Sharpness-Aware Minimization for Topic Models with High-Quality Document Representations

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

Abstract

Recent advanced frameworks in topic models have significantly enhanced the performance compared to conventional probabilistic approaches. Such models, mostly constructed from neural network architecture together with other advanced techniques such as contextual embedding, optimal transport distance and pre-trained language model, etc. have effectively improved the topic quality and document topic distribution. Despite the improvements, these methods lack considerations of effective optimization for complex objective functions that contain log-likelihood and additional regularization terms. In this study, we propose to apply an efficient optimization method to improve the generalization and performance of topic models. Our approach explicitly considers the sharpness of the loss landscape during optimization, which forces the optimizer to choose directions in the parameter space that lead to flatter minima, in which the models are typically more stable and robust to small perturbations in the data. Additionally, we propose an effective strategy to select the flatness region for parameter optimization by leveraging the optimal transport distance between doc-topic distributions and doc-cluster proportions, which can effectively enhance document representation. Experimental results on popular benchmark datasets demonstrate that our method effectively improves the performance of baseline topic models.

1 Introduction

Topic models (TMs) (Hofmann, 1999; Blei et al., 2003; Srivastava and Sutton, 2017; Wu et al., 2024a) are designed to uncover hidden topic structures within a corpus while also providing topic distributions for individual documents. Topic models are utilized across multiple fields in natural language processing (Van Linh et al., 2017; Le et al., 2018; Nguyen et al., 2019; Van Linh et al., 2022; Nguyen et al., 2021, 2022b). In recent years, several advanced topic models (Dieng et al., 2020; Zhao et al., 2021; Wang et al., 2022; Wu et al., 2023b; Pham et al., 2024b; Nguyen et al., 2025a), mostly based on neural networks, have emerged. These models not only enable efficient and flexible parameter inference through automatic gradient back-propagation but also improve the quality of topic-word distributions and document representations. In addition to the application of neural networks in topic modeling, several advanced techniques have been introduced to enhance model performance, such as integrating richer contextual information (Dieng et al., 2020; Bianchi et al., 2021a,b; Han et al., 2023; Pham et al., 2024b), leveraging contrastive learning strategies (Nguyen and Luu, 2021; Han et al., 2023), and applying Optimal Transport methods (Zhao et al., 2021; Wu et al., 2023b, 2024b), among others.

While modern neural topic models have successfully improved both the quality of discovered topics and the distribution of topics across documents, they have largely overlooked the issue of model optimization. Most recent models (Dieng et al., 2020; Wu et al., 2023b; Pham et al., 2024b) are built on the Variational Autoencoder (VAE) framework (Kingma and Welling, 2013), which relies on maximizing the log likelihood and regularizing the document-topic distribution. To further enhance model performance, additional objective constraints are often introduced, such as boosting topic diversity through Embedding Clustering Regularization (Wu et al., 2023b) or improving document representation by maximizing mutual information (Pham et al., 2024b). Although these techniques result in a more complex final objective function, the training process remains relatively straightforward, using standard gradient back-propagation.

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

In this paper, we apply an effective optimization for topic models that not only minimizes the loss function at specific model parameters, but also enhances the model's robustness to variations in those parameters within a local neighborhood, leading to flatter minima. The relationship between the flatness of minima and generalization has been widely studied from both theoretical and empirical perspectives (Keskar et al., 2017; Dziugaite and Roy, 2017; Jiang et al., 2020). Such sharpness-aware minimization techniques have shown promising results in other areas of machine learning, such as image classification, transfer learning, fine-tuning, and language modeling (Foret et al., 2021; Kwon et al., 2021; Sherborne et al., 2024). However, despite the complexity of objective functions in topic models, often involving additional regularization terms (Wu et al., 2023b; Pham et al., 2024b; Wu et al., 2024b), there has been little focus on improving optimization methods to boost performance in this domain. Our proposed method, Sharpness-Aware Minimization for Topic Modeling, can be seamlessly integrated into a variety of topic models, significantly improving both generalization and performance in terms of topic quality and document representations.

Specifically, we introduce a novel local neighborhood region for sharpness-aware minimization that effectively enhances the inference networks of topic models. Our method leverages the high clustering accuracy of pre-trained language models (Reimers and Gurevych, 2019; BehnamGhader et al., 2024). In detail, we use the Optimal Transport (OT) distance (Peyré and Cuturi, 2018) between the document-topic distributions from the topic model and the document-cluster proportions from pre-trained clustering to inform the sharpness-aware neighborhood. This approach introduces an awareness of regions with strong document representations during optimization, leading to flatter minima and improved inference performance. We call this method as **DREAM** which enhances **D**ocument **Re**presentations via Sharpness-**A**ware **M**inimization. We summarize the contributions of our study as follows:

- We propose to improve the performance of recent leading topic models with an effective optimization that simultaneously minimizes loss value and loss sharpness, leading to flatter minima and improved generalization.
- We introduce an innovation optimization method called DREAM. DREAM defines a neighborhood region specifically for sharpness-aware minimization in topic models, utilizing the OT distance between the document-topic distribution from the topic model and the document-cluster proportions from pre-trained clustering.
- We conduct extensive experiments on benchmark datasets, demonstrating that our method can effectively enhance the performance of several topic models.

2 Related Work

Topic modeling aims to uncover hidden topics within a corpus of documents. Traditionally, this problem has been addressed using graphical probabilistic methods (Hofmann, 1999; Blei et al., 2003). Beyond these standard approaches, various studies have extended topic modeling to specialized contexts, such as short texts (Tuan et al., 2020; Ha et al., 2019; Nguyen et al., 2022a; Mai et al., 2016) and streaming environments (Duc et al., 2017; Van Linh et al., 2022; Bach et al., 2021; Nguyen et al., 2022b, 2025b). More recently, the focus has shifted toward neural network-based models, which have demonstrated superior generalization and higher performance (Wu et al., 2024a; Srivastava and Sutton, 2017; Dieng et al., 2020; Wu et al., 2023b; Pham et al., 2024b).

Most neural topic models are built upon the VAE architecture (Kingma and Welling, 2013). In this framework, the encoder (inference) network generates the document's topic distribution, while the decoder (generative) network combines these topic proportions with the topic-word distribution to reconstruct the original data. Some approaches have focused on enhancing the inference network by incorporating document embeddings from pre-trained language models (PLMs) (Devlin et al., 2019; Brown et al., 2020) as input (Wu et al., 2023a; Han et al., 2023) or by imposing additional constraints using PLM representations in the objective function (Pham et al., 2024b). On the other hand, improvements to the generative process have been made through the use of word embeddings (Dieng et al., 2020; Wu et al., 2023b; Pham et al., 2024b), applying conditional transport (Wang et al., 2022), and leveraging optimal transport distance (Wu et al., 2023b; Pham et al., 2024b).

Another approach in neural topic modeling involves generating topics by clustering document representations directly (Grootendorst, 2022; Sia et al., 2020; Zhang et al., 2022). This method is efficient and yields coherent topics, but determining the topic proportions within a document is not straightforward. Additionally, recent research leverages large language models to generate topics as conceptual descriptions (Wang et al., 2023; Pham et al., 2024a), though these methods struggle to provide word distributions within topics or topic proportions within documents. Specifically, Wu et al. (2024b) propose a novel topic modeling approach based solely on Optimal Transport (Peyré and Cuturi, 2018), capturing the semantic relationships among documents, topics, and word embeddings.

3 Preliminaries

3.1 Topic Models

Let $\mathbf{X} = \{\mathbf{x}_d\}_{d=1}^{D}$ represent Bag-of-Words (BoW) vectors for D documents with a vocabulary of V words. Topic models aim to discover K hidden topics, where each topic k has a topic-word distribution $\beta_k \in \mathbb{R}^{V\times 1}$, forming the matrix $\beta \in \mathbb{R}^{V\times K}$ = $(\beta_1, \ldots, \beta_K)$. Given word embedding dimension L, we have the word embedding matrix $\mathbf{W} \in \mathbb{R}^{V\times L}$ with $\mathbf{w}_v \in \mathbb{R}^{L}$ represents the embedding for word v, and topic embedding matrix $\mathbf{T} \in \mathbb{R}^{K\times L}$ with $\mathbf{t}_k \in \mathbb{R}^{L}$ represents the embedding for topic k. Topic models also infer topic proportions $\theta_d \in \mathbb{R}^{K}$ for each document d.

Almost modern topic models represent β as a combination of topic and word embeddings. Typically, the matrix β is factorized into the product of word embeddings $\mathbf{W}$ and topic embeddings $\mathbf{T}$ (Dieng et al., 2020; Xu et al., 2022). However, more advanced models (Wu et al., 2023b; Pham et al., 2024b; Wu et al., 2024b) express β as:

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

where τ is a temperature hyperparameter. The word embeddings $\mathbf{W}$ are often initialized with pre-trained embeddings like GloVe (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013).

In VAE-based topic models, document topic proportions θ are inferred via an inference neural network. Specifically, the Bag-of-Words (BoW) representation of a document x_d is passed through the network to compute the parameters of a Gaussian distribution, where the mean is $\mu = h_\mu(x_d, \gamma)$ and the diagonal covariance matrix is $\Sigma = \mathrm{diag}(h_\Sigma(x_d, \gamma))$, with γ be the parameter of inference network. Using the reparameterization trick (Kingma and Welling, 2013), a latent variable z is sampled from the posterior $q(z|x_d) = \mathcal{N}(z|\mu, \Sigma)$, with a prior $p(z) = \mathcal{N}(z|\mu_0, \Sigma_0)$. The topic proportions θ are then achieved from z by using the softmax function, i.e., $\theta = \mathrm{softmax}(z)$. Topic models reconstruct the BoW representation from β and θ as: $\hat{\mathbf{x}}_{\mathrm{BoW}} \sim \mathrm{Multi}(\mathrm{softmax}(\beta\theta))$. The loss function for the model consists of a reconstruction loss and a regularization term as follows:

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

Recent advanced topic models often incorporate additional terms into their overall objective. For instance, ECRTM (Wu et al., 2023b) introduces Embedding Clustering Regularization to address the issue of topic collapse. NeuroMax (Pham et al., 2024b) employs Mutual Information Maximization with a pre-trained language model and Optimal Transport (OT) between topics to improve document representations. Conversely, FASTopic (Wu et al., 2024b) relies only on OT distance to model topics.

3.2 Sharpness-Aware Minimization

Let the overall loss function be $\mathcal{L}$, and the data batch be B. Sharpness-Aware Minimization (SAM) is a powerful technique designed to improve generalization by minimizing the worst-case loss within a neighborhood around the model parameters, guiding the training toward flatter minima (Foret et al., 2021). The SAM objective is expressed as:

$$\min_{w}\max_{\|\epsilon\|_2\le\rho}\mathcal{L}(w+\epsilon) \tag{2}$$

The perturbation ϵ is constrained within an ℓ_2 Euclidean ball of radius ρ. In the optimization algorithm, the minimax problem is solved by iteratively applying the following two-step procedure for $t = 0, 1, 2, \ldots$ as:

$$\begin{cases}\epsilon_t = \rho\dfrac{\nabla\mathcal{L}_B(\mathbf{w}_t)}{\|\nabla\mathcal{L}_B(\mathbf{w}_t)\|_2}\\ \mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t\left(\nabla\mathcal{L}_B(\mathbf{w}_t+\epsilon_t)\right)\end{cases} \tag{3}$$

where ∇L_B is the minibatch gradient, α_t is an appropriately learning rate.

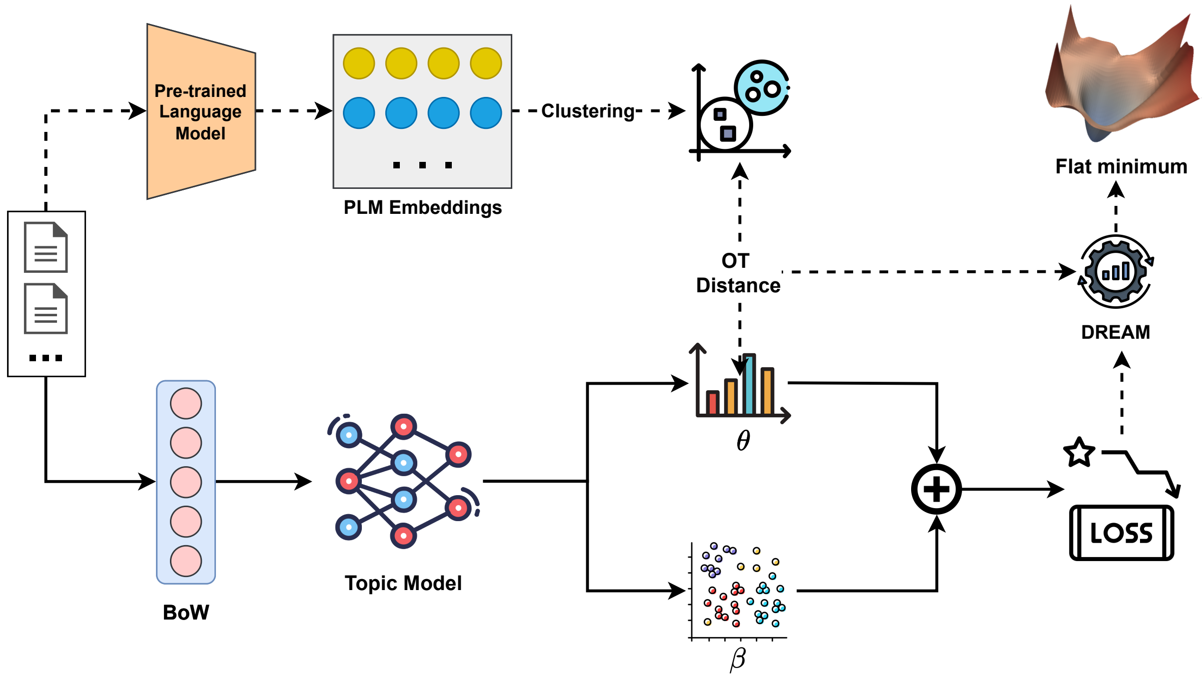

Figure 1: The overall workflow of DREAM when applied to standard topic models. The document dataset is also processed through a PLM-based clustering model to determine document cluster proportions. The OT distance between these proportions and the document-topic distribution is then used as the neighborhood radius in the DREAM optimizer.

4 Methodology

4.1 Sharpness-Aware Minimization for Topic Models

Recall the objective function of topic models as:

$$\mathcal{L}^{\mathrm{TM}} = \frac{1}{D}\sum_{i=1}^{D}\Big[-\left(\mathbf{x}_{i\mathrm{BoW}}\right)^{\top}\log(\mathrm{softmax}(\beta\theta_i)) + \mathrm{KL}(q(z|\mathbf{x}_i)\|p(z))\Big]. \tag{4}$$

While some approaches introduce additional regularizers, most topic models typically involve the following parameters: inference network parameters γ, topic embeddings $\mathbf{T}$, and word embeddings $\mathbf{W}$. Our goal is to find flat minima optimizers for these parameters.

Considering inference network parameters γ, we want to minimize the worst-case loss within a neighborhood of ρ radius as Equation 2 through the updates in 3. However, as discussed in a previous study, Friendly-SAM or F-SAM (Li et al., 2024), the minibatch gradient ∇L_B^{TM} can be decomposed into two components: the full gradient component and the remaining batch-specific stochastic gradient noise; and removing the full gradient component can lead to improved performance. Therefore, we propose to apply F-SAM, to update the inference network parameters γ as follows:

1. Denote $\mathbf{m}_t = \lambda\mathbf{m}_{t-1} + (1-\lambda)\nabla\mathcal{L}_B^{\mathrm{TM}}(\gamma_t)$. $\mathbf{m}_t$ is proven to be a good approximation of the full gradient (Li et al., 2024).

2. Iteratively apply the following two-step procedure:

$$\begin{cases} \epsilon_t = \rho\dfrac{\mathbf{d}_t}{\|\mathbf{d}_t\|} \text{ where } \mathbf{d}_t = \nabla\mathcal{L}_B^{\mathrm{TM}}(\gamma_t) - \sigma\mathbf{m}_t \\ \gamma_{t+1} = \gamma_t - \alpha_t\left(\nabla\mathcal{L}_B^{\mathrm{TM}}(\gamma_t+\epsilon_t)\right) \end{cases} \tag{5}$$

Where t is the iteration step, m is an approximation of the full gradient component of ∇L_B^{TM}, α is the learning rate, λ and σ are the hyperparameters. By removing the full gradient component (which is approximated by $\mathbf{m}_t$) from the minibatch gradient ∇L_B^{TM}, the optimizer can effectively improve generalization (Li et al., 2024).

Similarly, we apply the same procedure for updating $\mathbf{W}$ and $\mathbf{T}$, iterating through each optimization step to produce flat minima optimizers across all model parameters. During the early stages of training, the model tends to underfit, and its parameters are still far from reaching convergence. At this point, focusing on minimizing empirical loss is more critical than trying to find a locally flat region in the loss landscape (Jiang et al., 2020). So we only apply F-SAM after a number of I epochs. The full algorithm is described in the Appendix A.

4.2 Sharpness-Aware Minimization with high-quality clustering region

The neighborhood radius ρ in SAM can be adjusted to define a region that aligns with the specific objectives of the problem. In the study by Sherborne et al. (Sherborne et al., 2024), ρ is modified to represent a trust region, which helps keep the function output "close" to the previous distribution, thereby reducing catastrophic forgetting of pre-trained structures and enhancing fine-tuning. In this paper, we explore a novel specific region for topic models where the output of the inference network - the document topic distribution - achieves optimal performance.

Relying on the high representation of documents resulting from large language model embedding, our novel method, DREAM, constraints that the produced doc-topic distributions from topic models can achieve high clustering accuracy of this representation. We utilize Optimal Transport distance (Peyré and Cuturi, 2018) between the document-topic distribution from the topic model and the document-cluster proportions from pre-trained clustering to inform the SAM neighborhood. The OT distance is selected for its strong effectiveness in comparing distributions with different support sets, such as the doc-topic distributions and doc-cluster proportions in this case.

4.2.1 OT distance between doc-topic distribution and doc-cluster proportion

Let $\mathbf{X}_{PLM} \in \mathbb{R}^{D\times M}$ represent the pre-trained large language model embeddings for the datasets, where M denotes the size of the document embeddings. We apply a clustering method to partition the D documents into G clusters. The set of cluster centers is denoted as $(E_1, E_2, ..., E_G)$ with each $E_i \in \mathbb{R}^M$. We then construct a matrix $P \in \mathbb{R}^{D\times G}$ that demonstrates the cluster proportions of documents that:

$$P_{di} = \frac{p_{di}}{\sum_{g=1}^{G} p_{dg}} \tag{6}$$

where p_{di} is the distance between document d and the center of cluster i. We define two discrete measures for each topic distribution, θ_d, and cluster proportion, P_d, as follows: $\zeta = \sum_{k=1}^{K} \theta_{dk}\delta_{\mathbf{t}_k}$ and $\eta = \sum_{g=1}^{G} P_{dg}\delta_{E_g}$, where δ_x is the Dirac unit mass on x. The transportation cost between topic k and cluster center g is given by: $C_{TE} = \|\phi(\mathbf{t}_k) - E_g\|^2$, where ϕ is a learnable linear mapping from the topic embedding space to the cluster embedding space, parameterized by the weight matrix $W_\phi \in \mathbb{R}^{L\times M}$. For each document d, the optimal transport plan π^{d*} is the solution to the following optimization problem:

$$\begin{aligned} &\text{minimize } \langle C_{\mathrm{TE}}, \pi\rangle - \nu H(\pi) \\ &\text{s.t. } \pi \in \mathbb{R}^{K\times G} \\ &\pi \mathbb{1}_G = \theta_d, \pi^T \mathbb{1}_K = P_d \end{aligned} \tag{7}$$

with $\langle X, Y\rangle = \sum_{i,j} X_{ij}Y_{ij}$ for X, Y are the matrices of the same size; $H(\pi) = -\langle \pi, \log \pi - 1\rangle = -\sum_{i,j} \pi_{ij}(\log \pi_{ij} - 1)$ is the Shannon entropy of π (Cuturi, 2013); $\mathbb{1}_N$ is a vector of size N with all elements equal to 1. Subsequently, the Sinkhorn algorithm is employed to solve the optimization problem (Cuturi, 2013). For each d, the OT distance between topic distribution θ_d and cluster proportion P_d :

$$\mathrm{OT}_d = \sum_{k=1}^{K}\sum_{g=1}^{G} \|\phi(\mathbf{t}_k) - E_g\|^2 \pi_{kg}^{d*} \tag{8}$$

Finally, we have $\mathcal{L}_{\mathrm{OT}}$ be the average distance between doc-topic distributions and doc-cluster proportions over the whole dataset D as: $\mathcal{L}_{\mathrm{OT}} = \frac{1}{D}\sum_{d=1}^{D} \mathrm{OT}_d$

4.2.2 Sharpness-Aware Minimization with OT distance radius

DREAM leverages the OT distance to define the neighborhood of parameters in Sharpness-Aware Minimization. Specifically, the method replaces the neighborhood radius ρ in procedure 5 with the OT distance $\mathcal{L}_{\mathrm{OT}}$, which highlights regions of high document representation. We then constrain the maximization domain for ascent (i.e., $\gamma \to \gamma + \epsilon$) to parameters associated with these high-representation regions, i.e., $\max_{\|\epsilon\|_2 \le \mathcal{L}_{\mathrm{OT}}}$, as substituted in Equation 2. This ensures that the perturbation of γ occurs only within the parameter neighborhood relevant to high-quality doc-topic distribution. By doing so, DREAM incorporates high-clustering awareness alongside the sharpness-awareness objective for finding flatter minima. In contrast, the maximization region ρ in standard SAM is not sensitive to high-quality doc-topic distribution.

Additionally, SAM has the drawback of being sensitive to parameter scale. A practical solution to this issue is normalizing the perturbations based on the parameter scale, as introduced in ASAM (Adaptive Sharpness-Aware Minimization) (Kwon et al., 2021). For the inference network parameters, γ, the overall optimization process is as follows:

1. $\mathbf{m}_t = \lambda \mathbf{m}_{t-1} + (1-\lambda)\nabla \mathcal{L}_B^{\mathrm{TM}}(\gamma_t)$.

2. Iteratively apply the following two-step procedure:

$$\begin{cases} \epsilon_t = \mathcal{L}_{\mathrm{OT}} \dfrac{\gamma^2(\nabla \mathcal{L}_B^{\mathrm{TM}}(\gamma_t) - \sigma \mathbf{m}_t)}{\|\gamma(\nabla \mathcal{L}_B^{\mathrm{TM}}(\gamma_t) - \sigma \mathbf{m}_t)\|} \\ \gamma_{t+1} = \gamma_t - \alpha_t \left(\nabla \mathcal{L}_B^{\mathrm{TM}}(\gamma_t + \epsilon_t)\right) \end{cases} \tag{9}$$

By normalizing the perturbations relative to the scale of the parameters as described in procedure 9, DREAM ensures that all parameters, regardless of their scale, are perturbed in a balanced way. This adaptive approach allows the optimizer to focus on reducing sharpness in a more uniform manner across the network. Similarly, we apply the same update procedure for both $\mathbf{W}$ and $\mathbf{T}$. Additionally, like F-SAM, the proposed optimizer is employed only after a certain number of epochs, but with the following modifications:

1. In the early epochs, we optimize the combined loss, $\mathcal{L} = \mathcal{L}^{\mathrm{TM}} + \lambda_{\mathrm{OT}}\mathcal{L}_{\mathrm{OT}}$, using standard gradient descent and λ_{OT} is weight hyperparameter.

2. In the remaining epochs, we optimize $\mathcal{L}^{\mathrm{TM}}$ using the procedure outlined in 9.

The complete workflow of DREAM, as applied to standard topic models, is illustrated in Figure 1. The algorithm can be found in Appendix A.

4.3 Clustering Algorithm

To leverage the power of pre-trained language models for sharpness-aware minimization, our DREAM framework incorporates a clustering algorithm to determine document cluster proportions. The overall workflow for generating these cluster proportions is: first, the input document corpus is processed through a pre-trained language model (PLM) to obtain contextualized document embeddings. These embeddings, which capture rich semantic information, serve as the foundation for our clustering approach. Given the high dimensionality of these embeddings, we employ Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) for dimensionality reduction. UMAP is chosen for its ability to preserve the global and local structure of the high-dimensional data in a lower-dimensional space, effectively facilitating subsequent clustering. Subsequent to dimensionality reduction via UMAP, a clustering method (e.g., HDBSCAN (Campello et al., 2013), HAC (Murtagh and Contreras, 2012), KMeans (MacQueen, 1967) etc.) is applied to partition documents into groups based on their low-dimensional representations. As detailed in Section 4.2.1, these document-cluster proportions are then leveraged to compute the Optimal Transport (OT) distance. This OT distance, in turn, plays a pivotal role in defining the sharpness-aware neighborhood that guides the DREAM optimization process.

5 Experiments

5.1 Settings

Datasets. Our analysis employs some well-known datasets, including three standard datasets: **20 News Groups (20NG)** (Lang, 1995), a benchmark for topic modeling, **AGNews** (Zhang et al., 2015), which includes news articles from over 2,000 sources and **YahooAnswers** (Zhang et al., 2015), which contains questions and answers from the Yahoo! Answers platform. Additionally, we conduct experiments in two informal, short and noisy datasets: **SearchSnippets** (Phan et al., 2008) consisting of over 12,000 web search results divided into 8 different domains and **GoogleNews** (Yin and Wang, 2016), featuring titles from over 10,000 news articles organized into 152 clusters. The pre-processing steps and statistics of all datasets are described in Appendix B.2

Evaluation Metrics. We adopt the evaluation methodology outlined in (Wu et al., 2023b) to measure both topic quality and document-topic distributions. Topic quality is assessed through topic coherence and diversity metrics. For coherence, we utilize Cv15, 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[1] as an external reference. To

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

$K = 50$	20NG				YahooAnswers				AGNews			
	Cv	TD	Purity	NMI	Cv	TD	Purity	NMI	Cv	TD	Purity	NMI
ETM ‡	0.375	0.704	0.347	0.319	0.354	0.719	0.405	0.192	0.364	0.819	0.679	0.224
+ DREAM	0.376	0.699	0.404	0.379	0.365	0.661	0.507	0.260	0.377	0.692	0.766	0.303
ECRTM ‡	0.431	0.964	0.560	0.524	0.405	0.985	0.550	0.295	0.466	0.961	0.802	0.367
+ DREAM	0.442	0.855	0.574	0.539	0.412	0.872	0.573	0.319	0.464	0.831	0.831	0.374
NeuroMax ‡	0.435	0.912	0.623	0.570	0.404	0.979	0.588	0.331	0.385	0.952	0.804	0.410
+ DREAM	0.446	0.857	0.638	0.578	0.406	0.963	0.596	0.341	0.386	0.942	0.822	0.414
FASTopic ‡	0.427	0.980	0.583	0.528	0.390	0.878	0.589	0.353	0.379	0.960	0.831	0.352
+ DREAM	0.430	0.903	0.630	0.550	0.391	0.900	0.641	0.391	0.388	0.923	0.854	0.382

Table 1: Evaluation results on standard datasets, measured using Cv, TD, Purity, and NMI with $K = 50$. The green data indicates the DREAM-enhanced model performs better than its baseline counterpart, while the red data shows the opposite. ‡ Results resported in (Pham et al., 2024b).

$K = 100$	20NG				YahooAnswers				AGNews			
	Cv	TD	Purity	NMI	Cv	TD	Purity	NMI	Cv	TD	Purity	NMI
ETM ‡	0.369	0.573	0.394	0.339	0.353	0.624	0.428	0.208	0.371	0.773	0.674	0.204
+ DREAM	0.371	0.526	0.452	0.388	0.353	0.634	0.487	0.253	0.376	0.733	0.738	0.252
ECRTM ‡	0.405	0.904	0.555	0.494	0.389	0.903	0.563	0.311	0.416	0.981	0.812	0.428
+ DREAM	0.413	0.756	0.572	0.521	0.390	0.920	0.564	0.321	0.405	1.000	0.820	0.468
NeuroMax ‡	0.412	0.913	0.602	0.516	0.390	0.922	0.583	0.329	0.406	0.957	0.828	0.389
+ DREAM	0.415	0.781	0.633	0.554	0.393	0.769	0.595	0.337	0.409	0.973	0.833	0.412
FASTopic ‡	0.400	0.861	0.622	0.522	0.381	0.766	0.611	0.351	0.385	0.912	0.833	0.330
+ DREAM	0.404	0.800	0.643	0.547	0.385	0.739	0.642	0.385	0.387	0.817	0.852	0.353

Table 2: Evaluation results on standard datasets, measured using Cv, TD, Purity, and NMI with $K = 100$. The green data indicates the DREAM-enhanced model performs better than its baseline counterpart, while the red data shows the opposite. ‡ Results resported in (Pham et al., 2024b).

evaluate topic diversity, we calculate the ratio of unique words among the topic words, referred to as TD15. 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., 2021; Wang et al., 2022), where the most significant topic of each document determines its clustering assignment. While Cv15, Purity, and NMI reflect generalization with external and test data, TD is used to ensure that topics do not overlap too much.

Baseline models. We evaluate our novel optimizer by applying it to several advanced topic modeling frameworks. These include ETM (Dieng et al., 2020), a neural topic model that integrates word embeddings; ECRTM (Wu et al., 2023b), which enhances topic coherence and diversity through clustering regularization in the word embedding space; FASTopic (Wu et al., 2024b), which formulates the semantic relationships among documents, words, and topics as an Optimal Transport problem; and NeuroMax (Pham et al., 2024b) which regularizes doc-topic distributions with pretrained language model embeddings via maximizing mutual information.

5.2 Results in standard datasets

Tables 1 and 2 highlight the effectiveness of DREAM when applied to standard topic model baselines. Overall, DREAM consistently improves topic model performance. Notably, the proposed optimization significantly enhances the quality of document-topic distributions, as reflected in the superior Purity and NMI metrics. This improvement is evident not only in simpler models like ETM but also in cutting-edge models such as NeuroMax and FASTopic. By integrating high clustering-awareness with sharpness-awareness, DREAM effectively guides models to learn more accurate doc-

	$K = 50$								$K = 100$							
	SearchSnippets				**GoogleNews**				**SearchSnippets**				**GoogleNews**			
	Cv	TD	Purity	NMI	Cv	TD	Purity	NMI	Cv	TD	Purity	NMI	Cv	TD	Purity	NMI
ETM	0.397	0.594	0.688	0.389	0.402	0.916	0.366	0.560	0.389	0.448	0.692	0.365	0.398	0.677	0.554	0.713
+ DREAM	0.415	0.888	0.767	0.445	0.410	0.900	0.476	0.684	0.407	0.688	0.809	0.451	0.412	0.696	0.607	0.752
ECRTM	0.450	0.998	0.711	0.419	0.441	0.987	0.396	0.615	0.432	0.966	0.789	0.443	0.418	0.991	0.342	0.491
+ DREAM	0.463	1.000	0.751	0.431	0.450	0.820	0.472	0.650	0.439	0.987	0.820	0.519	0.433	0.879	0.653	0.776
NeuroMax	0.427	0.920	0.743	0.427	0.409	1.000	0.359	0.590	0.439	0.960	0.854	0.472	0.427	0.915	0.664	0.834
+ DREAM	0.426	0.965	0.784	0.463	0.437	0.979	0.376	0.643	0.452	0.969	0.856	0.477	0.434	0.956	0.705	0.859
FASTopic	0.395	0.710	0.792	0.481	0.446	0.440	0.351	0.659	0.386	0.634	0.807	0.458	0.438	0.369	0.458	0.722
+ DREAM	0.396	0.735	0.814	0.502	0.391	0.563	0.359	0.692	0.386	0.686	0.823	0.467	0.426	0.366	0.472	0.739

Table 3: Evaluation results on the two short and noisy datasets, measured by Cv, TD, Purity, and NMI with $K = 50$ and $K = 100$. The green data indicates the DREAM-enhanced model performs better than its baseline counterpart, while the red data shows the opposite.

ument representations, aligned with cluster proportions derived from large language model embeddings.

In addition to improving document-topic distribution, DREAM also enhances topic coherence across most datasets and methods, though this improvement is less pronounced. This is likely because doc-topic distribution is generated from a deeper inference network, while topic-word distribution uses a simpler combination of topic and word embeddings. Sharpness-aware minimization particularly benefits deep networks with rugged loss landscapes. However, DREAM shows lower Topic Diversity (TD) than the original models, despite some gains when the number of topics $K = 100$. The OT distance between doc-topic distributions and doc-cluster proportions may bring topics closer together, but the topic words, presented in the Appendix C, confirm that different topics are still being produced despite the lower TD.

5.3 Results in short and noisy data

To further validate the generalization and performance of the proposed optimization method, we conduct several experiments on two short and noisy text datasets, which are known to present challenges for topic models (Qiang et al., 2022; Lin et al., 2024; Nguyen et al., 2022b). The results are reported in Table 3. In these settings, the improvements achieved by DREAM are particularly noteworthy, especially regarding the quality of doc-topic distributions. The sparse and incomplete nature of the short and noisy text corpus, along with limited co-occurrence patterns, demands models with robust generalization capabilities. Consequently, DREAM demonstrates even greater advantages in this context. Additionally, DREAM shows improved performance in terms of Topic Coherence (Cv) and Topic Diversity (TD) metrics, outperforming all baseline models in TD of SearchSnippets dataset for both $K = 50$ and $K = 100$. These experimental settings underscore the need for effective optimization methods for topic models, particularly when dealing with informal data such as noisy datasets (e.g., search snippets) or very short data (e.g., news article titles).

Dataset	Method	Cv	TD
20NG	Top2Vec	0.441	0.356
	BERTopic	0.382	0.680
	ECRTM + DREAM	<u>0.442</u>	0.855
	NeuroMax + DREAM	**0.446**	<u>0.857</u>
	FASTopic + DREAM	0.430	**0.903**
AGNews	Top2Vec	0.384	0.121
	BERTopic	<u>0.389</u>	0.735
	ECRTM + DREAM	**0.464**	0.831
	NeuroMax + DREAM	0.386	**0.942**
	FASTopic + DREAM	0.388	<u>0.923</u>

Table 4: Performance comparison with clustering-based methods on 20NG and AGNews with $K = 50$. The **bold** values indicate the best performance, and the <u>underlined</u> values indicate the second-best performance for each metric.

5.4 Comparison with Clustering-Based approaches

To further demonstrate the efficacy of our DREAM approach, we conducted a comparative analysis against prominent clustering-based topic modeling

	YahooAnswers ($K = 50$)				YahooAnswers ($K = 100$)			
	Cv15	TD15	Purity	NMI	Cv15	TD15	Purity	NMI
ETM	0.354	**0.719**	0.405	0.192	0.353	0.624	0.428	0.208
+ F-SAM	0.356	0.696	0.473	0.248	**0.354**	0.583	0.472	0.239
+ DREAM	**0.365**	0.661	**0.507**	**0.260**	0.353	**0.634**	**0.487**	**0.253**
FASTopic	0.390	0.878	0.589	0.353	0.381	**0.766**	0.611	0.351
+ F-SAM	**0.392**	0.896	0.638	0.389	0.381	0.750	0.638	0.382
+ DREAM	0.391	**0.900**	**0.641**	**0.391**	**0.385**	0.739	**0.642**	**0.385**

Table 5: Evaluation results for ablation study, measured using Cv15, TD15, Purity, and NMI with $K = 50$ and $K = 100$ for the YahooAnswers dataset, using 2 original models ETM and FASTopic. The **bold** values indicate the best performance, and the underlined values indicate the second-best performance for each metric.

techniques, namely Top2Vec (Angelov, 2020) and BERTopic (Grootendorst, 2022). These models, representing a distinct paradigm in topic modeling, directly derive topics through clustering document embeddings, offering an efficient yet conceptually different approach from VAE-based methodologies. As clustering-based models do not inherently generate document-topic distributions, metrics such as Purity and NMI, which rely on such distributions, are not directly applicable for their evaluation. Consequently, our comparison focuses on topic quality as assessed by coherence (Cv) and diversity (TD).

The results, presented in Table 4, demonstrate that Top2Vec and BERTopic exhibit significantly lower performance compared to state-of-the-art VAE-based methods when enhanced with DREAM. These findings underscore that while clustering-based approaches offer computational efficiency, DREAM, by integrating clustering insights within a sharpness-aware optimization framework for VAE-based topic models, yields a more effective strategy for achieving high-quality and diverse topic representations.

5.5 Ablation study

In this section, we conduct experiments to assess the effectiveness of DREAM in comparison not only to the original models but also to the F-SAM optimizer, with results presented in Table 5. Overall, both F-SAM and DREAM effectively enhance the performance of the original models. Notably, while F-SAM relies solely on the original neighborhood radius hyperparameter, it still achieves improvements in doc-topic distribution quality; however, these enhancements in topic quality are less pronounced, similar to those seen with DREAM. This trend underscores the differing impact of sharpness-aware minimization on deep networks versus shallow networks. Since neither method employs a specific mechanism to control topic quality, their performances in terms of Topic Coherence (Cv) and Topic Diversity are not significantly different.

6 Conclusion

In conclusion, this paper presents a novel approach to enhancing topic model performance through an optimization strategy that minimizes both loss value and sharpness. Specifically, our proposed optimization, namely DREAM, conducts sharpness-aware minimization with a constraint with high-quality document representations. Extensive experiments on benchmark datasets show significant improvements of DREAM in topic quality and document-topic distribution across various topic models.

Limitations

While our proposed method has shown promising results, some limitations should be addressed in the future. Firstly, the effectiveness of the optimization process depends heavily on the quality of the pre-trained clustering, raising the question: how can we optimize clustering quality simultaneously with the topic model? This remains an open challenge for future investigation. Additionally, DREAM's reliance on pre-trained clustering currently limits its application to continuous environments. Further research is needed to explore how sharp-aware minimization can be effectively adapted for dynamic, streaming, and online topic models.

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 is funded by Hanoi University of Science and Technology (HUST) under project number T2024-TN-002.

References

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

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

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

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

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

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

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

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

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

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

Gintare Karolina Dziugaite and Daniel M. Roy. 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

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

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

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

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

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2020. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations*.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2017. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations*.

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

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

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. 2021. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*.

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

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*, 95:62–76.

Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. 2024. Friendly sharpness-aware minimization. In *Conference on Computer Vision and Pattern Recognition*.

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.

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

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

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

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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

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

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

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

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

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. In *Advances in Neural Information Processing Systems*, 34, pages 11974–11986. Curran Associates, Inc.

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

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

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.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2024a. Topicgpt: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984.

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

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

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

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

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.

Tom Sherborne, Naomi Saphra, Pradeep Dasigi, and Hao Peng. 2024. TRAM: Bridging trust regions and sharpness aware minimization. In *The Twelfth International Conference on Learning Representations*.

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

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

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.

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.

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

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

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

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

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

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

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*, 35, pages 31557–31570. Curran Associates, Inc.

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

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

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

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

Algorithm 1 Learning F-SAM topic model

Input: Document collection $\mathbf{X}$, pretrained word embedding $\mathbf{W}_{\text{pretrained}}$, number of topic K, total number of training epoch N, number of training epochs for the first stage I;
Output: Encoder network's parameter γ, word embedding $\mathbf{W}$, topic embedding $\mathbf{T}$;

```
Initialize W = W_pretrained
for t = 1, 2, ..., N do
    for each minibatch B do
        if t ≤ I then
            // Stage 1
            Estimate L_B^TM.
            Update W, T through regular gradient step.
            Update γ through regular gradient step.
        else
            Estimate L_B^TM
            Update W, T through F-SAM procedure (5).
            Update γ through F-SAM procedure (5).
        end if
    end for
end for
```

Algorithm 2 Learning DREAM topic model

Input: Document collection $\mathbf{X}$, pretrained word embedding $\mathbf{W}_{\text{pretrained}}$, number of topic K, the document cluster distribution matrix P, total number of training epoch N, number of training epochs for the first stage J;
Output: Linear projection weight W_ϕ, encoder network's parameter γ, word embedding $\mathbf{W}$, topic embedding $\mathbf{T}$;

```
Initialize W = W_pretrained
for t = 1, 2, ..., N do
    for each minibatch B do
        Update the average OT distance L_OT
        if t ≤ J then
            // Stage 1
            Estimate L = L^TM + λ_OT L_OT.
            Update W_φ through regular gradient step.
            Calculate π* using Sinkhorn algorithm.
            Update W, T through regular gradient step.
            Update γ through regular gradient step.
        else
            Estimate L_B^TM
            Calculate π* using Sinkhorn algorithm.
            Update W, T through DREAM procedure (9).
            Update γ through DREAM procedure (9).
        end if
    end for
end for
```

A Algorithm

The detailed training algorithms for the F-SAM topic model and the DREAM topic model are provided in Algorithms 1 and 2 respectively. It is important to note that the settings and parameters are generally applicable to most topic models; methods that introduce new parameters can be adapted similarly.

B Experiment Details

B.1 Implementation Details.

All experiments are conducted on a system equipped with a GeForce RTX 3090 GPU (24GB RAM), utilizing PyTorch 2.4.0+cu121 in a Python 3.12.3 environment. The model is trained for 200 epochs with a batch size of 200, employing the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.002. The OT weight hyperparameter λ_{OT} is selected from the interval $[0.01, 0.1, 0.5, 1.0, 5.0, 10.0]$, and the first-stage training lasts for 140 epochs. The F-SAM hyperparameters λ and σ are set to 0.9 and 0.005, respectively.

For the four topic modeling frameworks ETM, ECRTM, FASTopic, and NeuroMax, only ETM does not have any specific hyperparameters, while the others are configured as follows:

Dataset	# of texts	average text length	# of labels	vocab size
20NG	18,846	110.5	20	5,000
YahooAnswers	12,500	35.4	10	5,000
AGNews	12,500	20.1	4	5,000
SearchSnippets	12,294	14.4	8	4,618
GoogleNews	11,019	5.8	152	3,473

Table 6: Dataset statistics after preprocessing.

- **ECRTM**: Includes the Embedding Clustering Regularization (ECR) loss with the weight hyperparameter $\lambda^{\text{ECR}} \in [20, 40, 60, 100, 150, 200, 250]$.
- **NeuroMax**: Incorporates three loss functions, with their corresponding weight hyperparameters selected from the following ranges:
  - $\lambda^{\text{ECR}} \in [20, 40, 60, 100, 150, 200, 250]$
  - $\lambda^{\text{GR}} \in [1, 5, 10, 20, 50]$
  - $\lambda^{\text{InfoNCE}} \in [1, 10, 30, 50, 80, 100, 130]$
- **FASTopic**: Utilizes three main hyperparameters: $\epsilon_1 = 1/3$ (entropic regularization for document-topic relations), $\epsilon_2 = 1/2$ (entropic regularization for topic-word relations), and $\tau = 1.0$ (softmax temperature for semantic relations).

B.2 Dataset Statistics

Our experiments utilized some well-known datasets, including three standard datasets: **20 News Groups (20NG)** (Lang, 1995), **AGNews** (Zhang et al., 2015), and **YahooAnswers** (Zhang et al., 2015). Additionally, we conducted experiments on two informal datasets: **SearchSnippets** (Phan et al., 2008), which contains relatively short and noisy data, and **GoogleNews** (Yin and Wang, 2016), a collection of very short article titles.

For the standard datasets, we applied the pre-processing steps described in (Wu et al., 2023b) to generate bag-of-words representations. For the short and noisy text datasets, we utilized pre-processed versions available from the STTM library[2] (Qiang et al., 2022). Subsequently, we refined the datasets by removing words with a frequency of less than 3 and discarding any documents containing fewer than 2 terms. These pre-processing procedures were carried out using the TopMost tool[3]. The detailed statistics of all datasets after processing are presented in Table 6.

B.3 Pre-trained language model for Clustering Algorithm

We employed the stella-en-400M-v5 model [4] as the pre-trained language model for Clustering Algorithm. Clustering was then performed with UMAP, followed by clustering with HDBSCAN. This approach yielded optimal PLM clusters across different datasets: 20 clusters for 20NG, 8 for Yahoo Answers, 3 for AG News, 5 for Search Snippets, and 5 for Google News.

B.4 Scalability and Computational Cost

In this appendix, we clarify the issue of scalability in our approach. Although incorporating Optimal Transport (OT) in DREAM does increase the training time, it does not compromise scalability. Specifically, suppose that we have B documents in each batch of data, we would need to compute OT distance values (each doc has an OT distance between its topic distribution and cluster distribution). However, the OT

[2] https://github.com/qiang2100/STTM
[3] https://github.com/bobxwu/topmost
[4] https://huggingface.co/dunzhang/stella_en_400M_v5

Method	Dataset	Baseline	Baseline + OT	F-SAM	DREAM
ECRTM	20NG	1.332	1.361	2.470	2.480
NeuroMax	20NG	2.823	2.732	6.330	6.177
FASTopic	20NG	0.081	1.695	2.937	2.751
ECRTM	YahooAnswers	1.142	1.128	2.165	2.137
NeuroMax	YahooAnswers	1.842	1.794	3.960	3.892
FASTopic	YahooAnswers	0.058	1.386	2.433	2.212

Table 7: Training Time Comparison (seconds)

Model	Dataset	Cv15	TD15	Purity	NMI
FASTopic + DREAM (HDBSCAN)	20NG	0.430	0.903	0.630	0.549
FASTopic + DREAM (HAC)	20NG	0.417	0.925	0.611	0.551
FASTopic + DREAM (HDBSCAN)	YahooAnswers	0.391	0.900	0.641	0.391
FASTopic + DREAM (HAC)	YahooAnswers	0.375	0.929	0.639	0.375
FASTopic + DREAM (HDBSCAN)	AGNews	0.387	0.876	0.864	0.393
FASTopic + DREAM (HAC)	AGNews	0.392	0.863	0.856	0.379
FASTopic + DREAM (HDBSCAN)	GoogleNews	0.391	0.563	0.359	0.692
FASTopic + DREAM (HAC)	GoogleNews	0.446	0.500	0.353	0.703
FASTopic + DREAM (HDBSCAN)	SearchSnippets	0.396	0.735	0.814	0.502
FASTopic + DREAM (HAC)	SearchSnippets	0.402	0.792	0.814	0.482

Table 8: Impact of Clustering Methods on FASTopic + DREAM Performance

distance can be computed in parallel per batch through matrix operations (details of implementation are in the accompanying source code). Therefore, regardless of how large the dataset is, with a fixed batch size, DREAM still ensures scalability. Moreover, since the dimensions of both the transport plan and the cost matrix in DREAM are `num_of_cluster` $\times$ `num_of_topic`, their computational overhead is negligible.

Additionally, Table 7 reports the training times (in seconds) for four configurations: (i) the baseline models, (ii) the baselines with OT (phase 1 of our optimization algorithm), (iii) the baselines with FSAM, and (iv) the baselines with DREAM (phase 2 of our optimization algorithm). We observe that incorporating the OT distance does not slow down the training of models such as ECRTM and NeuroMax. In contrast, the F-SAM and DREAM configurations require approximately twice the training time, which is due to the SAM algorithm performing an additional perturbation coefficient update at each parameter update.

B.5 Impact of Clustering Algorithm

To further investigate the robustness of our DREAM method, we conducted experiments using an alternative clustering algorithm. While HDBSCAN was used in the primary experiments to generate document cluster proportions, we have conducted additional experiments using different clustering methods. We provide results in Table 8 below when using Hierarchical Agglomerative Clustering (HAC) instead of HDBSCAN. The results show negligible differences between these methods, indicating that the choice of clustering algorithm does not significantly impact the final results.

B.6 Pre-trained Clustering Details

In our study, we employ HDBSCAN as the pre-trained clustering algorithm owing to its notable advantages, including the ability to determine clusters without specifying their number in advance, a reduced parameter set, and efficient scalability to large datasets. To fine-tune its performance, we varied the primary parameter, `min_samples`, over the set $\{1, 2, 4\}$, and evaluated the resulting clusters using both Purity and Normalized

Dataset	`min_samples` = 1		`min_samples` = 2		`min_samples` = 4	
	Purity	NMI	Purity	NMI	Purity	NMI
GoogleNews	0.921	0.877	0.923	0.878	0.923	0.876
SearchSnippets	0.894	0.542	0.892	0.544	0.893	0.546
YahooAnswers	0.685	0.398	0.643	0.400	0.625	0.426
AGNews	0.857	0.483	0.861	0.494	0.842	0.528
20NG	0.767	0.581	0.749	0.590	0.713	0.580

Table 9: Clustering performance of HDBSCAN for different values of `min_samples`. We select the optimal configuration for HDBSCAN based on these results.

Mutual Information (NMI) metrics. The parameter configuration that produced the best Purity and NMI scores was selected for further experiments.

Table 9 summarizes the clustering performance across several datasets. The results clearly demonstrate the strong quality of the pre-trained clusters. Moreover, the enhanced performance observed when integrating these clusters within DREAM further confirms the effectiveness of our pre-trained clustering strategy.

C Examples of Topics

ECRTM + DREAM with 20NG (K = 50)

Topic #1 : nsa pgp denning inability chip toyota condemn publish tactics condemned
Topic #2 : turks homeland turkish proceeded greeks greece ethnic empire greek nazi
Topic #3 : entry output xterm window visual byte width guidelines file bytes
Topic #4 : sale shipping manuals cds sony offer email disks items speaker
Topic #5 : drive drives floppy scsi disks disk internal sony backup external
Topic #6 : max cliff vram vga diamond vesa simms simm eisa monitor
Topic #7 : detector detectors clinic livesey van gamma observatory sahak sensitivity amazed
Topic #8 : bos advance tor ext troy cal champs playoff duke grateful
Topic #9 : windows font dos logo fonts icon window beast xterm tiff
Topic #10: pitching hitter defensive innings puck scored score batting players talent
Topic #11: shaped israelis borders israeli brains lebanon beings deeply israel surrounding
Topic #12: tragedy serbs davidian neighbors father armenians troops secretary soviet bed
Topic #13: lebanese israels elias andi beyer jake optilink redundancy bosnians clayton
Topic #14: nhl hockey rangers **<u>devils</u>** winnipeg jets oilers detroit lemieux bruins
Topic #15: cease overwhelming volunteer consent oppose interpreted reactions horizontal applying removal
Topic #16: guns **<u>gun</u>** **<u>handgun</u>** firearms firearm violent weapons deaths tennessee criminals
Topic #17: modem ati linux upgrade desktop scanner upgrading interrupt editing sensor
Topic #18: victoria reserve tourist oxford temple oak columbus lincoln consultant significance
Topic #19: scsi bios controller drives jumper isa jumpers drive floppy disk
Topic #20: chastity shameful intellect skepticism helmet riding biker bikes drinking dod
Topic #21: serdar argic islam genocide tcp ohanus appressian bitmap convenient massacres
Topic #22: lobby circles muslim catholics libertarian moslem biblical courts youth distinction
Topic #23: sale shipping cds offer manuals sony air price speaker disks
Topic #24: captain abc gordon witnesses sexual harris wiretap rape palmer alien
Topic #25: jesus christ resurrection doctrine testament sin salvation pope lord heaven
Topic #26: money idea thing things profit bad talk really lot better
Topic #27: arbor ann port bmw ide telnet jews silicon bbs demo
Topic #28: baseball kids dreams loves miracle ball era hits ages exciting
Topic #29: lib mouse centris icons openwindows usr inet francis philadelphia sunos
Topic #30: linked church valley petaluma mhz bus duck melkonian cells civilians
Topic #31: morality atheists atheism absolute belief arrogance subjective moral evolution morals
Topic #32: spacecraft satellites mars launched lunar payload shuttle orbit launch orbital
Topic #33: foods ranch yeast survivors batf tear bds chronic davidians patients
Topic #34: phones penn regional russians storm newspapers stretch burned bull streets
Topic #35: contrib jpeg pub **<u>privacy</u>** anonymous export motif platforms gif graphics
Topic #36: radar roger stratus andre braves vnews islanders forwarded propulsion rochester
Topic #37: militia firearms firearm **<u>handgun</u>** possession constitution constitutional **<u>gun</u>** assault liberties
Topic #38: circuits wire wires zoology circuit voltage neutral wiring henry spencer
Topic #39: tires brake brakes tire rear valve wheels cars mileage suspension
Topic #40: malcolm sandvik rushdie marriage satan mormons married benedikt teachings rosenau
Topic #41: xxmessage xxdate nuntius useragent lciii ksand alink cookamunga csutexasedu solntzewpdsgicom
Topic #42: sale shipping cds manuals offer sony disks items speaker email
Topic #43: walker iran racist elizabeth athens clipper catholic yugoslavia mary bosnian
Topic #44: advertising billion rocks feds cuts station sought wings ottawa stayed
Topic #45: suck cubs cramer homosexual gregg gay rutgers cell ticket dakota
Topic #46: kent durham apps graphic penguins balls funny slick bang scared
Topic #47: msg superstition food tin driver objective newsreader reagan cnn poll
Topic #48: escrow omissions encryption toal conversations **<u>privacy</u>** trusted tapped voluntary initiative
Topic #49: schneider keith doug beaverton yankees nixon morgan kevin phil gardner
Topic #50: espn gerald **<u>devils</u>** europeans leafs jets helsinki hawks stadium traded

Table 10: Top 10 related words of 50 topics from 20NG. Some repeated words are **bold** and <u>underlined</u>. The topic diversity value of 0.855 in the ECRTM + DREAM model, though lower than the original model's 0.964, remains high enough to maintain a diverse range of topics. While some topic-words overlap - such as "gun" and "handgun" appearing in both Topic 16 and Topic 37 - this does not result in topic collapse. Instead, the two topics retain distinct focuses: one addresses crime, while the other discusses war.

FASTopic + DREAM with AGNews (K = 50)

```
Topic #1 : turkey annan ministers vows turkish ambassador calm chirac kofi constitution
Topic #2 : lives land apparently victim friends alert threatening cause schools believed
Topic #3 : photo size color font gates washingtonpostcom sans verdana serif helvetica
Topic #4 : intel chip ibm dell storage amd memory processor servers dual
Topic #5 : microsoft software internet computer music search online service web google
Topic #6 : hollywood movie satellite virgin film entertainment commercial venture blockbuster ebay
Topic #7 : ceo disney executive eisner owner walt owners marsh chairman sue
Topic #8 : enterprise application unveils feature platform halo infoworld solaris upgrade tools
Topic #9 : stewart martha retirement trump casino stern story fox charles hot
Topic #10: nba pacers guard bryant indiana detroit agent denver basketball spurs
Topic #11: cutting ups outsourcing workforce managers estimated eliminate invest roughly coffee
Topic #12: sex wife appointed refused son doctors stand resigned ruled resigns
Topic #13: project standards breakthrough approach challenges initiative projects allows progress
operate
Topic #14: killed people police bomb least attack afghan killing attacks dead
Topic #15: percent profit sales quarter shares target earnings ticker http href
Topic #16: good hot every longer story looks instead really want seems
Topic #17: economy interest rate rates august jobs mortgage debt economic fannie
Topic #18: want looks need needs become let getting instead turn good
Topic #19: court pay case trial charges judge cut federal union insurance
Topic #20: spam virus piracy theft lawsuits spyware sharing file peer mail
Topic #21: oil prices dollar stocks record crude barrel fuel high investors
Topic #22: king protest indonesia myanmar indonesian prince colombia thai ousted cambodia
Topic #23: president bush election presidential minister prime party john vote leader
Topic #24: first second win test back one day world australia won
Topic #25: million deal billion company buy inc business bid corp firm
Topic #26: ban trade law rules organization bill flu committee proposal climate
Topic #27: new said quot reuters year says wednesday tuesday monday thursday
Topic #28: cup tour golf title championship grand masters prix ryder formula
Topic #29: red sox boston series yankees league baseball houston victory astros
Topic #30: nfl yards quarterback touchdown bowl dolphins passes eagles packers colts
Topic #31: leave blood duty condition reaction insisted successor telling unable demanded
Topic #32: boxing harry button heavyweight retire knows moment great doesn never
Topic #33: mobile phone wireless sony video radio dvd phones cell electronics
Topic #34: never great nothing know doesn success age front seat harry
Topic #35: hostage arrested prison arrest accused hostages jail murder terrorism kidnapped
Topic #36: olympic gold athens medal olympics greece tennis champion greek phelps
Topic #37: talks nuclear afp darfur iran nations korea foreign peace sudan
Topic #38: holiday growth shopping spending survey consumers grow consumer retailers retail
Topic #39: manager pitcher anaheim mariners jays hander bobby lee carl steroids
Topic #40: champions england club madrid manchester arsenal spain chelsea liverpool striker
Topic #41: pensions impact influence momentum improve measure uncertainty grade savings natural
Topic #42: scientists study researchers human science water experts evidence children species
Topic #43: revealed serious status warn questions homeland contact spread remote kingdom
Topic #44: never great know thought young success nothing good turned things
Topic #45: season game team coach players sports play points football games
Topic #46: drug health ford drugs heart plant medical motor steel vioxx
Topic #47: space nasa flight prize station earth flights moon mars crew
Topic #48: hurricane storm ivan victims cuba islands tsunami typhoon frances flood
Topic #49: ohio tom practice virginia frank maryland chris ryan didn georgia
Topic #50: iraq iraqi baghdad troops palestinian israeli gaza army israel arafat
```

Table 11: Top 10 related words of 50 topics from AGNews. Some repeated words are **bold** and underlined. The topic diversity value of 0.923 in the FASTopic + DREAM model, although lower than the original model's 0.960, is still sufficiently high to preserve a broad range of topics. Some word overlap occurs, but these are common and insignificant words like "want" and "nothing", which do not impact the overall meaning of topics.