A Self-supervised Neural Topic Model Extended with Adversarial Data Augmentation

Anonymous ACL submission

Abstract

Neural topic models (NTMs) have advanced topic modeling through their flexibility, enabling self-supervised learning with contrastive samples at the document or topic representation level. However, prior tf-idf-based augmentation strategies provide limited guidance during training. To address this, we propose an adversarial framework with a trainable augmentation model that generates positive samples in the embedding space, leveraging contextualized word embeddings from large language models (LLMs). Experimental results demonstrate that our model surpasses previous approaches in topic coherence, highlighting the effectiveness of adversarial data augmentation in improving topic modeling performance.

1 Introduction

Topic modeling uses word co-occurrence patterns to extract latent topics from large text corpora, enabling applications such as text classification, clustering, regression, information retrieval, and recommendation systems (Mcauliffe and Blei, 2007; Zhao et al., 2021; Wei and Croft, 2006; Wang and Blei, 2011). Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a foundational conventional model, while neural topic models (NTMs), based on the variational autoencoder (VAE) (Welling and Kingma, 2014) framework, have gained prominence with advances in deep learning and GPUs. The flexibility and extensibility of NTMs have enabled extensions such as self-supervised learning, which leverage contrastive samples to improve topic representations and improve topic quality (Nguyen and Luu, 2021; Wu et al., 2022; Han et al., 2023).

Self-supervised NTMs use various strategies to build contrastive samples. CLNTM (Nguyen and Luu, 2021) generates contrastive bag-of-words (BoW) samples based on tf-idf values: positive samples replace unimportant words with reconstructed counterparts, while negative samples replace salient words. VICNTM (Xu et al., 2025) uses the same strategy to create only positive samples and employing Variance-Invariance-Covariance (VIC) regularization (Bardes et al., 2022) to act as implicit negative samples. However, this tf-idf-based strategy causes positive samples to become increasingly similar to anchor samples during training, limiting their effectiveness in guiding the learning process.

Adversarial data augmentation, widely used in computer vision to improve model generalization, generates informative positive samples by maximizing task loss while preventing collapse (Zhang et al., 2020; Tang et al., 2020; Suzuki, 2022). TeachAugment (Suzuki, 2022) introduces a teacher model to guide augmentation, requiring no prior knowledge or additional hyperparameters. This framework ensures that the generated positive samples remain challenging for the target model while still being recognizable by the teacher model.

Representation-level augmentation, which adds adversarial perturbations to anchor samples in the embedding space, is common in adversarial frameworks for text data (Miyato et al., 2017; Zhu et al., 2020). Chen et al. (2023) proposed an adversarial framework for low-resource text classification, generating hard positive samples by mixing embeddings of important words with unknown-word embeddings to improve robustness.

In this paper, we propose VICNTMxACE, an e**x**tension of VICNTM incorporating an **A**dversarial framework and **C**ontextualized **E**mbeddings, as illustrated in Fig. 1. To enhance VICNTM's performance, we optimize positive sample generation by applying Suzuki (2022)'s adversarial framework and adapting a representation-level augmentation strategy inspired by Chen et al. (2023). To augment anchor samples in the embedding space, we replace BoW representations with BERT-encoded word embed-

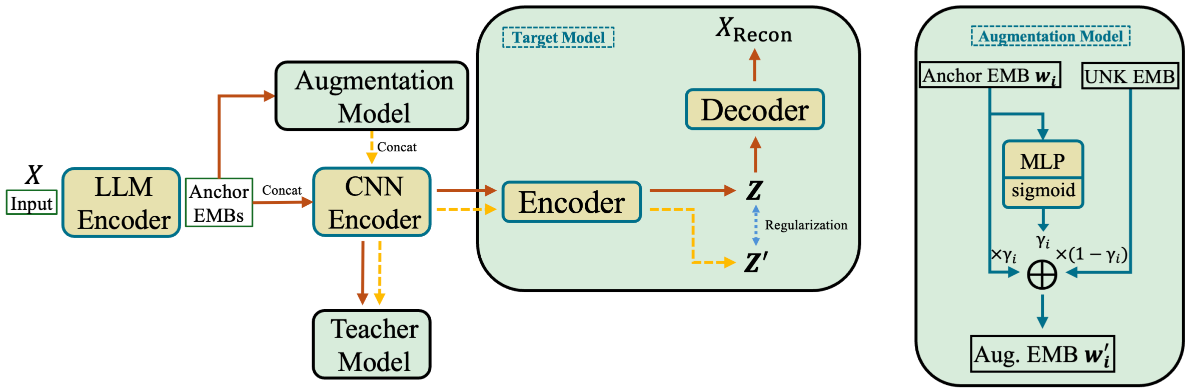

Figure 1: Illustration of our model. The left part of the figure depicts the structure of the model, with red (solid) lines representing the flow of anchor samples and yellow (dashed) lines indicating the flow of positive samples. The right part of the figure illustrates the structure of the augmentation model in detail.

dings (Devlin et al., 2019), compressed using a CNN encoder (Xu et al., 2023). Experiments on three widely used datasets demonstrate that our model outperforms baseline and state-of-the-art VAE-based models in topic coherence. An ablation study further verifies the effectiveness of each newly added component.

2 Related works

Research on NTMs has become an integral part of topic modeling. ProdLDA (Srivastava and Sutton, 2017) was the first NTM to use the VAE framework, with a logistic normal prior approximating the Dirichlet prior. Building on ProdLDA, SCHOLAR (Card et al., 2018) incorporated external information and improved topic quality by refining implementation details and leveraging word log-frequency.

Meanwhile, adversarial NTMs using generative adversarial networks (Goodfellow et al., 2014) generate negative samples via a generator and distinguish them with a discriminator (Wang et al., 2019, 2020; Hu et al., 2020), but Nguyen and Luu (2021) showed that mutual information between positive and anchor samples is more beneficial. Building on SCHOLAR, they proposed CLNTM, which uses tf-idf to generate both positive and negative samples. Avoiding the limitations of negative samples, Xu et al. (2025) adopted the same augmentation strategy to generate only positive samples and introduced regularizations between positive and anchor samples, as well as among samples within each group. Contrastive learning has also been utilized in other NTMs in various ways (Wu et al., 2022; Han et al., 2023).

Unlike the adversarial topic models, our model consists of an augmentation model, a self-supervised NTM as the target model, and a teacher model, which will be described in the next section.

3 Methodology

In this paper, we propose VICNTMxACE, an extension of the regularized self-supervised NTM, VICNTM (Xu et al., 2025), using an adversarial framework. Fig. 1 illustrates the model structure. For each minibatch of documents X, where each document consists of a sequence of tokens, anchor samples $\boldsymbol{X}$ are obtained with each sample represented as word embeddings $\{\mathbf{w}_0, \mathbf{w}_1, \cdots, \mathbf{w}_n\}$ via the LLM encoder, with n being the maximum number of tokens it can process. Positive samples $\boldsymbol{X}' = \alpha_{\boldsymbol{\phi}}(\boldsymbol{X})$ are generated through the augmentation model $\alpha_{\boldsymbol{\phi}}(\cdot)$, parameterized by $\boldsymbol{\phi}$. The anchor and positive word embeddings are compressed and concatenated into a single embedding, denoted as $\boldsymbol{X}_{\mathrm{c}} = g(\boldsymbol{X})$ and $\boldsymbol{X}'_{\mathrm{c}} = g(\boldsymbol{X}')$, respectively, which are then fed into the target model (VICNTM), consisting of an encoder and decoder parameterized by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathrm{enc}}, \boldsymbol{\theta}_{\mathrm{dec}}\}$. VIC regularization (Bardes et al., 2022) is applied to the inferred topic distributions $\boldsymbol{Z}$ and $\boldsymbol{Z}'$. Finally, the reconstructed BoW representations X_{recon} are used to compute the reconstruction error against the anchor BoW representations X_{BoW}. This model generates hard positive samples to enrich information and enhance NTM training, improving topic quality. The rest of this section details the adversarial framework, augmentation model, and target model.

Adversarial framework In this paper, we adopt TeachAugment (Suzuki, 2022) as the adversarial framework, which consists of three components: an augmentation model for generating positive samples from input samples, a target model trained on these positive samples, and a teacher model, implemented as the exponential moving average (EMA) of the target model. The adversarial framework is trained by alternately optimizing the target model to minimize its loss with a fixed augmentation model and optimizing the augmentation model to maximize the target model's loss while minimizing the teacher model's loss.

Augmentation model Similar to Chen et al. (2023), we build a noising network to produce informative positive samples by adding noise to the anchor samples. This is achieved by utilizing $\boldsymbol{e}_{\text{UNK}}$ and applying a multilayer perceptron (MLP) followed by a sigmoid function. As illustrated in Fig. 1, the weight γ_i determines the degree to which the anchor word embedding is retained. Each augmented embedding is computed as $\boldsymbol{w}_i' = \gamma_i \cdot \boldsymbol{w}_i + (1-\gamma_i) \cdot \boldsymbol{e}_{\text{UNK}}$.

Target model VICNTM incorporates VIC regularization into SCHOLAR(Card et al., 2018), utilizing the same sampling strategy as CLNTM (Nguyen and Luu, 2021). The model generates positive samples using the previously mentioned tf-idf-based sampling strategy, followed by applying VIC regularization to the anchor and positive latent topic representations. Given a minibatch with N documents, the model is trained by minimizing the reconstruction term, the Kullback-Leibler divergence term, and the regularization term, as shown below:

$$\begin{aligned} &\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{X}') = \mathcal{L}_{\text{NTM}} + \mathcal{L}_{\text{VICReg}}(\boldsymbol{Z}, \boldsymbol{Z}') \\ =\ & \Big(\textstyle\sum_i^N - \mathbb{E}_{q_{\theta_{\text{enc}}}(\boldsymbol{z}_i|\boldsymbol{x}_i)}[\log p_{\theta_{\text{dec}}}(\boldsymbol{x}_i|\boldsymbol{z}_i)] \\ +\ & \mathbb{KL}[q_{\theta_{\text{enc}}}(\boldsymbol{z}_i|\boldsymbol{x}_i) \| p(\boldsymbol{z}_i)] \Big) \\ +\ & \lambda s(\boldsymbol{Z}, \boldsymbol{Z}') + \mu[v(\boldsymbol{Z}) + v(\boldsymbol{Z}')] \\ +\ & \nu[c(\boldsymbol{Z}) + c(\boldsymbol{Z}')], \end{aligned} \tag{1}$$

where λ, μ, and ν are hyperparameters.

In this paper, the target model takes continuous representations from the augmentation model instead of discrete BoW representations. To leverage the rich information in word embeddings, we use the method by Xu et al. (2023), employing a CNN encoder to compress a sequence of 512 word embeddings (1024 dimensions each) into four embeddings of the same size. These are concatenated into a 4096-dimensional representation, which is then fed into the target model to infer its topic distribution.

Overall, our model is trained by optimizing the following min-max objective:

$$\begin{aligned} \max_{\boldsymbol{\phi}} \min_{\boldsymbol{\theta}} \mathbb{E}_{X\sim D} \quad & \Big[\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{X}_{\text{c}}, \boldsymbol{X}_{\text{c}}') \\ & - \mathcal{L}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}_{\text{c}}, \boldsymbol{X}_{\text{c}}') \Big]. \end{aligned} \tag{2}$$

The target model and augmentation model are trained in the same manner as in TeachAugment, with the teacher model parameterized by $\hat{\boldsymbol{\theta}}$ implemented as the EMA of the target model.

4 Experiments

Dataset	# Docs	Avg. Length	Split (%)
20NG	16469	$89_{\pm 152}$	48/12/40
IMDb	46304	$78_{\pm 54}$	50/25/25
Wiki	28590	$1320_{\pm 1057}$	70/15/15

Table 1: Dataset details.

We conducted experiments on three widely used datasets: 20Newsgroups (20NG) (Lang, 1995), IMDb movie reviews (IMDb) (Maas et al., 2011), and Wikitext-103 (Wiki) (Merity et al., 2017), to evaluate topic coherence (NPMI (Lau et al., 2014)) and topic diversity (TD (Dieng et al., 2020)) on the top ten words in each topic for topic numbers $K = 50$ and $K = 200$. Each experiment was run ten times with different random seeds. The datasets were preprocessed following Xu et al. (2025)'s approach, with modifications inspired by other approaches (Card et al., 2018; Xu et al., 2023). Table 1 summarizes the dataset details after preprocessing. The LLM encoder was implemented using BERT-large (Devlin et al., 2019), while the CNN encoder followed Xu et al. (2023)'s implementation. Hyperparameters, including batch size, the number of batches per update for the augmentation model, and the weights for the VIC regularization, were optimized using Optuna (Akiba et al., 2019). We compared our model against VAE-based approaches, including ProdLDA (Srivastava and Sutton, 2017), ECRTM (Wu et al., 2023), TSCTM (Wu et al., 2022), SCHOLAR (Card et al., 2018), CLNTM (Nguyen and Luu, 2021), and VICNTM (Xu et al., 2025).

Tables 2 and 3 present the results for NPMI and TD, respectively. Our model significantly outperformed other approaches on 20NG and IMDb in NPMI, the primary focus of this paper. However, the slightly lower NPMI on the Wiki dataset compared to other SCHOLAR-based models may result

Dataset	20NG		IMDb		Wiki	
K	50	200	50	200	50	200
ProdLDA	0.2347±0.0083	0.1739±0.0028	0.1075±0.0061	0.0735±0.0020	0.2554±0.0064	0.1916±0.0037
ECRTM	0.2354±0.0113	0.1630±0.0029	0.1048±0.0075	0.0605±0.0076	0.3799±0.0078	0.2457±0.0038
TSCTM	0.2469±0.0084	0.1571±0.0052	0.1262±0.0129	0.0787±0.0023	0.4250±0.0204	0.2075±0.0103
SCHOLAR	0.3519±0.0075	0.3122±0.0015	0.1551±0.0062	0.1274±0.0018	0.5138±0.0147	0.4571±0.0045
CLNTM	0.3530±0.0063	0.3115±0.0055	0.1568±0.0056	0.1255±0.0017	**0.5141**±0.0112	0.4564±0.0052
VICNTM	0.3543±0.0064	0.3148±0.0051	0.1558±0.0069	0.1272±0.0026	0.5090±0.0083	**0.4587**±0.0031
VICNTMxACE	**0.3632**±0.0046	**0.3452**±0.0055	**0.1678**±0.0065	**0.1353**±0.0065	0.5122±0.0149	0.4555±0.0047

Table 2: Results on NPMI when $K = 50$ and $K = 200$. Boldface indicates the optimal performance in each experiment.

Dataset	20NG		IMDb		Wiki	
K	50	200	50	200	50	200
ProdLDA	0.8858±0.0068	0.6892±0.0100	0.6694±0.0175	0.5809±0.0148	0.8364±0.0142	0.6248±0.0116
ECRTM	0.8790±0.0424	0.9544±0.0059	0.9616±0.0145	0.9409±0.1053	0.9806±0.0073	0.9118±0.0190
TSCTM	0.9302±0.0314	0.5508±0.0177	0.9772±0.0090	0.8570±0.0188	0.9878±0.0055	0.7871±0.0404
SCHOLAR	0.8874±0.0218	0.5037±0.0077	0.8778±0.0169	0.6895±0.0076	0.9912±0.0047	0.8221±0.0124
CLNTM	0.8904±0.0189	0.5084±0.0129	0.8592±0.0302	0.7033±0.0084	0.9876±0.0068	0.8223±0.0119
VICNTM	0.8878±0.0136	0.4998±0.0110	0.8712±0.0239	0.6947±0.0129	0.9842±0.0107	0.8242±0.0168
VICNTMxACE	0.8696±0.0162	0.2905±0.0137	0.8180±0.0650	0.1601±0.0310	0.9746±0.0294	0.7522±0.0231

Table 3: Results on TD when $K = 50$ and $K = 200$.

K	50		200	
	NPMI	TD	NPMI	TD
w/o TeachAugment	0.3528±0.0083	0.8736±0.0149	0.3427±0.0057	0.2910±0.0178
w/o word noising	0.3579±0.0075	0.8732±0.0238	0.3385±0.0073	0.2872±0.0204
w/o CNN	0.3577±0.0097	0.8766±0.0169	0.3341±0.0049	0.4142±0.0076
w/o LLM&CNN	0.3542±0.0068	0.8880±0.0159	0.3117±0.0052	0.5011±0.0065
VICNTMxACE	0.3632±0.0046	0.8696±0.0149	0.3452±0.0149	0.2905±0.0137

Table 4: Ablation study on the 20NG dataset.

from truncation, as the average document length exceeds 512 tokens. Although this may not apply to all datasets, due to the inverted pyramid structure, essential words are typically at the beginning of documents, while supplementary ones near the end may be truncated by the LLM tokenizer. CNN compression may further disregard less important words, reducing unique words and exacerbating this issue, especially as the number of topics increases, highlighting the need for an optimal K.

Additionally, Table 4 presents the results of the ablation study on the 20NG dataset. *w/o LLM&CNN*, which replaces embeddings with BoW representations, demonstrates that introducing the LLM encoder and the CNN encoder significantly improves NPMI. However, this improvement comes at the cost of reduced topic diversity, highlighting a clear trade-off between the two metrics. When the CNN encoder is replaced with an MLP encoder (*w/o CNN*), the performance reaches intermediate levels when $K = 200$. This suggests that the local feature extraction capability of the CNN encoder plays a crucial role in enhancing model performance. *w/o word noising* replaces the noising network in the augmentation model with an MLP $f(\cdot)$ so that $\boldsymbol{w}_i^{'} = f(\boldsymbol{w}_i)$, showing that the noising network contributes more when $K = 200$. *w/o TeachAugment* shows that when the number of topics is optimal, the TeachAugment framework contributes the most, demonstrating that adversarially generating positive samples is effective in our proposed model.

5 Conclusion

We proposed VICNTMxACE, a self-supervised NTM with adversarial data augmentation, leveraging word embeddings from an LLM encoder and a trainable augmentation model to generate positive samples. To our knowledge, this is the first adversarial framework applied to a VAE-based self-supervised NTM. Experiments show that our model outperforms its predecessor and other VAE-based NTMs in topic coherence, particularly on datasets with documents shorter than a given length. An ablation study further highlights that topic coherence benefits from adversarially generated informative positive samples and word embeddings.

6 Limitations

The introduction of the LLM encoder and the CNN encoder increases training time and computational resource requirements. Further improving topic coherence requires the document length being close to or shorter than the token limitation of the LLM. However, selecting an LLM with higher capacity would further increase computational costs. While we optimized several hyperparameters, those related to the CNN encoder remain unexplored. Furthermore, the positive examples generated by our model have not been demonstrated to be more informative than those generated by previous approaches. This will need to be explored in future work.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Adrien Bardes, Jean Ponce, and Yann Lecun. 2022. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *Proceedings of the 10th International Conference on Learning Representations*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022.

Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040.

Junfan Chen, Richong Zhang, Zheyan Luo, Chunming Hu, and Yongyi Mao. 2023. Adversarial word dilution as text data augmentation in low-resource regime. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12626–12634.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

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.

Xuemeng Hu, Rui Wang, Deyu Zhou, and Yuxuan Xiong. 2020. Neural topic modeling with cycle-consistent adversarial training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9018–9030.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on International Conference on Machine Learning*, pages 331–339.

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.

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

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:1280–1287.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of the 5th International Conference on Learning Representations*.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*.

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

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the 5th International Conference on Learning Representations*.

Teppei Suzuki. 2022. Teachaugment: Data augmentation optimization using teacher knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10904–10914.

Zhiqiang Tang, Yunhe Gao, Leonid Karlinsky, Prasanna Sattigeri, Rogerio Feris, and Dimitris Metaxas. 2020. Onlineaugment: Online data augmentation with less domain knowledge. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 313–329.

Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. 2020. Neural topic modeling with bidirectional adversarial training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 340–350.

Rui Wang, Deyu Zhou, and Yulan He. 2019. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098.

Xing Wei and W Bruce Croft. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185.

Max Welling and Diederik P Kingma. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, and Anh Tuan Luu. 2023. Effective neural topic modeling with embedding clustering regularization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 37335–37357.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2748–2760.

Weijie Xu, Wenxiang Hu, Fanyou Wu, and Srinivasan Sengamedu. 2023. Detime: Diffusion-enhanced topic modeling using encoder-decoder based llm. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 9040–9057.

Weiran Xu, Kengo Hirami, and Koji Eguchi. 2025. Self-supervised learning for neural topic models with variance-invariance-covariance regularization. *Zenodo*.

Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. 2020. Adversarial autoaugment. In *International Conference on Learning Representations*.

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

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.