluation metrics, indicating the potential benefits of integrating stronger encoders within the HiCOT framework in the future works.

4.5 Inference time

To evaluate the inference of our proposed method, HiCOT, we conducted experiments to assess the time required for inferring on the whole dataset against state-of-the-art (SOTA) methods, namely FASTopic and NeuroMax, as well as traditional methods, including ETM and ECRTM. As detailed in Table 6, HiCOT achieves significantly faster inference times than FASTopic while also demonstrating superior topic quality, as discussed in Section 4.2. Although traditional methods exhibit faster inference times, their topic quality metrics, as presented in Tables 1 and 2, are substantially lower compared to both the SOTA baselines and HiCOT. These findings underscore the effectiveness of our approach in mitigating the high inference costs typically associated with pretrained language models.

4.6 Visualization of Embedding Space

We visualize the learned topic and word embeddings using t-SNE (van der Maaten and Hinton, 2008) on the GoogleNews dataset with 50 topics. As shown in Figure 1, HiCOT effectively maintains

Dataset	NeuroMax	FASTopic	HiCOT	ETM	ECRTM
20NG	2.98	9.41	4.01	0.13	1.25
AGNews	2.53	4.87	3.43	0.11	0.94
IMDB	5.92	19.01	9.31	0.37	2.34
SearchSnippets	3.17	6.58	4.47	0.15	1.17
GoogleNews	2.90	5.91	4.15	0.12	1.19

Table 6: Inference time of various topic modeling methods across different datasets with 50 topics. All experiments are conducted on a NVIDIA RTX 3090 GPU.

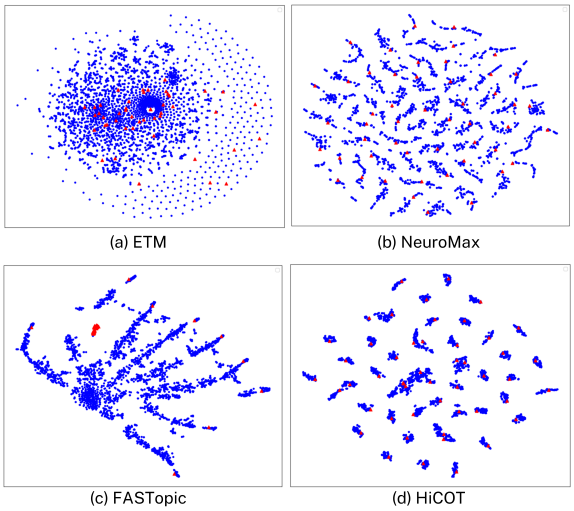

Figure 1: t-SNE visualization of word embeddings (●) and topic embeddings (▲) for GoogleNews dataset under 50 topics.

the structure of topic embeddings and prevents their collapse, whereas state-of-the-art baselines often suffer from embedding collapse or fail to capture clear word-topic relationships. This highlights the effectiveness of our method in clustering semantically coherent words. Additional visualizations for other datasets are provided in Appendix K.

4.7 Ablation study

We present an ablation study on the AGNews and Search Snippets datasets with 50 topics, to evaluate the contribution of each model component. Specifically, we define DT-SimpleNet as the model using $\mathcal{L}_{\text{TM}}$, $\mathcal{L}_{\text{DT}}$ and $\mathcal{L}_{\text{ECR}}$. We then systematically remove the contrastive loss, include $\mathcal{L}_{\text{CLT}}$ and $\mathcal{L}_{\text{CLC}}$, as well as the update of transport plan Q, subsequently evaluating their impact on model performance. Table 7 presents the results obtained. For document-topic distribution quality, evaluated by NMI and Purity, adding our components significantly improves performance, with DT-SimNet also exhibiting superior improvement. The model that integrates all components achieves the highest performance. Regarding topic quality, including coherence and diversity, the model remains competitive. In some datasets, our approach yields higher scores, while in others, the baseline is better.

	AGNews				SearchSnippets			
	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD
ECRTM	**0.466**	0.802	0.367	0.961	0.450	0.711	0.419	0.998
DT-SimNet	0.450	0.848	0.404	**0.992**	0.454	0.800	0.454	**1.000**
+ $\mathcal{L}_{\text{CLT}}$	0.448	0.855	0.409	**0.992**	0.460	0.815	0.471	**1.000**
+ $\mathcal{L}_{\text{CLC}}$	0.452	0.852	0.405	0.987	0.458	0.815	0.471	**1.000**
+ update Q	0.452	0.853	0.408	0.984	**0.463**	0.814	0.471	**1.000**
HiCOT	0.446	**0.857**	**0.412**	**0.992**	0.460	**0.818**	**0.478**	**1.000**

Table 7: Ablation study on AGNews and SearchSnippets datasets. The best and second-best results are highlighted in bold and underlined, respectively.

5 Conclusion

In conclusion, we introduce HiCOT, a novel neural topic modeling framework that integrates hierarchical clustering, contrastive learning, and optimal transport. By leveraging compact PLMs and optimal transport, HiCOT enhances document-topic alignment and captures semantic relationships between documents. Additionally, contrastive learning strengthens both topic-word associations and inter-topic interactions, resulting in more coherent topic representations. Comprehensive experiments on benchmark datasets demonstrate that HiCOT achieves superior performance in generating high-quality topics and document-topic distributions, providing a robust and scalable solution for neural topic modeling.

Limitations

While HiCOT has significantly improved topic quality, there are still some limitations. The selection of intervals for clustering updates needs to be carefully adjusted to ensure that model has sufficient time to learn stable semantic representations before applying clustering. However, excessively long intervals may slow convergence. An adaptive scheduling mechanism could further optimize this process. Additionally, the choice of positive and negative samples for contrastive learning remains an open research direction. These limitations suggest potential future studies for further improving the model.

Ethical Considerations

We comply with the ACL Code of Ethics and all relevant license terms. Our research in topic modeling is designed to enhance the field. When applied responsibly, it carries no significant societal risks.

Acknowledgments

This research has been supported by the NSF grant # 2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

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.

Tran Xuan Bach, Nguyen Duc Anh, Ngo Van Linh, and Khoat Than. 2021. Dynamic transformation of prior knowledge into bayesian models for data streams. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3742–3750.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *ACL-IJCNLP (Volume 2: Short Papers)*, pages 759–766. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, pages 1877–1901.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Marco Cuturi. 2013a. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 26. Curran Associates, Inc.

Marco Cuturi. 2013b. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Preprint*, arXiv:1306.0895.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Anh Nguyen Duc, Ngo Van Linh, Anh Nguyen Kim, and Khoat Than. 2017. Keeping priors in streaming bayesian learning. In *Advances in Knowledge Discovery and Data Mining*, pages 247–258, Cham. Springer International Publishing.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, 16.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *CoRR*, abs/2203.05794.

Cuong Ha, Van-Dang Tran, Linh Ngo Van, and Khoat Than. 2019. Eliminating overfitting of probabilistic topic models on short and noisy text: The role of dropout. *International Journal of Approximate Reasoning*, 112:85–104.

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. Association for Computational Linguistics.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339, San Francisco (CA). Morgan Kaufmann.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *Preprint*, arXiv:1405.4053.

Raymond Li, Felipe Gonzalez-Pizarro, Linzi Xing, Gabriel Murray, and Giuseppe Carenini. 2023. Diversity-aware coherence loss for improving neural topic models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1710–1722, Toronto, Canada. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

J MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.

Khai Mai, Sang Mai, Anh Nguyen, Ngo Van Linh, and Khoat Than. 2016. Enabling hierarchical dirichlet processes to work better for short texts at large scale. In *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II 20*, pages 431–442. Springer.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013*.

Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.

Duc Anh Nguyen, Kim Anh Nguyen, Canh Hao Nguyen, Khoat Than, et al. 2021. Boosting prior knowledge in streaming variational bayes. *Neurocomputing*, 424:143–159.

Ha Nguyen, Hoang Pham, Son Nguyen, Ngo Van Linh, and Khoat Than. 2022a. Adaptive infinite dropout for noisy and sparse data streams. *Machine Learning*, 111(8):3025–3060.

Quang Duc Nguyen, Tung Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025a. Glocom: A short text neural topic model via global clustering context. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *Preprint*, arXiv:2110.12764.

Tung Nguyen, Tue Le, Hoang Tran Vuong, Quang Duc Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025b. Sharpness-aware minimization for topic models with high-quality document representations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4507–4524.

Tung Nguyen, Trung Mai, Nam Nguyen, Linh Ngo Van, and Khoat Than. 2022b. Balancing stability and plasticity when learning topic models from short and noisy text streams. *Neurocomputing*, 505:30–43.

Tung Nguyen, Tung Pham, Linh Ngo Van, Ha-Bang Ban, and Khoat Than. 2025c. Out-of-vocabulary handling and topic quality control strategies in streaming topic models. *Neurocomputing*, 614:128757.

Tung Nguyen, Linh Ngo Van, Anh Nguyen Duc, and Sang Dinh Viet. 2025d. A framework for neural topic modeling with mutual information and group regularization. *Neurocomputing*, page 130420.

Van-Son Nguyen, Duc-Tung Nguyen, Linh Ngo Van, and Khoat Than. 2019. Infinite dropout for training bayesian models from data streams. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 125–134.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Gabriel Peyré and Marco Cuturi. 2018. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607.

Gabriel Peyré and Marco Cuturi. 2020. Computational optimal transport. *Preprint*, arXiv:1803.00567.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2024a. Topicgpt: A prompt-based topic modeling framework. 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 2956–2984.

Duy-Tung Pham, Thien Trang Nguyen Vu, Tung Nguyen, Linh Ngo Van, Duc Anh Nguyen, and Thien Huu Nguyen. 2024b. Neuromax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100.

Jipeng Qiang, Qian Zhenyu, Yun Li, Yunhao Yuan, and Xindong Wu. 2022. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Trans. Knowl. Data Eng.*, pages 1427–1445.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2019a. 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. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, page 399–408. Association for Computing Machinery.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE.

Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736. Association for Computational Linguistics.

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

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. *arXiv preprint arXiv:2305.12152*.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407.

Anh Phan Tuan, Bach Tran, Thien Huu Nguyen, Linh Ngo Van, and Khoat Than. 2020. Bag of biterms modeling for short texts. *Knowledge and Information Systems*, 62(10):4055–4090.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ngo Van Linh, Tran Xuan Bach, and Khoat Than. 2022. A graph convolutional topic model for short and noisy text streams. *Neurocomputing*, 468:345–359.

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022. Representing mixtures of word embeddings with mixtures of topic embeddings. In *The Tenth International Conference on Learning Representations, ICLR 2022*.

H. Wang, N. Prakash, N. Hoang, M. Hee, U. Naseem, and R. Lee. 2023. Prompting large language models for topic modeling. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liang-Ming Pan, and Anh Tuan Luu. 2023a. Infoctm: A mutual information maximization perspective of cross-lingual topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13763–13771.

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.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):1–30.

Xiaobao Wu, Thong Nguyen, Delvin Ce 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*.

Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024c. Fastopic: A fast, adaptive, stable, and transferable topic modeling paradigm. *arXiv preprint arXiv:2405.17978*.

Xiaobao Wu, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. 2024d. On the affinity, rationality, and diversity of hierarchical topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19261–19269.

Weijie Xu, Xiaoyu Jiang, Srinivasan Sengamedu Hanumantha Rao, Francis Iannacci, and Jinjin Zhao. 2023. vontss: vmf based semi-supervised neural topic modeling with optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2023*,

page 4433–4457. Association for Computational Linguistics.

Yishi Xu, Dongsheng Wang, Bo Chen, Ruiying Lu, Zhibin Duan, and Mingyuan Zhou. 2022. Hyperminer: Topic taxonomy mining with hyperbolic embedding. *Preprint*, arXiv:2210.10625.

Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 625–636.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 28. Curran Associates, Inc.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893. Association for Computational Linguistics.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2021. Neural topic model via optimal transport. In *9th International Conference on Learning Representations, ICLR 2021*.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2022. Neural topic model via optimal transport. *Preprint*, arXiv:2008.13537.

Algorithm 1 Learning HiCOT

Input: Document collection $\mathbf{X} = \{\mathbf{x}_d\}_{d=1}^{D}$, number of topics K, pretrained word embedding $\mathbf{W}_{\text{pretrained}}$, pretrained document embedding model f_{doc}, total number of training epochs N, threshold epoch for clustering $\mathbf{M}$, clustering update interval $\mathbf{I}$.
Output: Model parameters δ, word embedding $\mathbf{W}$, topic embedding $\mathbf{T}$.

```
Initialize W = W_pretrained
Random initialize T
Initialize document embedding E, with e_i = f_doc(x_i) for i = {1, 2, ··· , D}
for t = 1, 2, . . . , N do
    for each minibatch B do
        if t < M then
            // Stage 1
            Update Q following the Equation 7
            Update π as the solution of problem 5 by Sinkhorn's algorithm
            Update ψ using Sinkhorn algorithm to sovle problem 14
            Calculate L = L_TM + λ_DT L_DT + λ_ECR L_ECR
            Update W, T, E, δ with a gradient step based on the loss L.
        else
            // Stage 2
            if t = M then
                Use hierarchical clustering algorithm to perform clustering
            end if
            Update Q according to Equation 7
            Update π, ψ using Sinkhorn algorithm to sovle problem 5 and 14, respectively
            Calculate L_CL = λ_CLT L_CLT + λ_CLC L_CLC
            Calculate L = L_CL + L_TM + λ_DT L_DT + λ_ECR L_ECR
            Update W, T, E, δ with a gradient step based on the loss L.
            if t % I = 0 then
                Use hierarchical clustering algorithm to perform re-clustering
            end if
        end if
    end for
end for
```

A Algorithm

The detailed training algorithm for HiCOT is presented in Algorithm A.

B Related work

Topic Models and Neural Topic Models. The goal of topic modeling is to find hidden topics in a corpus of documents. Traditionally, this has been tackled with graphical probabilistic models (Hofmann, 1999; Blei et al., 2003), with extensions for specialized settings such as short texts (Tuan et al., 2020; Ha et al., 2019; Nguyen et al., 2022a; Mai et al., 2016) and streaming data (Duc et al., 2017; Nguyen et al., 2019, 2022b, 2025c, 2021). More recently, neural models have gained prominence for their superior generalization and performance (Wu et al., 2024a; Srivastava and Sutton, 2017; Dieng et al., 2020; Wu et al., 2023b; Pham et al., 2024b). Most neural topic models are based on the Variational Autoencoder (VAE) framework (Kingma and Welling, 2013), where an encoder infers a document's topic distribution and a decoder reconstructs the text using a topic-word distribution. Recent improvements include integrating pre-trained language model (PLM) or word embeddings (Bach et al., 2021; Van Linh et al., 2022; Reimers, 2019; Brown et al., 2020) into the encoder (Wu et al., 2023a; Han et al., 2023). Other approaches directly cluster document representations to form topics (Grootendorst, 2022; Sia et al., 2020; Zhang et al., 2022), though they often lack clear document-specific topic proportions. Additionally, large language models have been used to generate conceptual topic descriptions (Wang et al., 2023; Pham et al., 2024a), but these methods typically struggle with providing detailed word distributions. Notably, Wu et al. (2024b) propose an Optimal Transport-based framework (Peyré and Cuturi, 2018) that effectively captures the semantic relationships among documents, topics, and word embeddings.

Clustering methods. Clustering plays a key role in unsupervised learning by grouping data based on similarities. Traditional algorithms like KMeans (MacQueen, 1967), Hierarchical Agglomerative Clustering (HAC) (Murtagh and Contreras, 2012), and HDBSCAN (Campello et al., 2013) are widely used. In topic modeling, these methods are applied to cluster documents (based on topic distributions

20NG					
Model	Distance metric	C_V	Purity	NMI	TD
HiCOT + HAC	Euclidean	0.451	0.626	0.583	0.852
	Cosine	0.447	0.624	0.580	0.851
HiCOT + HDBSCAN	Euclidean	0.449	0.628	0.580	0.863
	Cosine	0.445	0.630	0.585	0.856
GoogleNews					
Model	Distance metric	C_V	Purity	NMI	TD
HiCOT + HAC	Euclidean	0.454	0.465	0.657	0.920
	Cosine	0.457	0.466	0.655	0.923
HiCOT + HDBSCAN	Euclidean	0.458	0.464	0.653	0.921
	Cosine	0.458	0.464	0.654	0.921

Table 8: Impact of Hierarchical Clustering Methods on the 20NG and GoogleNews Datasets with 50 Topics, Using Euclidean and Cosine Distance Metrics for Contrastive Learning.

from LDA (Blei et al., 2003) or neural models) (Wu et al., 2024a), words (using embeddings like Word2Vec (Mikolov et al., 2013) or BERT (Devlin, 2018), or directly for topic formation (clustering document embeddings, e.g., Top2Vec (Angelov, 2020), BERTopic (Grootendorst, 2022)). Recent advances integrate clustering with neural approaches, showing superior performance, particularly when clustering embeddings.

C Impact of Clustering Algorithm and Distance Metric on Contrastive Learning

To assess the impact of hierarchical clustering algorithms, we conduct experiments with HAC (Murtagh and Contreras, 2012) and HDBSCAN (Campello et al., 2013). In contrastive learning, we also evaluate different distance metrics, using same metric for $\mathcal{L}_{\text{CLT}}$ and $\mathcal{L}_{\text{CLC}}$ for simplicity. Our experiment considers two metrics: Euclidean and Cosine, on the 20NG and GoogleNews datasets, with $K = 50$ topics. As demonstrated in Table 8, the selection of hierarchical clustering algorithms and distance metrics within contrastive learning framework has minimal impact on model performance. This robustness highlights that the effectiveness of our approach mainly comes from its architectural and algorithmic innovations, rather than dependence on choosing clustering algorithms or distance metrics.

D Comparison with Other Neural Topic Models

D.1 Comparison with Clustering-based Neural Topic Models

50 Topics	**20NG**				**AGNews**				**IMDB**			
	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD
BERTopic	0.382	0.376	0.448	<u>0.680</u>	<u>0.390</u>	0.687	0.340	<u>0.735</u>	0.341	0.677	**0.086**	<u>0.505</u>
C-Top2Vec	<u>0.408</u>	<u>0.564</u>	<u>0.496</u>	0.577	0.371	<u>0.842</u>	<u>0.356</u>	0.417	<u>0.403</u>	<u>0.686</u>	0.049	0.149
HiCOT	**0.451**	**0.626**	**0.583**	**0.852**	**0.446**	**0.857**	**0.412**	**0.992**	**0.404**	**0.737**	<u>0.082</u>	**0.837**

Table 9: Evaluation results on standard datasets, assessed using C_V, TD, Purity and NMI under $K = 50$ topics. The highest-performing results are marked in bold, while the second-best values are underlined.

We conducted experiments on three standard datasets with 50 topics, comparing two clustering-based models: BERTopic (Grootendorst, 2022), which clusters document embeddings and discovers topics using

TF-IDF, and C-Top2Vec (Angelov and Inkpen, 2024), which leverages contextual token embeddings for multi-vector document representations. Our experimental results (see Table 9) demonstrate that HiCOT consistently achieves superior performance across all datasets. Specifically, HiCOT attains the highest scores across all metrics. This indicates that the topics generated by HiCOT not only enhance the quality of document-topic distributions but are also semantically coherent and sufficiently diverse, effectively mitigating topic collapse.

In contrast, although C-Top2Vec achieves the second-highest C_V score, it suffers from extremely low TD values. These low TD scores suggest that the model's topics lack the necessary diversity to capture a wide range of themes. Consequently, this leads to significant topic redundancy, reducing the model's ability to generate well-separated and meaningful topics. As illustrated in Table 20, this highlights the fundamental limitations of C-Top2Vec in maintaining both topic quality and diversity. These findings further underscore HiCOT's effectiveness as a robust solution for clustering-based neural topic modeling

D.2 Comparison with Auxiliary Loss-based Models

To further evaluate the effectiveness of our framework, we compare it against a recent approach that incorporates auxiliary training objectives to enhance topic coherence. Specifically, we consider the method proposed by Li et al. (2023), which introduces a novel auxiliary loss function designed to improve the performance of neural topic models.

We conducted experiments by integrating this auxiliary loss two baseline models: ETM and ECRTM (denoted as ETM + Aux and ECRTM + Aux, respectively). We then compare these models with their HiCOT-enhanced counterparts (ETM + HiCOT and ECRTM + HiCOT) on the 20NG dataset with $K = 50$ topics. The results are presented in Table 10. Our findings show that models augmented with HiCOT consistently outperform those using the auxiliary loss from Li et al. (2023) across all metrics. These results highlight the strength of HiCOT in enhancing topic modeling performance.

Model	**ETM**				**ECRTM**			
	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD
Baseline	0.375	0.347	0.319	0.704	0.431	0.560	0.524	**0.964**
+ Auxiliary Loss	0.374	0.420	0.394	0.681	0.427	0.593	0.547	0.922
+ HiCOT (ours)	**0.444**	**0.581**	**0.493**	**0.756**	**0.451**	**0.626**	**0.583**	0.852

Table 10: Comparison of ETM and ECRTM variants on the 20NG dataset with 50 topics. The highest-performing results are marked in bold.

D.3 Comparison with the CombinedTM VAE Baseline

We evaluated the performance of our proposed HiCOT against CombinedTM baseline (Bianchi et al., 2021) on three standard benchmark datasets: 20NG, AGNews, and IMDB. Experiments were conducted with topic numbers set to 50 and 100. As shown in Table 11, HiCOT consistently outperforms CombinedTM across all datasets and metrics. These results demonstrate the effectiveness of our approach relative to CombinedTM (Bianchi et al., 2021), highlighting its effectiveness within the class of VAE-based topic models.

D.4 Comparison with Short-Text Topic Models

We have conducted comparative experiments with GloCOM (Nguyen et al., 2025a) on the SearchSnippets and GoogleNews datasets (with $K = 50$ and $K = 100$ topics). It is pertinent to note that GloCOM is specifically designed for short-text topic modeling, leverages aggregation techniques tailored for this data type. As shown in Table 12, our proposed method, **without necessitating specialized short-text strategies, achieves good performance** in terms of overall topic quality on the SearchSnippets dataset compared to GloCOM. This result underscores the effectiveness of our approach.

Furthermore, our method has the potential to be integrated with techniques used in GloCOM. A potential direction for future work is incorporating clustering mechanisms, or designing objective functions based

Topics	Model	20NG				AGNews				IMDB			
		C_V	Purity	NMI	TD	C_V	Purity	NMI	TD	C_V	Purity	NMI	TD
50	CTM	0.426	0.565	0.465	0.792	**0.460**	0.801	0.324	0.950	0.395	0.687	0.051	0.709
	HiCOT	**0.451**	**0.626**	**0.583**	**0.852**	0.446	**0.857**	**0.412**	**0.992**	**0.404**	**0.737**	**0.082**	**0.837**
100	CTM	**0.424**	0.591	0.465	0.718	0.415	0.827	0.344	0.648	0.376	0.686	0.042	0.476
	HiCOT	**0.424**	**0.652**	**0.568**	**0.741**	**0.435**	**0.861**	**0.388**	**0.960**	**0.388**	**0.739**	**0.071**	**0.733**

Table 11: Performance comparison of HiCOT and CombinedTM on 20NG, AGNews, and IMDB datasets for 50 and 100 topics, using C_V, Purity, NMI and TD. The highest-performing results are marked in bold.

on Optimal Transport (OT) and contrastive learning. These enhancements have the potential to further improve our approach, and we intend to explore them in future research.

Topics	Model	SearchSnippets				GoogleNews			
		C_V	Purity	NMI	TD	C_V	Purity	NMI	TD
50	GloCOM	0.453	0.806	0.502	0.956	0.475	0.586	0.817	0.999
	HiCOT	0.460	0.818	0.478	1.000	0.454	0.465	0.657	0.920
100	GloCOM	0.443	0.822	0.501	0.920	0.450	0.761	0.900	0.944
	HiCOT	0.449	0.857	0.480	0.940	0.470	0.763	0.864	0.802

Table 12: Comparison between GloCOM and HiCOT on SearchSnippets and GoogleNews datasets.

E Text Classification

To assess extrinsic performance, we conduct text classification experiments as downstream tasks following (Wu et al., 2023b). Specifically, we utilize document-topic distributions generated by topic models as document features and train SVMs to predict class of each document. For evaluation, we employ three standard datasets with 50 topics. Performance is evaluated using Accuracy (Acc) and F1 score. As shown in Figure 2, HiCOT achieves performance comparable to recent state-of-the-art models, such as NeuroMax and FASTopic, while being better than other baseline methods across three datasets. Moreover, it notably enhances topic quality, as discussed in Section 4.2. These results highlight the effectiveness of HiCOT in downstream classification tasks.

F Topic Model Evaluation with Large Language Models

Following (Stammbach et al., 2023), we utilize two LLM-based evaluation tasks to assess topic model quality: **Rating Task** and **Intruder Detection Task**. In the rating task, an LLM assigns a coherence score to topic words on a scale from 1 to 3 ("1" = not very related, "2" = moderately related, "3" = very related), reflecting the semantic relatedness of the top 10 words per topic. With intruder detection task, we randomly select five words from the top 10 topic words and introduce an intruder word from a different topic (not among the top 50 words of the current topic), challenging the LLM to identify the outlier. In all experiments, we use the `gemini-2.0-flash` model.

For the rating task, Table 13 presents the mean of LLM ratings over three evaluations across 50 topics for each model and dataset. The results indicate that HiCOT achieves the highest average scores, underscoring HiCOT's superior topic coherence. For the intruder detection task, Figure 3 illustrates the distribution of detection accuracies across three test series per topic for 50 topics, with accuracies averaged over these series for each model. HiCOT exhibits the highest median accuracy, suggesting that our method generates more distinct and coherent topics.

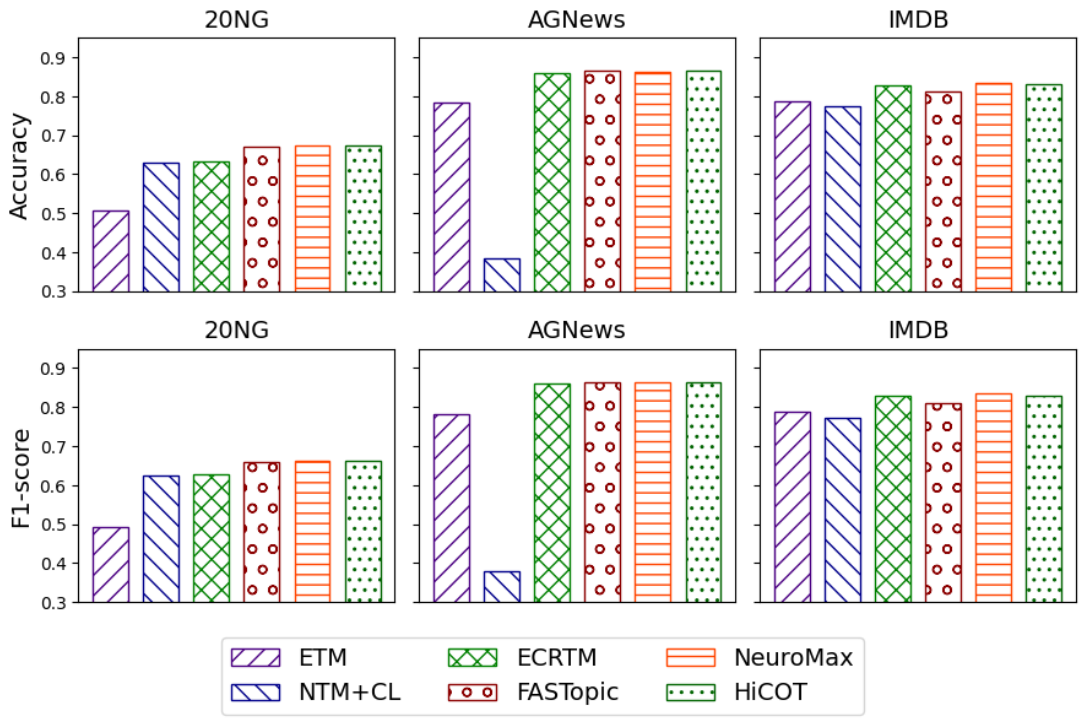

Figure 2: Comparison of text classification performance for HiCOT and baseline models across three standard datasets with 50 topics.

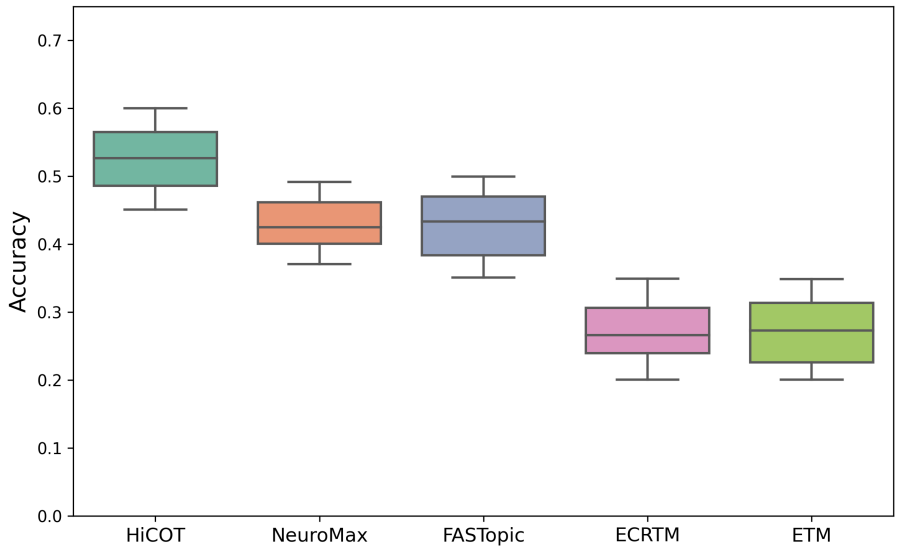

Figure 3: Distribution of average intruder detection accuracies across three test series per topic for 50 topics on the AGNews dataset.

Dataset	ETM	ECRTM	NeuroMax	FASTopic	HiCOT
20NG	2.01	1.96	1.99	1.01	2.04
SearchSnippets	1.78	1.45	1.80	1.83	1.88
GoogleNews	1.37	1.36	1.67	1.20	1.74

Table 13: Average scores for 50 topics across datasets and models, based on LLM ratings over three evaluation runs per topic.

Dataset	# of texts	average text length	# of labels	vocab size
20NG	18846	110.5	20	5000
AGNews	12500	20.1	4	5000
IMDB	50000	95.0	2	5000
SearchSnippets	12294	14.4	8	4618
GoogleNews	11019	5.8	152	3473

Table 14: Dataset statistics after preprocessing.

G Experiment Details

G.1 Implementation Details.

All experiments are performed on a system featuring a GeForce RTX 3090 GPU (24GB RAM), using PyTorch 2.1.0 with CUDA 12.1 in a Python 3.12 environment. The model is trained for 400 epochs with a batch size of 128, using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.002. The weight hyperparameters are selected from the following ranges:

- $\lambda_{\text{ECR}} \in [10, 30, 40, 50, 100, 200]$
- $\lambda_{\text{DT}} \in [0.5, 0.7, 1, 2, 5, 10]$
- $\lambda_{\text{CLT}}, \lambda_{\text{CLC}} \in [0.5, 1, 2, 5, 10]$

The clustering process begins at a specified threshold epoch $\mathbf{M}$ and is periodically updated at fixed intervals of $\mathbf{I}$. These parameters are searched in ranges as follows:

- Threshold epoch $\mathbf{M} \in [50, 70, 100, 150, 200, 250]$
- Clustering update interval $\mathbf{I} \in [30, 50, 70, 100, 150, 200]$

G.2 Dataset Statistics

Our evaluation was conducted on several widely used benchmark datasets: three standard corpora—**20 News Groups (20NG)** (Lang, 1995), **AGNews** (Zhang et al., 2015), and **IMDB** (Maas et al., 2011)—along with two informal datasets: **SearchSnippets** (Phan et al., 2008), containing short, noisy text snippets, and **GoogleNews** (Yin and Wang, 2016), composed of brief article headlines.

For standard datasets, we followed the preprocessing pipeline outlined in (Wu et al., 2023b) to construct bag-of-words representations. For short-text corpora, we utilized preprocessed versions provided by STTM library[2] (Qiang et al., 2022). Further refinement steps were applied, including the removal of words appearing fewer than 3 times and the exclusion of documents containing fewer than 2 tokens. All preprocessing was performed using the TopMost framework[3]. The final dataset statistics after processing are summarized in Table 14.

[2] https://github.com/qiang2100/STTM

[3] https://github.com/bobxwu/topmost

H Influence of coefficient

This sensitivity was part of our hyperparameter tuning process, as documented in the Appendix G.1. To provide the concrete details requested, we now present the specific results from this analysis. Table 15 illustrate the impact of varying each coefficient, showing results on a representative dataset 20NG (for λ_{DT}), IMDB (for λ_{CLT}), GoogleNews (for λ_{CLC}), and AGNews (for λ_{ECR}) respectively (all $K = 50$ topics). For example, on 20NG, setting $\lambda_{\mathrm{DT}} = 1$ yields a good performance, while performance slightly changes with other values. Similarly, for IMDB, $\lambda_{\mathrm{CLT}} = 1$ achieves strong results, with performance degrading at significantly lower or higher values. This detailed analysis demonstrates the model's sensitivity profile to these hyperparameters and provides clear empirical justification for the values selected and reported in our main experiments.

20NG					**AGNews**				
λ_{DT}	$\mathrm{C_V}$	Purity	NMI	TD	λ_{ECR}	$\mathrm{C_V}$	Purity	NMI	TD
0.5	0.449	0.621	0.568	0.771	10	0.469	0.816	0.336	0.617
0.7	0.447	0.636	0.579	0.775	30	0.470	0.832	0.344	0.912
1	0.451	0.626	0.583	0.852	40	0.478	0.849	0.375	0.776
2	0.451	0.644	0.587	0.768	50	0.446	0.857	0.412	0.992
5	0.438	0.649	0.585	0.868	100	0.436	0.754	0.282	0.987
10	0.437	0.635	0.581	0.849	200	0.416	0.675	0.218	0.968
IMDB					**GoogleNews**				
λ_{CLT}	$\mathrm{C_V}$	Purity	NMI	TD	λ_{CLC}	$\mathrm{C_V}$	Purity	NMI	TD
0.5	0.389	0.685	0.067	0.973	0.5	0.445	0.456	0.663	0.992
1	0.404	0.737	0.082	0.837	1	0.454	0.465	0.657	0.920
2	0.390	0.679	0.065	0.973	2	0.437	0.457	0.662	0.987
5	0.397	0.682	0.055	0.979	5	0.431	0.455	0.660	0.981
10	0.393	0.686	0.055	0.984	10	0.431	0.465	0.664	0.987

Table 15: Evaluation metrics across datasets with different hyperparameters.

I Training Time

HiCOT's integration of multiple techniques is crucial for enhancing topic quality and representation learning, but it also increases computational cost compared to less complex models. This design choice reflects a deliberate trade-off: the model prioritizes improvements in topic quality and enhanced representation learning, which inherently require a increase in computational complexity. However, we emphasize that this additional cost is well justified, particularly in light of the significant improvements observed in evaluation metrics and model's efficient inference capabilities. We measured average training time per epoch (in seconds) on two datasets: 20NG and AGNews, with 50 topics, as shown in Table 16. Despite this increase, HiCOT's training time remains reasonable, suggesting that the performance benefits of HiCOT come at an acceptable computational cost.

Datasets	**ETM**	**NTM+CL**	**ECRTM**	**FASTopic**	**NeuroMax**	**HiCOT**
20NG	0.263	1.104	1.332	2.161	2.823	3.033
AGNews	0.236	0.826	1.132	1.911	1.989	2.512

Table 16: Training time per epoch (seconds) for HiCOT and baselines on 20NG and AGNews with 50 topics.

J Statistical Significance Test

We conducted statistical paired t-tests comparing our HiCOT against FASTopic and NeuroMax on the AGNews dataset with 50 topics. Table 17 presents the p-values for TD, NMI, Purity, and C_V metrics. The result show that our improvement is statistically significant with all p-value < 0.05, validating our reported improvements.

Metric	HiCOT vs. FASTopic	HiCOT vs. NeuroMax
TD	0.0123	0.0087
NMI	0.0356	0.0194
Purity	0.0412	0.0335
C_V	0.0278	0.0449

Table 17: p-values from paired t-tests comparing HiCOT to FASTopic and NeuroMax on AGNews dataset with 50 topics.

We evaluated HiCOT across multiple random initializations to assess its stability. Due to space constraints, standard deviations for results in 4.2 are omitted from the main paper. Here, Table 18 and 19 present the means and standard deviations for C_V, Purity, NMI, and TD metrics, corresponding to Table 1 and 2 in the main paper. These results show HiCOT's consistent and strong performance, highlighting its robustness across initializations. Baseline results are reported in NeuroMax (Pham et al., 2024b) and GloCOM (Nguyen et al., 2025a).

K Expanded Visualization of Embedding Space

We extend the visualization of learned topic and word embeddings using t-SNE (van der Maaten and Hinton, 2008) to additional datasets, including SearchSnippets and GoogleNews, each with 50 topics. Figures 4 and 5 further demonstrate that HiCOT preserves topic structures and mitigates embedding collapse.

L Examples of Topics

Below are the discovered topics of HiCOT with the 20NG and IMDB datasets under 50 topics, as described in Tables 21 and 22.

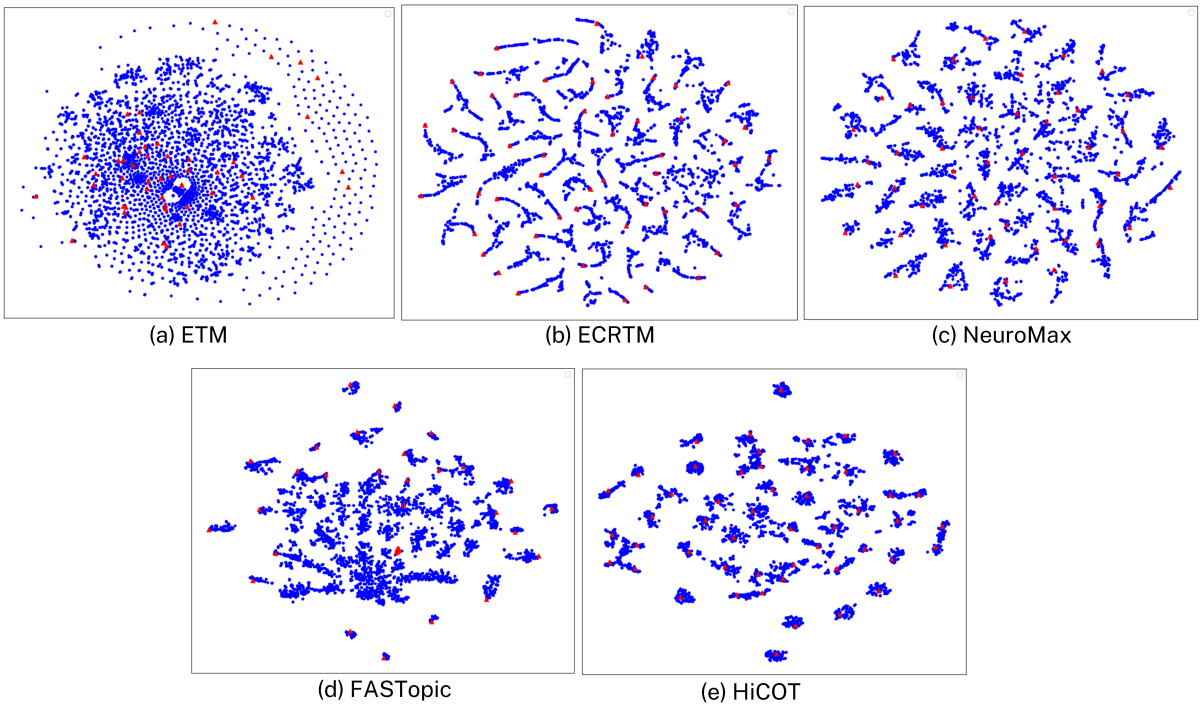

Figure 4: t-SNE visualization of word embeddings (●) and topic embeddings (▲) for SearchSnippets dataset with 50 topics.

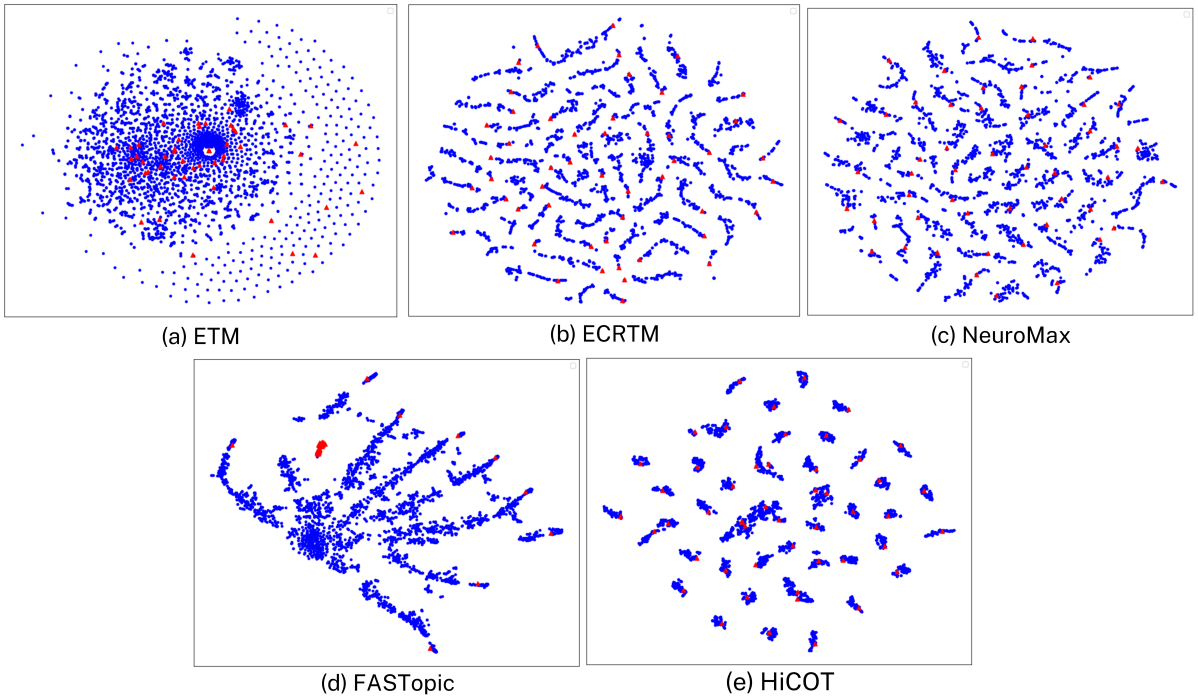

Figure 5: t-SNE visualization of word embeddings (●) and topic embeddings (▲) for GoogleNews dataset with 50 topics.

50 Topics	**20NG**				**AGNews**				**IMDB**			
	C_v	Purity	NMI	TD	C_v	Purity	NMI	TD	C_v	Purity	NMI	TD
ETM	0.375	0.347	0.319	0.704	0.364	0.679	0.224	0.819	0.346	0.660	0.038	0.557
NTM + CL	$0.437_{\pm 0.005}$	$0.582_{\pm 0.002}$	$0.491_{\pm 0.007}$	$0.802_{\pm 0.002}$	$0.440_{\pm 0.002}$	$0.322_{\pm 0.005}$	$0.100_{\pm 0.002}$	$0.441_{\pm 0.001}$	$0.396_{\pm 0.003}$	$0.657_{\pm 0.005}$	$0.044_{\pm 0.001}$	$0.617_{\pm 0.006}$
ECRTM	0.431	0.560	0.524	0.964	0.466	0.802	0.367	0.961	0.393	0.694	0.058	0.974
FASTopic	$0.427_{\pm 0.005}$	$0.583_{\pm 0.010}$	$0.528_{\pm 0.001}$	$0.980_{\pm 0.011}$	$0.379_{\pm 0.002}$	$0.831_{\pm 0.007}$	$0.352_{\pm 0.003}$	$0.960_{\pm 0.015}$	$0.371_{\pm 0.002}$	$0.683_{\pm 0.007}$	$0.055_{\pm 0.002}$	$0.969_{\pm 0.015}$
NeuroMax	0.435	0.623	0.570	0.912	0.385	0.804	0.410	0.952	0.402	0.709	0.061	0.936
HiCOT	$0.451_{\pm 0.012}$	$0.626_{\pm 0.006}$	$0.583_{\pm 0.009}$	$0.852_{\pm 0.007}$	$0.446_{\pm 0.008}$	$0.857_{\pm 0.003}$	$0.412_{\pm 0.001}$	$0.992_{\pm 0.004}$	$0.404_{\pm 0.003}$	$0.737_{\pm 0.009}$	$0.082_{\pm 0.004}$	$0.837_{\pm 0.006}$
100 Topics	**20NG**				**AGNews**				**IMDB**			
	C_v	Purity	NMI	TD	C_v	Purity	NMI	TD	C_v	Purity	NMI	TD
ETM	0.369	0.394	0.339	0.573	0.371	0.674	0.204	0.773	0.341	0.648	0.037	0.371
NTM + CL	$0.420_{\pm 0.003}$	$0.626_{\pm 0.013}$	$0.490_{\pm 0.008}$	$0.624_{\pm 0.004}$	$0.415_{\pm 0.005}$	$0.280_{\pm 0.003}$	$0.050_{\pm 0.002}$	$0.277_{\pm 0.002}$	$0.382_{\pm 0.003}$	$0.705_{\pm 0.012}$	$0.044_{\pm 0.009}$	$0.492_{\pm 0.005}$
ECRTM	0.405	0.555	0.494	0.904	0.416	0.812	0.428	0.981	0.373	0.694	0.049	0.887
FASTopic	$0.400_{\pm 0.002}$	$0.622_{\pm 0.001}$	$0.522_{\pm 0.008}$	$0.861_{\pm 0.014}$	$0.385_{\pm 0.009}$	$0.833_{\pm 0.015}$	$0.330_{\pm 0.004}$	$0.912_{\pm 0.008}$	$0.369_{\pm 0.003}$	$0.680_{\pm 0.007}$	$0.048_{\pm 0.004}$	$0.886_{\pm 0.008}$
NeuroMax	0.412	0.602	0.516	0.913	0.406	0.828	0.389	0.957	0.381	0.706	0.059	0.870
HiCOT	$0.424_{\pm 0.004}$	$0.652_{\pm 0.007}$	$0.568_{\pm 0.006}$	$0.741_{\pm 0.002}$	$0.435_{\pm 0.008}$	$0.862_{\pm 0.009}$	$0.388_{\pm 0.005}$	$0.960_{\pm 0.008}$	$0.388_{\pm 0.008}$	$0.739_{\pm 0.009}$	$0.071_{\pm 0.002}$	$0.733_{\pm 0.005}$

Table 18: Evaluation of topic models on 20NG, AGNews, and IMDB datasets with 50 and 100 topics. Each cell reports mean ± standard deviation.

50 Topics	**SearchSnippets**				**GoogleNews**			
	C_v	Purity	NMI	TD	C_v	Purity	NMI	TD
ETM	0.397	0.688	0.389	0.594	0.402	0.366	0.560	0.916
NTM + CL	$0.403_{\pm 0.002}$	$0.215_{\pm 0.008}$	$0.030_{\pm 0.006}$	$0.532_{\pm 0.011}$	$0.433_{\pm 0.009}$	$0.041_{\pm 0.003}$	$0.005_{\pm 0.001}$	$0.301_{\pm 0.008}$
ECRTM	0.450	0.711	0.419	0.998	0.441	0.396	0.615	0.987
FASTopic	$0.356_{\pm 0.021}$	$0.793_{\pm 0.006}$	$0.497_{\pm 0.009}$	$0.519_{\pm 0.030}$	$0.401_{\pm 0.015}$	$0.252_{\pm 0.016}$	$0.570_{\pm 0.031}$	$0.235_{\pm 0.014}$
NeuroMax	$0.427_{\pm 0.012}$	$0.743_{\pm 0.020}$	$0.427_{\pm 0.006}$	$0.920_{\pm 0.028}$	$0.409_{\pm 0.017}$	$0.359_{\pm 0.011}$	$0.590_{\pm 0.026}$	$1.000_{\pm 0.005}$
HiCOT	$0.460_{\pm 0.013}$	$0.818_{\pm 0.027}$	$0.478_{\pm 0.018}$	$1.000_{\pm 0.020}$	$0.454_{\pm 0.012}$	$0.465_{\pm 0.021}$	$0.657_{\pm 0.010}$	$0.920_{\pm 0.006}$
100 Topics	**SearchSnippets**				**GoogleNews**			
	C_v	Purity	NMI	TD	C_v	Purity	NMI	TD
ETM	0.389	0.691	0.365	0.448	0.398	0.554	0.713	0.677
NTM + CL	$0.406_{\pm 0.011}$	$0.217_{\pm 0.010}$	$0.020_{\pm 0.001}$	$0.394_{\pm 0.004}$	$0.432_{\pm 0.009}$	$0.039_{\pm 0.002}$	$0.005_{\pm 0.001}$	$0.367_{\pm 0.004}$
ECRTM	0.432	0.789	0.443	0.966	0.418	0.342	0.491	0.991
FASTopic	$0.350_{\pm 0.003}$	$0.801_{\pm 0.009}$	$0.466_{\pm 0.011}$	$0.463_{\pm 0.017}$	$0.366_{\pm 0.005}$	$0.237_{\pm 0.012}$	$0.459_{\pm 0.016}$	$0.100_{\pm 0.003}$
NeuroMax	$0.439_{\pm 0.007}$	$0.854_{\pm 0.018}$	$0.472_{\pm 0.009}$	$0.960_{\pm 0.014}$	$0.427_{\pm 0.013}$	$0.664_{\pm 0.008}$	$0.834_{\pm 0.016}$	$0.915_{\pm 0.022}$
HiCOT	$0.449_{\pm 0.003}$	$0.857_{\pm 0.008}$	$0.480_{\pm 0.011}$	$0.940_{\pm 0.010}$	$0.470_{\pm 0.008}$	$0.763_{\pm 0.005}$	$0.864_{\pm 0.012}$	$0.802_{\pm 0.009}$

Table 19: Evaluation of topic models on SearchSnippets and GoogleNews datasets with 50 and 100 topics. Each cell reports mean ± standard deviation.

Topic	Top 10 phrases
Topic #7	automotive concepts, subject automotive, experience riding, new motorcycles, eisa bus, dealer service, road suite, motorcycle tip, corporation lines, buying motorcycle
Topic #17	automotive concepts, subject automotive, dealer service, auto dealers, reducing dealer, performance cars, car buying, reduce dealer, dealer profit, corporation lines
Topic #20	automotive concepts, subject automotive, full auto, garage ama, road suite, drives feature, trucks read, road independence, experience riding, auto dealers
Topic #16	list biblical, follow teachings, online bible, gospel accounts, spiritual needs, daily verse, positive belief, teachings james, requires faith, spiritual world
Topic #24	list biblical, online bible, follow teachings, reading bible, spiritual world, gospel accounts, spiritual needs, daily verse, peoples spiritual, accept messiah
Topic #2	scripture arrogant, follow teachings, displaying ignorance, necessarily arrogant, agree definition, strong atheism, requires faith, reason believe, weak atheism, positive belief
Topic #4	strong atheism, introduction atheism, weak atheism, atheists organization, positive belief, reason believe, caused atheism, weak atheist, requires faith, share beliefs
Topic #3	keywords encryption, importance encryption, encryption management, encryption product, encryption providing, encryption differ, encryption method, strong encryption, approach encryption, cover encryption
Topic #32	importance encryption, encryption product, keywords encryption, encryption providing, encryption management, commercial encryption, nsa classified, trust nsa, encryption wiretap, encryption devices
Topic #5	organization portal, organization express, inc lines, organization netcom, corp internet, organization telecom, reply organization, corporation lines, organization ncr, organization polytechnic
Topic #14	corporation lines, inc lines, organization express, organization portal, organization netcom, corp internet, organization sdpa, forsale organization, corp distribution, incorporated lines
Topic #22	reply organization, organization portal, content communications, computers social, organization netcom, proposed newsgroup, organization express, corp internet, organization sdpa, group newsreader
Topic #25	reply organization, organization ncr, associated press, organization nyx, organization bell, organization bnr, organization portal, content communications, washington post, inc newsreader
Topic #38	federal agents, agencies messages, organization ncr, wiretap drug, organization nyx, secret service, organization california, organization bell, agencies tools, determine agencies
Topic #41	keywords frequently, join organized, beyond column, join reform, organization nyx, research cpr, organization ncr, keywords jpl, word processing, organization california
Topic #49	organization ncr, organization optilink, initiative congress, organization nyx, public spending, organization sdpa, policy research, administration saying, public interest, policy members

Table 20: Top 10 phrases of some topics from 50 topics from 20NG by C-Top2Vec. Repeated phrases are underlined. With the topic diversity value of 0.577, these topics semantically collapse towards each other with many repetitive phrases.

HiCOT with 20NG (K = 50)

Topic #1 : pitt marriage derek bonds morgan professor richardson rush timothy spending
Topic #2 : sahak anatolia turkish turkey armenians greeks armenian greek proceeded inhabitants
Topic #3 : escrow encryption encrypted phones chip enforcement rsa **privacy** tapped algorithm
Topic #4 : battery intellect beast tank discharge gear cad prince tanks cartridge
Topic #5 : nhl standings oilers espn playoff bruins penguins champs hockey traded
Topic #6 : simms centris lciii slots quadra vga powerbook vlb ram mac
Topic #7 : die oil anymore everybody minority lot sell going think weekend
Topic #8 : font fonts xterm terminal screen workstation salvation width christ patrick
Topic #9 : **widget** window sunos cursor display server bitmap client resource int
Topic #10: braves pitcher pitching partners bps votes serial characters hitter analog
Topic #11: jpeg **widget** gif graphics formats pub format sunos mac color
Topic #12: insurance taxes tax printer radius car health billion servicing scanner
Topic #13: int output null max file stream entry tiff byte files
Topic #14: **privacy** rsa pub encryption anonymous ftp cancer pgp requests networks
Topic #15: honda motorcycles rear levine ama tire bmw dog gardner wheels
Topic #16: cds manuals **sale** warranty rider disks buyer amp shipping excellent
Topic #17: armenians armenian **soldiers** tragedy neighbors president crowd turkey burned secretary
Topic #18: **firearms** handgun gun **weapons** arms crime possession constitutional violent firearm
Topic #19: suck duo games clipper melbourne sony alex game yamaha brad
Topic #20: god theists atheists **bible** christianity **gods** beliefs morality rosenau believers
Topic #21: gordon racism luis netherlands postscript science genocide yugoslavia ruling terrorism
Topic #22: windows dos diamond swap logo mouse microsoft driver novell drivers
Topic #23: henry rutgers pens nasa troy temple fax arbor museum calif
Topic #24: optilink yankees won switzerland toyota italy stadium angels suspension greece
Topic #25: min reagan det portal que micro danny livni benedikt howard
Topic #26: bike bikes son duke killed lock motorcycle hudson murder suicide
Topic #27: church captain sexual homosexual christians churches militia orientation orthodox arrogance
Topic #28: theodore satan gary kim sutherland keith serdar joy icon bryan
Topic #29: mormons nsa christian koresh alien nra classified consent dick incidents
Topic #30: pitchers batting pitches players mets defensive phillies morris scored season
Topic #31: gaza jake arabs israel arab palestine civilians inhabitants jerusalem **soldiers**
Topic #32: msg shameful skepticism foods sensitivity yeast diagnosed chinese patients symptoms
Topic #33: atf bds survivors reno fbi murders gun **weapons** **firearms** blast
Topic #34: helmet resurrection philadelphia moon detroit heaven toronto oakland scripture icons
Topic #35: bmp motif rec hawks lehigh beyer bbs ottawa petaluma ext
Topic #36: male nyx locks sox beer taste don men compression percent
Topic #37: max morality solntzewpdsgicom schneider cookamunga tourist cliff objective islam atheists
Topic #38: voltage amp circuit circuits wire wires wiring detector detectors cable
Topic #39: scsi ide controller drives bios jumper drive vlb floppy boot
Topic #40: jesus sins romans unto god messiah trinity **bible** **gods** sin
Topic #41: devils rangers islanders radar champions andre roger capitals stanley playoffs
Topic #42: revelation jewish catholics feelings jimmy faith clh utter occupied label
Topic #43: spacecraft zoology satellite jpl missions payload launch orbital mission atmosphere
Topic #44: cubs bnr eisa tin helsinki finland mathew newsreader isa beleive
Topic #45: blues accelerator shaped boots priest space mask interrupt span image
Topic #46: xxdate nuntius useragent xxmessage csutexasedu gerald olchowy lib gmt apr
Topic #47: covington georgia advance malcolm **sale** photography rgb monitor irq michael
Topic #48: catholic iran pat mary austria islamic elizabeth armenia chen stevens
Topic #49: baseball europeans lebanon water bosnians homeland occupation peace borders israelis
Topic #50: cramer clayton modem ticket abortion gay modems homosexuality tickets intergraph

Table 21: Top 10 related words of 50 topics from 20NG. Some repeated words are **bold** and underlined. The topic diversity value of 0.852 in the HiCOT model, though lower than FASTopic 0.980, remains high enough to maintain a diverse range of topics. While some topic-words overlap - such as "soldiers" appearing in both Topic 17 and Topic 31 - this does not result in topic collapse. Instead, the two topics retain distinct focuses: Turkey - Armenia war and Israel - Palestine war.

HiCOT with IMDB (K = 50)

Topic #1 : scary horror gore vampire terrible boring dracula bad **vampires** crappy
Topic #2 : novel jane timothy adaptation emma shakespeare novels **book** charlotte versions
Topic #3 : preview nicole hanks hbo premiere comedians bette stale screens edgy
Topic #4 : bergman diamond thrill realised masterful deranged preposterous hokey wrenching tasteless
Topic #5 : worst awful waste avoid terrible poorly wasted worse laughable amateurish
Topic #6 : pun unlikeable **book** races sue hack replace disney females childish
Topic #7 : season seasons episodes episode show abc series sitcom hooked network
Topic #8 : mute unlikable unreal formulaic sarah max coupled jean melodramatic matrix
Topic #9 : channels differently existent quotes dan sports dialogues respectively rank interaction
Topic #10: bollywood arnold batman cliche eastwood manipulative puppets tends werewolf showcase
Topic #11: drew nancy freeman romantic morgan jim comedic chemistry bruce definitely
Topic #12: stinks stunned stranded ridiculously crashes lately french heat france crash
Topic #13: ingredients ish caliber dance tap losers dancing phil priceless valley
Topic #14: amanda williams encourage filler sticking blast ryan bus tad complaints
Topic #15: godfather spoiled correctly desired ward neo fooled guessing viewings wizard
Topic #16: match stan jerry eddie baseball hardy bugs wrestling tag football
Topic #17: **superman** planet robot trek alien batman fiction space technology science
Topic #18: powell arthur hudson oscar dorothy chaplin **broadway** **fonda** henry moore
Topic #19: gems australian sean alex inspire aim monty savage derek scottish
Topic #20: society sexuality images desire cultural china nature catholic louis religious
Topic #21: waste worst crap terrible awful wasting please wasted worse avoid
Topic #22: amused murphy suspend jokes dave thumbs mail shouting freaks smiling
Topic #23: horror eerie haunted gothic carpenter creepy pitt murders caine curtis
Topic #24: twins widescreen lift remake kong models portrayals ford grand hitchcock
Topic #25: angela martha claude accomplish classical davis brazil titanic credited exposure
Topic #26: eddie stan hardy lucy taxi heist pearl lifeless monkey garden
Topic #27: sinatra **broadway** musicals musical powell sing kelly mgm dancers dance
Topic #28: chan jackie martial arts **dragon** kung ninja kong hong chinese
Topic #29: spielberg burton profanity entitled roberts talked racist ross barry biased
Topic #30: jesus christ bible documentary religious christian religion church catholic beliefs
Topic #31: pointless development uninteresting dimensional boring fails disjointed sex depth gay
Topic #32: marie ann mom charlie married marry mothers daughters finds dad
Topic #33: germans propaganda hitler nazis jews war soviet nazi documentary germany
Topic #34: stupid sucks crap gonna dumb guy laughing lame kid wanna
Topic #35: dinosaur dinosaurs shark alien **superman** seagal scientist bullets rubber **dragon**
Topic #36: andrews noir sheriff stewart eastwood police trail harry fbi clint
Topic #37: apocalypse khan april sometime vietnam nerd ken camp tacky war
Topic #38: santa kills lisa sheriff killed bat killing mom killer nurse
Topic #39: streep beautifully meryl deeply brilliantly **fonda** captures emotionally emotional passion
Topic #40: zombie zombies **vampires** gore slasher vampire gory werewolf nudity metal
Topic #41: stanley billed les predator welles murray quinn glenn ned sums
Topic #42: sheen hats preferred greatness alert turkey bush dropping lasting shoulders
Topic #43: fifth net subplots matthew aimed cheating robin remained refers exquisite
Topic #44: walken junior preachy karate bang foul severely buff favour acid
Topic #45: lion disney santa animals adults christmas bugs bears copy king
Topic #46: freddy hoot annie lip nails eva hilarity chuckle hitler rude
Topic #47: glover indian stuart continually abused jonathan baker roller fields nyc
Topic #48: eighties sixties seventies stumbled gary taylor wig irish korean comparing
Topic #49: documentary art cinema silent images artists documentaries color contemporary visual
Topic #50: parent schools juvenile teacher teaching posters sons overacting beverly adult

Table 22: Top 10 related words of 50 topics from IMDB. Some repeated words are **bold** and underlined. The topic diversity value of 0.837 in the HiCOT model, although lower than ECRTM 0.974, is still sufficiently high to preserve a broad range of topics. While some topic-words overlap - such as "book" appearing in both Topic 2 and Topic 6 - this does not result in topic collapse. Instead, the two topics retain distinct focuses: "classic literature and novel adaptations" and "criticism and controversy in books".