

S2WTM: Spherical Sliced-Wasserstein Autoencoder for Topic Modeling

Suman Adhya and Debarshi Kumar Sanyal

Indian Association for the Cultivation of Science

adhyasuman30@gmail.com, debarshi.sanyal@iacs.res.in

Abstract

Modeling latent representations in a hyperspherical space has proven effective for capturing directional similarities in high-dimensional text data, benefiting topic modeling. Variational autoencoder-based neural topic models (VAE-NTMs) commonly adopt the von Mises-Fisher prior to encode hyperspherical structure. However, VAE-NTMs often suffer from posterior collapse, where the KL divergence term in the objective function highly diminishes, leading to ineffective latent representations. To mitigate this issue while modeling hyperspherical structure in the latent space, we propose the Spherical Sliced Wasserstein Autoencoder for Topic Modeling (S2WTM). S2WTM employs a prior distribution supported on the unit hypersphere and leverages the Spherical Sliced-Wasserstein distance to align the aggregated posterior distribution with the prior. Experimental results demonstrate that S2WTM outperforms state-of-the-art topic models, generating more coherent and diverse topics while improving performance on downstream tasks.

 github.com/AdhyaSuman/S2WTM

1 Introduction

Topic modeling aims to uncover latent topics in a document corpus by representing each document as a probabilistic mixture of topics. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) typically assume a multinomial distribution over words. However, this assumption treats words as independent, ignoring important syntactic and semantic relationships. Moreover, classical topic models rely on computationally expensive inference techniques, such as Gibbs sampling and variational inference.

To address these limitations, Variational Autoencoder-based Neural Topic Models (VAE-NTMs) were introduced by Miao et al. (2016),

offering a more scalable and flexible alternative. VAE-NTMs typically use a Gaussian prior for the latent space. However, in high-dimensional settings, Gaussian distributions exhibit the “soap bubble effect” (Davidson et al., 2018), where probability mass concentrates near the surface of a hypersphere rather than around the mean. This effect undermines the utility of Euclidean distances, which tend to become uniform and less informative as dimensionality increases—a core aspect of the curse of dimensionality (Aggarwal et al., 2001). In contrast, cosine similarity, which measures the angular difference between vectors, remains a more robust metric in such regimes (Dhillon and Sra, 2003). These observations motivate the use of spherical latent space modeling, where data is embedded on the surface of a hypersphere and similarities are measured via angles rather than distances. This is particularly effective in large-scale applications involving directional data, such as text categorization and gene expression analysis (Dhillon and Sra, 2003; Banerjee et al., 2005), where the orientation of data vectors carries meaningful information. The von Mises-Fisher (vMF) distribution naturally models such data and has demonstrated superior performance over models with Gaussian priors in clustering and topic modeling tasks (Reisinger et al., 2010; Li et al., 2016; Batmanghelich et al., 2016). Recent advancements in topic modeling, such as vONTSS (Xu et al., 2023), have leveraged the \mathcal{S} -VAE framework and integrated supervised signals via optimal transport to enhance topic coherence and classification performance.

Despite these advancements, VAE-based models often suffer from *posterior collapse* (Bowman et al., 2016), wherein the latent representations fail to encode meaningful information. This issue occurs when the KL divergence term in the loss function approaches zero, causing the posterior distribution to converge to the uninformative prior. A detailed analysis of this phenomenon, including its formal

definition and decomposition via ELBO surgery, is provided in Appendix A.

To address these challenges, we propose S2WTM (Spherical Sliced-Wasserstein Autoencoder for Topic Modeling), an autoencoder-based topic model that models the latent space on the unit hypersphere. S2WTM introduces a flexible prior design by treating the choice of prior distribution as a hyperparameter, supporting three alternatives: the vMF distribution, a Mixture of vMF (MvMF) distributions, and the uniform distribution on the hypersphere. To align the aggregated posterior with the chosen prior while preserving spherical geometry, S2WTM employs the Spherical Sliced-Wasserstein (SSW) distance, effectively mitigating the posterior collapse problem.

Contributions: In summary, our work presents the following key contributions:

- We introduce S2WTM, a neural topic model leveraging the Spherical Sliced-Wasserstein autoencoder framework to align the aggregated posterior with a hyperspherical prior, ensuring effective latent space modeling.
- We explore three distinct choices of prior distributions defined on the unit hypersphere for modeling the latent space.
- We evaluate the performance of S2WTM using quantitative metrics, qualitative analysis, task-specific performance, and LLM-based human-comparable evaluations, demonstrating its advantages over state-of-the-art models.
- We empirically validate the benefits of moving beyond the conventional Euclidean latent space, highlighting performance improvements in topic modeling.

2 Preliminaries

This section introduces three selected distributions defined on a unit hypersphere. We then discuss optimal transport metrics, focusing on the Wasserstein distance, sliced Wasserstein distance, and spherical sliced-Wasserstein distance.

2.1 Distributions on the Unit Hypersphere

vMF distribution: The vMF distribution is defined on the unit $(K - 1)$ -sphere \mathcal{S}^{K-1} embedded in \mathbb{R}^K . The vMF distribution is parameterized

by its mean direction ($\boldsymbol{\mu} \in \mathcal{S}^{K-1}$), a unit vector indicating the central direction of the distribution, and its concentration ($\kappa \in \mathbb{R}^+ \cup \{0\}$), a non-negative scalar controlling the dispersion around $\boldsymbol{\mu}$. A larger κ indicates a higher concentration. For $\mathbf{x} \in \mathcal{S}^{K-1}$ with $\kappa \geq 0$ the vMF density function is: $\text{vMF}(\mathbf{x}; \boldsymbol{\mu}, \kappa) = c_p(\kappa) \exp(\kappa \boldsymbol{\mu}^\top \mathbf{x})$ where $c_K(\kappa)$ is the normalization constant: $c_K(\kappa) = \frac{\kappa^{\frac{K}{2}-1}}{(2\pi)^{\frac{K}{2}} I_{\frac{K}{2}-1}(\kappa)}$ with I_v denotes the modified Bessel function of the first kind of order v . The detailed sampling from the vMF distribution (Ulrich, 1984) is described in Appendix B.

MvMF distribution: The MvMF distribution is a convex combination of vMF distributions. If the mixture contains T vMF components, then the density is defined for all points $\mathbf{x} \in \mathcal{S}^{K-1}$ as $\text{MvMF}(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{t=1}^T \alpha_t f_t(\mathbf{x}|\boldsymbol{\psi}_t)$. Here, each component f_t is a vMF distribution with parameters $\boldsymbol{\psi}_t = (\boldsymbol{\mu}_t, \kappa_t)$, and $\boldsymbol{\Psi}$ represents the set of all such parameters. The mixture proportions $(\alpha_t)_{1 \leq t \leq T}$ satisfy $\alpha_t \geq 0$ and $\sum_{t=1}^T \alpha_t = 1$. The sampling technique from MvMF requires first selecting a component $i \sim \text{Cat}(\alpha_1, \alpha_2, \dots, \alpha_T)$, then generating a sample from the chosen vMF distribution.

Uniform distribution on the unit hypersphere:

The uniform distribution on \mathcal{S}^{K-1} assigns equal probability density to all directions. Sampling is performed by drawing from a standard Gaussian in \mathbb{R}^K and applying ℓ_2 normalization.

2.2 Spherical Sliced-Wasserstein Distance:

The Spherical Sliced-Wasserstein (SSW) distance is a variant of the Sliced-Wasserstein (SW) distance, a computationally efficient approximation of the Wasserstein distance. SSW adapts SW for distributions on the unit hypersphere or other spherical domains. We first review the Wasserstein and SW distances.

Definition 1 (Wasserstein Distance (Villani et al., 2009; Figalli and Villani, 2011)). Let $p \geq 1$. The p -Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} d(x, y)^p d\gamma(x, y) \quad (1)$$

where $\mathcal{P}_p(\mathbb{R}^d)$ is the set of Borel probability measures on \mathbb{R}^d with finite p -moments, and $\Gamma(\mu, \nu)$ are joint measures with marginals μ and ν .

Computing W_p is computationally expensive, scaling as $\mathcal{O}(n^3 \log n)$ for discrete measures with n samples (Peyré et al., 2019). However, W_p has a closed form for univariate distributions:

$$W_p^p(\mu, \nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p dt, \quad (2)$$

where F_μ^{-1} and F_ν^{-1} are the quantile functions. This leads to the SW, exploiting the efficiency of one-dimensional optimal transport.

Definition 2 (Sliced Wasserstein Distance (Julien et al., 2011; Bonneel et al., 2015)). For $p \geq 1$, the p -SW between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is

$$SW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p((\mathcal{R}\mu)^\theta, (\mathcal{R}\nu)^\theta) d\theta \quad (3)$$

where $(\mathcal{R}\eta)^\theta$ is the Radon transform, projecting η onto the line spanned by $\theta \in S^{d-1}$.

SW reduces computation to multiple one-dimensional problems. However, it requires many projections, especially in high dimensions. Critically, SW is not well-suited for spherical data. The SSW distance addresses this by incorporating spherical geometry.

Definition 3 (Spherical Sliced Wasserstein Distance (Bonet et al., 2023)). For $p \geq 1$, the p -SSW between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is

$$SSW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p((\tilde{\mathcal{R}}\mu)^\theta, (\tilde{\mathcal{R}}\nu)^\theta) d\theta \quad (4)$$

where $(\tilde{\mathcal{R}}\eta)^\theta$ is the spherical Radon transform, mapping η to a substructure of the sphere (e.g., a great circle).

3 Related Work

Reisinger et al. (2010) pioneered the idea of spherical latent space modeling in topic modeling by introducing the Spherical Admixture Model (SAM). SAM replaces the multinomial likelihood in LDA with the vMF distribution, representing documents as unit vectors on the hypersphere. This approach captures directional semantic relationships more effectively than traditional Euclidean representations and lays the groundwork for subsequent research in spherical topic models.

Li et al. (2016) advanced this line of work by proposing the mix-vMF topic model (MvTM) that integrates word embeddings using MvMF distributions, enabling the representation of complex topic structures and improving topic coherence.

Batmanghelich et al. (2016) further extended spherical topic modeling by introducing a nonparametric model that incorporates word embeddings via the vMF distribution and employs a Hierarchical Dirichlet Process (HDP) to automatically infer the number of topics.

More recently, Ennajari et al. (2022) proposed the Embedded Spherical Topic Model (ESTM) for supervised learning, which integrates semantic knowledge from word and knowledge graph embeddings within a hyperspherical latent space. ESTM improves both topic interpretability and predictive performance by modeling documents and topics on the sphere and leveraging variational inference techniques tailored to the vMF distribution.

In the neural setting, Xu et al. (2023) introduced vONT, an unsupervised topic model based on the \mathcal{S} -VAE framework, and vONTSS, its semi-supervised extension. While vONT models topics using latent variables constrained to the unit hypersphere, vONTSS further aligns latent representations with supervision signals using an optimal transport loss, thereby enhancing topic relevance and classification accuracy. However, like any VAE-based model, its reliance on KL divergence to regulate the posterior distribution to the prior distribution can lead to the posterior collapse problem (Bahuleyan et al., 2019), which may hinder the expressiveness of the latent space if not carefully mitigated.

In contrast, S2WTM addresses these limitations by replacing KL divergence with the Spherical Sliced-Wasserstein (SSW) distance, promoting alignment between the aggregated posterior and hyperspherical priors. This approach preserves spherical geometry and mitigates posterior collapse, distinguishing S2WTM from both classical and neural spherical topic models.

4 Proposed Methodology

The proposed S2WTM model is based on the Wasserstein Autoencoder framework (Tolstikhin et al., 2018), employing an encoder-decoder architecture. The framework is illustrated in Fig. 1, and a detailed description of the methodology is provided below.

4.1 Choice of Prior Distribution

To encourage a hyperspherical structure in the latent space, rather than assuming a standard Euclidean geometry, S2WTM regulates the encoder-

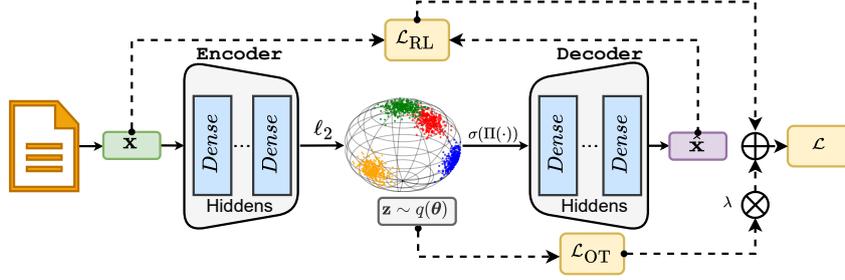


Figure 1: Proposed framework for S2WTM model.

generated posterior distribution to align with a prior distribution p_θ defined over the unit hypersphere \mathcal{S}^{K-1} . We consider three types of priors: (i) the vMF distribution, $p_\theta \sim \text{vMF}(\mu, \kappa)$; (ii) the Mixture of vMF (MvMF) distributions, $p_\theta \sim \text{MvMF}(\Psi)$; and (iii) the uniform distribution on the hypersphere, $p_\theta \sim \mathcal{U}(\mathcal{S}^{K-1})$. These priors offer varying inductive biases, enabling S2WTM to model diverse latent topic structures while adapting to the characteristics of different datasets. The alignment between the aggregated posterior and the chosen prior is enforced using the SSW distance.

4.2 Encoder Network

The encoder processes a document representation, denoted as $\mathbf{x} \in \mathbb{R}^V$ (where V represents the vocabulary size), and maps it to a low-dimensional representation $\mathbf{z} \in \mathcal{S}^{K-1}$, with K being the number of topics to be learned. The transformation sequence is as follows: [Linear(V, H') \rightarrow Dropout(p) \rightarrow ReLU \rightarrow Linear(H', H'') \rightarrow Dropout(p) \rightarrow ReLU \rightarrow Linear(H'', K) \rightarrow L2Norm(\cdot)]. Here, Linear(M, N) maps an input vector of size M to N , Dropout(p) applies dropout with probability p , ReLU is an activation function defined as $\text{ReLU}(x) = \max(0, x)$, and L2Norm normalizes the output onto the unit hypersphere \mathcal{S}^{K-1} .

S2WTM employs a deterministic encoder, which learns the *aggregate posterior* distribution: $q(\theta) = \int q(\theta|\mathbf{x})p(\mathbf{x}) d\mathbf{x}$ where $p(\mathbf{x})$ denotes the empirical data distribution. The encoder generates $\mathbf{z} \sim q(\theta)$ such that $\mathbf{z} \in \mathcal{S}^{K-1}$.

4.3 Decoder Network

Given the encoder’s generated latent representation \mathbf{z} , it is transformed into the document-topic distribution vector through a projection network followed by Softmax operation. Finally, the decoder maps it to $\hat{\mathbf{x}} \in \mathbb{R}$, representing a probability distribution over vocabulary words. The transformation follows: [Linear(K, H) \rightarrow Dropout(p) \rightarrow

ReLU \rightarrow Linear(H, V) \rightarrow Softmax(\cdot)]. Here, Softmax is applied to the output to generate the document-word distribution $\hat{\mathbf{x}}$.

The decoder’s weight matrix defines the *topic-word matrix*, $\beta \in \mathbb{R}^{K \times V}$. Following [Srivastava and Sutton \(2017\)](#), we relax the simplex constraint on β , which is an empirically validated approach that improves the quality of topics.

4.4 Training Objective

S2WTM optimizes an objective comprising a reconstruction loss and a regularization term. Unlike VAEs, where regularization is applied to the posterior distribution $q(\theta|\mathbf{x})$, S2WTM regularizes the aggregated posterior $q(\theta)$ using the SSW. In VAEs, “ELBO surgery” ([Hoffman and Johnson, 2016](#)) decomposes the regularization term into two non-negative components: (1) the mutual information between the input variable \mathbf{x} and its latent representation θ , and (2) the KL divergence between the prior $p(\theta)$ and the aggregated posterior $q(\theta)$. Minimizing the KL term reduces mutual information, often leading to posterior collapse (uninformative latent representations) (see Appendix A). To mitigate this, S2WTM replaces KL regularization with SSW, ensuring effective alignment between the aggregated posterior and the prior while preserving hyperspherical geometry. The objective function for S2WTM is defined as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{RL}} + \lambda \mathcal{L}_{\text{OT}} \\ &= \inf_{q(\theta|\mathbf{x})} \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q(\theta|\mathbf{x})} [c(\mathbf{x}, \hat{\mathbf{x}})] + \lambda \text{SSW}_2^2(q\theta, p\theta). \end{aligned} \quad (5)$$

Here $p(\mathbf{x})$ is the data distribution. The reconstruction loss $c(\mathbf{x}, \hat{\mathbf{x}})$, is calculated as the cross-entropy between the input document representation \mathbf{x} and decoder generated document-word distribution $\hat{\mathbf{x}}$: $\mathcal{L}_{\text{RL}} = -\sum_{i=1}^V x_i \log \hat{x}_i$ where V is the vocabulary size. The regularization term employs SSW_2^2

defined as:

$$\text{SSW}_2^2(q_\theta, p_\theta) \approx \frac{1}{M} \sum_{i=1}^M W_2^2\left(\tilde{\mathcal{R}}_i q_\theta, \tilde{\mathcal{R}}_i p_\theta\right) \quad (6)$$

where M is the number of random projections, $\tilde{\mathcal{R}}_i$ denotes the i^{th} spherical radon transform. With an increasing number of projections, the SSW distance more accurately captures the structure of the data distribution. In practice, the reconstruction loss often significantly outweighs the regularization term $\text{SSW}_2^2(q_\theta, p_\theta)$. To balance these terms, a scaling factor λ is introduced for the regularization loss. We have treated λ as a hyperparameter.

5 Experimental Settings

We conducted our experiments using **OCTIS** (Terragni et al., 2021a), a publicly available framework to compare and optimize the topic models.

5.1 Datasets

We used seven publicly available datasets. Four of them, namely 20Newsgroups (**20NG**) (Mitchell, 1997), BBCNews (**BBC**) (Greene and Cunningham, 2006), **M10**, and **DBLP** (Pan et al., 2016), were obtained in pre-processed form from OCTIS. The remaining three datasets – SearchSnippets (**SS**), PascalFlickr (**Pascal**), and Biomedicine (**Bio**) – were sourced from (Qiang et al., 2022) and pre-processed using OCTIS. Further details about these baselines are provided in Appendix D. Dataset statistics are shown in Table 1.

| Dataset | #Docs | #Labels | #Words |
|---------------|-------|---------|--------|
| 20NG | 16309 | 20 | 1612 |
| BBC | 2225 | 5 | 2949 |
| M10 | 8355 | 10 | 1696 |
| SS | 12295 | 8 | 2000 |
| Pascal | 4834 | 20 | 2630 |
| Bio | 19448 | 20 | 2000 |
| DBLP | 54595 | 4 | 1513 |

Table 1: Statistics of the datasets used.

5.2 Baselines

We evaluate S2WTM against a range of traditional and neural topic models. Traditional baselines include **LDA** (Blei et al., 2003), **LSI** (Dumais, 2004), and **NMF** (Zhao and Tan, 2017), which are well-established for topic extraction.

Neural baselines include **ETM** (Dieng et al., 2020), which integrates word embeddings; **DVAE-TM** and **DVAE-RSVI-TM** (Burkhardt and Kramer, 2019), based on Dirichlet VAEs; **ProdLDA** (Srivastava and Sutton, 2017), which adopts a product-of-experts approach; **CombinedTM** and **ZeroshotTM** (Bianchi et al., 2021b,a), which leverage SBERT embeddings (Reimers and Gurevych, 2019) to incorporate contextualized document representations; **WTM** (Nan et al., 2019), based on Wasserstein autoencoders; **vONT** (Xu et al., 2023), a vMF-VAE-based model; and **ECRTM** (Wu et al., 2023), which enhances topic diversity through embedding clustering. See Appendix D for baseline details.

5.3 Automatic Evaluation of Topics

We assess topic models based on topic quality and distinctiveness. Topic quality, reflecting interpretability and coherence, is measured using NPMI (Lau et al., 2014) and CV (Röder et al., 2015), as they strongly correlate with human judgment. We exclude UCI (Newman et al., 2010) and UMass (Mimno et al., 2011) due to their weaker correlation with human assessments (Hoyle et al., 2021).

Topic diversity assesses the degree to which topics are distinct from each other. To evaluate this, we employ IRBO (Bianchi et al., 2021a), as well as its word embedding-based extension, wI-C (Centroid) (Terragni et al., 2021b). Higher NPMI and CV scores indicate better topic coherence, while higher IRBO and wI-C scores denote greater topic diversity.

| Dataset | M | p_θ | N_{batch} | p_{drop} | λ |
|---------------|------|----------------------------------|--------------------|-------------------|-----------|
| 20NG | 4000 | vMF | 1024 | 0.5 | 8.526 |
| BBC | 8000 | vMF | 256 | 0.05 | 5.567 |
| M10 | 2000 | $\mathcal{U}(\mathcal{S}^{K-1})$ | 64 | 0.5 | 7.065 |
| SS | 1000 | MvMF | 128 | 0.5 | 3.838 |
| Pascal | 500 | MvMF | 64 | 0.5 | 0.879 |
| Bio | 1000 | $\mathcal{U}(\mathcal{S}^{K-1})$ | 1024 | 0.5 | 3.701 |
| DBLP | 1000 | $\mathcal{U}(\mathcal{S}^{K-1})$ | 512 | 0.2 | 7.018 |

Table 2: Hyperparameter values of S2WTM.

5.4 Hyperparameter Tuning

In the S2WTM, hyperparameter tuning was carried out for each dataset using Bayesian Optimization in OCTIS, maximizing the NPMI score. The tuned parameters, listed in Table 2, include the number of projections (M), prior distribution (p_θ), batch

| Model | 20NG | | BBC | | M10 | | SS | | Pascal | | Bio | | DBLP | |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | NPMI | CV |
| LDA | 0.092 | 0.599 | 0.076 | 0.565 | -0.047 | 0.369 | -0.066 | 0.362 | -0.072 | 0.356 | 0.019 | 0.444 | 0.015 | 0.348 |
| LSI | 0.006 | 0.457 | 0.064 | 0.539 | 0.001 | 0.390 | -0.062 | 0.314 | -0.045 | 0.293 | -0.026 | 0.320 | 0.009 | 0.334 |
| NMF | 0.118 | 0.648 | 0.065 | 0.555 | 0.050 | 0.448 | <u>0.019</u> | 0.462 | -0.042 | 0.378 | 0.100 | 0.537 | <u>0.016</u> | 0.354 |
| ETM | 0.066 | 0.564 | 0.070 | 0.581 | 0.018 | 0.374 | 0.000 | 0.420 | -0.020 | 0.310 | -0.062 | 0.182 | -0.059 | 0.160 |
| DVAE-TM | <u>0.155</u> | <u>0.748</u> | -0.032 | 0.530 | -0.054 | 0.381 | -0.175 | 0.357 | 0.000 | 0.421 | 0.113 | 0.540 | -0.271 | 0.373 |
| DVAE-RSVI-TM | 0.146 | 0.750 | -0.051 | 0.523 | -0.052 | 0.412 | -0.192 | 0.411 | -0.019 | 0.422 | 0.100 | 0.537 | -0.269 | 0.356 |
| ProdLDA | 0.107 | 0.660 | 0.010 | 0.639 | 0.027 | 0.481 | -0.009 | 0.560 | -0.023 | 0.414 | 0.107 | 0.594 | -0.065 | 0.472 |
| ZeroshotTM | 0.103 | 0.653 | 0.038 | 0.673 | 0.041 | 0.481 | 0.017 | <u>0.565</u> | <u>0.005</u> | 0.428 | 0.133 | 0.604 | -0.062 | 0.474 |
| CombinedTM | 0.107 | 0.655 | 0.017 | 0.683 | <u>0.059</u> | 0.490 | 0.018 | 0.531 | -0.002 | 0.421 | <u>0.133</u> | <u>0.608</u> | -0.065 | <u>0.485</u> |
| WTM | 0.046 | 0.505 | -0.006 | 0.454 | -0.052 | 0.298 | -0.013 | 0.405 | -0.089 | 0.298 | 0.052 | 0.434 | -0.044 | 0.202 |
| vONT | 0.045 | 0.505 | -0.001 | 0.468 | -0.053 | 0.301 | -0.015 | 0.407 | -0.090 | 0.302 | 0.052 | 0.442 | -0.043 | 0.204 |
| ECRTM | -0.089 | 0.416 | <u>0.170</u> | <u>0.804</u> | -0.445 | 0.516 | -0.333 | 0.423 | -0.414 | <u>0.510</u> | -0.421 | 0.510 | -0.248 | 0.377 |
| S2WTM | 0.167 | 0.723 | 0.252 | 0.863 | 0.101 | <u>0.492</u> | 0.146 | 0.683 | 0.045 | 0.572 | 0.191 | 0.663 | 0.133 | 0.558 |

Table 3: Median coherence scores over five runs per metric, with the highest values in **bold** and the second-highest values underlined.

| Model | 20NG | | BBC | | M10 | | SS | | Pascal | | Bio | | DBLP | |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | IRBO | wI-C |
| LDA | 0.970 | 0.845 | 0.968 | 0.844 | 0.949 | 0.838 | 0.960 | 0.842 | 0.902 | 0.824 | 0.976 | 0.844 | 0.856 | 0.833 |
| LSI | 0.911 | 0.840 | 0.899 | 0.841 | 0.820 | 0.829 | 0.789 | 0.833 | 0.792 | 0.820 | 0.814 | 0.833 | 0.510 | 0.804 |
| NMF | 0.970 | 0.844 | 0.963 | 0.845 | 0.956 | 0.840 | 0.944 | 0.842 | 0.931 | 0.828 | 0.972 | 0.843 | <u>0.892</u> | 0.831 |
| ETM | 0.828 | 0.829 | 0.969 | 0.844 | 0.451 | 0.797 | 0.957 | 0.842 | 0.207 | 0.750 | 0.127 | 0.761 | 0.021 | 0.734 |
| DVAE-TM | 0.986 | <u>0.850</u> | 0.979 | 0.849 | 1.000 | 0.840 | 1.000 | 0.842 | 0.976 | 0.833 | 0.993 | 0.845 | 0.669 | 0.813 |
| DVAE-RSVI-TM | 0.987 | <u>0.850</u> | 0.994 | 0.849 | 1.000 | 0.840 | 1.000 | 0.842 | 0.978 | 0.833 | 0.996 | 0.845 | 0.546 | 0.827 |
| ProdLDA | 0.991 | <u>0.850</u> | 1.000 | 0.848 | 0.997 | <u>0.842</u> | 1.000 | 0.845 | 0.987 | 0.833 | 0.997 | <u>0.846</u> | 1.000 | <u>0.845</u> |
| ZeroshotTM | 0.991 | <u>0.850</u> | 1.000 | 0.849 | 1.000 | <u>0.842</u> | 1.000 | 0.844 | 0.987 | <u>0.834</u> | 0.996 | 0.845 | 1.000 | <u>0.845</u> |
| CombinedTM | 0.992 | <u>0.850</u> | 1.000 | 0.848 | 0.999 | <u>0.842</u> | 1.000 | <u>0.845</u> | 0.987 | 0.833 | 0.993 | <u>0.846</u> | 1.000 | <u>0.845</u> |
| WTM | 0.787 | 0.831 | 0.938 | 0.843 | 0.960 | 0.839 | <u>0.995</u> | 0.844 | 0.898 | 0.824 | 0.976 | 0.844 | 0.891 | 0.842 |
| vONT | 0.887 | 0.835 | <u>0.983</u> | 0.844 | 0.847 | 0.830 | 0.933 | 0.840 | 0.873 | 0.819 | 0.819 | 0.829 | 0.313 | 0.779 |
| ECRTM | 0.996 | 0.843 | 1.000 | <u>0.850</u> | 1.000 | 0.836 | 1.000 | 0.840 | 1.000 | 0.814 | 1.000 | 0.839 | 1.000 | <u>0.845</u> |
| S2WTM | <u>0.994</u> | 0.881 | 1.000 | 0.854 | <u>0.999</u> | 0.857 | 1.000 | 0.851 | <u>0.994</u> | 0.868 | <u>0.998</u> | 0.859 | 1.000 | 0.847 |

Table 4: Median diversity scores over five runs per metric, with the highest values in **bold** and the second-highest values underlined.

size (N_{batch}), dropout probability (p_{drop}), and scaling factor (λ). The number of topics was set to match the number of class labels in the dataset, and training was conducted for 100 epochs.

6 Results and Discussions

We categorize our findings into the following sections: (1) quantitative evaluation (Section 6.1), (2) extrinsic evaluation (Section 6.3), (3) qualitative evaluation (Section 6.2), and (4) LLM-based topic quality evaluations (Section 6.4).

6.1 Quantitative Evaluation

In the quantitative evaluation, we have evaluated the topic models using automatic topic evaluation scores.

Experimental Setup: For each dataset, we set the topic count K equal to the number of ground-truth labels, which are available for all datasets

used. The values of K for **20NG**, **BBC**, **M10**, **SS**, **Pascal**, **Bio**, and **DBLP** are 20, 5, 10, 8, 20, 20, and 4, respectively. For the robustness of the results, we have reported the median value over 5 random runs for a given model, a given dataset, and a given topic count.

Findings: Table 3 presents the coherence scores for all models across the evaluated datasets. S2WTM consistently achieves the highest coherence scores in both NPMI and CV across all datasets, with two exceptions. On the **M10** dataset, S2WTM ranks second in CV score behind ECRTM. On the **20NG** dataset, it ranks third in CV score, following DVAE-RSVI-TM and DVAE-TM. However, S2WTM still achieves the highest NPMI score on both M10 and 20NG, indicating strong overall topic coherence. These results highlight the consistent superiority of S2WTM in coherence compared to existing models.

Table 4 reports the diversity scores for all models.

| Model | Topics |
|--------------|---|
| LDA | game, team, year, player, play, win, good, hit, season, run card, drive, disk, system, work, problem, driver, machine, run, memory key, president, chip, government, make, encryption, security, option, press, phone |
| LSI | entry, team, game, send, graphic, season, mail, play, year, list drive, atheist, window, graphic, scsi, bit, administration, program, master, government key, encryption, bit, chip, call, launch, secret, block, war, widget |
| NMF | game, team, season, play, year, player, win, good, score, draft drive, disk, controller, hard, bio, support, card, feature, scsi, rom internet, privacy, information, secure, computer, security, mail, network, address, service |
| ETM | game, play, win, team, year, player, good, hit, season, score drive, disk, card, system, run, bit, window, scsi, problem, monitor key, chip, encryption, government, bit, clipper, message, system, algorithm, phone |
| DVAE-TM | game, team, playoff, score, season, shot, player, penalty, make, hockey scsi, controller, ide, disk, boot, drive, people, mhz, card, connector encryption, key, clipper, chip, escrow, entry, enforcement, privacy, encrypt, serial |
| DVAE-RSVI-TM | playoff, game, baseball, team, hockey, fan, pen, dog, score, season scsi, controller, ide, disk, left, meg, boot, bus, card, mhz encryption, key, clipper, chip, population, enforcement, secure, serial, encrypt, agency |
| ProdLDA | game, goal, play, score, wing, ranger, period, tie, playoff, blue card, drive, driver, board, problem, controller, bus, disk, scsi, port chip, phone, make, key, bit, clipper, algorithm, block, conversation, secret |
| ZeroshotTM | year, game, team, good, season, play, hit, player, league, average drive, scsi, disk, card, boot, problem, bus, board, transfer, ide chip, agency, law, key, algorithm, clipper, secure, security, encryption, secret |
| CombinedTM | game, good, year, team, play, season, player, win, hit, league card, drive, driver, scsi, bus, problem, board, memory, port, cpu chip, key, phone, law, warrant, illegal, clipper, encryption, cop, police |
| WTM | game, team, player, play, year, good, win, season, make, time drive, work, price, good, sell, make, buy, monitor, card, system key, government, system, encryption, make, program, time, space, chip, number |
| vONT | game, year, good, time, make, team, car, win, play, player drive, problem, card, run, system, disk, scsi, driver, bus, window key, chip, encryption, government, bit, system, make, clipper, time, phone |
| ECRTM | baseball, fan, pitcher, player, expansion, pitch, league, draft, suck, apple scsi, card, ide, pin, ranger, modem, port, ram, mouse, disk illegal, transmit, patient, warrant, taxis, fund, restriction, secret, budget, crack |
| S2WTM | game, team, win, score, player, playoff, goal, play, stat, season drive, card, scsi, ide, bus, controller, driver, disk, system, ram encryption, secure, chip, encrypt, phone, secret, communication, clipper, agency, security |

Table 5: Three representative topics from the 20NG dataset ($K = 20$), with unrelated words highlighted in blue.

S2WTM achieves the highest wI-C diversity score across all datasets. Although ECRTM generally performs well in terms of IRBO scores, due to its embedding clustering regularization approach, S2WTM performs comparably on the **20NG**, **M10**, **Pascal**, and **Bio** datasets, and outperforms ECRTM on the remaining ones.

6.2 Qualitative Evaluation

As topic models are unsupervised, their evaluation should go beyond automated coherence scores and include manual inspection (Hoyle et al., 2021).

Experimental Setup: We performed a qualitative analysis using the **20NG** dataset, training all models with 20 topics. Three topics were randomly selected from those generated by S2WTM. To compare these topics across models, we applied the

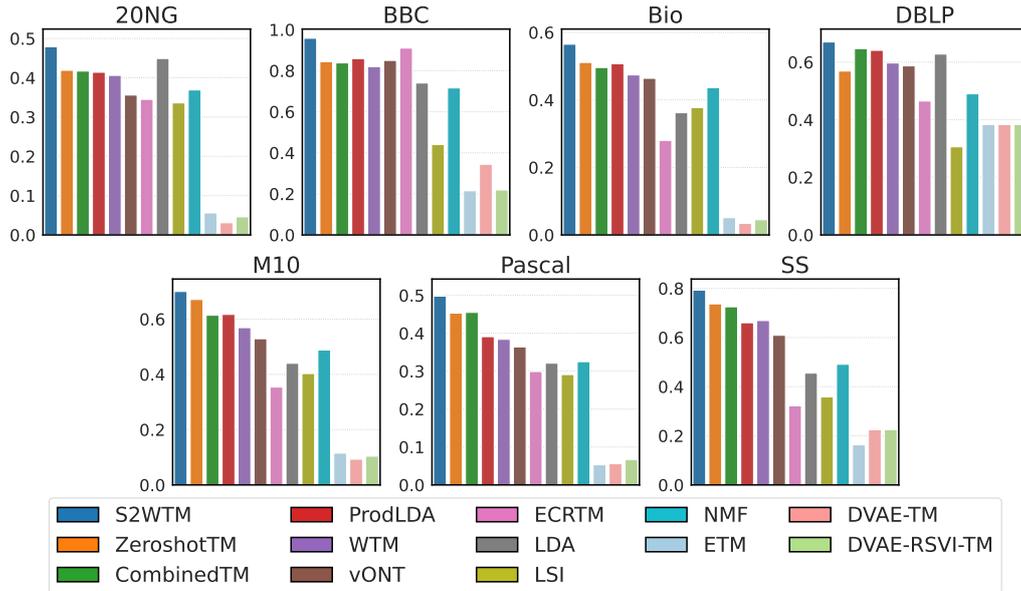


Figure 2: Median document classification accuracy of the models for all datasets.

topic alignment algorithm proposed by Adhya et al. (2023) (see Appendix H), which identifies the best-matching topics between different models. Table 5 presents the aligned topics, ensuring that each row represents semantically similar topics across models. The table also includes NPMI scores for each topic and a match score indicating its similarity to the corresponding topic in S2WTM. Unrelated words in a topic are highlighted manually.

Findings: The three topics in the table broadly represent *sports*, *hardware*, and *encryption/security*. Across all three, S2WTM consistently generates highly coherent and semantically meaningful topics. The *sports* topic includes terms like “game,” “team,” “win,” “score,” “player,” etc., forming a well-defined theme without unrelated words. The *hardware* topic consists solely of relevant terms such as “drive,” “card,” “scsi,” “ide,” “bus,” etc., avoiding noise present in other models. Similarly, the *encryption/security* topic includes “encryption,” “secure,” “chip,” “encrypt,” “phone,” etc., forming a clear, interpretable theme.

6.3 Extrinsic Evaluation

To further evaluate the quality of the topic models, we incorporated an extrinsic task focused on document classification. This task assesses how well the learned topic representations can support downstream applications.

Experimental Setup: We partitioned each of the seven datasets into 70% train, 15% validation, and

15% test. Each topic model was trained using a number of topics equal to the number of labels in the corresponding dataset. The resulting *document-topic* representations (i.e., the topic vector for each document) were then used to train a linear SVM classifier. Fig. 2 shows the median accuracy over five runs for all models across all the datasets.

Findings: As shown in Fig. 2, S2WTM consistently achieved the highest classification accuracy across all datasets, indicating that it learns more discriminative *document-topic* representations compared to other models. The superior performance of S2WTM is further supported by its higher NMI and Purity scores (detailed in Table 8 in the Appendix), suggesting that the learned topic distributions align well with ground truth labels and form coherent clusters. These results collectively demonstrate the effectiveness of S2WTM for both predictive and clustering-based downstream tasks.

6.4 LLM-based Topic Quality Assessment

LLM-based Topic Quality Assessment serves as a proxy for human interpretability, addressing the limitations of automated topic quality evaluations.

Experimental Setup: Following Stambach et al. (2023), we employ two LLM-based evaluation tasks: *rating* and *intrusion detection*. In the rating task, an LLM scores topic coherence from 1 to 3 (“1” = not very related, “2” = moderately related, “3” = very related).

In the intrusion detection task, an unrelated word

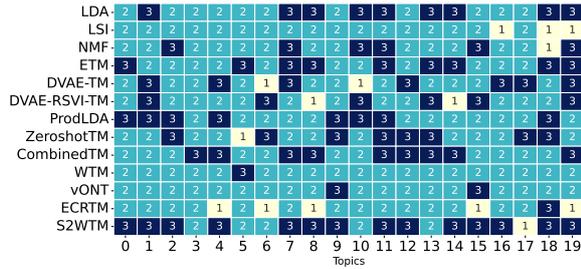


Figure 3: LLM-assigned topic ratings for all models on 20NG.

is inserted among the top topic words, and the LLM must identify it. This assumes that intruders are easier to detect in well-defined topics but harder in incoherent ones (Chang et al., 2009). While LLM-based evaluations have limitations (Li et al., 2024), they provide valuable insights into topic quality. We used GPT-4 for these evaluations. All models were trained on the 20NG dataset with 20 topics.

Findings in rating task: Fig. 3 shows the LLM’s scores for the rating task, demonstrating that S2WTM generates the highest-quality topics among the evaluated models. S2WTM achieves the highest mean score of 2.7 and a median score of 3.0, indicating consistently high ratings for its topics.

Findings in intrusion detection task: We performed three test series for each of the 20 topics generated by each model. For each series, we computed the accuracy of correctly detecting the intruder word. The statistics of these accuracy scores across the three tests are shown in Fig. 4, which again demonstrates the high quality of S2WTM-generated topics compared to existing models.

7 Beyond Euclidean Assumptions: Empirical Insights

We aim to evaluate the impact of modeling the latent space on a hypersphere instead of in Euclidean space and its effect on topic quality.

Experimental Setup: To assess this, we modified the objective function and prior distribution of S2WTM. Specifically, we replaced the Spherical Sliced-Wasserstein distance with the standard Sliced-Wasserstein distance for distribution matching and substituted the hyperspherical prior with a Dirichlet distribution. All other parameters remained unchanged. Table 6 reports median topic quality scores over five random runs.

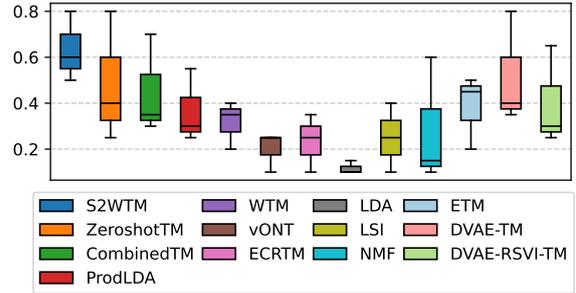


Figure 4: Intruder word detection accuracy over three test runs per model on 20NG.

Findings: Comparing the results of S2WTM in Tables 3 and 4, NPMI improves by 54.6% on 20NG and over 100% on other datasets. CV gains range from 15.9% to 51.2%, except for M10, where it remains similar. IRBO and wI-C scores improve by 1.0% to 52% and 2.3% to 5.9%, respectively, highlighting the benefits of modeling the latent space on a hypersphere.

| Dataset | Coherence | | Diversity | |
|---------|---------------|---------------|---------------|---------------|
| | NPMI | CV | IRBO | wI-C |
| 20NG | 0.108 (0.167) | 0.623 (0.723) | 0.984 (0.994) | 0.821 (0.881) |
| BBC | 0.081 (0.252) | 0.592 (0.863) | 0.941 (1.000) | 0.835 (0.854) |
| M10 | 0.038 (0.101) | 0.494 (0.492) | 0.964 (0.999) | 0.826 (0.857) |
| SS | 0.024 (0.146) | 0.549 (0.683) | 0.940 (1.000) | 0.809 (0.851) |
| Pascal | 0.007 (0.045) | 0.425 (0.572) | 0.887 (0.994) | 0.826 (0.868) |
| Bio | 0.094 (0.191) | 0.572 (0.663) | 0.972 (0.998) | 0.818 (0.859) |
| DBLP | 0.004 (0.133) | 0.369 (0.558) | 0.658 (1.000) | 0.838 (0.847) |

Table 6: Median scores over five runs without the spherical latent space assumption. Corresponding S2WTM scores are shown in parentheses.

8 Conclusion

S2WTM is a Spherical Sliced-Wasserstein Autoencoder-based topic model that models the latent space as hyperspherical instead of Euclidean while avoiding posterior collapse seen in VAE-based models. We evaluated S2WTM on seven publicly available datasets for topic modeling assessment. In most experiments, S2WTM consistently achieved superior topic coherence and diversity compared to competitive topic models from the literature. Manual evaluation of selected topics further confirmed that S2WTM produces more coherent topics than alternative models. Additionally, in extrinsic evaluations, S2WTM outperformed existing models across all datasets. Finally, our LLM-based evaluation, serving as a proxy for human judgment, further demonstrated that S2WTM generates higher-quality topics.

Limitations

While S2WTM performs well, computing SSW requires multiple spherical Radon transforms, leading to increased computational cost as the number of projections grows.

Additionally, in Section 6.4, we employ LLMs for topic quality evaluation. However, LLMs are known to exhibit biases (Li et al., 2024; Guo et al., 2024; Kotek et al., 2023). To mitigate this as much as possible, we conducted a manual evaluation of randomly selected topics (see Section 6.2).

Ethics Statement

The topic words presented in Table 5 are generated by topic models trained on the **20NG** dataset. These models learn patterns from data without explicit supervision, and their outputs may reflect biases present in the dataset. The authors have no intention to cause harm or offense to any community, religion, country, or individual. Efforts have been made to ensure that the generated topics are analyzed objectively, and any unintended biases in the model outputs are acknowledged as part of the limitations of automated topic modeling.

References

- Suman Adhya, Avishek Lahiri, and Debarshi Kumar Sanyal. 2023. [Do neural topic models really need dropout? analysis of the effect of dropout in topic modeling](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2220–2229, Dubrovnik, Croatia. Association for Computational Linguistics.
- Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. [On the surprising behavior of distance metrics in high dimensional space](#). In *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Hareesh Bahuleyan, Lili Mou, Hao Zhou, and Olga Vechtomova. 2019. [Stochastic Wasserstein autoencoder for probabilistic sentence generation](#). 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)*, pages 4068–4076, Minneapolis, Minnesota. Association for Computational Linguistics.
- Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. [Clustering on the unit hypersphere using von mises-fisher distributions](#). *Journal of Machine Learning Research*, 6(46):1345–1382.
- Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. [Nonparametric spherical topic modeling with word embeddings](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 537–542, Berlin, Germany. Association for Computational Linguistics.
- Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. [Pre-training is a hot topic: Contextualized document embeddings improve topic coherence](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766. 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(Jan.):993–1022.
- Clément Bonnet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh Tan Pham. 2023. [Spherical sliced-wasserstein](#). In *The Eleventh International Conference on Learning Representations*.
- Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. 2015. [Sliced and Radon Wasserstein Barycenters of Measures](#). *Journal of Mathematical Imaging and Vision*, 1(51):22–45.
- Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. [Generating sentences from a continuous space](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.
- Sophie Burkhardt and Stefan Kramer. 2019. [Decoupling sparsity and smoothness in the Dirichlet Variational Autoencoder topic model](#). *Journal of Machine Learning Research*, 20(131):1–27.
- Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. [Reading tea leaves: How humans interpret topic models](#). In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.
- Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. 2018. [Hyperspherical variational auto-encoders](#). In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 856–865. Association For Uncertainty in Artificial Intelligence (AUAI).

- Inderjit S Dhillon and Suvrit Sra. 2003. [Modeling data using directional distributions](#). Technical report, Cite-seer.
- Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. [Topic modeling in embedding spaces](#). *Transactions of the Association for Computational Linguistics*, 8:439–453.
- Susan T. Dumais. 2004. [Latent semantic analysis](#). *Annual Review of Information Science and Technology*, 38(1):188–230.
- Hafsa Ennajari, Nizar Bouguila, and Jamal Bentahar. 2022. [Embedded spherical topic models for supervised learning](#). In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1650–1656.
- Alessio Figalli and Cédric Villani. 2011. [Optimal Transport and Curvature](#), pages 171–217. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Derek Greene and Pádraig Cunningham. 2006. [Practical solutions to the problem of diagonal dominance in kernel document clustering](#). In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, page 377–384, New York, NY, USA. Association for Computing Machinery.
- Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. [Bias in large language models: Origin, evaluation, and mitigation](#). *Preprint*, arXiv:2411.10915.
- Matthew D Hoffman and Matthew J Johnson. 2016. [ELBO surgery: yet another way to carve up the variational evidence lower bound](#). In *Workshop in advances in approximate Bayesian inference, NIPS*, volume 1.
- Thomas Hofmann. 1999. [Probabilistic latent semantic indexing](#). In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99*, page 50–57, New York, NY, USA. Association for Computing Machinery.
- Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. [Is automated topic model evaluation broken? the incoherence of coherence](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.
- Rabin Julien, Gabriel Peyré, Julie Delon, and Bernot Marc. 2011. [Wasserstein Barycenter and its Application to Texture Mixing](#). In *Proc. SSVM'11*, pages 435–446, Israel. Springer.
- Yuri Kinoshita, Kenta Oono, Kenji Fukumizu, Yuichi Yoshida, and Shin-ichi Maeda. 2023. [Controlling posterior collapse by an inverse lipschitz constraint on the decoder network](#). In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- Hadas Kotek, Rikker Dockum, and David Sun. 2023. [Gender bias and stereotypes in large language models](#). In *Proceedings of The ACM Collective Intelligence Conference, CI '23*, page 12–24, New York, NY, USA. Association for Computing Machinery.
- Jey Han Lau, David Newman, and Timothy Baldwin. 2014. [Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539. Association for Computational Linguistics.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. [LLMs-as-judges: A comprehensive survey on llm-based evaluation methods](#). *Preprint*, arXiv:2412.05579.
- Ximing Li, Jinjin Chi, Changchun Li, Jihong Ouyang, and Bo Fu. 2016. [Integrating topic modeling with word embeddings by mixtures of vMFs](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 151–160, Osaka, Japan. The COLING 2016 Organizing Committee.
- Yishu Miao, Lei Yu, and Phil Blunsom. 2016. [Neural variational inference for text processing](#). In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA. PMLR.
- David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. [Optimizing semantic coherence in topic models](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Tom Mitchell. 1997. [Twenty Newsgroups](#). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5C323>.
- Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. [Topic modeling with Wasserstein autoencoders](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381. Association for Computational Linguistics.
- David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. [Automatic evaluation of topic coherence](#). In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

- Shirui Pan, Jia Wu, Xingquan Zhu, Chengqi Zhang, and Yang Wang. 2016. **Tri-party deep network representation**. In *IJCAI-16 - Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, pages 1895–1901, United States of America. Association for the Advancement of Artificial Intelligence (AAAI).
- Gabriel Peyré, Marco Cuturi, et al. 2019. **Computational optimal transport: With applications to data science**. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.
- Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2022. **Short text topic modeling techniques, applications, and performance: A survey**. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):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.
- Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney. 2010. **Spherical topic models**. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML’10*, page 903–910, Madison, WI, USA. Omnipress.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. **Exploring the space of topic coherence measures**. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, page 399–408. Association for Computing Machinery.
- Akash Srivastava and Charles Sutton. 2017. **Autoencoding variational inference for topic models**. In *International Conference on Learning Representations*.
- Dominik Stammach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. **Revisiting automated topic model evaluation with large language models**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics.
- Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021a. **OCTIS: Comparing and optimizing topic models is simple!** In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics.
- Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021b. **Word embedding-based topic similarity measures**. In *Natural Language Processing and Information Systems*, pages 33–45, Cham. Springer International Publishing.
- Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. 2018. **Wasserstein autoencoders**. In *International Conference on Learning Representations*.
- Gary Ulrich. 1984. **Computer generation of distributions on the m-sphere**. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(2):158–163.
- Cédric Villani et al. 2009. *Optimal transport: old and new*, volume 338. Springer.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. **A similarity measure for indefinite rankings**. *ACM Transactions on Information Systems*, 28(4).
- Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023. **Effective neural topic modeling with embedding clustering regularization**. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37335–37357. PMLR.
- Weijie Xu, Xiaoyu Jiang, Srinivasan Sengamedu Hanumantha Rao, Francis Iannacci, and Jinjin Zhao. 2023. **vONTSS: vMF based semi-supervised neural topic modeling with optimal transport**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4433–4457, Toronto, Canada. Association for Computational Linguistics.
- Renbo Zhao and Vincent Y. F. Tan. 2017. **Online non-negative matrix factorization with outliers**. *IEEE Transactions on Signal Processing*, 65(3):555–570.

A ELBO Surgery and Posterior Collapse

In VAEs, *posterior collapse* refers to the phenomenon where the approximate posterior $q(\mathbf{z}|\mathbf{x})$ becomes indistinguishable from the prior $p(\mathbf{z})$, resulting in latent variables that carry little to no information about the input \mathbf{x} . This can be formally characterized as follows:

Definition 4 (ϵ -Posterior Collapse (Kinoshita et al., 2023)). For a given parameter $\theta \in \mathbb{R}^K$, a dataset $\mathbf{x} \in \mathbb{R}^{m \times n}$, and a closeness criterion $d(\cdot, \cdot)$, the ϵ -posterior collapse is defined for a given $\epsilon \geq 0$, as

$$d(q_\phi(z_i|x_i), p(z_i)) < \epsilon, \quad \forall i = 1, \dots, n.$$

Typically, the closeness criterion $d(\cdot, \cdot)$ is chosen as the Kullback–Leibler divergence $\text{KL}(\cdot||\cdot)$, consistent with its use in the Evidence Lower Bound (ELBO) objective:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \text{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$$

To understand the effect of the KL divergence term, [Hoffman and Johnson \(2016\)](#) introduced a decomposition known as *ELBO surgery*, which analyzes the expected KL term under the data distribution $p(\mathbf{x})$:

$$\begin{aligned} & \mathbb{E}_{p(\mathbf{x})} [\text{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))] \\ &= \mathbb{E}_{p(\mathbf{x})} \left[\int q(\mathbf{z}|\mathbf{x}) \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{z} \right] \\ &= \mathbb{E}_{p(\mathbf{x})} \left[\int q(\mathbf{z}|\mathbf{x}) \left(\log \frac{q(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} + \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right) d\mathbf{z} \right] \\ &= \underbrace{\mathbb{E}_{p(\mathbf{x})} [\text{KL}(q(\mathbf{z}|\mathbf{x})\|q(\mathbf{z}))]}_{I_q(\mathbf{x};\mathbf{z})} + \underbrace{\text{KL}(q(\mathbf{z})\|p(\mathbf{z}))}_{\text{Marginal KL}} \end{aligned}$$

Thus, we obtain the decomposition:

$$\mathbb{E}_{p(\mathbf{x})} [\text{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))] = I_q(\mathbf{x};\mathbf{z}) + \text{KL}(q(\mathbf{z})\|p(\mathbf{z}))$$

This decomposition reveals that the KL term implicitly includes a *mutual information* component $I_q(\mathbf{x};\mathbf{z})$, which measures how much information about \mathbf{x} is encoded in the latent variable \mathbf{z} . Minimizing the KL term too aggressively can therefore inadvertently reduce mutual information, leading to *posterior collapse*, where the encoder learns to ignore \mathbf{x} and produces non-informative latent representations.

B Sampling from the vMF Distribution

Sampling from the vMF distribution ([Ulrich, 1984](#)) involves generating two components:

- A *tangential component* $\mathbf{v} \sim \mathcal{S}^{d-2}$.
- A *radial component* t sampled from $f_{\text{radial}}(t|\kappa, p)$.

An intermediate sample $\mathbf{z}' \in \mathbb{R}^p$ is then computed as:

$$\mathbf{z}' = t\mathbf{e}_1 + \sqrt{1-t^2}\mathbf{v} \quad (7)$$

where $\mathbf{e}_1 = (1, 0, \dots, 0)$ is a unit vector. The final sample is obtained via the Householder transformation:

$$\mathbf{z} = \mathbf{H}\mathbf{z}' \quad (8)$$

where \mathbf{H} reflects vectors to align \mathbf{e}_1 with the desired mean direction $\boldsymbol{\mu}$.

C Datasets Description

We used seven publicly available datasets from diverse domains, including news, academia, biomedical literature, and web search, to evaluate topic modeling approaches.

1. **20NewsGroups (20NG)** is a popularly used text dataset for topic modeling, consisting of documents collected from 20 different online newsgroups. It contains a total of 16,309 pre-processed documents. Each document is assigned a label corresponding to its respective newsgroup.
2. **BBCNews (BBC)** ([Greene and Cunningham, 2006](#)) is a collection of 2,225 news articles published by the British Broadcasting Corporation (BBC). Each article is assigned to one of five predefined categories: *tech*, *business*, *entertainment*, *sports*, or *politics*.
3. **M10** ([Greene and Cunningham, 2006](#)) is a subset of the CiteSeer^X digital library, consisting of 8,355 academic documents spanning 10 different research topics.
4. **SearchSnippets (SS)** ([Qiang et al., 2022](#)) is a dataset constructed from web search transactions, consisting of 12,295 short text documents after pre-processing. These documents are derived from predefined search phrases and are categorized into eight thematic domains: *business*, *computers*, *culture-arts*, *education-science*, *engineering*, *health*, *politics-society*, and *sports*.
5. **PascalFlickr (Pascal)** ([Qiang et al., 2022](#)) is a dataset consisting of 4,834 image captions (after pre-processing) sourced from Pascal and Flickr image collections. These captions are categorized into 20 different thematic classes.
6. **Biomedicine (Bio)**¹ ([Qiang et al., 2022](#)) is derived from biomedical literature made available through the BioASQ challenge. It contains 19,448 documents after pre-processing, each assigned to one of 20 biomedical categories.
7. **DBLP** ([Pan et al., 2016](#)) is a bibliographic dataset in computer science, curated by selecting conferences from four research domains:

¹The **SS**, **Pascal**, and **Bio** datasets, along with their corresponding labels, are available at: <https://github.com/qiang2100/STTM/tree/master/dataset>

| Dataset | M=250 | M=500 | M=1000 | M=2000 | M=4000 | M=8000 |
|---------------|------------------|------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| 20NG | (0.1456, 5.903) | (0.1453, 6.869) | (0.1519, 7.467) | (0.1518, 10.489) | (0.1670, 13.515) | (0.1649, 22.128) |
| BBC | (0.2139, 3.440) | (0.2219, 5.474) | (0.2501, 7.649) | (0.2371, 9.102) | (0.2437, 11.175) | (0.2523, 20.239) |
| M10 | (0.0642, 5.800) | (0.0772, 6.185) | (0.0976, 9.764) | (0.1010, 11.665) | (0.0990, 17.658) | (0.1007, 32.467) |
| SS | (0.1142, 7.322) | (0.1165, 8.132) | (0.1456, 9.065) | (0.1455, 11.765) | (0.1450, 18.792) | (0.1445, 33.772) |
| Pascal | (0.0140, 4.922) | (0.0446, 6.641) | (0.0406, 7.601) | (0.0427, 9.326) | (0.0420, 12.755) | (0.0435, 21.093) |
| Bio | (0.1767, 4.856) | (0.1872, 5.465) | (0.1905, 7.346) | (0.1904, 10.642) | (0.1884, 17.374) | (0.1897, 31.194) |
| DBLP | (0.1257, 10.629) | (0.1261, 11.361) | (0.1333, 12.074) | (0.1303, 16.006) | (0.1300, 21.938) | (0.1329, 37.242) |

Table 7: NPMI and Per-Epoch Training Time (in seconds) for Different Numbers of Projections M . The best NPMI values are highlighted.

databases, data mining, artificial intelligence, and computer vision. The dataset comprises 54,595 documents after pre-processing.

The datasets – **20NG**, **BBC**, **DBLP** and **M10** are available in the **OCTIS** framework. For the remaining three datasets – **SS**, **Pascal**, and **Bio** – we performed additional pre-processing to ensure consistency with the OCTIS format and maintain comparability across all datasets. The details of our pre-processing steps are described in Section C.1.

C.1 Preprocessing

Using OCTIS, we convert all text to lowercase, remove punctuation, apply lemmatization, filter out words with fewer than three characters, and discard documents containing fewer than three words.

D Baseline Configurations

We reproduced all baseline models based on their original papers, using either official implementations or OCTIS. For LDA (Blei et al., 2003), LSI (Dumais, 2004), NMF (Zhao and Tan, 2017), ETM (Dieng et al., 2020), ProLDA (Srivastava and Sutton, 2017), ZeroShotTM (Bianchi et al., 2021b), and CombinedTM (Bianchi et al., 2021a), we used the default parameter settings available in OCTIS.

For DVAE-TM, DVAE-RSVI-TM (Burkhardt and Kramer, 2019)², WTM (Nan et al., 2019)³, vONT (Xu et al., 2023)⁴, and ECRTM (Wu et al., 2023)⁵. These models were integrated into the OCTIS framework to facilitate a standardized evaluation, ensuring fair and reproducible comparisons with other baseline models.

²DVAE-TM and DVAE-RSVI-TM: <https://github.com/mayanknagda/neural-topic-models>

³WTM: https://github.com/zll17/Neural_Topic_Models/blob/master/models/WTM.py

⁴vONT: <https://github.com/xuweijieshuai/vONTSS>

⁵ECRTM: <https://github.com/BoxWu/ECRTM>

E Computing Infrastructure

Our experiments were conducted on a system with an Intel® Core® i7-10700K processor, 32 GB of RAM, an NVIDIA GeForce GTX 1660 SUPER GPU with 6 GB of VRAM, CUDA 12.2, and the Ubuntu 22.04 operating system.

F Trade-off Between Performance and Computational Cost in Terms of Number of Projections

The computational complexity of SSW grows linearly with the number of projections M , as noted by Bonet et al. (2023). Their empirical analysis confirms that increasing M proportionally increases runtime. Following this, we investigate the trade-off between performance and computational cost for S2WTM by varying M and measuring its impact on model quality.

Table 7 reports the median NPMI scores across five random runs, paired with the corresponding per-epoch training times, for different values of M . We highlight the best NPMI scores for each dataset.

The results suggest that while increasing M generally improves performance, the gains plateau beyond a certain point. Across most datasets, performance stabilizes around $M = 1000$, offering a strong balance between topic quality and computational efficiency. Although theoretically determining the optimal M remains an open research question, our empirical findings suggest that setting $M \approx 1000$ yields favorable results without incurring excessive computational costs. These observations are consistent with the linear scaling behavior reported by Bonet et al. (2023).

G Clustering Alignment Metrics: NMI and Purity

To complement the document classification results reported in Section 6.3, we evaluate how well the

| Model | 20NG | | BBC | | M10 | | SS | | Pascal | | Bio | | DBLP | |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | NMI | Purity |
| LDA | 0.425 | 0.461 | 0.617 | 0.757 | 0.157 | 0.376 | 0.219 | 0.469 | 0.303 | 0.313 | 0.255 | 0.371 | 0.099 | 0.533 |
| LSI | 0.282 | 0.326 | 0.250 | 0.231 | 0.259 | 0.395 | 0.111 | 0.342 | 0.295 | 0.274 | 0.364 | 0.410 | 0.044 | 0.449 |
| NMF | 0.298 | 0.360 | 0.602 | 0.808 | 0.240 | 0.449 | 0.152 | 0.423 | 0.326 | 0.343 | 0.338 | 0.421 | 0.090 | 0.505 |
| ETM | 0.026 | 0.091 | 0.022 | 0.270 | 0.017 | 0.157 | 0.007 | 0.228 | 0.059 | 0.106 | 0.015 | 0.064 | 0.231 | 0.383 |
| DVAE-TM | 0.227 | 0.230 | 0.414 | 0.431 | 0.192 | 0.269 | 0.251 | 0.225 | 0.263 | 0.258 | 0.260 | 0.267 | 0.182 | 0.383 |
| DVAE-RSVI-TM | 0.238 | 0.237 | 0.308 | 0.374 | 0.164 | 0.254 | 0.269 | 0.225 | 0.270 | 0.264 | 0.267 | 0.262 | 0.107 | 0.383 |
| CombinedTM | 0.406 | 0.418 | 0.650 | 0.838 | 0.437 | 0.614 | 0.503 | 0.725 | 0.463 | 0.459 | 0.407 | 0.496 | 0.240 | 0.647 |
| ZeroshotTM | 0.383 | 0.421 | 0.637 | 0.844 | 0.460 | 0.671 | 0.509 | 0.737 | 0.465 | 0.459 | 0.406 | 0.511 | 0.168 | 0.569 |
| ProdLDA | 0.369 | 0.414 | 0.685 | 0.859 | 0.420 | 0.617 | 0.418 | 0.660 | 0.407 | 0.397 | 0.398 | 0.508 | 0.230 | 0.641 |
| WTM | 0.370 | 0.366 | 0.718 | 0.853 | 0.340 | 0.543 | 0.431 | 0.678 | 0.401 | 0.375 | 0.347 | 0.473 | 0.188 | 0.598 |
| vONT | 0.328 | 0.351 | 0.721 | 0.851 | 0.351 | 0.529 | 0.402 | 0.629 | 0.420 | 0.363 | 0.342 | 0.440 | 0.179 | 0.572 |
| ECRTM | 0.336 | 0.345 | 0.716 | 0.861 | 0.142 | 0.355 | 0.084 | 0.322 | 0.315 | 0.317 | 0.167 | 0.281 | 0.059 | 0.465 |
| S2WTM | 0.437 | 0.469 | 0.729 | 0.874 | 0.464 | 0.680 | 0.547 | 0.749 | 0.471 | 0.521 | 0.557 | 0.644 | 0.254 | 0.686 |

Table 8: Clustering performance (NMI and Purity) of the models across all the datasets.

learned topic distributions align with ground-truth class labels using two clustering metrics: Normalized Mutual Information (NMI) and Purity. While classification accuracy reflects predictive utility, these unsupervised metrics assess the structure and separability of the learned representations.

As shown in Table 8, S2WTM consistently achieves the highest scores across all datasets. This indicates that the learned topic distributions align well with the true labels and form coherent, well-separated clusters, further confirming the strength of S2WTM for clustering-based downstream tasks.

H Topic Alignment Algorithm

We align topics between two given models using the following two-step strategy as prescribed by (Adhya et al., 2023):

1. Construct Similarity Matrix:

- (a) Let $P = \{P[1], P[2], \dots, P[K]\}$ and $Q = \{Q[1], Q[2], \dots, Q[K]\}$ be the topic lists from the two models.
- (b) Compute the Rank-biased Overlap (RBO) similarity (Webber et al., 2010) for each topic pair:

$$a_{i,j} = \text{RBO}(P[i], Q[j]), \\ \forall i, j \in \{1, \dots, K\}.$$

- (c) Construct the similarity matrix $\mathbf{A} = (a_{ij})_{1 \leq i, j \leq K}$, where each entry $a_{ij} \in [0, 1]$, with 0 indicating no overlap and 1 indicating exact overlap.

2. Iterative Topic Pairing:

- (a) While there are unaligned topics in P and Q :

- i. Identify the pair (i^*, j^*) with the highest similarity score:

$$(i^*, j^*) = \arg \max_{i,j} a_{ij}$$

- ii. Align the topics $P[i^*]$ and $Q[j^*]$.
- iii. Exclude the selected topics from further consideration:

$$P \leftarrow P \setminus \{P[i^*]\}, \quad Q \leftarrow Q \setminus \{Q[j^*]\}$$

ensuring that each topic is aligned only once.