Neural Topic Modeling based on Cycle Adversarial Training and Contrastive Learning

Boyu Wang[1] **Linhai Zhang**[1] **Deyu Zhou**[1*]
Yi Cao[2] **Jiandong Ding**[2]
[1] School of Computer Science and Engineering, Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, China
[2] Huawei Technologies Co., Ltd., China
{wby1999, lzhang472, d.zhou}@seu.edu.cn
{caoyi23, dingjiandong2}@huawei.com

Abstract

Neural topic models have been widely used to extract common topics across documents. Recently, contrastive learning has been applied to variational autoencoder-based neural topic models, achieving promising results. However, due to the limitation of the unidirectional structure of the variational autoencoder, the encoder is enhanced with the contrastive loss instead of the decoder, leading to a gap between model training and evaluation. To address the limitation, we propose a novel neural topic modeling framework based on cycle adversarial training and contrastive learning to apply contrastive learning on the generator directly. Specifically, a self-supervised contrastive loss is proposed to make the generator capture similar topic information, which leads to better topic-word distributions. Meanwhile, a discriminative contrastive loss is proposed to cooperate with the self-supervised contrastive loss to balance the generation and discrimination. Moreover, based on the reconstruction ability of the cycle generative adversarial network, a novel data augmentation strategy is designed and applied to the topic distribution directly. Experiments have been conducted on four benchmark datasets and results show that the proposed approach outperforms competitive baselines.

1 Introduction

Topic modeling, uncovering the semantic structures within a collection of documents, has been widely used in various natural language processing (NLP) tasks (Zhou et al., 2017; Yang et al., 2018, 2019; Zhou et al., 2021; Wang et al., 2022). Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a probabilistic graphical model, is one of the most popular topic models due to its interpretability and effectiveness. However, the parameter estimation methods for LDA and its variants, such as collapsed Gibbs sampling (Griffiths and Steyvers, 2004), are model-specific and require specialized derivations.

To tackle such disadvantages, neural topic models have been proposed with a flexible training process, which can be divided into two categories, variational autoencoder (VAE) based and generative adversarial network (GAN) based. VAE-based neural topic models regard the encoded latent vector as the topic distribution of the input document, then employ the decoder to reconstruct the word distribution (Miao et al., 2016; Srivastava and Sutton, 2017; Miao et al., 2017; Card et al., 2018; Wang et al., 2021). To address the limitation that VAE-based neural topic models cannot approximate Dirichlet distribution precisely, Wang et al. (2019) propose an adversarial topic model, in which the topic distribution is sampled from the Dirichlet prior distribution directly and transformed into the word distribution by the generator. In order to uncover topic distribution and infer document topic simultaneously, Bidirectional Adversarial Topic model (Wang et al., 2020) and Topic Modeling with Cycle-consistent Adversarial Training (Hu et al., 2020) have been proposed in turn.

Recently, a neural topic model named CLNTM has been proposed to apply contrastive learning to VAE-based neural topic models (Nguyen and Luu, 2021). A data augmentation strategy is proposed to replace the salient and non-salient parts of the document representation according to word frequency information to construct positive and negative examples. Although achieving promising results, it has such a disadvantage. As shown in Figure 1, due to the limitation of the unidirectional structure of VAE, the encoder is optimized through the contrastive loss, instead of the decoder which generates topic-word distribution, leading to the gap between model training and evaluation.

Therefore, in this paper, we consider discovering topics based on cycle adversarial training and contrastive learning. As illustrated in the lower left part of Figure 1, by incorporating cycle adversarial training, topic distribution θ and word

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 9720–9731

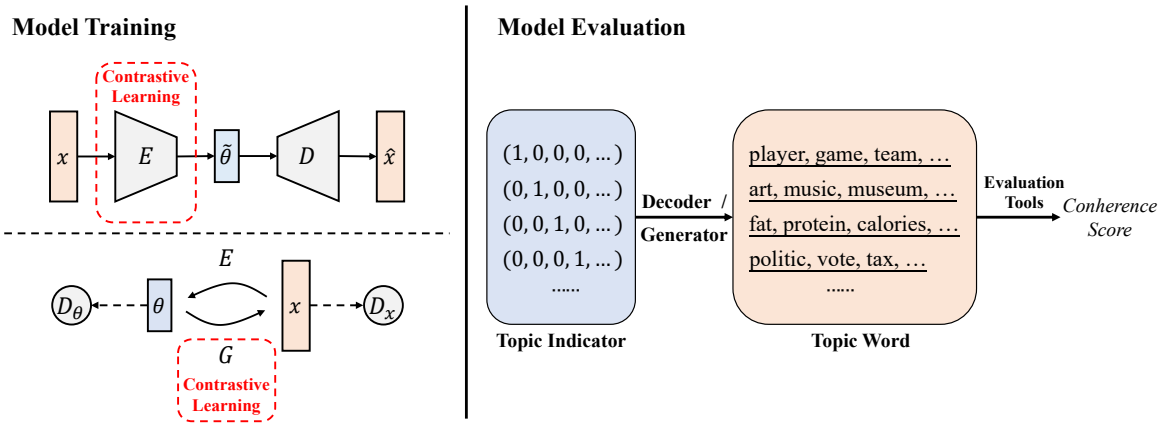

Figure 1: Difference between CLNTM (Nguyen and Luu, 2021) and the proposed approach.

distribution $\boldsymbol{x}$ will be transformed bidirectionally, breaking the structural limitations of VAE-based neural topic models. However, it is not straightforward to combine contrastive learning and cycle adversarial training. On the one hand, it is crucial to construct applicable positive samples for topic distributions. On the other hand, it is hard to improve the learning of topic-word distribution while maintaining the bidirectional mapping ability of cycle adversarial training.

To overcome the above challenges, we propose a novel Neural Topic Modeling framework based on Adversarial training and Contrastive Learning (NTM-ACL). A self-supervised contrastive loss is employed to make the generator capture similar topic information between positive pairs. The generation of topic-word distribution is improved directly, which mitigates the gap between model training and evaluation. Meanwhile, a discriminative contrastive loss is designed to cooperate with supervised contrastive loss to avoid the adversarial training being undermined by the unbalance between generation and discrimination. Moreover, data augmentation is applied to construct positive samples of topic distribution with the reconstruction ability of cycle generative adversarial network structure. The minimum items in the reconstructed distribution are substituted for corresponding items in the original distribution, which hasn't been explored before. We conduct extensive experiments to fully exploit the effectiveness of our proposed model.

In a nutshell, the main contributions of our paper can be summarized as follows:

- We propose NTM-ACL, a novel neural topic modeling framework where contrastive learning is directly applied to the generation of topic-word distribution.
- We propose a novel data augmentation strategy for topic distribution based on the reconstruction ability of cycle adversarial training. To the best knowledge, we are the first to apply data augmentation to construct positive samples of topic distribution.
- We conduct extensive experiments and experimental results show that NTM-ACL outperforms several competitive baselines on four benchmark datasets.

2 Related Work

Our work is mainly related to two lines of research, including neural topic models and contrastive learning.

2.1 Neural Topic Model

Inspired by VAE, Miao et al. (2016) proposed Neural Variational Document Model (NVDM) for text modeling, employing Gaussian as the prior distribution of latent topics. Following that, (Srivastava and Sutton, 2017; Card et al., 2018) proposed to approximate Dirichlet distribution with a logistic normal prior distribution. To break the limitation of VAE, Wang et al. (2019) proposed an Adversarial Topic Model (ATM), which consists of a generator and a discriminator. The generator maps the topic distribution randomly sampled from the Dirichlet prior distribution to the word distribution, and the discriminator judges whether the word distribution comes from real documents or is generated by the generator. The two modules are trained adversarially against each other. In order to realize topic mining and document topic inference simultaneously,

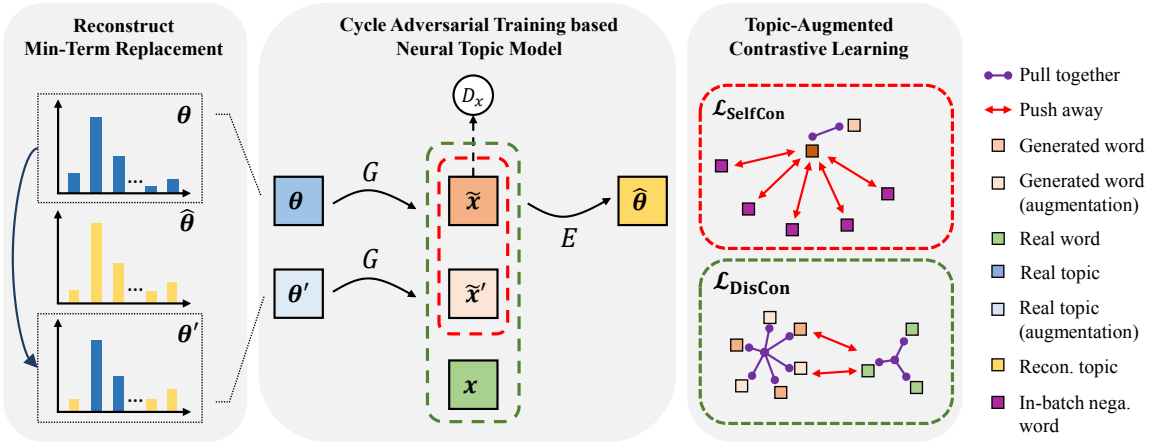

Figure 2: The architecture of the proposed model, NTM-ACL.

Bidirectional Adversarial Topic (BAT) (Wang et al., 2020) constructs two-way adversarial training on the basis of ATM. Hu et al. (2020) propose Topic Modeling with Cycle-consistent Adversarial Training (ToMCAT) to realize the transformation between a topic distribution and word distribution, inspired by Cycle-GAN (Zhu et al., 2017).

2.2 Contrastive Learning

Contrastive learning, as a self-supervised learning method, improves the transforming ability of models without large-scale labeled data and becomes a popular technique in computer vision domain (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Grill et al., 2020; Zhao et al., 2021). Chen et al. (2020a) proposed SimCLR, applied image transformations to generate two positive samples for each image randomly, and used the normalized temperature-scaled cross-entropy loss (NT-Xent) as the training loss to make positive pair close in the representation space.

With the success of contrastive learning in computer vision tasks, recent studies attempt to extend it to other domains. Khosla et al. (2020) extended the self-supervised approach to the fully-supervised setting, allowing models to effectively leverage label information. Jeong and Shin (2021) proposed ContraD to incorporate a contrastive learning scheme into GAN. In natural language processing, contrastive learning is widely applied to various tasks, such as sentence embedding, text classification, information extraction, and stance detection (Gao et al., 2021; Yan et al., 2021; Zhang et al., 2022; Wu et al., 2022; Chuang et al., 2022; Liang et al., 2022). In neural topic modeling, contrastive learning has been used to improve the VAE-based neural topic model by adding a contrastive objective to the training loss and taking a more principled approach to creating positive and negative samples (Nguyen and Luu, 2021).

3 Method

The overall architecture of the proposed NTM-ACL is shown in Figure 2, which consists of three parts: 1) **Cycle Adversarial Training based Neural Topic Model**, which includes the generator, the encoder, and discriminators to transform topic distribution and word distribution bidirectionally; 2) **Topic-Augmented Contrastive Learning**, which includes the Self-supervised contrastive loss and the Discriminative contrastive loss to enhance the generator without affecting the adversarial training; 3) **Reconstruct Min-Term Replacement**, which is based on reconstruction ability of cycle generative adversarial network to create positive samples of topic distributions.

3.1 Problem Setting

We denote corpus as $\mathcal{D}$, which consists of M documents $\{\boldsymbol{x}_i\}_{i=1}^{M}$. Given the document $\boldsymbol{x}_i \in \mathbb{R}^V$ where V is the vocabulary size, the first purpose of topic modeling is **topic inference**, inferring the corresponding topic distribution $\boldsymbol{\theta}_i \in \mathbb{R}^K$ where K is the number of topics.

To formalize topic modeling, we use X to stand for word distribution set where the document is represented in normalized Term Frequency Inverse Document Frequency (TF-IDF), use Θ to stand for topic distribution set where topic distribution is sampled from a Dirichlet distribution with parameter $\boldsymbol{\alpha} \in \mathbb{R}^K$.

During the training process, we need to learn two

mapping functions, *generator* G and *encoder* E. G transforms samples from Θ into X while E is the reverse function of G. After G is well-trained, the indicator vector of each topic is input to get the topic-word distribution. This is another purpose of topic modeling, referred as **topic discovery**. The one-hot vector $\boldsymbol{I}_k \in \mathbb{R}^K$ denotes the indicator vector of the k-th topic, where the value at the k-th index is 1.

3.2 Cycle Adversarial Training

Following (Hu et al., 2020), NTM-ACL consists of two mapping functions, *generator* G: $\Theta \rightarrow X$, *encoder* E: $X \rightarrow \Theta$ and their related discriminators, $D_{\boldsymbol{x}}$ and $D_{\boldsymbol{\theta}}$. They are all implemented in the structure of a three-layer multi-layer perceptron (MLP), with a H-dim hidden layer using LeakyReLU as an active function and batch normalization, followed by an output layer using softmax. The cycle adversarial training objective is composed of adversarial loss and cycle consistency loss. We apply Wasserstein GAN (WGAN) (Arjovsky et al., 2017) adversarial losses to train G and corresponding discriminator $D_{\boldsymbol{x}}$:

$$\begin{aligned}\mathcal{L}_{\text{adv}}\ (G, D_{\boldsymbol{x}}) =& \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[D_{\boldsymbol{x}}(\boldsymbol{x})\right] - \\ & \mathbb{E}_{\boldsymbol{\theta}\sim p_{\text{data}}(\boldsymbol{\theta})}\left[D_{\boldsymbol{x}}(G(\boldsymbol{\theta}))\right],\end{aligned} \tag{1}$$

in which G tries to generate word distributions similar to samples in X, while $D_{\boldsymbol{x}}$ aims to distinguish generated samples and real samples. G aims to minimize this objective against an adversary $D_{\boldsymbol{x}}$ that tries to maximize it.

To further constrain the relationship between origin distribution and target distribution, we additionally use cycle-consistency losses, encouraging G and E to reconstruct the origin distribution. Cycle-consistency losses are implemented as follows:

$$\begin{aligned}\overrightarrow{\mathcal{L}_{\text{cyc}}}(G, E) &= \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta})}\left[\|E(G(\boldsymbol{\theta})) - \boldsymbol{\theta}\|_1\right], \\ \overleftarrow{\mathcal{L}_{\text{cyc}}}(G, E) &= \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\|G(E(\boldsymbol{x})) - \boldsymbol{x}\|_1\right],\end{aligned} \tag{2}$$

where $\|\cdot\|_1$ denotes L1 norm. Combining adversarial loss and cycle-consistency loss, the objective of cycle adversarial training is:

$$\begin{aligned}\mathcal{L}_{\text{Cyc-adv}} = & \mathcal{L}_{\text{adv}}\ (G, D_x) + \mathcal{L}_{\text{adv}}\ (E, D_\theta) + \\ & \lambda_1 \overrightarrow{\mathcal{L}_{\text{cyc}}}(G, E) + \lambda_2 \overleftarrow{\mathcal{L}_{\text{cyc}}}(G, E),\end{aligned} \tag{3}$$

where λ_1 and λ_2 control the importance of the losses respectively.

3.3 Data Augmentation

In this subsection, we describe how to apply data augmentation to construct positive samples of topic distribution, which takes advantage of the structural features of cycle adversarial training.

Given distribution $\boldsymbol{\theta}_i = \{\theta_{i1}, \theta_{i2}, \cdots, \theta_{iK}\}$, the reconstructed distribution $\hat{\boldsymbol{\theta}}_i$ created by the cycle of G and E is similar to the original distribution: $\boldsymbol{\theta}_i \rightarrow G(\boldsymbol{\theta}_i) \rightarrow E(G(\boldsymbol{\theta}_i))$=$\hat{\boldsymbol{\theta}}_i \approx \boldsymbol{\theta}_i$. Meanwhile, we hypothesize that items with the maximum value in $\boldsymbol{\theta}_i$ indicate the salient topic information, which has a significant effect on generating the word distribution $\tilde{\boldsymbol{x}}_i$ through G. On the contrary, items with the minimum values in $\boldsymbol{\theta}_i$ have limited effects. After making slight modifications to them, G can still generate a word distribution similar to $\tilde{\boldsymbol{x}}_i$.

Based on the above assumptions, we propose a data augmentation strategy for topic distribution named Reconstruct Min-Term Replacement (RMR). For the reconstructed topic distribution $\hat{\boldsymbol{\theta}}_i = \{\hat{\theta}_{i1}, \hat{\theta}_{i2}, \cdots, \hat{\theta}_{iK}\}$, we select the minimum p items from it. The indices of these items in $\hat{\boldsymbol{\theta}}_i$ are denoted as $\{a_1, a_2, \cdots, a_p\}$. We replace the value at the corresponding index in $\boldsymbol{\theta}_i$:

$$\theta_{ia_j} = \hat{\theta}_{ia_j} (1 \leq j \leq p), \tag{4}$$

For the topic distribution $\boldsymbol{\theta}_i$, we denote its data-augmented distribution as $\boldsymbol{\theta}'_i$. Correspondingly, the topic distribution set Θ after data augmentation is denoted as Θ'.

3.4 Topic-Augmented Contrastive Learning

In this subsection, we will introduce Topic-Augmented Contrastive Learning, which enhances G while keeping the balance of generation and discrimination. This part mainly consists of two training objectives, Self-supervised contrastive loss, and Discriminative contrastive loss.

Self-supervised contrastive loss We follow the setting in SimCLR(Chen et al., 2020a), use Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent Loss) to calculate the Self-supervised contrastive loss $\mathcal{L}_{\text{SelfCon}}$. $\mathcal{L}_{\text{SelfCon}}$ helps improving the mapping ability of G, capturing similar topic information to generate better topic-word distribution. Self-supervised contrastive loss pulls word distributions of positive topic distribution pairs together while pushing away distance between the word distributions corresponding to the negative sample pairs, which is shown in the upper right part

in Figure 2. Given representation $\boldsymbol{r}_i$, its positive sample is denoted as $\boldsymbol{r}_i^+$, and the set of its negative samples is recorded as $\boldsymbol{r}^-$, the NT-Xent Loss between $\boldsymbol{r}_i$, $\boldsymbol{r}_i^+$ and $\boldsymbol{r}^-$ is:

$$l(\boldsymbol{r}_i, \boldsymbol{r}_i^+, \boldsymbol{r}^-) = -log\frac{exp(\boldsymbol{r}_i \cdot \boldsymbol{r}_i^+/\tau)}{exp(\boldsymbol{r}_i \cdot \boldsymbol{r}_i^+/\tau) + \sum_{j=1}^{|\boldsymbol{r}^-|} exp(\boldsymbol{r}_i \cdot \boldsymbol{r}_j^-/\tau)}, \tag{5}$$

where τ is a temperature hyperparameter.

Assuming that topic set Θ of the current training batch contains N samples, we get Θ' after data augmentation and the number of training samples is expanded to $2N$. Two training sets are transformed to $\tilde{X}$ and $\tilde{X}'$ respectively. For word distribution $\tilde{\boldsymbol{x}}_i$ in $\tilde{X}$, we can find its positive sample $\tilde{\boldsymbol{x}}_i'$ in $\tilde{X}'$. The remaining $2N-2$ word distributions form the set of negative samples, denoted as X_i^-:

$$X_i^- = [\tilde{X} \setminus \tilde{\boldsymbol{x}}_i; \tilde{X}' \setminus \tilde{\boldsymbol{x}}_i'] \tag{6}$$

Based on the above description, we define $\mathcal{L}_{\text{SelfCon}}$ as follow:

$$\mathcal{L}_{\text{SelfCon}} = \frac{1}{2N}\sum_{i=1}^{N}\left[l\left(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_i', X_i^-\right) + l\left(\tilde{\boldsymbol{x}}_i', \tilde{\boldsymbol{x}}_i, X_i^-\right)\right], \tag{7}$$

Discriminative contrastive loss The Self-supervised contrastive loss $\mathcal{L}_{\text{SelfCon}}$ can make the generator better perceive the similarity between two topics, then generate topic-word distributions that are more in line with the corresponding topics. However, only improving the mapping ability leads to an imbalance between generation and discrimination, which undermines the performance of cycle adversarial training. Therefore, we additionally design a discriminative contrastive loss $\mathcal{L}_{\text{DisCon}}$, leveraging category information of real samples and generated samples to keep the balance of generation and discrimination.

It is obvious that samples in X belong to the real category, while samples in $\tilde{X}$ and $\tilde{X}'$ belong to the generated category. For any $\boldsymbol{x}_i$ in X, we denote $U_i = [X \setminus \boldsymbol{x}_i; \tilde{X}; \tilde{X}']$. The main purpose of discriminative contrastive loss is not to focus on the similarity between the positive sample pair but make samples of the same category closer. We define the discriminative contrastive loss between $\boldsymbol{x}_i$ and U_i as:

$$l_{Dis}(\boldsymbol{x}_i, U_i) = -\frac{1}{|X \setminus \boldsymbol{x}_i|}\sum_{\boldsymbol{x}_+ \in X \setminus \boldsymbol{x}_i} log\frac{exp(\boldsymbol{x}_i \cdot \boldsymbol{x}_+/\tau)}{\sum_{j=1}^{|U_i|} exp(\boldsymbol{x}_i \cdot \boldsymbol{x}_j/\tau)}, \tag{8}$$

where $\boldsymbol{x}_+$ stands for samples of the same category as $\boldsymbol{x}_i$.

For the whole batch, we define discriminative contrastive loss $\mathcal{L}_{\text{DisCon}}$ as:

$$\mathcal{L}_{\text{DisCon}} = \frac{1}{N}\sum_{i=1}^{N} l_{Sup}(\boldsymbol{x}_i, U_i) \tag{9}$$

Overall Training Objective Summing up $\mathcal{L}_{\text{Cyc-adv}}$, $\mathcal{L}_{\text{SelfCon}}$ and $\mathcal{L}_{\text{DisCon}}$, the overall training objective of our model is:

$$\mathcal{L} = \mathcal{L}_{\text{Cyc-adv}} + \lambda_3\mathcal{L}_{\text{SelfCon}} + \lambda_4\mathcal{L}_{\text{DisCon}}, \tag{10}$$

where λ_3 and λ_4 control the relative significance of Self-supervised contrastive loss and Discriminative contrastive loss respectively. At each training iteration, the parameters of G and E are updated once after parameters of $D_{\boldsymbol{\theta}}$ and $D_{\boldsymbol{x}}$ have been updated 5 times.

4 Experiments

4.1 Datasets

We conduct experiments on four datasets: NYTimes[1](**NYT**), Grolier[2](**GRL**), DBPedia[3](**DBP**) and 20Newsgroups[4](**20NG**). We follow the same processing as (Wang et al., 2019). The statistics of the processed datasets are shown in Table 1.

Dataset	#Documents	Vocabulary Size
NYTimes	99,992	12,604
Grolier	29,762	15,276
DBPedia	99,991	9,005
20Newsgroups	11,258	2,000

Table 1: Dataset statistics.

4.2 Baselines

We compare NTM-ACL with the following baselines:

[1]http://archive.ics.uci.edu/ml/datasets/Bag+of+Words
[2]https://cs.nyu.edu/~roweis/data
[3]http://wikidata.dbpedia.org/develop/datasets
[4]http://qwone.com/~jason/20Newsgroups

Dataset	Metric	Method								
		LDA*	NVDM*	ProdLDA*	Scholar*	ATM*	BAT*	ToMCAT	CLNTM	NTM-ACL
NYT	C_A	0.215	0.077	0.184	0.195	0.229	0.236	0.253	0.216	**0.255**
	C_P	0.323	-0.537	0.126	0.045	0.333	0.375	0.381	0.068	**0.398**
	NPMI	0.081	-0.146	0.016	-0.029	0.081	0.095	0.094	-0.019	**0.098**
GRL	C_A	0.196	0.072	0.148	0.206	0.220	0.211	0.248	0.242	**0.252**
	C_P	0.197	-0.519	-0.065	0.215	0.258	0.231	0.280	0.206	**0.310**
	NPMI	0.053	-0.123	-0.019	0.059	0.066	0.061	0.082	0.061	**0.091**
DBP	C_A	0.276	0.139	0.265	0.301	0.293	0.236	0.334	0.319	**0.340**
	C_P	0.352	-0.297	0.215	0.237	0.340	0.375	0.411	0.248	**0.419**
	NPMI	0.103	-0.117	0.021	0.066	0.110	0.095	0.140	0.071	**0.146**
20NG	C_A	0.186	0.112	0.178	0.178	0.183	0.199	0.213	**0.240**	0.217
	C_P	0.282	-0.063	0.071	0.212	0.257	0.296	0.323	**0.350**	0.327
	NPMI	0.064	-0.050	-0.044	0.043	0.038	0.056	0.068	0.065	**0.069**

Table 2: Average topic coherence scores (C_A, C_P, and NPMI) of 5 settings of topic number (20, 30, 50, 75, 100) on 4 datasets. Bold values indicate best-performing models under corresponding settings. Results with * are reported in (Hu et al., 2020).

- LDA (Blei et al., 2003), a probabilistic graphical model, which is one of the most popular conventional models, we used the implementation of GibbsLDA++[5].
- NVDM (Miao et al., 2016), a VAE-based neural topic model that employs Gaussian prior for topic distributions.
- ProdLDA (Srivastava and Sutton, 2017), a VAE-based neural topic model that employs logistic normal prior to approximate Dirichlet prior.
- Scholar (Card et al., 2018), a VAE-based neural topic model that integrates metadata on the basis of ProdLDA.
- ATM (Wang et al., 2019), the first GAN-based nerual topic model.
- BAT (Wang et al., 2020), an adversarially trained bidirectional neural topic model.
- ToMCAT (Hu et al., 2020), an adversarial neural topic model with cycle-consistent objective.
- CLNTM (Nguyen and Luu, 2021), the first attempt to combine contrastive learning with a VAE-based topic model.

[5]http://gibbslda.sourceforge.net/

4.3 Implementation Details and Evaluation

We set the Dirichlet parameter $\boldsymbol{\alpha}$ to $\frac{1}{K}$. The dimension H of the hidden layer is set to 100. The number of replacement items p changes dynamically according to the number of topics K. To be specific, set $p = \lfloor \frac{K}{4} \rfloor$. For the training objective, we set λ_1, λ_2, λ_3, and λ_4 to be 2, 0.2, 1e-3, and 1e-3 respectively, aligning the magnitudes of different losses. During training, we set the batch size to 256 for NYTimes and Grolier, 1,024 for DBPedia, and 64 for 20Newsgroups. The training epoch is set to 150. We use Adam optimizer to update the model parameters, whose learning rate is 1e-4 and the momentum term is 0.5.

Following the previous work (Wang et al., 2020), we evaluate the performance of NTM-ACL and baselines using topic coherence measures highly correlated with human subjective judgments. For each topic, we select the top 10 topic words based on probability to represent the topic. C_A (Aletras and Stevenson, 2013), C_P (Röder et al., 2015), and NPMI (Aletras and Stevenson, 2013) are three topic coherence measures we use to evaluate models. We apply the Palmetto [6] tool to calculate coherence scores. We refer readers to (Röder et al., 2015) for more details of topic coherence measures.

4.4 Experiment Results

To make a robust comparison of NTM-ACL with baselines, we set topic numbers as 20, 30, 50, 75, and 100 on each dataset. Then we calculate the

[6]https://github.com/AKSW/Palmetto

average topic coherence score of 5 settings. The experimental results are presented in Table 2.

Compared with baselines of diverse structures, NTM-ACL performs better on most datasets and topic coherence measures, illustrating the effectiveness of our proposed approach. VAE-based neural topic models perform poorly due to the assumption of prior. Compared with GAN-based neural topic models, especially ToMCAT, which is also based on cycle adversarial training, NTM-ACL achieves State-of-Art results on all datasets, demonstrating the effectiveness of Topic-Augmented Contrastive Learning. CLNTM, as the first method to incorporate contrastive learning with topic modeling, performs worse compared with NTM-ACL except for the 20Newsgroups dataset in terms of C_A and C_P score. This result illustrates that combining contrastive learning with cycle adversarial training is more effective to improve the performance of topic discovery by eliminating the gap between model training and evaluation. To calculate coherence scores, Palmetto uses Wikipedia as a reference corpus, while CLNTM uses the training corpus itself as a reference, leading to the result reporting difference between ours and (Nguyen and Luu, 2021).

NYTimes			
Baseball	**Politics**	**Music**	**Rugby**
inning	voter	song	yard
homer	republican	album	quarterback
run	campaign	music	game
hit	abortion	band	touchdown
yankees	vote	musical	patriot
Grolier			
Nobel	**Philosophy**	**Nature**	**Egypt**
chemistry	philosophy	water	egypt
physics	philosopher	air	dynasty
chemist	knowledge	pressure	emperor
physicist	reason	surface	king
nobel	philosophical	weather	empire

Table 3: Top 4 topics discovered by NTM-ACL on NYTimes and Grolier.

Based on the C_P score, we select 4 group topics with the best coherence score from 50 topic results of NYTimes and Grolier respectively. Every topic is represented in the form of top-5 topic words. As shown in Table 3, the best topics of the NYTimes are related to sports, politics, and music news, while the topics of Grolier reflect science and culture.

5 Analysis and Discussion

5.1 Ablation Study

We conduct an ablation study on the relative contributions of different training objectives to topic modeling performance. We compare our full model with the following ablated variants: 1) **Self-supervised only** removes $\mathcal{L}_{\text{DisCon}}$ in contrastive learning objective. 2) **Discriminative only** removes $\mathcal{L}_{\text{SelfCon}}$ in contrastive learning objective. 3) **w/o Adversarial Loss** removes adversarial loss for word distribution and only relies on contrastive learning to distinguish samples. 4) **w/o Cycle-Consistency Loss** removes Cycle-consistency losses. We perform experiments on the Grolier dataset. The results are shown in Table 4.

From Table 4, we can obtain the following observations: 1) The removal of **Adversarial Loss** and **Cycle-Consistency Loss** both lead to performance drops, indicating that reserving the full objective of cycle adversarial training is a necessary condition for the proposed method. 2) **Self-supervised only** creates an imbalance between generation and discrimination, causing damage to model performance. 3) Although **Discriminative only** achieves a higher C_P score, the overall performance decreases compared to NTM-ACL, indicating the effectiveness of Self-supervised contrastive loss to improve topic-word generation.

Models	C_A	C_P	NPMI
Full model	**0.252**	0.310	**0.091**
Self-supervised only	0.215	0.294	0.085
Discriminative only	0.215	**0.322**	0.068
w/o Adversarial Loss	0.208	0.177	0.033
w/o Cycle-Consistency Loss	0.232	0.254	0.063

Table 4: Performance of different ablated variants compared with the full model.

5.2 Different Data Augmentation Strategies

To fully exploit the effectiveness of the proposed Reconstruct Min-Term Replacement strategy, we design two simple data augmentation strategies for comparison: 1) **Noise Added (NA)**, topic distribution $\boldsymbol{\theta}_i$ is added with the noise distribution which is of the same dimension, randomly sampled from a Gaussian distribution with expectation 0 and variance 0.01. 2) **Zero Masked (ZM)**, when getting the indices $\{a_1, a_2, \cdots, a_p\}$, the value at the corresponding index is set to 0. We apply different data

augmentation strategies and keep other experimental settings unchanged. The results are shown in Table 5.

Dataset	Metric	Strategy		
		NA	ZM	RMR
NYT	C_A	0.251	0.251	**0.255**
	C_P	0.393	0.390	**0.398**
	NPMI	0.097	0.096	**0.098**
GRL	C_A	0.246	0.249	**0.252**
	C_P	0.277	0.284	**0.310**
	NPMI	0.079	0.083	**0.091**
DBP	C_A	0.334	0.338	**0.340**
	C_P	0.418	0.413	**0.419**
	NPMI	0.141	0.142	**0.146**
20NG	C_A	0.215	0.213	**0.217**
	C_P	0.324	0.315	**0.327**
	NPMI	0.067	0.065	**0.069**

Table 5: Effectiveness of different data augmentation strategies.

It can be observed that most coherence scores increase compared to GAN-based neural topic models, indicating the robustness of our contrastive learning approach. On the other hand, the results of NTM-ACL are the highest among the three strategies, which is proved to be a more suitable strategy for topic distribution.

5.3 Effect of Replacement Number

The number of replacement items p is one of the important hyperparameters for the RMR. For different K, the number of topics, it is inappropriate if p is set to a fixed value. In this subsection, we compare the dynamic setting to fixed numbers (1, 5, 15) on four datasets, using C_P coherence measure. The results are shown in Figure 3.

It can be observed that the data augmentation strategy with dynamic replacement numbers achieves the best performance. When the number p is too small, the difference between the positive pair is too slight. When the number p is too large, the similarity between positive samples cannot provide sufficient information for Self-supervised contrastive loss.

5.4 Training Strategy

In this subsection, we explore different training strategies, making contrastive learning and cycle adversarial training work respectively in different stages of 150 epochs. We abbreviate contrastive learning as CL and cycle adversarial training as CA. We perform experiments on the Grolier dataset with the following strategies: 1) **Disentangle**, separating CL from CA. The first 50 epochs use CL/CA to update model parameters, and the next 100 epochs only employ CA/CL. Also, we design an alternation strategy in that CL is used to update the parameters after 5 epochs of CA only. 2) **Warm up**, using CL to warm up in the first 50 epochs, the next 100 epochs adopt the original strategy or are divided into two stages equally, i.e. CL+CA→CA. The results are shown in Table 6.

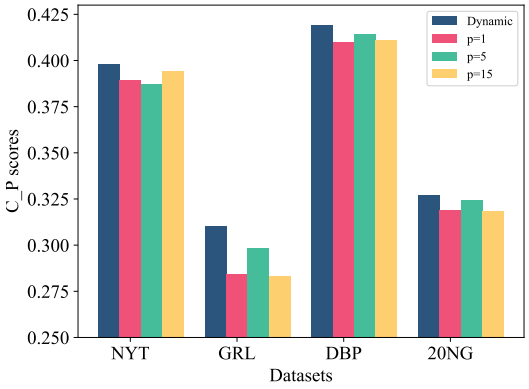

Figure 3: Dynamic replacement number compared with fixed number setting (1, 5, 15).

Training Strategy		C_A	C_P	NPMI
Disentangle	CL→CA	0.248	0.307	0.089
	CA→CL	0.245	0.280	0.083
	Alternation	0.239	0.274	0.080
Warm up	CL→CL+CA	0.247	0.291	0.081
	CL→CL+CA→CA	0.250	0.289	0.086
NTM-ACL		**0.252**	**0.310**	**0.091**

Table 6: Comparison between different training strategies on Grolier.

From the results, we can observe that NTM-ACL achieves the best performance. The result of CL→CA is second only to the best result. Using CL after CA undermines the stable symmetric structure, instead of further improving mapping ability, which should be avoided in future studies.

6 Conclusion

In this paper, we have proposed NTM-ACL, a novel topic modeling framework based on cycle adversarial training and contrastive learning. Self-supervised contrastive loss improves the generation of topic-word distribution that is used for the evaluation of topic modeling, while Discriminative con-

trastive loss keeps the balance of generation and discrimination. Moreover, a novel data augmentation strategy is designed to create positive samples of topic distributions based on the reconstruction ability of cycle adversarial training. The experimental results show that the proposed method outperforms competitive baselines of different structures.

Limitations

In this section, we describe the limitation of our proposed method in terms of data augmentation and the way to combine contrastive learning with cycle adversarial training. First, our data augmentation strategy relies on the reconstruction ability of cycle adversarial training. We believe that more data augmentation strategies for topic distribution will be studied. Second, with the symmetrical structure of cycle adversarial training, it is worth exploring how to optimize the *encoder* E and *generator* G through contrastive learning simultaneously. We can extend the proposed framework to a conjugated structure in future work. Moreover, although it has been explored that contrastive learning and cycle adversarial training working synchronously performs better, we believe that more sophisticated training strategies will be designed to further improve the performance of topic modeling.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments and we thank Huawei for supporting this project. This work is funded by the National Natural Science Foundation of China (62176053). This work is supported by the Big Data Computing Center of Southeast University.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

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, Melbourne, Australia. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

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, Online. Association for Computational Linguistics.

Jongheon Jeong and Jinwoo Shin. 2021. Training gans with stronger augmentations via contrastive discriminator. *arXiv preprint arXiv:2103.09742*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, pages 2410–2419. PMLR.

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, pages 1727–1736, New York, New York, USA. PMLR.

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*, WSDM '15, pages 399–408, New York, NY, USA. ACM.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

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, Online. Association for Computational Linguistics.

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

Tao Wang, Linhai Zhang, Chenchen Ye, Junxi Liu, and Deyu Zhou. 2022. A novel framework based on medical concept driven attention for explainable medical code prediction via external knowledge. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1407–1416, Dublin, Ireland. Association for Computational Linguistics.

Yiming Wang, Ximing Li, Xiaotang Zhou, and Jihong Ouyang. 2021. Extracting topics with simultaneous word co-occurrence and semantic correlation graphs: neural topic modeling for short texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 18–27.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3898–3907, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Yang Yang, ZHOU Deyu, and Yulan He. 2018. An interpretable neural network with topical information for relevant emotion ranking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3423–3432.

Yang Yang, Deyu Zhou, Yulan He, and Meng Zhang. 2019. Interpretable relevant emotion ranking with event-driven attention. 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 177–187.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903, Dublin, Ireland. Association for Computational Linguistics.

Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. 2021. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10623–10633.

Deyu Zhou, Jiale Yuan, and Jiasheng Si. 2021. Health issue identification in social media based on multitask hierarchical neural networks with topic attention. *Artificial Intelligence in Medicine*, 118:102119.

Deyu Zhou, Xuan Zhang, and Yulan He. 2017. Event extraction from Twitter using non-parametric Bayesian mixture model with word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 808–817, Valencia, Spain. Association for Computational Linguistics.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We discuss the limitations of our work after the conclusion section.

☒ A2. Did you discuss any potential risks of your work?
As a research domain with a lot of practice, the topic model does not show obvious potential risks. So we don't discuss this aspect specifically.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 4.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Previous research extensively validated the datasets we use to ensure data security.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.

C ☒ Did you run computational experiments?

Our model does not use pre-trained language models and requires few computing resources.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Our work does not involve human annotators.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.