DACTM: Enhancing Neural Topic Models through Denoising Autoencoders with Intelligent Semantic Masking and Contrastive Learning

1st YiLong Kang
Southwest University of Science and Technology
Mianyang, China
adekang@163.com

2nd Jie Yin
National Science Library (Chengdu), Chinese Academy of Sciences Sciences
Chengdu, China
caesar.yin@qq.com

3rd LiJuan Peng*
Southwest University of Science and Technology
Mianyang, China
plj@swust.edu.cn

4th ShiDi Xie
Southwest University of Science and Technology
Mianyang, China
1165474497@qq.com

5th ChunJiang Liu
National Science Library (Chengdu), Chinese Academy of Sciences Sciences
Chengdu, China
liucj@clas.ac.cn

6th HaiYun Xu
School of Business, Shandong University of Technology
Shandong, China
xuhaiyunnemo@gmail.com

***Abstract*—Topic modeling is a key technique for uncovering latent thematic structure from large-scale text corpora. However, existing neural topic models largely rely on sparse bag-of-words inputs, which makes it difficult to adequately capture intra-document semantic dependencies; moreover, they are prone to latent topic mode collapse, resulting in weak document–topic associations and limited interpretability. To address these issues, we propose DACTM, a neural topic model enhanced by semantic awareness and contrastive learning. On the encoder side, DACTM introduces a Semantic-Aware Module that constructs a document-specific sparse semantic subgraph to improve document–topic alignment. On the decoder side, DACTM incorporates topic-level contrastive regularization to impose geometric constraints in the topic embedding space, which (i) increases semantic separation among topics and (ii) aligns topic representations with genuine word clusters, thereby alleviating mode collapse. Experiments on six benchmark datasets covering both long and short texts demonstrate that DACTM consistently outperforms representative neural topic models in terms of topic coherence, clustering performance, and the quality of learned document representations.**

***Index Terms*—neural topic models, topic modeling , contrastive learning**

I. Introduction

Topic modeling is a prominent research direction in natural language processing (NLP). It aims to uncover latent, abstract topics from large-scale text collections and to produce high-level semantic representations of documents, thereby facilitating efficient understanding and organization of unstructured data. Topic models have been widely applied to a variety of text analysis tasks, including text mining [1], [2], recommender systems [3], [4], streaming learning [5], [6], and trend analysis [7], [8].

* Corresponding author.

Compared with traditional Bayesian approaches, neural topic models (NTMs) [9] leverage the expressive power of neural networks to alleviate the computational bottlenecks of classical methods when scaling to large corpora. Most NTMs adopt the variational autoencoder (VAE) framework [10], which maps each document into a latent topic space (encoder) and reconstructs the original text representation from this space (decoder). Consequently, improving the encoder's contextual sensitivity and strengthening the semantic structure of the decoder have become central directions for advancing neural topic modeling.

On the encoder side, to overcome the sparsity of bag-of-words (BoW) representations, prior work often incorporates pretrained language models (PLMs) [11] as external knowledge sources, e.g., using BERT [12] or GPT [13] to obtain contextualized document representations. Although these large models strengthen the ability to capture linguistic patterns, their substantial parameter counts and computational overhead considerably limit the practicality of topic models in latency-sensitive or resource-constrained settings. On the decoder side, a representative recent line of work [14] adopts an embedding-based generation paradigm, modeling the topic–word distribution as a Softmax-normalized function of the negative Euclidean distance between topic and word embeddings in the latent space. While this approach effectively injects lexical semantics into the generation process, it primarily emphasizes local word–topic alignment and overlooks global topic–topic geometric constraints. Without explicit semantic repulsion among topics, learned topic embeddings can become entangled in the latent space, which ultimately harms interpretability.

To address these limitations, we propose DACTM, a neural topic modeling framework with semantic awareness and contrastive regularization. To mitigate the encoder's limited

contextual sensitivity and its heavy reliance on external models, we design a Semantic-Aware Module (SAM). Unlike approaches that directly incorporate large BERT-style encoders, SAM exploits the sparsity inherent in each document by dynamically constructing a sparse masked subgraph and efficiently modeling fine-grained word–word co-occurrence relations through masked attention. For the decoder, we further introduce Topic-level Contrastive Regularization (TCR) over the global topic embedding matrix. Specifically, we treat Dropout as a stochastic perturbation in the latent topic space to generate augmented views, and optimize a contrastive objective to explicitly increase the semantic separation among topic vectors. This encourages approximately orthogonal topic representations, thereby improving topic diversity and interpretability. Our contributions are summarized as follows:

- We propose a semantic-aware encoder module, SAM, which replaces the reliance on PLMs with data-driven dynamic sparse attention. By constructing a document-specific sparse semantic subgraph, SAM recovers fine-grained intra-document co-occurrence structure and contextual semantics.
- We introduce Topic-level Contrastive Regularization (TCR) over the global topic embedding matrix. TCR imposes geometric constraints in the topic embedding space and explicitly increases semantic separation among topics, alleviating mode collapse while improving topic diversity and interpretability.
- We present an end-to-end topic model, DACTM, and conduct extensive experiments on multiple benchmark datasets. The results demonstrate that DACTM consistently outperforms strong state-of-the-art baselines.

II. Related Work

Traditional topic modeling methods, such as probabilistic graphical models including latent Dirichlet allocation (LDA) and algebraic approaches such as non-negative matrix factorization (NMF) and latent semantic analysis (LSA), primarily rely on bag-of-words (BoW) representations to extract topics. In recent years, neural topic models (NTMs) [15]–[17] have emerged as a promising alternative: they leverage neural networks to improve modeling flexibility and topic quality, while alleviating the inefficient parameter inference of classical topic models. Most NTMs are built upon the variational autoencoder (VAE) framework, where the encoder infers document–topic proportions and the decoder reconstructs the original text representation.

Within the VAE-based topic modeling framework, two major research directions aim to improve its two components: the encoder and the decoder. On the encoder side, pretrained language models are frequently incorporated as external knowledge sources. For instance, contextual document embeddings from SBERT [18] have been used as additional inputs to complement BoW representations and capture valuable semantic signals. Han et al. [11] further leverage SBERT to generate term weights and integrate them into the reconstruction loss, thereby down-weighting irrelevant words.

On the decoder side, Dieng et al. [19] propose incorporating word embeddings, such as Word2Vec [20], to better capture lexical semantics and thereby induce topics composed of more semantically related words. Xu et al. [21] introduce HyperMiner, a hyperbolic-embedding-based approach for topic taxonomy mining. By representing topics in a hyperbolic (e.g., hyperspherical) space, HyperMiner leverages non-Euclidean geometry to capture hierarchical relations among topics and construct a topic taxonomy effectively. Zhao et al. [15] propose the NSTM model, which measures reconstruction error using the optimal transport distance between the inferred document–topic distribution and the input document. Wu et al. [14] apply a clustering-based regularization strategy to encourage each topic embedding to act as a centroid of a distinct cluster of word embeddings, thereby alleviating topic collapse.

Beyond VAEs, alternative architectures have also been explored. Grootendorst et al. [22] group document embeddings to form topics by introducing a class-based TF–IDF variant. Zhang et al. [23] cluster high-quality sentence embeddings with an appropriate term selection strategy to obtain more coherent and diverse topics. However, these clustering-based methods cannot directly infer document-level topic proportions, which limits their generative capability and transferability. Wu et al. [24] propose FASTopic, which improves topic coherence and interpretability via a dual semantic relation reconstruction paradigm. Despite its high efficiency, FASTopic is constrained by train–test inconsistency and remains sensitive to the choice of pretrained language model.

III. Background

Let $\mathcal{X} = \{\mathbf{x}_d\}_{d=1}^{D}$ denote a corpus of D documents represented by bag-of-words (BoW) vectors, defined over a vocabulary of size V. The goal of topic modeling is to uncover K latent topics from $\mathcal{X}$. We define the word embedding matrix as $\mathbf{W} \in \mathbb{R}^{V \times L}$ and the topic embedding matrix as $\mathbf{T} \in \mathbb{R}^{K \times L}$, where L is the embedding dimension.

Following recent neural topic models such as ETM, we do not parameterize the topic–word distribution $\boldsymbol{\beta}$ via a simple matrix factorization. Instead, to capture semantic geometry, we model $\boldsymbol{\beta}$ based on distances between topic embeddings and word embeddings in the latent space. Concretely, the probability of word v under topic k, denoted by β_{kv}, is defined as a Softmax-normalized negative Euclidean distance:

$$\beta_{kv} = \frac{\exp\left(-\|\mathbf{t}_k - \mathbf{w}_v\|^2/\tau\right)}{\sum_{v'=1}^{V} \exp(-\|\mathbf{t}_k - \mathbf{w}_{v'}\|^2/\tau)} \tag{1}$$

where $\mathbf{t}_k$ is the embedding of topic k, $\mathbf{w}_v$ is the embedding of word v, and τ is a temperature hyperparameter controlling the sharpness of the distribution. This formulation encourages each topic to concentrate probability mass on words that are semantically close to the topic center in the embedding space.

Another core objective is to infer the topic proportions for each document $\mathbf{x}_d$, denoted by $\boldsymbol{\theta}_d \in \Delta^{K-1}$. In VAE-based topic models, the intractable true posterior is approximated by a variational distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$. An inference network

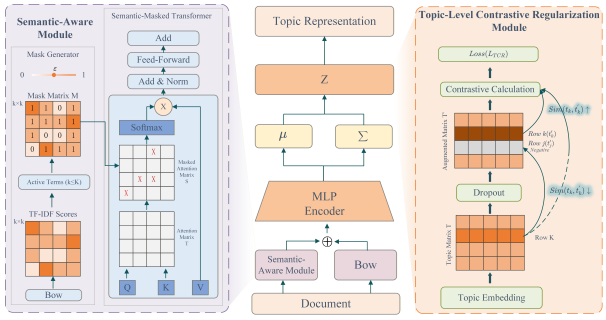

Fig. 1. DACTM consists of a semantic-aware encoder and a contrastively regularized decoder. On the encoder side, the Semantic-Aware Module (SAM) generates a semantic mask to constrain self-attention, strengthening semantic modeling of BoW documents. On the decoder side, Topic-level Contrastive Regularization (TCR) applies contrastive learning to the topic embeddings, mitigating topic collapse and improving interpretability.

(e.g., an MLP) outputs the parameters of a Gaussian posterior, namely the mean $\boldsymbol{\mu}$ and a diagonal covariance $\boldsymbol{\Sigma}$. Using the reparameterization trick, we sample the latent variable as

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

Subsequently, we obtain the topic proportions $\boldsymbol{\theta}_d$ by applying a Softmax function to $\mathbf{z}$, thereby approximating a Logistic-Normal prior distribution. Finally, the BoW representation is reconstructed via the generative distribution $\hat{\mathbf{x}} \sim \mathrm{Multinomial}(\mathrm{softmax}(\boldsymbol{\beta\theta}))$. The standard loss function of VAE-based topic models is the evidence lower bound (ELBO), consisting of a reconstruction term and a regularization term:

$$\mathcal{L}_{\mathrm{TM}} = \frac{1}{D}\sum_{i=1}^{D}\Big[-(\mathbf{x}_{i\mathrm{BoW}})^{\top}\log\big(\mathrm{softmax}(\beta\theta_i)\big) + \mathrm{KL}\big(q(z|\mathbf{x}_i)\,\|\,p(z)\big)\Big] \tag{3}$$

IV. Proposed Method

The high sparsity of BoW inputs makes it difficult for the encoder to capture word–word dependencies, while the lack of explicit separation in the latent topic space can lead to collapsed topic representations. DACTM addresses these two issues at the document level and the topic level, respectively: on the encoder side, we enhance document representations via structured semantic constraints, and on the decoder side, we increase separation among topic embeddings through contrastive regularization. Together, these components improve the interpretability and stability of the learned topics. As shown in Figure 1, our framework is built upon a VAE and is trained end-to-end by jointly optimizing both types of constraints. The following sections detail how these components cooperate within a unified end-to-end learning procedure.

A. Semantic-Aware Module

The Semantic-Aware Module (SAM) consists of three core components: (1) a Dynamic Sparse Mask Generator (DSMG), (2) a Semantic-Masked Transformer layer (SMT), and (3) Semantic Contrastive Optimization (SCO). Specifically, DSMG is designed to perceive the sparsity of the input document and dynamically generate a document-specific mask matrix (i.e., an effective word subgraph), which guides the masked-attention computation in SMT to filter out noise. SMT models fine-grained intra-document word co-occurrence patterns and semantic dependencies. SCO further guides DSMG and SMT during training to explore more robust and discriminative semantic representations.

a) Dynamic Sparse Mask Generator (DSMG): We treat document-level lexical associations as a dynamic state that is strongly conditioned on the current input content. Accordingly, we design a statistics-driven dynamic mask generator that represents each document's salient lexical substructure with a binary mask matrix, thereby isolating interference from absent words and low-frequency noisy terms. Unlike the independence assumption in BoW, and to avoid the expensive full attention computation over the entire vocabulary, we exploit BoW sparsity to construct a semantic window of controllable size within each document, and use it as a structural constraint for subsequent attention computation.

At the document level, given the BoW input vector of the i-th document $\mathbf{x}^{(i)} \in \mathbb{R}^V$, we first perform threshold filtering to remove long-tail noisy terms:

$$\widetilde{\mathbf{x}}^{(i)} = \mathrm{Thresh}\Big(\mathbf{x}^{(i)}, \varepsilon\Big) \tag{4}$$

where $\mathrm{Thresh}(\cdot)$ sets terms with frequency lower than the threshold ε to zero. Next, to obtain a deterministic set of salient terms, we select the top-K words with the highest TF–IDF scores from $\widetilde{\mathbf{x}}^{(i)}$ to form an effective index set $\mathcal{I}^{(i)}$, and sort them in descending order to obtain an effective word sequence $\mathbf{s}^{(i)} = [v_1, \ldots, v_k]$, where k denotes the effective length for the current document. Over this sequence space, we construct a document-specific mask matrix $\mathbf{M}^{(i)}$ following a co-occurrence-based statistical prior: any two words within the effective length k are assumed to share a context and thus exhibit a potential semantic dependency. Formally, the mask is defined as

$$\mathbf{M}^{(i)}_{p,q} = \begin{cases} 1, & \text{if } 1 \le p \le k \text{ and } 1 \le q \le k, \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where p and q are position indices in the sequence. In this way, $\mathbf{M}^{(i)}$ forms a fully connected subgraph over the effective semantic region while setting the padded region to zero, thereby dynamically cropping a compact semantic window for each document and providing an explicit structural constraint and computation boundary for subsequent Transformer attention.

b) Semantic-Masked Transformer layer (SMT): Given the effective word set and its mask matrix produced by DSMG for the i-th document, we employ a Transformer layer to further model fine-grained semantic relations among salient words. Notably, SMT does not introduce positional encodings, so the attention computation does not depend on token order and instead treats the input as a set, focusing

on pairwise word relations. Before each attention block, we apply layer normalization to stabilize training and to mitigate the dominance of certain high-frequency words or high-norm embeddings on attention weights:

$$E^{*(i)} = \mathrm{LayerNorm}\Big(E'^{(i)}\Big) = \frac{E'^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{6}$$

where $E'^{(i)}$ and $E^{*(i)}$ denote the initial and normalized hidden representations of the effective word sequence for the i-th document, respectively, L is the sequence length, and D is the embedding dimension.

We then use masked attention to precisely model dependencies between effective word pairs. To ensure gradients can propagate effectively under the sparse structure, we inject the binary mask $\mathbf{M}^{(i)}$ as a hard structural constraint into the attention computation:

$$Q^{(i)} = E^{*(i)}W^Q,\ K^{(i)} = E^{*(i)}W^K,\ V^{(i)} = E^{*(i)}W^V \tag{7}$$

$$T^{(i)} = Q^{(i)}\big(K^{(i)}\big)^\top,$$

$$S^{(i)} = T^{(i)} \odot M^{(i)} + \big(1 - M^{(i)}\big) \odot (-\infty) \tag{8}$$

$$\mathrm{MaskedScores}^{(i)} = \frac{S^{(i)}}{\sqrt{d}},$$

$$\widetilde{E}^{(i)} = \mathrm{Softmax}\big(\mathrm{MaskedScores}^{(i)}\big)V^{(i)} \tag{9}$$

where $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ are learnable projection matrices, $\mathbf{M}^{(i)}$ is the binary sparse mask from the previous stage, and $\mathbf{T}^{(i)}$ is the raw attention score matrix. By adding $-\infty$ at positions where $\mathbf{M}^{(i)} = 0$, the corresponding probabilities become exactly zero after Softmax, so information aggregation is performed only over effective word pairs that co-occur within the document. SMT adopts multi-head attention and can be stacked to extract higher-order semantic interaction features.

c) Semantic Contrastive Optimization (SCO): A hard mask constraint alone is insufficient to make the attention projection parameters $\mathbf{W}^Q$ and $\mathbf{W}^K$ fully adapt to the sparse semantic structure. To this end, we introduce a dynamic semantic contrastive loss $\mathcal{L}_{\mathrm{DC}}$ to explicitly strengthen semantic paths over the masked subgraph. For the i-th document, let L denote the number of effective words and $\mathbf{M}^{(i)}$ be the corresponding mask. Let $\mathbf{P}'^{(i)} = \mathrm{Softmax}\big(\mathbf{A}'^{(i)}\big)$ denote the masked attention probability matrix, where $\mathbf{A}'^{(i)}$ is obtained by the masked attention in Eq. (8)–(9). We treat word pairs with $\mathbf{M}^{(i)}_{p,q} = 1$ as the positive set $\mathcal{P}^{(i)}_p$ for anchor word p, and uniformly sample r negatives from positions with $\mathbf{M}^{(i)}_{p,q} = 0$ to form the negative set $\mathcal{N}^{(i)}_p$. The dynamic contrastive loss is defined over each anchor p as

$$\mathcal{L}^{(i)}_{\mathrm{DC}} = -\sum_{p=1}^{L} \log \frac{\sum_{q \in \mathcal{P}^{(i)}_p} \exp\big(\mathbf{P}'^{(i)}_{p,q}/\tau\big)}{\sum\limits_{q \in \mathcal{P}^{(i)}_p} \exp\big(\mathbf{P}'^{(i)}_{p,q}/\tau\big) + \sum\limits_{q \in \mathcal{N}^{(i)}_p} \exp\big(\mathbf{P}'^{(i)}_{p,q}/\tau\big)} \tag{10}$$

where τ is a temperature coefficient. This loss increases the attention response for co-occurring word pairs while suppressing non-co-occurring pairs, encouraging the learned attention weights to align with the sparse structure.

B. Topic-Level Contrastive Regularization

A long-standing challenge in neural topic modeling is *mode collapse*, where multiple latent topics converge to similar semantic meanings, resulting in highly redundant topics and poor interpretability. Motivated by contrastive learning, we propose Topic-level Contrastive Regularization (TCR), which is directly applied to the decoder's global topic embedding matrix and explicitly encourages different topics to be separated in semantic space.

Let $\mathbf{T} \in \mathbb{R}^{K \times D}$ denote the learnable topic embedding matrix, where K is the number of topics and D is the embedding dimension. We first ℓ_2-normalize the embeddings onto the unit hypersphere. Next, we treat Dropout as a minimal data augmentation strategy to construct perturbed views of the topic space. Concretely, during training we apply a standard dropout mask to $\mathbf{T}$ to obtain an augmented topic matrix $\widetilde{\mathbf{T}}$. For a specific topic k, its original embedding $\mathbf{t}_k$ and the perturbed version $\widetilde{\mathbf{t}}_k$ form a positive pair, while all other perturbed topic embeddings $\{\widetilde{\mathbf{t}}_j\}_{j \neq k}$ serve as negative samples. The objective maximizes the similarity of the positive pair and minimizes the similarity to negative samples, which can be formulated as:

$$\mathcal{L}_{\mathrm{TCR}} = -\sum_{k=1}^{K} \log \frac{\exp\Big(\mathrm{sim}\Big(\hat{\boldsymbol{\phi}}_k, \hat{\boldsymbol{\phi}}'_k\Big)/\tau\Big)}{\sum_{j=1}^{K} \exp\Big(\mathrm{sim}\Big(\hat{\boldsymbol{\phi}}_k, \hat{\boldsymbol{\phi}}'_j\Big)/\tau\Big)} \tag{11}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature hyperparameter that controls the strength of penalties on hard negatives. By minimizing $\mathcal{L}_{\mathrm{TCR}}$, the model is explicitly driven to push different topic embeddings apart in semantic space, thereby suppressing topic collapse and improving topic diversity without introducing rigid orthogonality constraints or additional discriminators. In particular, rigid orthogonality constraints explicitly enforce the topic embedding matrix to satisfy orthogonality, assuming topics are completely independent in geometric space.

C. Objective Function

Our inference procedure and the base topic modeling loss follow standard VAE-based neural topic models by optimizing the evidence lower bound (ELBO), denoted as $\mathcal{L}_{\mathrm{TM}}$. In addition, inspired by Wu et al. [16], we introduce Embedding Clustering Regularization (ECR) to alleviate semantic mismatch between topics and words. The ECR loss is formulated as an optimal transport distance:

$$\mathcal{L}_{\mathrm{ECR}} = \sum_{i=1}^{V}\sum_{j=1}^{K} \|\mathbf{w}_i - \mathbf{t}_j\|^2\ \pi^*_{ij} \tag{12}$$

where π^* is the solution to the following Sinkhorn optimization problem:

$$\begin{aligned} \underset{\pi}{\text{minimize}} \quad & \langle C_{\mathrm{WT}}, \pi \rangle - \nu H(\pi) \\ \text{s.t.} \quad & \pi \in \mathbb{R}^{V \times K} \\ & \pi 1_K = \frac{1}{V} 1_K, \quad \pi^T 1_V = \frac{1}{K} 1_K \end{aligned} \tag{13}$$

where $\mathbf{C}_{WT}$ is the distance matrix between word embeddings and topic embeddings, $H(\pi)$ is an entropy regularization term, and π^* is efficiently obtained via the Sinkhorn–Knopp algorithm.

In summary, beyond the base topic modeling objective, we employ three auxiliary losses to control different aspects of representation learning: $\mathcal{L}_{\mathrm{ECR}}$ captures precise word–topic geometry, $\mathcal{L}_{\mathrm{TCR}}$ regularizes topic–topic discriminability, and $\mathcal{L}_{\mathrm{DC}}$ enhances the encoder's contextual sensitivity. DACTM is trained with the following weighted objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{TM}} + \lambda_{\mathrm{ECR}} \mathcal{L}_{\mathrm{ECR}} + \lambda_{\mathrm{TCR}} \mathcal{L}_{\mathrm{TCR}} + \lambda_{\mathrm{DC}} \mathcal{L}_{\mathrm{DC}} \tag{14}$$

where λ_{ECR}, λ_{TCR}, and λ_{DC} are hyperparameters that balance the contributions of different terms.

V. Experiments

A. Settings

a) Datasets: We evaluate our model on three long-text datasets: 20 Newsgroups (20NG) [25], a standard benchmark for topic modeling; AGNews [23], which contains news articles collected from over 2,000 sources; and IMDB [26], a corpus of movie reviews. We further consider three short-text datasets: SearchSnippets [27], which includes over 12,000 web search snippets from 8 domains; GoogleNews [28], which contains 10,000 news headlines categorized into 152 topics; and Yahoo Answers (Yahoo) [29], a collection of questions and answers from the Yahoo! community covering 10 topic categories. This diverse suite enables a comprehensive evaluation across different text types and structural complexities.

b) Baselines: We compare DACTM with several representative topic modeling frameworks, including LDA [30], a classical probabilistic topic model; ETM [19], a neural topic model with word embeddings; NTM+CL [31], which uses contrastive learning to model relationships between similar and dissimilar documents; NSTM [15], which employs a Sinkhorn distance to measure discrepancies between document–word and document–topic distributions; ECRTM [14], which improves topic coherence and discriminability via clustering regularization; FASTopic [24], which models document–word–topic relations using optimal transport; and NeuroMax [32], which refines topic distributions via pretrained embeddings and mutual information maximization.

c) Evaluation metrics: Following the evaluation framework of [14], we assess models from two perspectives: topic quality and document–topic distributions. For topic quality, we report (1) topic coherence ($\mathrm{C_V}$) [33], which measures the semantic coherence of topics. Empirically, $\mathrm{C_V}$ is often more consistent with human judgments than NPMI, UCI, and UMass. It is computed using a Wikipedia-based reference corpus, which helps reduce bias introduced by small in-domain reference corpora and improves reproducibility; and (2) topic diversity (TD), defined as the proportion of unique words among the top words across topics. For document–topic distributions, we evaluate document clustering performance using normalized mutual information (NMI) and Purity [34]. NMI measures the agreement between clustering assignments and ground-truth labels, while Purity quantifies the extent to which each cluster is dominated by a single class; together, they provide a complementary assessment of clustering quality.

d) Implementation details: To ensure fairness and reproducibility, we use the same model architecture and training pipeline across all datasets. Method-specific hyperparameters, such as τ and ε, are selected via grid search. Unless otherwise specified, we set the regularization weights in the overall objective to $\lambda_{\mathrm{ECR}} = 100$, $\lambda_{\mathrm{DC}} = 0.5$, and $\lambda_{\mathrm{TCR}} = 50$.

B. Topic and Doc-Topic Distribution Quality

We evaluate topic quality and the effectiveness of document–topic distributions on six benchmark datasets. Table I and Table II report results on short-text and long-text datasets with $K{=}50$ and $K{=}100$ topics, evaluated by topic coherence (CV), topic diversity (TD), and clustering-oriented metrics (NMI and Purity).

Across the short-text benchmarks, DACTM shows a clear advantage in clustering quality. For $K{=}50$, it achieves the best Purity on all three datasets (SearchSnippets/GoogleNews/Yahoo) and also delivers the best NMI on SearchSnippets and GoogleNews, indicating more discriminative document representations under sparse contexts. For $K{=}100$, DACTM remains the strongest overall: it ranks first on CV for all three datasets and yields the best TD across the board, while maintaining top Purity on GoogleNews and Yahoo and competitive NMI. These results suggest that DACTM better balances semantic coherence and representation separability even when the topic granularity increases.

On long-text datasets, DACTM remains competitive and demonstrates robust scaling from $K{=}50$ to $K{=}100$. For $K{=}50$, it attains the best CV on 20NG and AGNews and achieves strong clustering performance, with the highest Purity on 20NG and near-best results on AGNews and IMDB. When increasing to $K{=}100$, DACTM consistently obtains the best CV on 20NG and AGNews and the best NMI on 20NG and AGNews, while also achieving the best Purity on 20NG and AGNews. Although some baselines (e.g., NeuroMax) can be competitive on specific metrics (notably IMDB NMI/Purity), DACTM provides a more consistent improvement pattern across datasets and metrics.

Overall, the gains in NMI and Purity validate that DACTM learns higher-quality document–topic distributions that better support clustering, while the improvements in CV together with consistently high TD indicate that these discriminative representations do not come at the cost of topic interpretability or diversity. The advantage is particularly pronounced on short texts, where sparse word co-occurrence makes doc-topic inference more challenging, highlighting the effectiveness of our semantic-aware modeling.

TABLE I

RESULTS ON THE SHORT-TEXT BENCHMARKS WITH K=50 AND K=100 TOPICS, EVALUATED BY CV, TD, NMI, AND PURITY. BEST RESULTS ARE IN BOLD AND SECOND-BEST ARE UNDERLINED.

	SearchSnippets				GoogleNews				Yahoo			
Model	CV	TD	NMI	Purity	CV	TD	NMI	Purity	CV	TD	NMI	Purity
$K = 50$ ***Topics***												
ETM	0.397	0.594	0.389	0.688	0.364	0.916	0.560	0.366	0.354	0.557	0.192	0.405
NTM+CL	0.437	0.532	0.030	0.215	0.433	0.301	0.005	0.041	0.387	0.429	0.021	0.547
NSTM	0.422	0.492	0.319	0.475	0.402	0.598	0.474	0.341	0.390	0.658	0.241	0.395
ECRTM	0.450	**0.998**	0.419	0.711	**0.466**	0.987	0.615	0.396	0.405	0.985	0.295	0.550
FASTopic	0.356	0.519	0.497	0.793	0.379	0.235	0.570	0.252	0.416	0.938	**0.354**	0.579
NeuroMax	0.427	0.920	0.427	0.743	0.385	**1.000**	0.590	0.359	0.404	0.979	0.331	0.588
DACTM	**0.454**	0.986	**0.512**	**0.795**	0.457	0.999	**0.621**	**0.401**	**0.422**	**0.994**	0.345	**0.593**
$K = 100$ ***Topics***												
ETM	0.389	0.448	0.365	0.691	0.398	0.677	0.713	0.554	0.353	0.624	0.208	0.428
NTM+CL	0.406	0.394	0.020	0.217	0.432	0.367	0.005	0.039	0.387	0.659	0.242	0.405
NSTM	0.399	0.547	0.416	0.762	0.415	0.774	0.538	0.384	0.368	0.374	0.031	0.215
ECRTM	0.405	0.966	0.443	0.789	0.418	0.991	0.491	0.342	0.389	0.903	0.311	0.563
FASTopic	0.400	0.463	0.466	0.801	0.366	0.100	0.459	0.237	0.397	0.869	0.311	**0.591**
NeuroMax	0.412	0.960	**0.472**	**0.854**	0.427	0.915	0.834	**0.664**	0.390	0.922	0.329	0.583
DACTM	**0.436**	**0.995**	0.695	0.841	**0.447**	**0.997**	**0.861**	0.642	**0.424**	**0.996**	**0.332**	0.587

TABLE II

RESULTS ON THE LONG-TEXT BENCHMARKS WITH K=50 AND K=100 TOPICS, EVALUATED BY CV, TD, NMI, AND PURITY. BEST RESULTS ARE IN BOLD AND SECOND-BEST ARE UNDERLINED.

	20NG				AGNews				IMDB			
Model	CV	TD	NMI	Purity	CV	TD	NMI	Purity	CV	TD	NMI	Purity
$K = 50$ ***Topics***												
ETM	0.375	0.704	0.319	0.347	0.364	0.819	0.224	0.679	0.346	0.557	0.038	0.660
NTM+CL	0.437	0.802	0.491	0.582	0.440	0.441	0.100	0.322	0.396	0.617	0.044	0.657
NSTM	0.391	0.473	0.363	0.383	0.411	0.877	0.324	0.772	0.334	0.175	0.040	0.658
ECRTM	0.431	0.964	0.524	0.560	0.466	0.961	0.367	0.802	0.393	0.974	0.058	0.694
FASTopic	0.427	**0.980**	0.528	0.583	0.379	0.960	0.352	**0.831**	0.371	0.969	0.055	0.683
NeuroMax	0.435	0.912	**0.570**	**0.623**	0.385	0.952	**0.410**	0.804	0.402	0.936	**0.061**	**0.709**
DACTM	**0.456**	0.964	0.564	**0.641**	**0.476**	**0.999**	0.351	0.822	**0.428**	**0.980**	0.055	0.690
$K = 100$ ***Topics***												
ETM	0.369	0.573	0.339	0.394	0.371	0.674	0.204	0.674	0.341	0.371	0.037	0.648
NTM+CL	0.420	0.624	0.490	0.626	0.415	0.277	0.050	0.280	0.382	0.492	0.044	0.705
NSTM	0.391	0.473	0.363	0.383	0.421	0.832	0.359	0.764	0.340	0.255	0.039	0.659
ECRTM	0.405	0.904	0.494	0.555	0.416	0.981	0.428	0.812	0.373	0.887	0.049	0.694
FASTopic	0.400	0.861	0.522	0.622	0.385	0.912	0.330	0.833	0.369	0.886	0.048	0.680
NeuroMax	0.412	**0.913**	0.516	0.602	0.406	0.957	0.389	0.828	0.381	0.870	**0.059**	**0.706**
DACTM	**0.448**	0.869	**0.538**	**0.645**	**0.460**	**0.998**	0.344	**0.835**	**0.411**	**0.921**	0.045	0.681

C. Visualization of Embedding Space

We use t-SNE to visualize the learned topic embeddings and word embeddings on the Yahoo dataset with $K = 50$ topics. As shown in Figure 2, DACTM preserves a well-structured topic embedding space and effectively prevents topic collapse, whereas strong baselines often suffer from embedding collapse or fail to capture a clear word–topic structure. This observation highlights the effectiveness of our approach in clustering semantically coherent words and inducing well-separated topics.

D. The Impact of Pre-trained Embeddings

To investigate whether DACTM can be further improved by incorporating external knowledge—and to verify whether it relies on such knowledge to handle sparsity—we compare DACTM with two variants that integrate pretrained embeddings: +SBERT, which fuses sentence embeddings from BERT, and +LLM2VEC, which incorporates recent embedding representations derived from large language models. As shown in Figure 3, introducing stronger pretrained embeddings does not yield notable performance gains, and even leads to degradations on some metrics. These results provide strong evidence that the proposed Semantic-Aware Module can ef-

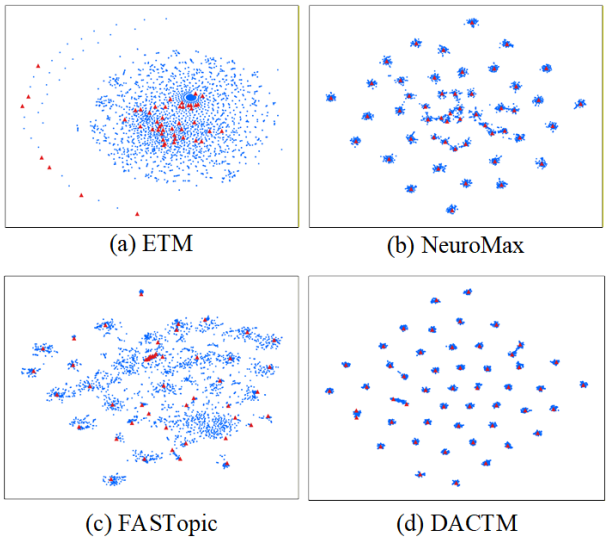

Fig. 2. t-SNE visualization of word embeddings (∗) and topic embeddings (▲) on the Yahoo dataset with $K = 50$ topics.

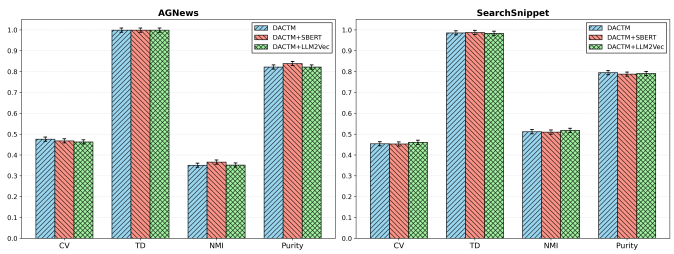

Fig. 3. different document encoders.

ficiently capture intrinsic semantic structure and contextual relations directly from the corpus. Overall, DACTM is a self-contained and efficient framework that avoids dependence on external pretrained models, making it more lightweight and adaptable for domain-specific applications.

E. Ablation Study

To verify the effectiveness of the key components in DACTM, we conduct ablation experiments on one long-text dataset (20NG) and one short-text dataset (Yahoo), using ECRTM as a strong baseline. Table III reports the results with K=50 topics.

Removing Topic-level Contrastive Regularization (W/O TCR) leads to consistent performance drops across metrics, with the most pronounced degradation on topic diversity (TD). On Yahoo, TD decreases sharply from 0.994 to 0.658, accompanied by declines in NMI and CV, indicating that conventional regularization alone is insufficient to prevent mode collapse. In contrast, TCR introduces a contrastive "repulsion" in the topic embedding space, explicitly increasing semantic separation among topics and improving topic independence and coverage.

When removing the Semantic-Aware Module (W/O SAM), the decrease in CV is relatively small, but clustering quality

TABLE III
ABLATION RESULTS ON 20NG AND YAHOO WITH K=50 TOPICS.

Model	20NG				Yahoo			
	CV	TD	NMI	Purity	CV	TD	NMI	Purity
ECRTM	0.431	0.964	0.524	0.560	0.405	0.985	0.295	0.550
DACTM	0.456	0.964	0.564	0.641	0.420	0.994	0.345	0.593
w/o SAM	0.455	0.904	0.604	0.532	0.410	0.924	0.300	0.526
w/o TCR	0.448	0.860	0.549	0.621	0.390	0.658	0.241	0.395

deteriorates noticeably. For example, Purity on 20NG drops from 0.641 to 0.532. This suggests that global word-frequency statistics are inadequate for capturing intra-document latent semantic structure. By enabling sparse contextual modeling with a Transformer under a document-specific mask, SAM better captures local co-occurrence relations and yields more accurate document–topic distributions, and this advantage also manifests in short-text clustering.

CONCLUSION

We propose DACTM, a novel neural topic modeling framework. By introducing a dynamic sparse mask generator, SAM restores fine-grained intra-document contextual dependencies from BoW inputs without increasing computational overhead. By further imposing contrastive regularization on the topic embedding matrix, DACTM explicitly promotes geometric diversity in the latent space. Extensive experiments on both long-text and sparse short-text benchmarks show that DACTM consistently outperforms state-of-the-art methods in topic coherence, topic diversity, and document clustering quality. Moreover, our analyses reveal that DACTM is highly self-sufficient: it achieves strong performance without relying on heavy pretrained encoders such as SBERT. This makes DACTM an efficient and robust solution for topic modeling in domain-specific and resource-constrained scenarios. In future work, we will explore extending the proposed dynamic sparse masking mechanism to hierarchical topic modeling in order to capture more complex semantic taxonomies.

ACKNOWLEDGMENTS

The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences. This research was funded by the National Natural Science Funding of China(No. 72274113), in part by National Science Library (Chengdu), Chinese Academy of Sciences.

REFERENCES

[1] F. Shahriar and M. A. Bashar, "Automatic monitoring social dynamics during big incidences: A case study of covid-19 in bangladesh," *arXiv preprint arXiv:2101.09667*, 2021.
[2] S. Khan, F. Ahmed, and M. Mubeen, "A text-mining research based on lda topic modelling: A corpus-based analysis of pakistan's un assembly speeches (1970–2018)," *International Journal of Humanities and Arts Computing*, vol. 16, no. 2, pp. 214–229, 2022.
[3] C. Wang, Y. Ma, Y. Shi, and G. Chen, "A type-2 fuzzy review topic-based model for personalized recommendation.," *Electronic Commerce Research*, vol. 25, no. 5, 2025.

[4] M. Kawai, H. Sato, and T. Shiohama, "Topic model-based recommender systems and their applications to cold-start problems," *Expert Systems with Applications*, vol. 202, p. 117129, 2022.

[5] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 937–946, 2009.

[6] D. A. Nguyen, K. A. Nguyen, C. H. Nguyen, K. Than, *et al.*, "Boosting prior knowledge in streaming variational bayes," *Neurocomputing*, vol. 424, pp. 143–159, 2021.

[7] D. Sharma, B. Kumar, and S. Chand, "A trend analysis of machine learning research with topic models and mann-kendall test," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 2, pp. 70–82, 2019.

[8] W. Liu, H. Zhang, Z. Shi, Y. Wang, J. Chang, and J. Zhang, "Risk topics discovery and trend analysis in air traffic control operations—air traffic control incident reports from 2000 to 2022," *Sustainability*, vol. 15, no. 15, p. 12065, 2023.

[9] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," *arXiv preprint arXiv:1703.01488*, 2017.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] S. Han, M. Shin, S. Park, C. Jung, and M. Cha, "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*, pp. 1802–1817, 2023.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[14] X. Wu, X. Dong, T. T. Nguyen, and A. T. Luu, "Effective neural topic modeling with embedding clustering regularization," in *International Conference on Machine Learning*, pp. 37335–37357, PMLR, 2023.

[15] H. Zhao, D. Phung, V. Huynh, T. Le, and W. Buntine, "Neural topic model via optimal transport," *arXiv preprint arXiv:2008.13537*, 2020.

[16] X. Wu, F. Pan, T. Nguyen, Y. Feng, C. Liu, C.-D. Nguyen, and A. T. Luu, "On the affinity, rationality, and diversity of hierarchical topic modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 19261–19269, 2024.

[17] Q. D. Nguyen, T. Nguyen, D. A. Nguyen, L. N. Van, S. Dinh, and T. H. Nguyen, "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)*, pp. 1109–1124, 2025.

[18] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[19] A. B. Dieng, F. J. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[21] Y. Xu, D. Wang, B. Chen, R. Lu, Z. Duan, M. Zhou, *et al.*, "Hyperminer: Topic taxonomy mining with hyperbolic embedding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31557–31570, 2022.

[22] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[23] Z. Zhang, M. Fang, L. Chen, and M.-R. Namazi-Rad, "Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics," *arXiv preprint arXiv:2204.09874*, 2022.

[24] X. Wu, T. Nguyen, D. Zhang, W. Y. Wang, and A. T. Luu, "Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 84447–84481, 2024.

[25] K. Lang, "Newsweeder: Learning to filter netnews," in *Machine learning proceedings 1995*, pp. 331–339, Elsevier, 1995.

[26] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

[27] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "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*, pp. 91–100, 2008.

[28] J. Yin and J. Wang, "A model-based approach for text clustering with outlier detection," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 625–636, IEEE, 2016.

[29] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[30] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 2006.

[31] T. Nguyen, T. Le, H. T. Vuong, Q. D. Nguyen, D. A. Nguyen, L. N. Van, S. Dinh, and T. H. Nguyen, "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)*, pp. 4507–4524, 2025.

[32] D.-T. Pham, T. T. N. Vu, T. Nguyen, L. Van Ngo, D. A. Nguyen, and T. H. Nguyen, "Neuromax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7758–7772, 2024.

[33] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408, 2015.

[34] C. D. Manning, *Introduction to information retrieval*. Syngress Publishing,, 2008.