Topic Modeling for Short Texts via Optimal Transport-Based Clustering

Tu Vu[1]*, Manh Do[2]*, Tung Nguyen[2]*
Linh Ngo Van[2]†, Sang Dinh[2], Thien Huu Nguyen[3]

[1]Bytedance Inc
[2]Hanoi University of Science and Technology (HUST), Vietnam
[3]University of Oregon, USA

Abstract

Discovering topics and learning document representations in topic space are two crucial aspects of topic modeling, particularly in the short-text setting, where inferring topic proportions for individual documents is highly challenging. Despite significant progress in neural topic modeling, effectively distinguishing document representations as well as topic embeddings remains an open problem. In this paper, we propose a novel method called **En**hancing Global **C**lustering with **O**ptimal **T**ransport in Topic Modeling (**EnCOT**). Our approach utilizes an abstract global clusters concept to capture global information and then employs the Optimal Transport framework to align document representations in the topic space with global clusters, while also aligning global clusters with topics. This dual alignment not only enhances the separation of documents in the topic space but also facilitates learning of latent topics. Through extensive experiments, we demonstrate that our method outperforms state-of-the-art techniques in short-text topic modeling across commonly used metrics.[1]

1 Introduction

Topic models aim to uncover a set of latent topics from a document collection by analyzing word co-occurrence patterns. Each topic represents a coherent semantic concept and is characterized by related words. Additionally, topic models estimate the topic distribution within each document (topic proportions), shedding light on their underlying meanings. Traditional approaches to topic modeling rely on either probabilistic graphical models (Hofmann, 1999; Blei et al., 2003; Blei and Lafferty, 2006a) or non-negative matrix factorization techniques (Kim et al., 2015; Shi et al., 2018). Recently, Neural Topic Models (NTMs) have emerged as a powerful alternative, leveraging advances in deep learning. Unlike conventional methods, NTMs utilize deep neural networks to model distributions, allowing efficient and flexible parameter inference through automatic gradient back-propagation, as exemplified in (Kingma and Welling, 2013a; Srivastava and Sutton, 2017). This adaptability enables researchers to customize model architectures to suit a wide range of application scenarios.

Conventional topic models often face significant challenges when applied to short texts. The primary issue lies in their word co-occurrence information to infer latent topics. In short texts, this information is highly sparse due to the limited context available, making it difficult for topic models to extract meaningful patterns (Duc et al., 2017; Tuan et al., 2020; Bach et al., 2023; Nguyen et al., 2022a,b). This issue, commonly referred to as data sparsity (Yan et al., 2013), hampers the models' ability to generate high-quality topics and has therefore become a focal point of interest within the research community. Numerous studies have been proposed to tackle the issue of data sparsity in short texts. (Wu et al., 2020) develop NQTM, which applies vector quantization to doc-topic distributions based on the concept from (Van Den Oord et al., 2017). (Wang et al., 2021) propose leveraging word co-occurrence and semantic correlation graphs to enhance the learning signals for short texts. (Zhao et al., 2021b) integrate entity vector representations into a neural topic model (NTM) for short texts, learning these vectors from manually curated knowledge graphs. Building upon NQTM, (Wu et al., 2022) introduce TSCTM, a contrastive learning method designed to capture topic semantics more effectively by modeling similarity relationships among short texts.

A significant limitation of previous studies is their inability to disentangle doc-topic representations, leading to poor performance in downstream

*These authors contributed equally to this work.
†Corresponding author: linhnv@soict.hust.edu.vn
[1]Our code is publicly available at: `https://github.com/manhdo249/EnCOT`.

Findings of the Association for Computational Linguistics: ACL 2025, pages 7666–7680
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

tasks such as document clustering, document classification. Without clear separation in these representations, it becomes challenging to identify meaningful groupings of documents. This drawback underscores the need for methods that enhance the interpretability and structure of doc-topic distributions to better support tasks requiring high-quality clustering and classification. To combat the challenges of data sparsity and improve topic coherence, (Nguyen et al., 2025a) introduce GloCOM, a Neural Topic Model that integrates Global Clustering Context. The model begins with clustering documents and then creating global contexts by merging short texts within each cluster, thereby enhancing word co-occurrence statistics. Additionally, GloCOM leverages pre-trained language model (PLM) embeddings for clustering, which significantly improves the semantic of document clusters and the quality of inferred topics. As demonstrated with **NMI** and **Purity** score, this PLM-enhanced global clustering contributes to superior topic coherence and topic diversity, highlighting the model's effectiveness in semantic richness and clustering precision.

Despite its innovations, GloCOM and existing neuron topic modeling approaches face two significant challenges. First, none of these methods directly disentangle document representations by explicitly designing loss functions to achieve this. While GloCOM is the first model to leverage clustering for document separation, its improved performance primarily relies on clustering and data aggregation techniques rather than a targeted disentangling strategy. Second, like document separation, this model lacks explicit regularization on topic representations. In GloCOM, the loss function does not directly account for enhancing topic quality. This highlights the need for a method that actively separates topics, ensuring distinct and coherent topic representations.

Building on these two findings, we propose an approach to tackle the identified challenges. Clustering serves as an intuitive and effective method for data point separation by naturally grouping similar data points within a cluster while segregating dissimilar ones into distinct clusters. This approach aligns closely with the concept of proximity in feature space, where similar entities are grouped together due to shared attributes. With respect to document representation in topic space, clustering ensures that closely related documents within the same topic are grouped, while unrelated documents are assigned to separate clusters. Similarly, topics can be hierarchically organized, with broader themes encompassing more specific subtopics. For instance, "Physics" and "Biology" might collectively form the higher-level topic "Science," while "Politics," "Sports," and "Technology" can be grouped under "News." This hierarchical arrangement offers a more natural and accurate representation of the data, capturing both broad and nuanced relationships within the dataset. For choosing a clustering method, we rely on Optimal Transport (OT). OT treats documents and topics as distributions and computes the minimal "transport cost" required to map them to cluster centers (centroids) - a novel abstract global clusters concept. This approach allows OT to dynamically adapt to the semantic relationships within the data, ensuring that documents with similar content are grouped together towards a same centroid and topics with overlapping themes are aligned to a share centroid. By aligning documents to centroids, OT enhances document representation, leading to better separation and coherence. Likewise, aligning topics with centroids ensures that topics are enriched with higher-level thematic structures, allowing for a hierarchical understanding of the corpus. This dual application of OT not only enhances the separation of documents but also improves the coherence and diversity of topics.

The contributions of this paper are summarized as follows:

- In order to enhance document separation, we propose using OT as a clustering framework that simultaneously learns cluster centers (centroids) and document representations.
- We present a general methodology that applicable to various neural topic models. This approach encourages similar topics to cluster together while pushing dissimilar topics apart, resulting in more coherent and diverse topics.
- Our experimental results show that **EnCOT** significantly enhances state-of-the-art neural topic modeling method and boosts existing baselines by a substantial margin, particularly in short-text scenario.

2 Preliminaries

In this section, we provide an overview of topic modeling, including the problem settings and notations. We then introduce the background of a state-of-the-art method - Global Clustering COntexts for Topic Models (GloCOM) (Nguyen et al., 2025a), which serves as the foundation for illustrating how our approach can be applied to various models.

2.1 Notations

In this paper, we denote D is the number documents in the corpus, V is the size of vocabulary, K is number of hidden topics, L is the word and topic embedding dimension and G is number of clusters. $\mathbf{X} = \{x^d\}_{d=1}^{D}$ is a collection of D documents, where x^d represents the Bag-of-Words (BoW) vector for document d. x^d_{PLM} is the embedding of a document via a pre-trained language model. The clustering algorithm applied to x^d_{PLM} produces G clusters. $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G \in \mathbb{R}^L$ are the cluster centroids. We denote $\mathbf{x}^d_{emb} \in \mathbb{R}^L$ is another document embedding, which is used to align with cluster centroids. We denote $\mathcal{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_V) \in \mathbb{R}^{V \times L}$ as the word embedding matrix and $\mathcal{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_K) \in \mathbb{R}^{K \times L}$ as the topic embedding matrix. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{V \times K}$ is the topic-word distributions matrix of all K topics, where $\boldsymbol{\beta}_k \in \mathbb{R}^{V \times 1}$ is the topic-word distribution of topic k. For each document $\theta_d \in \Delta^K$ is its doc-topic proportion where $\Delta^K = \{\mathbf{x} \in \mathbb{R}^K \mid x_k \geq 0, \sum_{k=1}^{K} x_k = 1\}$ is the simplex. We define $\mathcal{LN}(.)$ and $\mathcal{N}(.)$ is the logistic-normal distribution and normal distribution, respectively. $\mathbb{I}$ is the indicator function.

2.2 GloCOM

GloCOM utilizes pre-trained language model embeddings (Reimers, 2019; BehnamGhader et al., 2024) to capture the semantic of documents and present them for clustering. After that, it concatenates short documents (local documents) within the same cluster to form a global document x^g, with g is a cluster containing document x^d. The aggregated document is $\tilde{x}^d = x^d + \eta x^g$, where η is the augmentation coefficient. The reconstruction loss is computed based on aggregated documents.

The formal process to generate documents in GloCOM is as follows (the graphical model is depicted in Figure 1):

1. Calculate $\boldsymbol{\beta}$ as:

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

2. For each cluster g, generate $\theta^g \sim \mathcal{LN}(\mathbf{0}, \mathbf{I})$.

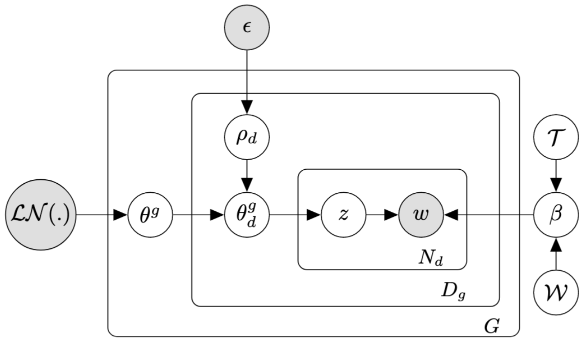

Figure 1: The probabilistic graphical model of GloCOM. (Nguyen et al., 2025a)

3. For each document d in cluster g:
   (a) Draw an adaptive variable: $\rho_d \sim \mathcal{N}(1, \epsilon I)$, where ϵ is a hyper-parameter.
   (b) Generate topic distribution:
   $$\theta^g_d = \text{softmax}(\theta^g \odot \rho_d) \tag{2}$$
   (c) For each word in document d:
      i. Draw a topic index: $z_{dn} \sim \text{Multinomial}(\theta^g_d)$
      ii. Draw the word: $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$

The Neural Topic Model loss of GloCOM is follow:

$$\mathcal{L}_{\text{TM}} = \sum_{d}^{D} \sum_{g}^{G} \mathbb{I}\left[x_d \in g\right] \mathcal{L}^d(\phi, \gamma, w, t) \tag{3}$$

where the lower bound for document d is:

$$\begin{aligned}\mathcal{L}^d(\phi, \gamma, w, t) = &-(\tilde{x}^d)^T \log(\text{softmax}(\beta\theta^g_d)) \\ &- D_{KL}\big(q_\phi(\theta^g|x^g)||p(\theta^g)\big) \\ &- D_{KL}\big(q_\gamma(\rho_d|x^d)||p(\rho_d|\epsilon)\big).\end{aligned} \tag{4}$$

Besides $\mathcal{L}_{\text{TM}}$ loss, GloCOM employs Embedding Clustering Regulazation (ECR) (Wu et al., 2023a). The ECR loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{ECR}} = \sum_{j=1}^{V} \sum_{k=1}^{K} ||\mathbf{w}_j - \mathbf{t}_k||^2 \pi^*_{\epsilon,jk} \\ \text{where } \boldsymbol{\pi}^*_\epsilon = \underset{\phi \in \mathbb{R}_+^{V \times K}}{\text{argmin}}\ \mathcal{L}_{OT_\epsilon}(\boldsymbol{W}, \boldsymbol{T}).\end{aligned} \tag{5}$$

The GloCOM final loss function is:

$$\mathcal{L}_{\text{GloCOM}} = \mathcal{L}_{\text{TM}} + \lambda_{ECR} * \mathcal{L}_{\text{ECR}} \tag{6}$$

where λ_{ECR} is a hyper-parameter.

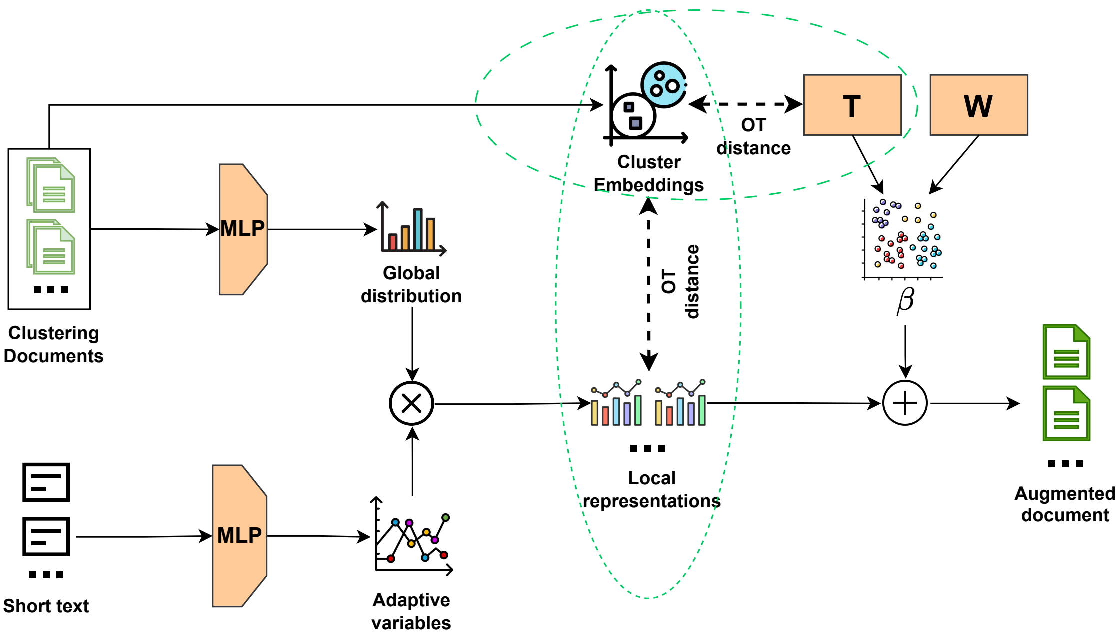

Figure 2: The demonstration of EnCOT integrated in GloCOM architecture. To enhance representations of documents in topic space as well as topic embeddings, EnCOT introduces OT losses to align documents and topic with clusters.

The GloCOM loss comprises a standard topic modeling loss along with a regularization term, ECR, which has been shown to effectively prevent topic collapsing problem (Wu et al., 2023a).

3 Methodology

Figure 2 illustrates the architecture of the GloCOM integrated with our novel EnCOT. In the GloCOM framework, short texts are first clustered using embeddings derived from a pre-trained language model (PLM) and global documents are created by concatenating texts within each cluster. The short texts are processed through a Multi-Layer Perceptron (MLP) to generate an adaptive variable. In parallel, the global documents are transformed through another MLP to produce a global distribution. The document representation is obtained as the dot product of the adaptive variable and the global distribution, integrating information from the global clusters. Word embeddings ($\mathcal{W}$) and topic embeddings ($\mathcal{T}$) are employed to construct topics β following the methodology described in (Wu et al., 2023a). The augmented documents are central to compute the reconstruction loss, serving as essential components within the GloCOM framework (Nguyen et al., 2025a). The newly introduced elements— EnCOT with two OT losses—are highlighted in green ellipses. As depicted in the figure, the OT losses are applied directly to the representations of topic embeddings and document-topic vectors, ensuring that these representations remain distinct, as discussed in the previous analysis.

3.1 Optimal Transport as Clustering

Optimal Transport (OT) is a mathematical approach focused on converting one mass distribution into another while minimizing an associated cost. This concept can be imagined as relocating quantities of mass from one arrangement to another, where each arrangement represents a probability distribution. The quantity of mass corresponds to the weight of the distribution, and its location defines the position in the space. The objective is to determine the most efficient way to transfer mass between configurations at minimal cost, which is generally calculated based on a distance metric, such as the Euclidean distance. The transport plan (**Tr**) outlines the mapping of mass from the source distribution to the target distribution, specifying the amount to be transferred between corresponding points. In clustering, OT naturally quantifies similarity among data points by reducing the transport cost between clusters. The transport plan organizes data points by moving them between distributions in a manner that groups similar points together. The cost function aligns data points based on their similarities, ensuring the clustering respects both geometric relationships and distributional characteristics. Further details on OT notation, loss functions, and algorithms are provided in Appendix B.

Now, consider the representation of documents $\theta_d^g \in \Delta^K$ in topic space. We treat a collection of D documents is an uniform distribution while the mass for each document is $1/D$. The probability measure of documents is:

$$f_D = \sum_{d=1}^{D} \frac{1}{D} * \delta_{\mathbf{x}_{emb}^d} \in P(\Omega_D) \tag{7}$$

where document $\mathbf{x}_{emb}^d$ is viewed a point in Ω_D.

Similarly, the clusters is an uniform distribution hence the mass for each cluster is $1/G$. There are G clusters with the centroids $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g\} \in \mathbb{R}^{L\times 1}$. The probability measure of centroids is:

$$f_C = \sum_{g=1}^{G} \frac{1}{G} * \delta_{\boldsymbol{\mu}_g} \in P(\Omega_G) \tag{8}$$

where centroid $\boldsymbol{\mu}_g$ is considered as a point in Ω_G.

To align documents with clusters, we define the cost to move the mass from a document d to a centroid g as Euclidean distance:

$$\text{Cost}(d, g) = ||\boldsymbol{x}_{emb}^d - \boldsymbol{\mu}_g||^2 \tag{9}$$

where $\boldsymbol{x}_{emb}^d$ is computed as:

$$\boldsymbol{x}_{emb}^d = \theta_d^g * \mathcal{T} \tag{10}$$

The OT loss between documents and clusters is defined as:

$$\begin{aligned} \mathcal{L}_{\text{OT}}^{DG} &= \sum_{d=1}^{D}\sum_{g=1}^{G} ||\mathbf{x}_{emb}^d - \boldsymbol{\mu}_g||^2 \pi_{\epsilon,jk}^* \\ \text{where } \boldsymbol{\pi}_\epsilon^* &= \underset{\phi\in\mathbb{R}_+^{V\times K}}{\text{argmin}}\ \mathcal{L}_{OT_\epsilon}(\boldsymbol{X}_{emb}, \boldsymbol{\mu}). \end{aligned} \tag{11}$$

We impose a OT loss on the representation of documents in order to group them into different clusters. By this way, the documents with the similar representations are pulled into same groups while dissimilar documents are push far away.

With regard to the topics, we use the same technique to design the alignment between topics and clusters. We consider K topics as an uniform distribution, hence the mass for each topic is $1/K$. The probability measure of topics is:

$$f_T = \sum_{j=1}^{K} \frac{1}{K} * \delta_{\mathbf{t}_j} \in P(\Omega_T) \tag{12}$$

where topic $\mathbf{t}_j$ is considered as a point in Ω_T. We define the cost to transform as between a topic $\mathbf{t}_j$ to a centroid $\boldsymbol{\mu}_g$ is Euclidean distance:

$$\text{Cost}(t, g) = ||\boldsymbol{t}_j - \boldsymbol{\mu}_g||^2 \tag{13}$$

where $\boldsymbol{t}_j$ is the topic embedding representation of topic t. The OT loss between topics and clusters is defined as:

$$\begin{aligned} \mathcal{L}_{\text{OT}}^{TG} &= \sum_{j=1}^{T}\sum_{g=1}^{G} ||\boldsymbol{t}_j - \boldsymbol{\mu}_g||^2 \pi_{\epsilon,jk}^* \\ \text{where } \boldsymbol{\pi}_\epsilon^* &= \underset{\phi\in\mathbb{R}_+^{V\times K}}{\text{argmin}}\ \mathcal{L}_{OT_\epsilon}(\mathcal{T}, \boldsymbol{\mu}). \end{aligned} \tag{14}$$

The topic-cluster OT loss encourages similar topics to converge while promoting the separation of dissimilar ones.

The final EnCOT loss is the sum of document-cluster OT loss and topic-cluster OT loss, which is defined as:

$$\mathcal{L}_{\text{EnCOT}} = \lambda_{OT}^{DG} * \mathcal{L}_{\text{OT}}^{DG} + \lambda_{OT}^{TG} * \mathcal{L}_{\text{OT}}^{TG} \tag{15}$$

where λ_{OT}^{DG} and λ_{OT}^{TG} are hyper-parameters of document-cluster alignment and topic-cluster alignment.

3.2 Overall Objective Function

To this end, the overall loss function of GloCOM equipped with EnCOT is:

$$\begin{aligned} &\mathcal{L}_{\text{GloCOM}-\text{EnCOT}} = \mathcal{L}_{\text{GloCOM}} + \mathcal{L}_{\text{EnCOT}} \\ &= \mathcal{L}_{\text{GloCOM}} + \lambda_{OT}^{DG} * \mathcal{L}_{\text{OT}}^{DG} + \lambda_{OT}^{TG} * \mathcal{L}_{\text{OT}}^{TG} \end{aligned} \tag{16}$$

The overall loss function comprises two components: $\mathcal{L}_{\text{GloCOM}}$ and $\mathcal{L}_{\text{EnCOT}}$. We retain the original $\mathcal{L}_{\text{GloCOM}}$ as described in the subsection 2.2. The $\mathcal{L}_{\text{EnCOT}}$ loss consists of two terms, $\mathcal{L}_{\text{OT}}^{DG}$ and $\mathcal{L}_{\text{OT}}^{TG}$, which are our innovative losses imposing directly on desired representations. This enhancement leads to improved performance in both document-topic and topic-word distributions within GloCOM.

3.3 Training Procedure

In this section, we outline GloCOM training steps with EnCOT loss:

Model	GoogleNews				SearchSnippets				StackOverflow				Biomedical			
$K = 50$	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
ProdLDA	0.437	0.991	0.201	0.384	0.406	0.546	0.731	0.435	0.388	0.588	0.117	0.151	0.469	0.520	0.136	0.177
ETM	0.402	0.916	0.366	0.560	0.397	0.594	0.688	0.389	0.367	0.766	0.418	0.280	0.450	0.723	0.406	0.273
ECRTM	0.441	0.987	0.396	0.615	0.450	0.998	0.711	0.419	0.381	0.941	0.197	0.192	0.468	0.987	0.414	0.315
FASTopic	0.446	0.440	0.351	0.659	0.395	0.710	0.792	0.481	0.317	0.222	0.408	0.486	0.418	0.482	0.456	0.369
NQTM	0.408	0.959	0.536	0.716	0.436	0.922	0.435	0.150	0.382	0.933	0.392	0.238	0.471	0.915	0.191	0.109
TSCTM	0.437	0.988	0.552	0.761	0.424	0.993	0.724	0.386	0.378	0.911	0.572	0.418	0.484	0.972	0.480	0.341
KNNTM	0.435	0.986	0.579	0.795	0.425	0.995	0.768	0.429	0.380	0.922	0.636	0.490	0.490	0.972	0.526	0.380
GloCOM	**0.475**	0.999	0.586	0.817	0.453	0.956	0.806	0.502	0.390	0.962	0.653	0.588	0.490	0.998	0.546	0.437
GloCOM-EnCOT	0.45	**1.0**	**0.613**	**0.848**	0.454	**1.0**	**0.839**	**0.53**	**0.391**	**1.0**	**0.675**	**0.622**	**0.491**	**1.0**	**0.557**	**0.447**
Model	**GoogleNews**				**SearchSnippets**				**StackOverflow**				**Biomedical**			
$K = 100$	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
ProdLDA	0.435	0.611	0.611	0.600	0.424	0.679	0.766	0.415	**0.382**	0.466	0.098	0.090	0.463	0.465	0.079	0.050
ETM	0.398	0.677	0.554	0.713	0.389	0.448	0.692	0.365	0.369	0.444	0.475	0.331	0.452	0.476	0.404	0.268
ECRTM	0.418	0.991	0.342	0.491	0.432	0.966	0.789	0.443	0.375	0.993	0.172	0.179	0.444	0.974	0.124	0.113
FASTopic	0.438	0.369	0.458	0.722	0.386	0.634	0.807	0.458	0.309	0.186	0.495	0.514	0.440	0.457	0.495	0.375
NQTM	0.397	0.898	0.706	0.788	0.438	0.638	0.334	0.077	0.379	0.818	0.417	0.255	0.460	0.572	0.142	0.056
TSCTM	0.448	0.941	0.754	0.835	0.430	0.894	0.757	0.384	0.380	0.620	0.563	0.386	**0.485**	0.806	0.487	0.330
KNNTM	0.441	0.959	0.797	0.870	0.421	0.948	0.800	0.421	0.381	0.663	0.611	0.436	0.483	0.848	0.530	0.362
GloCOM	**0.450**	0.944	0.761	0.900	**0.443**	0.920	0.822	0.501	**0.382**	0.804	0.658	0.585	0.462	0.997	0.536	0.422
GloCOM-EnCOT	0.42	**1.0**	**0.801**	**0.911**	0.42	**1.0**	**0.839**	**0.516**	0.377	**1.0**	**0.679**	**0.609**	0.473	**1.0**	**0.557**	**0.445**

Table 1: Topic quality (C_V, TD), and doc-topic quality (Purity, NMI) with $K = 50$ and $K = 100$. The **bold** values indicate the best performance. The underline values indicate the second best performance. **GloCOM-EnCOT** is **GloCOM** trained with our **EnCOT**.

Algorithm 1 GloCOM framework with EnCOT loss.

Input: Input corpus $\mathbf{X}$, Topic number K, epoch number N, and clusters G.

Output: K topic-word distributions $\boldsymbol{\beta}_k$, N doc-topic distributions θ_d^g, cluster centroids $\boldsymbol{\mu}_g$.

```
for epoch from 1 to N do
    For a random batch of B documents do
    L_batch ← 0;
    for each local doc x^d and its respective global doc x^g in the batch do
        Compute the adaptive variable p_d;
        Compute the global topic distribution θ^g;
        Compute the local topic distribution θ_d^g by Eq. 2;
        Compute document embedding x_emb by Eq. 10
        L_batch ← L_batch + L_GloCOM-EnCOT by Eq. 16;
    end for
    Update model parameters with ∇L_batch;
end for
```

Overall, our training procedure follows the same structure as GloCOM. With the introduction of two new OT losses, the cluster centroids $\boldsymbol{\mu}_g$ are dynamically updated in each batch. These centroids act as intermediaries to improve the representation of documents and topics. Consequently, during the inference phase, they are excluded from the prediction of the topic distribution θ_d^g for held-out documents.

We have demonstrated how to apply EnCOT to a state-of-the-art model, GloCOM. The training procedure of the original GloCOM algorithm requires minimal modifications, as we only need to add the OT losses directly to the typical neural topic model loss. Hence, EnCOT can be easily adapted to any NTM to enhance its performance.

4 Experiments

4.1 Settings

Datasets. We conduct experiments with well-known datasets, including four datasets: **GoogleNews**, **SearchSnippets**, **StackOverflow** and **Biomedical**. The datasets are derived from GloCOM (Nguyen et al., 2025a). The datasets contain short documents from different sources. The dataset statistics and pre-process details are in Appendix C.

Model	GoogleNews				SearchSnippets				StackOverflow				Biomedical			
$K = 50$	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
ETM	0.402	0.916	0.366	0.56	0.397	0.594	0.688	0.389	0.367	0.766	0.418	0.28	0.45	0.723	0.406	0.273
ETM-EnCOT	**0.406**	**0.919**	**0.382**	**0.593**	**0.41**	**0.66**	**0.753**	**0.441**	**0.369**	**0.769**	**0.566**	**0.415**	**0.454**	**0.779**	**0.432**	**0.302**
ECRTM	0.441	0.987	0.396	0.615	0.450	0.998	0.711	0.419	0.381	0.941	0.197	0.192	0.468	0.987	0.414	0.315
ECRTM-EnCOT	**0.453**	**1.0**	**0.5**	**0.719**	**0.456**	**1.0**	**0.764**	**0.451**	**0.386**	**1.0**	**0.228**	**0.213**	**0.47**	**1.0**	**0.442**	**0.358**
GloCOM	0.475	0.999	0.586	0.817	0.453	0.956	0.806	0.502	0.39	0.962	0.653	0.588	0.49	0.998	0.546	0.437
GloCOM-EnCOT	0.45	**1.0**	**0.613**	**0.848**	**0.454**	**1.0**	**0.839**	**0.53**	**0.391**	**1.0**	**0.675**	**0.622**	**0.491**	**1.0**	**0.557**	0.447
Model	**GoogleNews**				**SearchSnippets**				**StackOverflow**				**Biomedical**			
$K = 100$	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
ETM	0.398	0.677	0.554	0.713	0.389	0.448	0.692	0.365	0.369	0.444	0.475	0.331	0.452	0.476	0.404	0.268
ETM-EnCOT	**0.402**	**0.78**	**0.677**	**0.8**	**0.401**	**0.565**	**0.712**	**0.375**	**0.372**	**0.447**	**0.572**	**0.429**	**0.464**	**0.554**	**0.412**	**0.274**
ECRTM	0.418	0.991	0.342	0.491	0.432	0.966	0.789	0.443	0.375	0.993	0.172	0.179	0.444	0.974	0.124	0.113
ECRTM-EnCOT	**0.419**	**0.995**	**0.425**	**0.568**	**0.438**	**0.996**	**0.8**	**0.452**	**0.377**	**1.0**	**0.201**	**0.192**	**0.446**	**0.977**	**0.194**	**0.205**
GloCOM	0.45	0.944	0.761	0.9	0.443	0.92	0.822	0.501	0.382	0.804	0.658	0.585	0.462	0.997	0.536	0.422
GloCOM-EnCOT	0.42	**1.0**	**0.801**	**0.911**	0.42	**1.0**	**0.839**	**0.516**	0.377	**1.0**	**0.679**	**0.609**	**0.473**	**1.0**	**0.557**	**0.445**

Table 2: EnCOT enhancements topic quality and doc-topic quality with $K = 50$ and $K = 100$, $G = 30$.

Model	GoogleNews				SearchSnippets				StackOverflow				Biomedical			
	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
$K = 50$																
EnCOT	**0.45**	1.0	**0.613**	0.848	0.454	1.0	0.839	**0.53**	0.391	1.0	0.675	**0.622**	0.491	1.0	0.557	**0.447**
EnCOT$_{\mathrm{w/oTG}}$	0.448	1.0	0.612	0.849	0.454	1.0	0.839	0.518	0.391	1.0	0.676	0.601	0.493	1.0	0.558	0.446
$K = 100$																
EnCOT	**0.42**	1.0	**0.801**	**0.911**	**0.42**	1.0	0.839	**0.516**	**0.377**	1.0	**0.679**	**0.609**	**0.473**	**1.0**	**0.557**	**0.445**
EnCOT$_{\mathrm{w/oTG}}$	0.419	1.0	0.79	0.907	0.419	1.0	0.839	0.513	0.375	1.0	0.677	0.602	0.471	0.988	0.542	0.438
$K = 150$																
EnCOT	**0.409**	1.0	**0.821**	**0.901**	**0.42**	1.0	**0.839**	**0.512**	**0.371**	**0.97**	**0.675**	**0.601**	**0.453**	1.0	**0.528**	**0.442**
EnCOT$_{\mathrm{w/oTG}}$	0.404	1.0	0.802	0.893	0.406	1.0	0.824	0.499	0.367	0.968	0.669	0.597	0.451	1.0	0.523	0.429
$K = 200$																
EnCOT	**0.407**	1.0	**0.827**	**0.897**	**0.405**	1.0	**0.839**	**0.513**	**0.373**	**0.632**	**0.656**	**0.612**	0.443	1.0	**0.533**	**0.437**
EnCOT$_{\mathrm{w/oTG}}$	0.406	1.0	0.817	0.886	0.402	1.0	0.823	0.497	0.363	0.624	0.632	0.602	0.443	1.0	0.519	0.426

Table 3: **GloCOM** ablation study without $\mathcal{L}_{OT}^{TG}$ in different K.

Evaluation Metrics. We adopt the evaluation methodology outlined in (Wu et al., 2023a) to measure both topic quality and document-topic distributions. Topic quality is assessed through topic coherence (**TC**) and topic diversity (**TD**). For topic coherence, we utilize $C_V 15$, where 15 represents the top words in each topic. These metrics are well-established in topic modeling and show strong alignment with human judgment (Röder et al., 2015). The coherence calculations are based on a version of the Wikipedia corpus[2] as an external reference. To evaluate topic diversity, we calculate the ratio of unique words among the topic words, referred to as TD. For document-topic distribution quality, we use Normalized Mutual Information (NMI) and Purity (Manning et al., 2008) in the document clustering task for the test data, following the approach in (Zhao et al., 2021a; Wang et al., 2022a). To summary, we use C_V, TD, Purity, and NMI as our main metrics. In addition, we evaluate our model using the more recent LLMScore metric (Stammbach et al., 2023), which leverages ChatGPT to assess topic quality. The LLMScore results are provided in Appendix F.

Baseline models. We evaluate our novel model attaching with recent advanced topic modeling frameworks ETM, ECRTM and GloCOM. Besides, we compare our results with other state-of-the-art models, including the conventional neural topics models and short-text topic models. For conventional neural topic models, we consider ProdLDA

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

(Srivastava and Sutton, 2017), a pioneering VAE-based topic model; ETM (Dieng et al., 2020) incorporates word embeddings; ECRTM (Wu et al., 2023a), based on ETM with regularization between word and topic embeddings; FASTopic (Wu et al., 2024b), a state-of-the-art model for identifying topics via word, topic, and document embeddings. For short-text topic models, we include NQTM (Wu et al., 2020), a neural topic model dedicated to short text problems with vector quantization for topic distribu- tions; TSCTM (Wu et al., 2022), a NQTM improvement with an additional contrastive loss on topic distributions; kNNTM (Lin et al., 2024), a recent state-of-the-art short text neural topic model that augments a document with its neighbors via the kNN algorithm. Except for kNNTM [3], we use the implementation of the other models provided by TopMost (Wu et al., 2023b) and fine-tune these baselines on various datasets.

4.2 Topic and Doc-topic Distribution Quality

Table 1 illustrates the key metrics evaluating topic quality and document-topic distribution quality. For document-topic distribution quality, GloCOM-EnCOT achieves superior performance compared to all baselines, as measured by Purity and NMI. It significantly outperforms neural topic models (ProdLDA, ETM, ECRTM, FASTopic), short-text specific models (NQTM, TSCTM, KNNTM) and also its base model GloCOM, establishing state-of-the-art results by a large margin. Regarding topic quality, GloCOM-EnCOT also demonstrates superior performance in learning high-quality topics. Notably, it achieves a topic diversity score of TD at **1.0**, the maximum possible value, reflecting its ability to fully separate topics. These superior performance of GloCOM-EnCOT across all datasets confirm the effectiveness of our method in learning high-quality document-topic distributions and topics.

4.3 EnCOT with different methods

We evaluate the effectiveness of EnCOT across three baseline models: ETM (Dieng et al., 2020), ECRTM (Wu et al., 2023a), and GloCOM (Nguyen et al., 2025a). The EnCOT term is incorporated into the loss functions of these baseline methods to enhance their performance. The results are recorded in Table 2. As demonstrated, EnCOT significantly improves the baselines across different metrics and datasets. The topic quality, as reflected by TD values, consistently reaches 1.0 across most settings, highlighting the model's ability to generate diverse topics. Additionally, document representation quality is improved, as indicated by significant increases in Purity and NMI scores.

4.4 Ablation Study

In this section, we conduct experiments to evaluate the effectiveness of $\mathcal{L}_{OT}^{TG}$, the component of EnCOT. Since G centroids are generated after the document clustering step by imposing $\mathcal{L}_{OT}^{DG}$ loss, the topic-cluster loss $\mathcal{L}_{OT}^{TG}$ comes after when centroids are available. We evaluate GloCOM-EnCOT without topic-cluster loss with different K. The results are depicted in Table 3. It is naturally assumed that topics themselves can be organized into different clusters, where each cluster contains similar topics. Moreover, clustering topics becomes more effective when the number of topics is sufficiently large. When both the number of topics and clusters are small, topics tend to be inherently close together, making topic grouping less efficient. Our experimental results in Table 3, with different values of $K \in \{50, 100, 150, 200\}$, demonstrate that a high K (i.e., $K = 150$ or $K = 200$) makes the effect of grouping/clustering topics more apparent. Specifically, with $K = 200$, ablating leads to a significant reduction in compared to original EnCOT, indicating that $\mathcal{L}_{OT}^{TG}$ contributes to learning topics in a hierarchical manner. This also implies that topic-cluster and document-cluster collaborate to enhance the representations of both documents and topics, ultimately improving the performance of EnCOT.

We further analyze the sensitivity of EnCOT to varying hyperparameter values. Additional details are provided in Appendix D.

5 Conclusion

In this paper, we introduce a novel approach called Enhancing Global Clustering with Optimal Transport in Topic Modeling (EnCOT), designed to simultaneously address the challenges of document separation and topic separation. EnCOT leverages clustering concepts within the Optimal Transport framework to achieve this distinction effectively. Comprehensive experiments validate the effectiveness of EnCOT, demonstrating its ability to achieve robust neural topic modeling by ensuring clear separation between topics and between documents.

[3]We will publish the code for the kNNTM models alongside our codebase

Additionally, EnCOT consistently delivers state-of-the-art performance in generating high-quality topics and document-topic distributions.

Limitations

While our approach achieves outstanding performance in short text topic modeling, it does have certain limitations. First, the G centroids contain valuable information for aligning documents and topics, but this information is not utilized during inference, leaving its potential unexplored. Future studies could focus on leveraging these centroids to better analyze the corpus. Moreover, our method faces challenges when applied to other contexts, such as dynamic topic modeling, online learning, and streaming learning. Adapting our approach to effectively capture topic relationships within temporal data presents an important direction for future research.

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

Tran Xuan Bach, Nguyen Duc Anh, Ngo Van Linh, and Khoat Than. 2023. Dynamic transformation of prior knowledge into bayesian models for data streams. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3742–3750.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

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.

David Blei and John Lafferty. 2006a. Correlated topic models. *Advances in neural information processing systems*, 18:147.

David M. Blei and John D. Lafferty. 2006b. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, page 113–120.

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

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

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.

Cuturi M Sinkhorn Distances. 2013. Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300.

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: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II 21*, pages 247–258. Springer.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

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.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.

Ou Jin, Nathan N. Liu, Kai Zhao, Yong Yu, and Qiang Yang. 2011. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM International Conference* *on Information and Knowledge Management*, page 775–784.

Hannah Kim, Jaegul Choo, Jingu Kim, Chandan K Reddy, and Haesun Park. 2015. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 567–576.

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

Diederik P. Kingma and Max Welling. 2013b. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Hoa M Le, Son Ta Cong, Quyen Pham The, Ngo Van Linh, and Khoat Than. 2018. Collaborative topic model for poisson distributed ratings. *International Journal of Approximate Reasoning*, pages 62–76.

Chenliang Li, Yu Duan, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2017. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Trans. Inf. Syst.*, pages 1–30.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 165–174.

Ximing Li, Jihong Ouyang, You Lu, Xiaotang Zhou, and Tian Tian. 2015. Group topic model: organizing topics into groups. *Information Retrieval Journal*, pages 1–25.

Yang Lin, Xinyu Ma, Xin Gao, Ruiqing Li, Yasha Wang, and Xu Chu. 2024. Combating label sparsity in short text topic modeling via nearest neighbor augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13762–13774.

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*, pages 431–442. Springer.

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

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 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1109–1124, Albuquerque, New Mexico. Association for Computational Linguistics.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. In *Advances in Neural Information Processing Systems*, pages 11974–11986.

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*, pages 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.

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.

Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *24th International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 2270–2276. AAAI Press/International Joint Conferences on Artificial Intelligence.

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

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 world wide web conference*, pages 1105–1114.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

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*.

Jian Tang, Ming Zhang, and Qiaozhu Mei. 2013. One theme in all views: modeling consensus topics in multiple contexts. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 5–13.

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.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ngo Van Linh, Nguyen Kim Anh, Khoat Than, and Chien Nguyen Dang. 2017. An effective and interpretable method for document classification. *Knowledge and Information Systems*, pages 763–793.

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.

Cédric Villani et al. 2009. *Optimal transport: old and new*, volume 338. Springer.

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

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022b. Representing mixtures of word embeddings with mixtures of topic embeddings. *arXiv preprint arXiv:2203.01570*.

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

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, page 261–270.

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

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. *arXiv preprint arXiv:2211.12878*.

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 Thanh Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024b. Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xiaobao Wu, Fengjun Pan, and Anh Tuan Luu. 2023b. Towards the topmost: A topic modeling system toolkit. *arXiv preprint arXiv:2309.06908*.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31.

Yi Xu, Dongsheng Wang, Bo Chen, Ruiying Lu, Zhibin Duan, Mingyuan Zhou, et al. 2022. Hyperminer: Topic taxonomy mining with hyperbolic embedding. In *Advances in Neural Information Processing Systems*, pages 31557–31570.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 233–242.

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. IEEE.

Delvin Ce Zhang and Hady Lauw. 2022. Meta-complementing the semantics of short texts in neural topic models. In *Advances in Neural Information Processing Systems*, pages 29498–29511.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2020. Neural topic model via optimal transport. *arXiv preprint arXiv:2008.13537*.

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

Xiaowei Zhao, Deqing Wang, Zhengyang Zhao, Wei Liu, Chenwei Lu, and Fuzhen Zhuang. 2021b. A neural topic model with word vectors and entity vectors for short texts. *Information Processing & Management*, 58(2):102455.

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

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2105–2114.

A Related Work

Neural Topic Modeling. Topic models have been widely adopted across various fields, such as text mining (Van Linh et al., 2017), recommender systems (Le et al., 2018), and streaming learning (Nguyen et al., 2019, 2022a, 2025c). Traditional topic models such as LDA (Blei et al., 2003) and probabilistic LSI (Hofmann, 1999) are built on generative probabilistic frameworks. Although enhancements have been proposed (Blei and Lafferty, 2006b; Li et al., 2015; Nguyen et al., 2022b), these models remain less efficient and fall short in performance when compared to modern neural network-based methods, particularly those utilizing VAE architectures (Kingma and Welling, 2013b). Recent advancements in topic modeling include the integration of pre-trained language models (Han et al., 2023; Pham et al., 2024b; Nguyen et al., 2025d), the adoption of optimal transport metrics (Zhao et al., 2021a; Nguyen et al., 2025b), and the use of contrastive loss techniques (Nguyen and Luu, 2021). Other approaches refine the generative process by incorporating pre-trained embeddings (Dieng et al., 2020; Xu et al., 2022) or leveraging optimal transport distances (Wang et al., 2022a; Wu et al., 2023a, 2024b). However, these methods continue to face challenges with short-text data due to issues like data sparsity. Some other studies such as TopicGPT utilize large language models to generate more human-readable topic descriptions (Pham et al., 2024a). However, this approach differs significantly from traditional topic modeling frameworks, which primarily focus on inferring topic-word distributions. Due to these fundamental differences in topic representation, comparing the quality of generated topics between TopicGPT and other methods is impractical - especially since TopicGPT lacks a standardized method for topic quality evaluation.

Topic Modeling for Short Text . Conventional short text topic models (Li et al., 2016, 2017; Yin and Wang, 2014) typically assume that each text is associated with only a few topics, while Biterm Topic Models (Yan et al., 2013; Cheng et al., 2014; Mai et al., 2016; Tuan et al., 2020) leverage word co-occurrence patterns for topic inference. To address data sparsity, aggregation-based methods (Hong and Davison, 2010; Tang et al., 2013; Quan et al., 2015) have also been introduced. However, these approaches face challenges, including difficulties in inferring individual document topics (Weng et al., 2010) and significant computational demands (Zuo et al., 2016). Clustering methods, which rely on term frequency representations, have similarly proven inadequate for capturing the semantics of short texts (Jin et al., 2011). Recently, neural short text topic models have demonstrated superior performance and generalization over traditional methods (Wu et al., 2024a). Some approaches utilize pre-trained embeddings (Dieng et al., 2020; Bianchi et al., 2021; Van Linh et al., 2022; Nguyen et al., 2021) or word co-occurrence graphs (Zhu et al., 2018; Wang et al., 2021), while others target variable-length corpora (Zhang and Lauw, 2022). Techniques like topic distribution quantization (Wu et al., 2020, 2022) have been effective in mitigating data sparsity, with kNNTM (Lin et al., 2024) emerging as a leading method for addressing label sparsity in short texts. GloCOM (Nguyen et al., 2025a) represents a state-of-the-art method that leverages a data aggregation and clustering approach.

B Optimal Transport

Optimal Transport (OT) (Villani et al., 2009) is a mathematics framework to measure the dissimilarity between probability distributions. In comparison with others measures such as Kullback–Leibler divergence (KL) or Jensen-Shannon Divergence (JS) which require two distributions share the same support, OT does not require that condition. This feature enables OT widely used in machine learning, especially in topic models (Zhao et al., 2020; Wang et al., 2022b; Wu et al., 2023a).

Formally, consider distributions are discrete. Given a complete separable metrics space (Ω, d), where $d : \Omega \times \Omega \to \mathbb{R}$ is the metrics on the space Ω, let $P(\Omega)$ denote the set of all Borel probability measures on Ω. Given to sets $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_N)$, $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ...\boldsymbol{y}_M)$ of N and M sample points in Ω, their empirical probability measures are defined as $f = \sum_{i=1}^{N} \alpha_i \delta_{\boldsymbol{x}_i} \in P(\Omega)$ and $g = \sum_{j=1}^{M} \beta_j \delta_{\boldsymbol{y}_j} \in P(\Omega)$, respectively, where $\delta_{\boldsymbol{x}}$ is the Dirac unit mass on the position of $\boldsymbol{x}$ in Ω, α_i and β_j are the weight on the unit mass on $\boldsymbol{x}_i$, $\boldsymbol{y}_j$ respectively. Since f, g are probability distributions, the weights vectors $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ...\alpha_N)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_M)$ lie in the simplexes $\Theta_N := \{\alpha_i \geq 0 \forall i = 1, ..., N | \sum_{i=1}^{N} \alpha_i = 1\}$ and $\Theta_M := \{\beta_j \geq 0 \forall j = 1, ..., M | \sum_{j=1}^{M} \beta_j = 1\}$. The empirical joint probability measure of $(\boldsymbol{X}, \boldsymbol{Y})$ is denoted as:

$$h = \sum_{i=1}^{N} \sum_{j=1}^{M} \gamma_{ij} (\delta_{\boldsymbol{x_i}}, \delta_{\boldsymbol{y_j}}) \tag{17}$$

whose marginal measures w.r.t $\boldsymbol{X}$ and $\boldsymbol{Y}$ are f and g, respectively. The weight matrix $[\gamma_{ij}]$ is a $N \times M$ non-negative matrix with row and column marginals $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$. More concrete, $\sum_{i=1}^{N} \gamma_{ij} = \beta_j \quad \forall j = 1 \ldots M$ and $\sum_{j=1}^{M} \gamma_{ij} = \alpha_i \quad \forall i = 1 \ldots N$. The set of all the feasible weight matrixes is defined as the transportation polytope $U(\boldsymbol{\alpha}, \boldsymbol{\beta})$ of $\boldsymbol{\alpha}, \boldsymbol{\beta}$:

$$\begin{aligned} &U(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \\ &\{\boldsymbol{T} \in \mathbb{R}_{+}^{N \times M} | \boldsymbol{T} \mathbf{1}_M = \boldsymbol{\alpha}, \boldsymbol{T}^T \mathbf{1}_N = \boldsymbol{\beta}\}. \end{aligned} \tag{18}$$

An element t_{ij} of a feasible $\boldsymbol{T}$ can be seen as the amount of mass transported from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$. The distance between $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ is measured by a metric d raised to the power p. Matrix $\boldsymbol{D}$ is the pairwise distances between elements in $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$\boldsymbol{D} := [d(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij} \in \mathbb{R}^{N \times M}. \tag{19}$$

The cost of transporting f to g given a transport $\boldsymbol{T}$ is the Frobenius dot product between $\boldsymbol{T}$ and $\boldsymbol{D}$, which is $\langle \boldsymbol{T}, \boldsymbol{D} \rangle = tr(\boldsymbol{T}^T \boldsymbol{D})$.

Given $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{D}$, the OT distance between empirical probability measures f and g is a linear programing problem:

$$d_W(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{D}) = \min_{\boldsymbol{T} \in U(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \boldsymbol{T}, \boldsymbol{D} \rangle. \tag{20}$$

The solution to obtain the optimal $\boldsymbol{T}$ is quite computationally expensive. Cuturi (Distances, 2013) introduced an entropy constraint to the transportation polytope, converting the original problem to an entropy regularized optimal transportation problem, resulting in *Sinkhorn distance*, i.e:

$$\begin{aligned} &d_S^{\lambda}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{D}) = \langle \boldsymbol{T}^{\lambda}, \boldsymbol{D} \rangle \\ \text{s.t.} \quad &\boldsymbol{T}^{\lambda} = \operatorname*{argmin}_{\boldsymbol{T} \in U(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \boldsymbol{T}, \boldsymbol{D} \rangle - \frac{1}{\lambda} h(\boldsymbol{T}) \end{aligned} \tag{21}$$

where $h(\boldsymbol{T}) = -\sum_{i=1}^{N} \sum_{j=1}^{M} t_{ij} \log t_{ij}$ is the entropy of $\boldsymbol{T}$. The optimal $\boldsymbol{T}^{\lambda}$ that minimizes (21) is:

$$\boldsymbol{T}^{\lambda} = diag(\boldsymbol{\kappa}_1) \exp^{-\lambda \boldsymbol{D}} diag(\boldsymbol{\kappa}_2) \tag{22}$$

where $\exp^{-\lambda \boldsymbol{D}}$ is the element-wise exponential of the matrix $-\lambda \boldsymbol{D}$, $\boldsymbol{\kappa}_1 \in \mathbb{R}^N$, $\boldsymbol{\kappa}_2 \in \mathbb{R}^M$ are the non-negative scaling factors, which can be effectively solved after some Sinkhorn iterations. Hence, the computational cost is greatly reduced, comparing with the original problem.

C Datasets

In the experiments, we use four datasets containing short documents. The specification of each dataset is as follows:

- **GoogleNews** includes 11,109 article titles related to 152 events, originally published and processed by (Yin and Wang, 2016).

Dataset	G	C_V	TD	Purity	NMI
Google News	10	0.419	**1.0**	0.737	0.885
	20	0.417	**1.0**	0.742	0.887
	30	**0.422**	**1.0**	0.719	0.884
	40	0.417	**1.0**	**0.749**	**0.888**
	50	0.412	**1.0**	0.723	0.883
Search Snippets	10	0.411	**1.0**	**0.839**	0.507
	20	0.412	**1.0**	0.826	0.503
	30	0.420	0.997	**0.839**	0.508
	40	0.413	**1.0**	**0.839**	**0.509**
	50	**0.426**	**1.0**	**0.839**	0.505

(a) Different numbers of clusters G.

Dataset	λ_{ECR}	C_V	TD	Purity	NMI
Google News	20	0.406	**1.0**	**0.719**	**0.884**
	30	**0.422**	**1.0**	**0.719**	**0.884**
	60	0.417	**1.0**	0.710	0.873
	90	0.409	**1.0**	0.706	0.869
	120	0.406	**1.0**	0.684	0.863
Search Snippets	20	0.412	0.987	**0.839**	**0.513**
	30	**0.420**	0.997	**0.839**	0.508
	60	0.414	**1.0**	**0.839**	0.506
	90	0.417	**1.0**	**0.839**	0.506
	120	0.404	**1.0**	**0.839**	0.506

(b) Different weight λ_{ECR}.

Dataset	λ_{OT}^{DG}	C_V	TD	Purity	NMI
Google News	0.01	0.421	**1.0**	0.718	0.883
	0.1	0.409	**1.0**	0.701	0.878
	0.5	0.409	**1.0**	0.704	0.874
	1.0	**0.415**	**1.0**	0.703	0.882
	10	0.412	**1.0**	**0.752**	**0.888**
Search Snippets	0.01	**0.420**	0.997	**0.839**	0.508
	0.1	0.416	**1.0**	**0.839**	0.508
	0.5	0.405	**1.0**	**0.839**	**0.514**
	1.0	0.416	**1.0**	**0.839**	0.508
	10	0.407	**1.0**	**0.839**	0.510

(c) Different weight λ_{OT}^{DG}.

Dataset	λ_{OT}^{TG}	C_V	TD	Purity	NMI
Google News	0.01	0.404	**1.0**	**0.765**	0.893
	0.1	0.417	**1.0**	0.760	**0.891**
	0.5	**0.422**	**1.0**	0.729	0.873
	1.0	**0.422**	**1.0**	0.708	0.867
	10	0.414	**1.0**	0.703	0.863
Search Snippets	0.01	**0.417**	0.993	**0.839**	0.508
	0.1	0.413	**1.0**	**0.839**	0.508
	0.5	0.409	**1.0**	**0.839**	0.511
	1.0	0.408	**1.0**	**0.839**	0.512
	10	0.402	**1.0**	**0.839**	**0.517**

(d) Different weight λ_{OT}^{TG}.

Table 4: C_V, TD, Purity and NMI with different settings. The **bold** values are the best.

- **SearchSnippets** includes 12,340 snippets extracted from web searches, categorized into 8 groups by (Phan et al., 2008).
- **StackOverflow** is the dataset from Kaggle challenge[4]. We sample 20,000 question titles from 20 categories by (Xu et al., 2017).
- **Biomedical** is a subset of PubMed data provided by BioASQ [5], with 20,000 paper titles randomly selected from 20 categories by (Xu et al., 2017).

We reproduce settings established by (Nguyen et al., 2025a). We first obtain preprocessed versions of four datasets provided by the STTM library [6]. For each dataset, we remove words with a frequency below 3. After that, we filter out all documents with a term length of less than 2. These pre-processing steps are implemented using TopMost[7]. For global clustering, we use pre-trained language model all-MiniLM-L6-v2[8] to embed documents into a semantic representation. Then, these embeddings are clustered into a chosen number of groups using DBSCAN (Ester et al., 1996) algorithm. Table 6 provides an overview of the dataset statistics after pre-processing.

D Additional Results

We conduct experiments on the sensitivity of our proposed GloCOM-EnCOT with different values of hyper-parameters. We vary the number of clusters $G \in \{10, 20, 30, 40, 50\}$ and the weight of Optimal Transport losses $\lambda_{\mathrm{OT}}^{DG}, \lambda_{\mathrm{OT}}^{TG} \in \{0.01, 0.1, 0.5, 1, 10\}$. We choose dataset GoogleNews and SearchSnippets to evaluate the changes of hyper-parameters. The details are reported in Table 4. We also report the performance of GloCOM-EnCOT with different topics K and the weight of Embedding Cluster Regularization loss $\lambda_{ECR} \in \{20, 30, 60, 90, 120\}$. As demonstrated in Table 4, the metrics exhibit minimal sensitivity to variations in hyper-parameter values. This indicates that our models are robust, user-friendly, and

[4]https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow

[5]http://participants-area.bioasq.org/

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

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

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

Discovered Topic Examples
Topic 0: studio visual visualstudio vsnet debugger solution projects breakpoint project solutions intellisense debugging winforms ide debug
Topic 1: hibernate hql jpa manytomany criteria onetomany cascade mapping flush relation associated persisting joined unidirectional association
Topic 2: blog pages rss posts wordpress feed page comment theme homepage cms publishing sidebar template layout
Topic 3: excel vba workbook worksheet spreadsheet sheet cells formulas macro sheets oledb worksheets poi automation cell
Topic 4: widget widgets qobject qtableview qwidget handler window qtablewidget qtreewidget qtextedit qprocess qtwebkit qmenu qdialog mouse
Topic 5: linq iqueryable linqtosql lambda datacontext ienumerable submitchanges orderby joins subquery distinct dataset datasets aggregate deferred
Topic 6: error errors fatal fails wrong cause found missing problems always trying fail credit unable anymore
Topic 7: mac cocoa leopard objective snow osx nstableview nsview nsoutlineview nsstring macports nsarraycontroller macosx nstextview macbook
Topic 8: haskell monad monads ghc parsec ffi ghci foldr bytestring infinite functional tail comprehension composition pure
Topic 9: apache rewrite mod htaccess rewriterule rewriting permalink httpd www httpdconf virtualhost hosts xampp rule localhost
Topic 10: ajax jquery xmlhttprequest toolkit json div responses polling reloading chat response partial push extender request
Topic 11: oracle pl sqlplus oci apex dblink procedures blobs procedure plan ref stored odpnet rollback inserted
Topic 12: magneto adminhtml cck description forms checkbox customize form rate paypal cart currency outlook panels contact
Topic 13: scala actors immutable val implicit iterable traits abstract tuples reflection yield iterator inference trait derived
Topic 14: category categories taxonomy menu products grouped caml product filters detail stock exposed price shown grouping
Topic 15: matlab matrix plot mex figure axes plotting vector simulink plots struct dimensional vectors matrices solving
Topic 16: svn subversion repository branch revision tortoisesvn repositories externals branches repo trunk commit tortoise commits branching
Topic 17: spring bean beans aop webflow inject flow freemarker propertyplaceholderconfigurer security aspectj applicationcontext flex aspect jdbctemplate
Topic 18: bash shell stderr stdout echo bashrc script stdin awk pipe commands scripting prompt crontab scripts
Topic 19: generated may locking around connected world great happens gives visible expected facebook become coming less

Table 5: Top 15 related words of 20 discovered topics from StackOverflow. **No repeated words** are found.

Dataset	#documents	avg length	vocab
GoogleNews	11,019	5.753	3,473
SearchShippets	12,294	14.426	4,618
StackOverflow	16,378	4.4988	2,226
Biomedical	19,433	7.430	3,867

Table 6: Datasets statistics after pre-processing.

support stable training. The topic quality identified by GloCOM-EnCOT is diverse, as reflected by TD values consistently reaching 1.0 across most settings. Additionally, the document-topic representations are strong, with Purity scores exceeding 0.70 on GoogleNews and 0.83 on SearchSnippets, suggesting that EnCOT effectively enhances document-topic coherence and quality.

E Topic-word visualization

Table 5 presents the top words for each topic. The high quality of these topics is evident, as they distinctly represent specific domains such as Visual Studio IDE, Object-Relational Mapping, Blogging, and Excel. Within each topic, the words exhibit strong semantic coherence, while those across different topics display clear diversity and distinction.

F EnCOT with LLMScore

We follow setting in (Stammbach et al., 2023) to evaluate LLMscore metric for ECRTM, KNNTM and ENCOT in four datasets. The results is in Table 7.

The results show that EnCOT outperforms existing models under the LLMScore metric, highlighting its state-of-the-art topic quality as evaluated by ChatGPT.

	ECRTM	KNNTM	EnCOT
GoogleNews	2.08	2.24	**2.46**
SearchShippets	2.22	2.52	**2.88**
StackOverflow	12.0	2.48	**2.82**
Biomedical	2.04	2.24	**2.44**

Table 7: Comparison of EnCOT with ECRTM and KNNTM using the LLMScore metric. Bold values indicate the best performance.