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

Anonymous ACL submission

Abstract

Neural topic models (NTMs) have become an increasingly important component of topic modeling due to their flexibility and extensibility, which have facilitated various advancements, including the incorporation of self-supervised learning. Self-supervised NTMs construct contrastive samples either in the document representation space or the topic representation space, aiming to optimize the relationship between anchor and contrastive samples. However, previous approaches often rely on tf-idf-based augmentation strategies, which produce contrastive samples with limited informativeness, constraining their effectiveness in enhancing topic quality. To address this limitation, we propose an extension of the predecessor model into an adversarial framework, where positive samples are dynamically generated in the embedding space by a trainable augmentation model. Our approach further integrates contextualized word embeddings extracted from large language models (LLMs), enhancing the semantic richness of the generated samples. Extensive experiments demonstrate that our model consistently outperforms existing methods in terms of topic coherence, validating the effectiveness of adversarial learning for self-supervised NTMs.

1 Introduction

Topic modeling statistically leverages word co-occurrence patterns in each document to extract latent structural information called topics from large text corpora. The topics can be used for various downstream 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 the most representative conventional topic model. With the development of deep learning and advances in hardware like GPUs, neural topic models (NTMs) built on variational autoencoder (VAE) (Welling and Kingma, 2014) framework have become an increasingly important part of topic modeling. The greater flexibility and extensibility of NTMs have led to various extensions, including those that integrate self-supervised learning. These models construct contrastive samples to learn better topic representations, ultimately improving topic quality (Nguyen and Luu, 2021; Wu et al., 2022; Han et al., 2023).

Self-supervised NTMs have employed various strategies to build contrastive samples. CLNTM (Nguyen and Luu, 2021) generates contrastive bag-of-words (BoW) samples by modifying the values of unimportant and salient words, identified based on their words' term frequency-inverse document frequency (tf-idf) values. Specifically, positive samples are constructed by replacing the values of unimportant words with their counterparts in the reconstructed BoW representation, while negative samples are created by replacing the values of salient words with their reconstructed conuterparts. VICNTM (Xu et al., 2025) adopts 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 faces the problem that the positive samples become closer to the anchor samples as training progresses, providing limited guidance to the training process.

Adversarial data augmentation, widely used in computer vision to improve model generalization, generates informative positive samples. These methods typically use augmentation models to create samples that maximize task loss for the target classification model while applying constraints to prevent collapse (Zhang et al., 2020; Tang et al., 2020; Suzuki, 2022). TeachAugment (Suzuki, 2022), on the other hand, introduces a teacher model to guide the augmentation process. Un-

like previous approaches, TeachAugment requires 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. Furthermore, it can be applied to unsupervised models.

For text data augmentation, representation-level augmentation methods, which generate adversarial samples by adding adversarial perturbations to anchor samples in the embedding space, are commonly employed in adversarial frameworks (Miyato et al., 2017; Zhu et al., 2020). However, the noise derived from gradients used to minimize the model objective often lacks interpretability and flexibility. To address these issues, Chen et al. (2023) proposed an adversarial framework for training text classification models in low-resource scenarios. In this framework, hard positive samples are generated by weighted mixing embeddings of important words with unknown-word embeddings, thereby improving model 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. Our motivation is to enhance the performance of VICNTM by optimizing the generation of positive samples. To achieve this, we apply the adversarial framework proposed by Suzuki (2022) to the self-supervised NTM, utlizing the representation-level augmentation strategy inspired by Chen et al. (2023) as the augmentation model within this adversarial framework. Since we aim to augment the anchor samples in the embedding space, we replace BoW representations used in VICNTM with word embeddings. Drawing inspiration from Xu et al. (2023), we represent each input as word embeddings encoded by BERT (Devlin et al., 2019), which are then compressed by a CNN encoder. Experimental results on three widely used datasets demonstrate that our model outperforms baseline and state-of-the-art VAE-based models in terms of topic coherence, quantitatively. Additionally, we conducted an ablation study to further verify the effectiveness of each newly added component. Our contributions are summarized as follows:

1. We introduce VICNTMxACE, an enhanced version of VICNTM that incorporates a convenient adversarial framework to optimize the generation of positive samples, thereby boosting model performance. To the best of our knowledge, this is the first study to explore self-supervised NTMs in an adversarial setting.

2. Unlike traditional BoW-based augmentation methods, we introduce a word embedding space augmentation strategy that generates positive samples in the embedding space. This strategy aligns with the adversarial framework's requirement for trainable and informative hard positive samples.

3. We also introduce a CNN encoder to compress the word embeddings of each document into a unified document representation, enabling effective input for the NTM.

4. Extensive experiments demonstrate that our approach surpasses the state-of-the-art VAE-based NTMs in terms of topic coherence, further validated by both standard metrics and LLM-based evaluations.

2 Related works

Research on NTMs has become an integral part of topic modeling. ProdLDA (Srivastava and Sutton, 2017) was the first NTM to implement topic modeling using the VAE framework, with a logistic normal prior approximating the Dirichlet prior. By incorporating the log-frequency of words and refining implementation details from ProdLDA, SCHOLAR (Card et al., 2018), a general NTM capable of incorporating external information, achieved significantly improved topic quality compared to its predecessor.

Meanwhile, NTMs based on adversarial frameworks have been explored using generative adversarial networks (Goodfellow et al., 2014), where a generator creates negative samples and a discriminator distinguishes them from true samples (Wang et al., 2019, 2020; Hu et al., 2020). However, Nguyen and Luu (2021) showed that representation learning benefits more from the mutual information between positive and anchor samples than from negative samples. Building on SCHOLAR, they proposed CLNTM, which leverages both positive and negative samples generated based on the tf-idf values of anchor 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

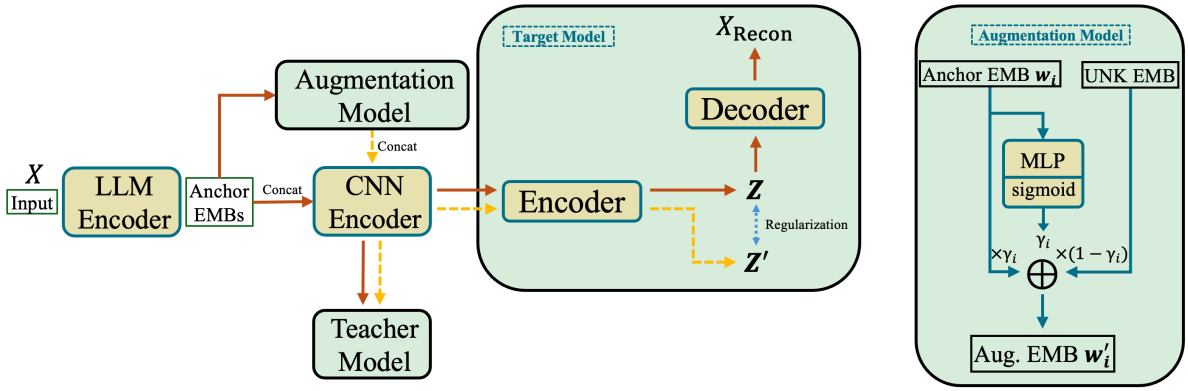

Figure 1: Illustration of our model. The left part of the figure depicts the structure of the model, with red (solid) lines representing the flow of anchor samples and yellow (dashed) lines indicating the flow of positive samples. The right part of the figure illustrates the structure of the augmentation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

among samples within each group. Contrastive learning has also been utilized in other NTMs to improve topic quality in various ways (Wu et al., 2022; Han et al., 2023).

Unlike the adversarial topic models, our model consists of a trainable augmentation model, a self-supervised NTM as the target model, and a teacher model. These components will be described in detail 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 $\boldsymbol{X}^{'} = \alpha_{\phi}(\boldsymbol{X})$ are generated through the augmentation model $\alpha_{\phi}(\cdot)$, parameterized by ϕ. The anchor and positive word embeddings are compressed and concatenated into a single embedding, denoted as $\boldsymbol{X}_{\text{c}} = g(\boldsymbol{X})$ and $\boldsymbol{X}_{\text{c}}^{'} = g(\boldsymbol{X}^{'})$, respectively, which are then fed into the target model (VICNTM), consisting of an encoder and decoder parameterized by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}\}$. VIC regularization (Bardes et al., 2022) is applied to the topic distributions $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ inferred by the encoder $\boldsymbol{\theta}_{\text{enc}}$. Finally, the BoW representations X_{recon} reconstructed by the decoder $\boldsymbol{\theta}_{\text{dec}}$ are used to compute the reconstruction error against the anchor BoW representations X_{BoW}. This model aims to generate hard positive samples that provide richer information to better guide the training process of the NTM, thereby improve topic quality. The remainder of this section presents the adversarial framework, followed by detailed descriptions of the target model and the augmentation model. Finally, we explain how word embeddings of each samples are fed into the NTM.

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 for image classification, and a teacher model, implemented as the exponential moving average (EMA) of the target model. Originally, TeachAugment generates positive image representations through a trainable augmentation model that applies geometric augmentation and color augmentation to the input images. The adversarial framework is trained by alternately optimizing the target model to minimize its classification 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. This alternating optimization ensures that the generated positive samples remain challenging for the target model while still being recognizable by the teacher model. In our work, we extend this framework to text data augmentation by introducing a tailored augmentation model. Additionally, both anchor samples and positive samples are simultaneously fed into the target and teacher models.

Augmentation model To generate positive sam-

ples for text classification in an adversarial framework, Chen et al. (2023) proposed Adversarial Word Dilution, which neutralize strongly positive words by introducing a dilution network for each class. This network produces a dilution weight for each word embedding in a given text sequence, creating a diluted word embedding through a weighted combination of the original word embedding and the embedding for an unknown token ([UNK]) extracted from the LLM encoder. Inspired by Chen et al. (2023), we design a noising network to generate informative positive samples by injecting noise into anchor samples. Our approach leverages the unknown token embedding, denoted as $\boldsymbol{e}_{\text{UNK}}$, and applies a multilayer perceptron (MLP) followed by a sigmoid activation to produce a weight coefficient γ_i for each word. This coefficient determines the retention degree of the original word embedding in the augmented representation. As illustrated in Fig. 1, the augmented embedding is computed as: $\boldsymbol{w}_i^{'} = \gamma_i \cdot \boldsymbol{w}_i + (1 - \gamma_i) \cdot \boldsymbol{e}_{\text{UNK}}$, where $\boldsymbol{w}_i$ represents the original word embedding and $\boldsymbol{w}_i^{'}$ is the generated positive word embedding. This mechanism ensures that the augmented samples remain semantically relevant while introducing sufficient perturbations for effective adversarial training.

Target model Our target model is VICNTM, a self-supervised NTM that integrates VIC regularization into SCHOLAR(Card et al., 2018). VICNTM leverages the same sampling strategy as CLNTM (Nguyen and Luu, 2021) to generate positive samples. Originally, the model employed the previously mentioned tf-idf-based sampling strategy, which were subsequently processed with VIC regularization to refine the latent topic representations of both anchor and positive samples. VIC regularization consists of three components:

- **Variance regularization**: Ensures diversity among latent topic representations within a mini-batch, preventing representation collapse. The variance regularization is defined as:

$$v(\boldsymbol{Z}) = \frac{1}{d}\sum_{j=1}^{d} \max(0, \tau - \sqrt{\text{Var}(\boldsymbol{Z}^j) + \epsilon}),$$

  where $\text{Var}(\boldsymbol{Z}^j)$ denotes the variance of the j-th dimension across the mini-batch.

- **Invariance regularization**: Minimizes the difference between the latent topic representations of anchor samples and their corresponding positive samples. The invariance regularization is defined as:

$$s(\boldsymbol{Z}, \boldsymbol{Z'}) = \frac{1}{n}\sum_{i=1} \|\boldsymbol{z}_i - \boldsymbol{z'}_i\|_2^2.$$

- **Covariance regularization**: Minimizes linear correlations across different topics, enhancing topic disentanglement and preventing redundancy in the learned representations. The covariance regularization is defined as:

$$\begin{aligned} c(\boldsymbol{Z}) &= \tfrac{1}{d}\textstyle\sum_{i\neq j}[C(\boldsymbol{Z})_{i,j}^2], \\ C(\boldsymbol{Z}) &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(\boldsymbol{z} - \bar{\boldsymbol{z}})(\boldsymbol{z} - \bar{\boldsymbol{z}})^\top, \\ \bar{\boldsymbol{z}} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\boldsymbol{z}_i. \end{aligned}$$

Given a minibatch of N documents, the model optimizes the NTM loss $\mathcal{L}_{\text{NTM}}$, which includes a reconstruction term and a Kullback-Leibler divergence term, alongside the VIC regularization term $\mathcal{L}_{\text{VICReg}}(\boldsymbol{Z}, \boldsymbol{Z'})$. The complete objective is defined as follows:

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

where λ, μ, and ν are the hyperparameters for invariance, variance, and covariance terms, respectively.

In this paper, the target model takes continuous representations output from the augmentation model, replacing the traditional discrete BoW representations. To effectively leverage the rich semantic information embedded in word vectors, we adopt the method proposed by Xu et al. (2023), which employs a CNN encoder to compress the sequence of word embeddings in a token-wise manner. Specifically, a sequency of 512 word embedding, each with a dimensionality of 1024, is compressed into a sequency of 4 embeddings of the same dimension. These compressed embeddings are subsequently concatenated to form a 4096-dimensional document representation, which is then fed into the target model to infer its topic distribution. The input to the target and the teacher models are denoted as $\boldsymbol{X}_{\text{c}}$ and $\boldsymbol{X}_{\text{c}}^{'}$, respectively, as aforementioned.

To achieve adversarial learning, our model is trained by optimizing the following min-max objective:

$$\max_{\phi} \min_{\theta} \mathbb{E}_{X \sim D} \Big[\mathcal{L}_{\theta}(\boldsymbol{X}_{\mathrm{c}}, \boldsymbol{X}'_{\mathrm{c}}) - \mathcal{L}_{\hat{\theta}}(\boldsymbol{X}_{\mathrm{c}}, \boldsymbol{X}'_{\mathrm{c}}) \Big], \quad (2)$$

where the target model (parameterized by θ) and the augmentation model (parameterized by ϕ) are trained alternately, following the procedure in TeachAugment. The teacher model, parameterized by $\hat{\theta}$, is implemented as the EMA of the target model, providing stable supervision during adversarial training.

4 Experiments

4.1 Setup

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 of top ten words in each topic for two different topic settings, $K = 50$ and $K = 200$. The datasets were preprocessed following the approach in Xu et al. (2025), with additional modifications inspired by Card et al. (2018) and Xu et al. (2023). The detailed statistics of the preprocessed datasets are summarized in Table 1.

We compared our model against several state-of-the-art 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).

For model implementation, we utlized BERT-large (Devlin et al., 2019) as the LLM encoder and adopted the CNN encoder from Xu et al. (2023). Hyperparameters, including batch size, the number of batches per update for the augmentation model, and the weights for the VIC regularization terms, were optimized using Optuna (Akiba et al., 2019). Each experiment was repeated with ten times with different random seeds to ensure statistiacal reliability.

We evaluate the quality of the learned topics using the following metrics:

- **Topic coherence**: We assessed topic coherence using well-established automated metrics, as detailed below.
  - **NPMI** (Lau et al., 2014): We evaluate the top ten words of each topic internally (using test data).
  - C_V (Röder et al., 2015): We evaluate each topic externally using a collection of Wikipedia articles as the reference corpus, employing the Palmetto tool [1].
- **LLM-based Evaluation Metrics**: To evaluate the semantic consistency and interpretability of the learned topics, we conducted LLM-based evaluation using Llama-3.1-8B-Instruct [2] to perform the two tasks below. The prompts templates used for each task are detailed in A.
  - **Intruder detection**: Following the approach in Stammbach et al. (2023), we prompt the LLM to identify an out-of-topic word among the top-ranked words of a topic. Specifically, for each topic, we select the top five words and introduce one additional word from a different topic that does not belong to the current topic distribution. The LLM is then asked to identify the intruder word. This task is evaluated using detection accuracy.
  - **Rating**: Following the approach in Stammbach et al. (2023), we prompt the LLM to rate the semantic coherence of the top ten words in each topic on a scale of 1 to 3.
- **Topic diversity**: Addtionally, we assess topic diversity (TD, Dieng et al. (2020)) to evaluate the breadth of discoverd topics. Detailed results and analysis are provided in B.

4.2 Results

In this section, we present the experimental results of our proposed model, VICNTMxACE, across

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

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

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	0.3839±0.0138	0.3243±0.0039	**0.4211**±0.0097	0.3516±0.0076	0.4717±0.0139	0.4062±0.0065
ECRTM	0.3629±0.0206	0.3158±0.0051	0.3427±0.0124	0.2608±0.0185	0.4548±0.0294	0.3679±0.0137
TSCTM	0.3237±0.0192	0.2998±0.0108	0.3616±0.0174	0.2815±0.0046	0.4693±0.0140	0.3542±0.0108
SCHOLAR	0.3975±0.0090	0.3700±0.0059	0.3975±0.0090	0.3700±0.0059	**0.5324**±0.0189	**0.5126**±0.0050
CLNTM	0.3948±0.0104	0.3688±0.0041	0.3813±0.0102	0.3550±0.0075	0.5200±0.0141	0.5052±0.0100
VICNTM	0.4014±0.0118	0.3724±0.0044	0.3690±0.0114	0.3541±0.0068	0.5238±0.0216	0.5122±0.0079
VICNTMxACE	**0.4067**±0.0123	**0.4024**±0.0108	0.3863±0.0114	**0.3869**±0.0340	0.5241±0.0191	0.5051±0.0102

Table 3: Results on C_V 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	0.3140±0.0550	0.2205±0.0199	0.3160±0.0548	0.2385±0.0519	0.4060±0.0822	0.3550±0.0369
ECRTM	0.3340±0.0481	0.2665±0.0362	0.2940±0.0795	0.2360±0.0296	0.4660±0.0647	0.3740±0.0347
TSCTM	0.2480±0.0828	0.2125±0.0279	0.2860±0.0558	0.2015±0.0246	0.4900±0.0886	0.2995±0.0353
SCHOLAR	0.3840±0.0440	0.3335±0.0333	0.3800±0.0525	0.3230±0.0396	0.6000±0.0625	0.5540±0.0258
CLNTM	**0.3960**±0.0788	0.3395±0.0316	0.4140±0.0718	0.3345±0.0393	0.6000±0.0442	**0.5705**±0.0244
VICNTM	0.3520±0.0634	0.3185±0.0302	0.4000±0.0516	0.3300±0.0196	0.5740±0.0859	0.5565±0.0342
VICNTMxACE	0.3740±0.0724	**0.3795**±0.0144	**0.4900**±0.0634	**0.5005**±0.0480	**0.6040**±0.0832	0.5455±0.0234

Table 4: Results on intruder detection task 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	2.3320±0.0612	2.1455±0.0134	2.5760±0.610	2.3725±0.0298	2.6530±0.0442	2.5315±0.0315
ECRTM	2.2980±0.0846	2.1515±0.0201	2.4100±0.0492	2.1130±0.0656	2.6540±0.0517	2.4090±0.0311
TSCTM	2.2040±0.0440	2.1145±0.0211	2.5460±0.0706	2.1975±0.0247	2.8040±0.0556	2.4465±0.0573
SCHOLAR	2.5440±0.0617	2.4565±0.0430	2.6660±0.0550	2.5380±0.0337	**2.9080**±0.0424	**2.9045**±0.0126
CLNTM	2.5160±0.0711	2.4545±0.0215	2.6680±0.0509	2.5235±0.0215	2.9000±0.0267	2.8965±0.0242
VICNTM	2.5340±0.0626	2.4640±0.0265	2.6780±0.0476	2.5415±0.0232	2.8980±0.0319	2.9000±0.0252
VICNTMxACE	**2.5480**±0.0535	**2.5360**±0.0464	**2.7200**±0.0680	**2.6490**±0.0825	2.9060±0.0299	**2.9045**±0.0083

Table 5: Results on rating task when $K = 50$ and $K = 200$. Boldface indicates the optimal performance in each experiment.

three datasets. The evaluation is conducted using both traditional topic coherence metrics (NPMI and C_V) and LLM-based methods (intruder detection and topic rating).

Tables 2 and 3 present the results of NPMI and C_V for the top ten words of each topic. Among the SCHOLAR-based NTMs (SCHOLAR, CLNTM, VICNTM, VICNTMxACE), our proposed VICNTMxACE consistently achieves the best performance on 20NG and IMDb for most settings, surpassing all other baselines. On the Wiki dataset, the slightly lower performance compared to other SCHOLAR-based models may due to the average document length exceeding the 512-token limit (as shown in Table 1), which may truncate critical information during encoding. Nonetheless, VICNTMxACE maintains competitive performance on this dataset.

To further evaluate the semantic consistency and interpretability of the learned topics, we performed LLM-based evaluations using intruder detection and rating tasks, as shown in Tables 4 and 5. Consistent with traditional topic coherence metrics, SCHOLAR-based models outperformed the baselines, with VICNTMxACE achieving the best results across most of the settings. In the intruder detection task, our model achieves the highest detection accuracy for most settings, indicating that the topics generated by VICNTMxACE are more distinguishable and semantically consistent. For the rating task, LLM assigned higher ratings to the top ten words of each topic produced by our model, suggesting a stronger alignment with semantic coherence. Notably, recent studies have demonstrated that LLM-based evaluations correlate more closely with human judgement than traditional automated metrics(Stammbach et al., 2023; Yang et al., 2025). Hence, the superior performance of VICNTMxACE in these tasks highlights its ability to generate topics that are not only coherent but also contextually meaningful.

Overall, VICNTMxACE consistently surpasses state-of-the-art baselines in both traditional and LLM-based evaluations, demonstrating the effectiveness of our adversarial learning strategy in improving topic coherence and interpretability.

4.3 Ablation study

To verify the contributions of each newly introduced component, we conducted an ablation study on the 20NG dataset. The results are summarized in Table 6. We first evaluate the model without both the LLM encoder and the CNN encoder, replacing them with traditional BoW representations (*w/o LLM&CNN*). The results show a noticeable drop in NPMI, indicating that the introduction of the LLM encoder and CNN encoder improves topic coherence. Next, we examine the impact of the CNN encoder by replacing it with an MLP encoder (*w/o CNN*). The results indicate that the performance degrades, especially when $K = 200$, highlighting the importance of local feature extraction provided by the CNN encoder. This suggests that the local feature extraction capability of the CNN encoder plays a crucial role in enhancing model performance. We also evaluate the effectiveness of the noising network in the augmentation model by replacing it with a simple MLP mapping $f(\cdot)$, such that the augmented embedding is computed as $\boldsymbol{w}_i^{'} = f(\boldsymbol{w}_i)$ (*w/o word noising*). The results show that the original noising network achieves better NPMI, indicating its role in generating harder positive samples that improve model performance. Finally, we assess the impact of the TeachAugment framework by removing the adversarial training mechanism (*w/o TeachAugment*). We observe that when the number of topics is optimally set, the adversarial generation of positive samples through TeachAugment significantly boosts topic coherence. This validates the effectiveness of our proposed augmentation strategy in refining topic quality.

Overall, the ablation study highlights the importance of each component in achieving optimal performance. The LLM encoder and CNN encoder enhance semantic representation, the noising network introduces harder positives for robust learning, and TeachAugment enables effective adversarial augmentation.

5 Conclusion

In this paper, we propose VICNTMxACE, a self-supervised NTM enhanced with adversarial data augmentation, building upon VICNTM. To generate richer and more informative positive samples, we integrate word embeddings extracted from an LLM encoder and introduce a trainable augmentation model. To the best of our knowledge, this is the first application of an adversarial framework in the context of self-supervised NTMs. Extensive experiments across multiple datasets demonstrate that VICNTMxACE consistently outperforms its predecessor (VICNTM) as well as other state-of-VAE-based NTMs. Our model achieves significant improvements in both traditional topic coherence

K	50	200
w/o TeachAugment	$0.3528_{\pm 0.0083}$	$0.3427_{\pm 0.0057}$
w/o word noising	$0.3579_{\pm 0.0075}$	$0.3385_{\pm 0.0073}$
w/o CNN	$0.3577_{\pm 0.0097}$	$0.3341_{\pm 0.0049}$
w/o LLM&CNN	$0.3542_{\pm 0.0068}$	$0.3117_{\pm 0.0052}$
VICNTMxACE	$0.3632_{\pm 0.0046}$	$0.3452_{\pm 0.0149}$

Table 6: Ablation study in terms of NPMI on the 20NG dataset.

metrics (NPMI and C_V) and LLM-based evaluations (intruder detection and rating tasks), validating its capability to generate semantically coherent and interpretable topics. The results of the ablation study further confirm the effectiveness of each component, highlighting the contributions of the LLM-based word embeddings, the CNN encoder, and the adversarially generated positive samples. These components collectively enhance the semantic richness of the learned topics, leading to improved topic quality. Overall, VICNTMxACE effectively leverages adversarial learning for data augmentation, leading to more semantically coherent and interpretable topic representations.

6 Limitations

The introduction of the LLM encoder and the CNN encoder increases overall training time and computational resource requirements. For reference, training a baseline model such as VICNTM on the 20NG dataset with an NVIDIA A40 GPU (48GB memory) typically takes around 5 minutes, whereas our method requires approximately six times longer due to the additional computational complexity. 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.

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*, pages 399–408.

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

Dominik Stammbach, 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.

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. *Knowledge and Information Systems*.

Xiaohao Yang, He Zhao, Dinh Phung, Wray Buntine, and Lan Du. 2025. Llm reading tea leaves: Automatically evaluating topic models with large language models. *Transactions of the Association for Computational Linguistics*, 13:357–375.

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*.

A Prompt Templates for LLM-based Evaluation

This section presents the prompt templates used for the LLM-based evaluation tasks, including intruder detection and rating tasks. All evaluations were conducted with Llama-3.1-8B-Instruct.

A.1 Intruder detection prompt

For the intruder detection task, the LLM is prompted to identify the word that does not belong semantically to the group of top topic words. The following template was used:

> **System prompt:** You are a helpful assistant evaluating the top words of a topic model output for a given topic. Select which word is the least related to all other words. If multiple words do not fit, choose the word that is most out of place. Reply with a single word. Do not provide me explanations.
>
> **User prompt:** processor, quadra, drink, motherboard, port, apple

The six words consist of five words of the top ten words from a single topic and one additional word sampled from another topic, ensuring it is not within the top 50 words of the current topic. The LLM's response is evaluated based on whether it correctly identifies the intruder. The final reported accuracy is averaged across 10 different random seeds.

A.2 Rating prompt

For the rating task, the LLM is prompted to assess the semantic coherence of the top ten words in each topic on a scale of 1 to 3. The following template was used:

> **System prompt:** You are a helpful assistant evaluating the top words of a topic model output for a given topic. Please rate how related the following words are to each other on a scale from 1 to 3 ("1" = not very related, "2" = moderately related, "3" = very related). Reply with a single number, indicating the overall appropriateness of the topic. Do not provide me explanations.
>
> **User prompt:** processor, board, quadra, simms, monitor, mhz, port, apple, motherboard, centris

The average rating for each model is computed over topics, and the final reported score is averaged across 10 different random seeds.

B Topic Diversity

In this section, we present the results of topic diversity across three datasets when when $K = 50$ and $K = 200$, as shown in Table 7. We observed that our model achieves comparable TD performance to other SCHOLAR-based NTMs when $K = 50$, but experiences a noticeable decline when $K = 200$. We identify two primary reasons for this observation:

- **Document truncation by LLM tokenizer**: Although this may not apply universally to all datasets, the inverted pyramid structure commonly found in documents places essential information at the beginning, while supplementary details tend to be near the end. Due to the token limit of the LLM encoder, these less critical but diverse words are often truncated, resulting in reduced topic diversity.

- **Compression by CNN**: The CNN encoder further compresses document representations by focusing on local patterns, which, while enhancing topic coherence, may also discard less informative words. This effect is amplified as the number of topics increases, leading to a narrower range of unique terms and a drop in TD.

We believe that selecting an optimal number of topics that aligns with the dataset's intrinsic structure is crucial for achieving better topic diversity. To validate this, we conducted an additional evaluation on 20NG with $K = 20$, which matches its 20 well-defined categories. As shown in Table 8, our model with $K = 20$ achieves topic diversity that is either superior to or comparable with other state-of-the-art models. Furthermore, our model consistently outperforms the baselines in terms of

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

Table 7: Results on TD when $K = 50$ and $K = 200$.

$K = 20$	TD	NPMI
ProdLDA	$0.9590_{\pm 0.0126}$	$0.2628_{\pm 0.0131}$
ECRTM	$0.9290_{\pm 0.0497}$	$0.3394_{\pm 0.0391}$
TSCTM	$0.9825_{\pm 0.0098}$	$0.3670_{\pm 0.0247}$
SCHOLAR	$0.9845_{\pm 0.0154}$	$0.3962_{\pm 0.0177}$
CLNTM	$0.9825_{\pm 0.0106}$	$0.3894_{\pm 0.0127}$
VICNTM	$0.9795_{\pm 0.0169}$	$0.3944_{\pm 0.0109}$
VICNTMxACE	$0.9825_{\pm 0.0138}$	$0.3977_{\pm 0.0100}$

Table 8: Results on TD and NPMI when $K = 20$ on the 20NG dataset.

NPMI, highlighting its ability to generate more coherent and diverse topics. This is in contrast to the results for $K = 50$ and $K = 200$ presented in Table 7. These findings highlight the importance of aligning the number of topics with the dataset characteristics to maximize diversity. Under such conditions, our model demonstrates strong performance in terms of topic diversity while maintaining topic coherence.