Self-Supervised Learning of Contextualized Neural Topic Models with VIC Regularization

Anonymous ACL submission

Abstract

In modern society, the widespread use of the Internet has led to the generation of massive amounts of textual data, creating a growing demand for advanced text mining techniques to efficiently extract valuable information. One such technique is topic modeling, which analyzes large document collections to uncover underlying latent topics. This approach has applications in document retrieval, classification, and beyond. Recently, research on neural topic models, which leverage neural networks for topic extraction, has gained attention, particularly with the integration of contextual embeddings from sentence embedding. Self-supervised learning, which uses pseudo-labels derived from the data itself, has shown promise in this domain. Variance-Invariance-Covariance (VIC) Regularization, originally introduced for multimodal analysis, has been shown to be effective for neural topic models using only word-based embeddings, however, its applicability to neural topic models incorporating contextual embeddings remains unexplored. This study proposes a self-supervised neural topic model incorporating VIC Regularization and contextual embeddings. Our experimental results indicate improved topic coherence compared to conventional neural topic models.

1 Introduction

In recent years, the widespread use of the Internet has led to the generation of massive amounts of textual data, making efficient text data processing and the extraction of valuable information increasingly important. One prominent technique for this purpose is topic modeling, which uncovers useful information from large document collections. A representative model in this field is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which assumes that each document is composed of multiple latent topics drawn from a document-specific distribution, with words generated according to these topics. By estimating topics from observed words, LDA facilitates the semantic analysis of entire documents. However, the necessity to define and derive inference algorithms for each modeling objective poses a significant challenge. To address these issues, neural variational inference was proposed (Miao et al., 2017), while a logistic normal prior was introduced for neural topic models (Srivastava and Sutton, 2017), integrating deep neural networks with traditional topic models. However, these models struggle to capture semantic relationships and complex patterns within documents. To overcome this, a neural topic model with contrastive learning (Nguyen and Luu, 2021), which leverages semantic relationships through a novel sampling method, was proposed. Nevertheless, these models rely on word-level embeddings that disregard the sequential structure and contextual information in documents. To address this limitation, Contextualized Topic Model (CTM) (Bianchi et al., 2021) was introduced, combining Bag-of-Words (BoW) embeddings with context-aware document embeddings.

Traditional machine learning models often rely on supervised learning with large amounts of labeled data. However, creating labeled datasets is labor-intensive and costly, particularly for large-scale text corpora. Self-supervised learning, which uses pseudo-labels generated from the data itself, presents a promising alternative. Contrastive learning methods, such as SimCLR (Chen et al., 2020) and SwAV (Caron et al., 2020), have shown remarkable performance in various downstream tasks. Meanwhile, Variance-Invariance-Covariance (VIC) Regularization (VICReg) (Bardes et al., 2022) was introduced to enhance self-supervised learning by applying three distinct regularization terms: variance, invariance, and covariance. While VIC Regularization has been shown to improve BoW-based neural topic models (VICNTM) (Xu et al., 2025), its potential when applied to CTM remains unexplored.

This study proposes a VIC-regularized contextualized neural topic model that integrates both BoW and contextual embeddings. We generate positive samples for contrastive learning using tf-idf (term frequency-inverse document frequency) based sampling and replace the traditional contrastive loss with VIC Regularization terms. We evaluate model performance using a topic coherence metric and demonstrate that our approach improves topic coherence without sacrificing predictive performance. Additionally, we find that selecting an appropriate number of topics further enhances model effectiveness.

2 Related Work

2.1 Topic Model

Topic modeling is an analytical method for discovering meaningful information from large collections of documents. In each document, multiple latent topics are probabilistically generated, and words appearing in the document are assumed to be generated from these topics. By estimating the probability distribution of topics for each document and the probability of word generation for each topic based on observed words, it is possible to analyze topic similarities and document semantics. A representative model of this topic model is Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

2.2 Neural Topic Model

Neural topic models integrate neural networks with topic modeling to overcome the computational challenges posed by the increasing number of parameters in traditional models like LDA. Notable examples include the Neural Topic Model (NTM) (Miao et al., 2017), which is based on the Variational Autoencoder (VAE) (Kingma and Welling, 2014), and ProdLDA (Srivastava and Sutton, 2017), which replaces Dirichlet priors with logistic normal priors. Furthermore, SCHOLAR (Card et al., 2018) was developed as an extension of ProdLDA. In this model, latent variables are sampled from a multivariate normal distribution with parameters sampled from a logistic normal prior, and these latent variables are subsequently mapped to multinomial parameters via a softmax function.

2.3 Neural Topic Model with Context

Traditional neural topic models have primarily focused on the words present in document data. However, document embeddings based on the BoW approach disregard contextual information, making it difficult to distinguish between words that have different meanings depending on the context. To address this issue, contextual embeddings have recently been introduced to capture semantic and contextual relationships within document data. Contextual embeddings utilize pre-trained models such as BERT (Devlin et al., 2018) and its extension SBERT (Reimers and Gurevych, 2019) to represent words and sentences in a context-dependent manner. SBERT employs triplet loss (Schroff et al., 2015) during fine-tuning, which brings the anchor and positive samples (documents with the same labels) closer while pushing the anchor and negative samples (documents with different labels) apart. Notable implementations include all-mpnet-base-v2[1], based on MPNet (Song et al., 2020), and all-distilroberta-v1[2], based on RoBERTa (Liu et al., 2019). Specifically, all-mpnet-base-v2 has been fine-tuned using more than one billion document samples, including Reddit comments (Henderson et al., 2019). In recent years, large language models (LLM) have seen remarkable advancements. One prominent model is LLaMA (Large Language Model Meta AI) (Touvron et al., 2023), which consists of tens of billions of parameters and achieves state-of-the-art performance. These technological advancements enable deeper, meaning-based document analysis and are expected to improve the quality of topic distributions in neural topic models.

2.4 Self-supervised Learning

Self-supervised learning is a learning method that does not require explicit labels for the training data but instead generates pseudo-labels automatically from the data itself. From the perspective of not providing explicit labels for the training data, self-supervised learning can be considered a type of unsupervised learning. This approach has the advantage of avoiding the annotation costs, which often hinder the processing of large-scale datasets. Contrastive learning, which is a type of self-supervised learning, is a learning method that encourages sim-

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[2] https://huggingface.co/sentence-transformers/all-distilroberta-v1

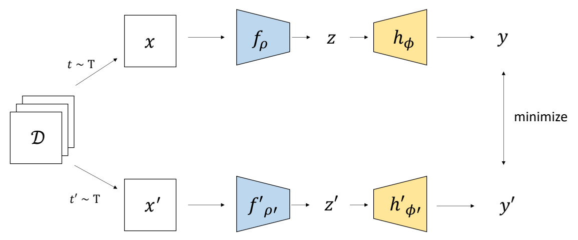

Figure 1: The process of learning positive samples in VICReg is illustrated.

ilar data points to have similar embedding vectors in the latent space while ensuring that dissimilar data points have distinct embedding vectors. Consequently, contrastive learning requires similar data samples. To generate these positive samples, data augmentation techniques are applied to the original data. Data augmentation involves applying specific transformations to the data to create new samples that closely resemble the original data. In this framework, the original data serves as the anchor, the augmented data as the positive sample, and a randomly selected data point as the negative sample, enabling the model to learn to distinguish between them.

2.5 Variance-Invariance-Covariance Regularization

This subsection provides a brief overview of VICReg (Bardes et al., 2022), a general theoritical framework in deep learning. The VICReg aims to minimize a loss function composed of three regularization terms: the variance term, the invariance term, and the covariance term. The process of learning positive samples in VICReg is illustrated in Figure 1. The input data $\mathcal{D}$ is transformed into $\boldsymbol{x} = t(i)$ and $\boldsymbol{x}' = t'(i)$ following distributions T. These inputs are passed through encoders f_ρ and $f'_{\rho'}$ with parameters ρ and ρ', producing embedding vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$. Next, these embeddings are processed by expanders h_ϕ and $h'_{\phi'}$ with parameters ϕ and ϕ', resulting in latent vectors $\boldsymbol{y}$ and $\boldsymbol{y}'$. For a mini-batch $Y = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n]$ and $Y' = [\boldsymbol{y}'_1, \dots, \boldsymbol{y}'_n]$, the model learns to bring the embeddings $\boldsymbol{y}_i$ and $\boldsymbol{y}'_i$ closer while applying the loss function composed of the three terms: variance, invariance, and covariance. First, let ϵ be a small scalar value, and define $\mathcal{S}(x, \epsilon) = \sqrt{\text{Var}(x) + \epsilon}$. The variance term is then expressed as follows:

$$v(Y) = \frac{1}{d} \sum_{j=1}^{d} \max(0, \gamma - \mathcal{S}(y^j, \epsilon)) \quad (1)$$

where y^j represents the j-th component of the d-dimensional latent space Y that contains all latent vectors. This term uses a hinge loss to maintain the standard deviation of each component of the embedding vectors in the mini-batch above a certain threshold. By using equation (1), the embeddings of the samples within the mini-batch are encouraged to have distinct values from one another. Next, the invariance term is shown as follows:

$$s(Y, Y') = \frac{1}{n} \sum_{i} ||\boldsymbol{y}_i - \boldsymbol{y}'_i||_2^2 \quad (2)$$

which is the mean squared distance between the embedding vectors. This encourages the paired embedding vectors $\boldsymbol{y}_i$ and $\boldsymbol{y}'_i$ to be close to each other. Finally, the covariance term is shown as follows:

$$c(Y) = \frac{1}{d} \sum_{i \neq j} [\mathcal{C}(Y)]_{i,j}^2 \quad (3)$$

where $\mathcal{C}(Y)$ denotes the covariance matrix for the mini-batch Y. By using equation (3), the off-diagonal elements of $\mathcal{C}(Y)$ are minimized to suppress correlations between different dimensions in the embedding space. This prevents the collapse of information caused by high correlations across dimensions (i.e., outputting redundant information). Therefore, the overall VICReg loss function is given by the following equation:

$$\begin{aligned} l(Y, Y') = {} & \lambda s(Y, Y') \\ & + \mu \left[v(Y) + v(Y') \right] \\ & + \nu \left[c(Y) + c(Y') \right] \end{aligned} \quad (4)$$

3 Proposed Method

3.1 A Brief Explanation of Neural Topic Model

In neural topic models, given $\boldsymbol{x}$ as the encoder input, $\boldsymbol{z}$ as the encoder output, and $p(\boldsymbol{z})$ as the prior distribution of $\boldsymbol{z}$, the decoder is represented by ϕ with $p_\phi(\boldsymbol{x}|\boldsymbol{z})$ as the network that generates documents from topics. The encoder is represented

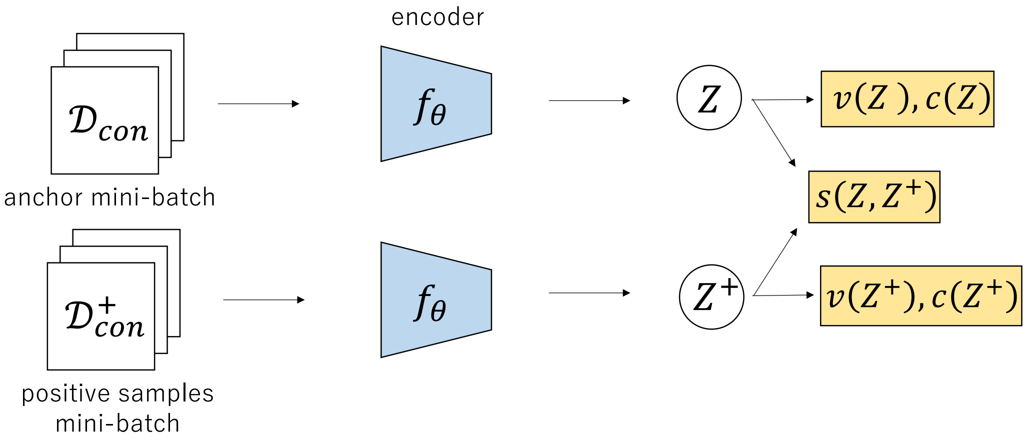

Figure 2: The architecture of proposed method.

by θ with $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ as the variational distribution, and the following objective function is minimized:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}) = &-\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\phi(\boldsymbol{x}|\boldsymbol{z})\right] \\ &+ \mathbb{KL}\left[q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})\right] \end{aligned} \tag{5}$$

3.2 Baseline Neural Topic Model

In this study, we propose a model that applies VIC Regularization and the tf-idf sampling method,as proposed in CLNTM (Nguyen and Luu, 2021)', to the existing CTM (Bianchi et al., 2021) framework. While CLNTM is based on SCHOLAR (Card et al., 2018), CTM is based on ProdLDA (Srivastava and Sutton, 2017). For training, the input dataset $\mathcal{D}$ is used to obtain document embeddings $\boldsymbol{x_w}$ using BoW and contextual embeddings $\boldsymbol{x_c}$ using SBERT (Reimers and Gurevych, 2019) for each document. These embeddings are concatenated to form the document embeddings $\boldsymbol{x}$, and the anchor dataset $\mathcal{D}_{con}$ is generated from these embeddings.

3.3 Architecture

In this study, we propose **VICCTM**, a contextualized neural topic model that incorporates VIC Regularization into CLNTM-based neural topic model to improve topic coherence. Thus, the objective of this study is to examine the effectiveness of VIC Regularization in a contextualized neural topic model based on CLNTM by minimizing the loss function:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta}, \phi) = &-\sum_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\phi(\boldsymbol{x}|\boldsymbol{z})\right]\right. \\ &+\mathbb{KL}\left(q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \mid\mid p(\boldsymbol{z})\right)] \\ &+ \lambda s(Z, Z^+) \\ &+ \mu\big[v(Z) + v(Z^+)\big] \\ &+ \nu\big[c(Z) + c(Z^+)\big] \end{aligned} \tag{6}$$

Here, let Z be the set of latent vectors z obtained by passing the anchor $\boldsymbol{x}$ through the encoder $f_{\boldsymbol{\theta}}$, and Z^+ be the set of latent vectors $\boldsymbol{z}^+$ obtained by passing the positive samples $\boldsymbol{x}^+$ through the same encoder $f_{\boldsymbol{\theta}}$. Note that the notation $\boldsymbol{x}$ indicates the document embedding obtained by concatenating BoW-based embeddings $\boldsymbol{x_w}$ and contextual embeddings $\boldsymbol{x_c}$, as described in Section 3.2. Similarly, $\boldsymbol{x}^+$ is obtained in the same manner. The term $s(Z, Z^+)$ represents the invariance term, $v(Z) + v(Z^+)$ represents the variance term, and $c(Z) + c(Z^+)$ represents the covariance term. The hyperparameters λ, μ, and ν control the importance of each loss component. In the subsequent experiments, these hyperparameters will be explored using Bayesian optimization. The proposed learning model is illustrated in Figure 2. In the proposed method, positive samples are first generated for each mini-batch using a tf-idf based method, as to be described in Section 3.4. Next,

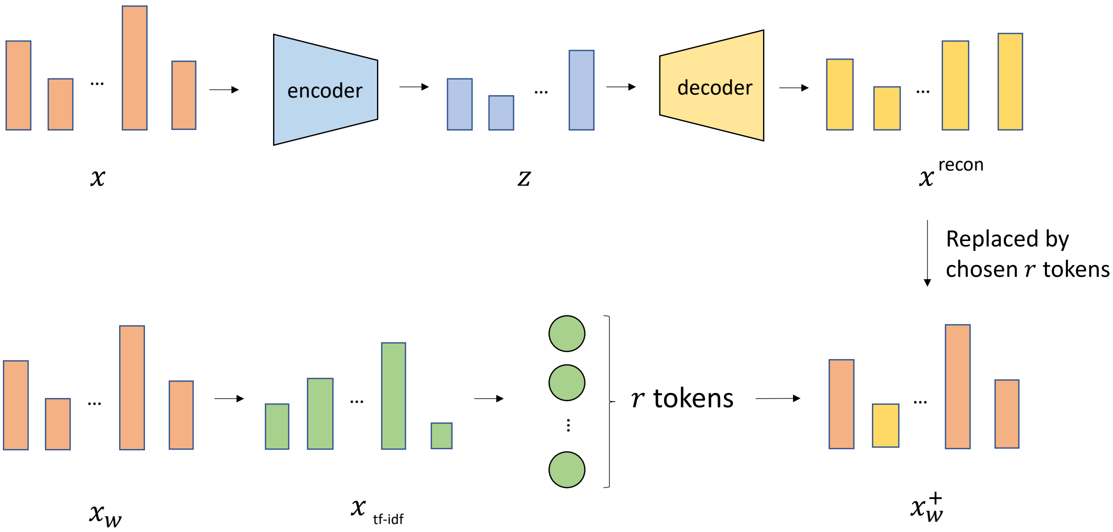

Figure 3: Positive example sampling method using tf-idf.

the anchor set $\mathcal{D}_{con}$ and the positive set $\mathcal{D}_{con}^+$ are passed through the encoder f_θ and embedded into the latent vectors Z and Z^+. By minimizing the loss function in equation (6), the latent topic distributions of the anchor and positive samples are expected to become similar. Additionally, the variance of the embedding vectors in each mini-batch is maintained above a certain threshold, and the covariance between the dimensions in the embedding space of each mini-batch is expected to approach zero.

3.4 Sampling Methodolgy

In this study, positive sample sampling is performed using the tf-idf based method proposed at CLNTM. Tf-idf is a statistical measure that indicates the importance of each word token in a document based on its occurrence patterns. The term frequency (tf) represents how frequently a word token appears in a document, while the inverse document frequency (idf) measures how many documents contain that word token, taking the reciprocal of that value. The product of these two values is the tf-idf score. The tf-idf score has the characteristic that word tokens with lower scores are less relevant to the topic of the document in which they appear. Additionally, when generating positive samples, it is essential to augment the data without deviating from the existing topics. Therefore, positive samples are generated by replacing word tokens with low tf-idf scores in a document with other word tokens. The sampling method for positive examples used in this study is illustrated in Figure 3. First, the tf-idf values of all word tokens in the input data are computed. Next, the document $\boldsymbol{x}$ is passed through the encoder and decoder to obtain the reconstructed document $\boldsymbol{x}^{\text{recon}}$. In this reconstructed document, r word tokens with the lowest tf-idf values are selected for replacement. These selected word tokens are then replaced with their corresponding word tokens from the original document $\boldsymbol{x_w}$, resulting in the positive example $\boldsymbol{x}_w^+$. The obtained $\boldsymbol{x}_w^+$ is then concatenated with contextual embeddings $\boldsymbol{x_c}$ to form the positive dataset $\mathcal{D}_{con}^+$.

4 Experiments Setup

4.1 Datasets

In the experiment, we applied preprocessing in the same manner as VICNTM (Xu et al., 2025), which originates from CLNTM (Nguyen and Luu, 2021). This preprocessing was performed on three different datasets by removing stopwords (single-character words) and word types that appeared fewer than 100 times:

- **20Newsgroups (20NG)** (Lang, 1995): This dataset consists of approximately 13,000 news articles. This was split into training, validation, and test sets with proportions of 48%, 12%, and 40%, respectively.

- **IMDb movie reviews (IMDb)** (Maas et al., 2011): This dataset consists of movie reviews collected from IMDb, which includes about 43,000 movie reviews. The dataset was split of 70%, 15%, and 15%, respectively.

- **Wikitext-103 (Wiki)** (Merity et al., 2017): This dataset contains approximately 28,000 articles, consisting of Wikipedia articles that meet the "Good" or "Featured" criteria. The dataset was split into 50%, 25%, and 25% of ratio.

Furthermore, the number of word tokens to be replaced when generating positive samples was set to $r = 15$, which was found to achieve optimal performance (Nguyen and Luu, 2021). Additionally, the minimum effective word token count was set to 30, and documents with fewer than 30 word tokens were excluded.

4.2 Evaluation Metrics

For model evaluation, we used the NPMI (Normalized Pointwise Mutual Information) (Chang et al., 2009) (Newman et al., 2010) as a metric for measuring topic coherence, following the approaches of CLNTM (Nguyen and Luu, 2021) and VICNTM (Xu et al., 2025), which serves as the base model in this study. NPMI is a metric for measuring topic coherence based on word co-occurrence frequencies within the corpus. In the experiment, we calculated the NPMI of the top ten word types for each topic on the test set using the model trained on the training set. The NPMI calculation formula is:

$$\begin{aligned} \mathrm{PMI}(v, v') &= \log \frac{P(v, v')}{P(v)P(v')} \\ \mathrm{NPMI}(v, v') &= \frac{PMI(v, v')}{-\log P(v, v')} \end{aligned} \tag{7}$$

For each topic, let v and v' be any two word types from the set of the top ten word types. With $P(\cdot)$ representing the probability, the pointwise mutual information (PMI) is first computed. Next, PMI is normalized to mitigate the influence of word type rarity, resulting in the NPMI calculation formula. In addition to the evaluation metrics used in previous studies, we also evaluated the model's predictive performance using perplexity, a metric that has long been employed in topic modeling research. The formula for calculating perplexity is:

$$\text{perplexity} = \exp\left(-\frac{1}{m}\sum_{d}\sum_{t}\log P(w_{d,t})\right) \tag{8}$$

where m be the total number of word tokens in the test set, perplexity is defined based on the likelihood $P(w_{d,t})$ that the model assigns to the t-th word token of the d-th document in the test set.

4.3 Detailed Settings

The experiment was conducted by evaluating the model using NPMI, which measures topic coherence, and perplexity, which measures predictive performance, for topic numbers $k = 20$, $k = 50$, and $k = 200$. As a baseline, we used CTM (Bianchi et al., 2021). For contextual embeddings, we utilized the pre-trained SBERT-based model **all-mpnet-base-v2**[3]. To prevent overfitting, we calculated the model's NPMI on the validation set after each epoch. If no improvement was observed for 30 consecutive epochs, we considered the training converged and applied early stopping. To exclude the effect of locally optimal gradients when reconstructing $\boldsymbol{x}$, convergence detection was started only after 150 epochs. Each mini-batch consisted of 70 documents randomly sampled from the training set. The learning rate was set to 0.001, and the document embedding dimension was $\dim(\boldsymbol{x}) = 1068$, consisting of a BoW embedding dimension of $\dim(\boldsymbol{x_w}) = 300$ and a contextual embedding dimension of $\dim(\boldsymbol{x_c}) = 768$. The hyperparameters λ, μ, ν were optimized using Bayesian optimization[4]. For this, batches were randomly sampled from the training data, and 100 trials of hyperparameters searches were performed.

5 Results

5.1 Evaluation Results and Analysis

Table 1 summarizes the NPMI and perplexity results for all experiments conducted with topic numbers $k = 20, k = 50$, and $k = 200$. The reported values represent the mean and sample standard deviation obtained from 10 runs with different random seeds. From the results, we observe that the proposed method consistently improves topic coherence, as measured by NPMI, compared to

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[4] https://optuna.org/

Table 1: For each dataset, we present the NPMI and perplexity values along with their sample standard deviations for topic numbers $k = 50$ and $k = 200$. A higher NPMI indicates better topic coherence, while a lower perplexity indicates better predictive performance of the model.

Dataset	$k = 20$		$k = 50$		$k = 200$	
	NPMI	Perplexity	NPMI	Perplexity	NPMI	Perplexity
20NG						
CTM	0.395 (±0.016)	1532 (±34)	0.355 (±0.007)	1656 (±25)	0.286 (±0.005)	2878 (±78)
VICCTM	**0.404** (±0.014)	1540 (±39)	**0.362** (±0.007)	1675 (±44)	**0.287** (±0.004)	2863 (±44)
IMDb						
CTM	0.175 (±0.009)	1861 (±7)	0.158 (±0.004)	2092 (±10)	**0.133** (±0.002)	3615 (±30)
VICCTM	**0.176** (±0.008)	1860 (±14)	**0.160** (±0.005)	2099 (±9)	0.131 (±0.003)	3608 (±44)
Wiki						
CTM	0.498 (±0.020)	3661 (±40)	0.495 (±0.010)	3373 (±36)	0.446 (±0.004)	3167 (±16)
VICCTM	**0.501** (±0.019)	3644 (±78)	**0.498** (±0.010)	3357 (±46)	**0.450** (±0.004)	3157 (±15)

the baseline CTM model across various datasets and topic numbers. Specifically, when $k = 20$ and $k = 50$, the proposed model outperforms CTM, demonstrating higher NPMI values across all datasets. For $k = 200$, the model maintains better performance in two of the datasets, while in the remaining dataset, it achieves comparable results. The perplexity scores also show that the proposed model generally exhibits better predictive performance than CTM, with lower perplexity values across most of the experimental settings. The results suggests that integrating CTM with VIC Regularization helps the model capture document-topic relationships more effectively. Furthermore, focusing on the case where $k = 20$, we observe a notable improvement in topic coherence with the proposed method. This experiment suggests that selecting an appropriate number of topics is crucial for maximizing the performance of the proposed method. If the number of topics is not suitable, overfitting or information dispersion may occur during model training, potentially degrading the quality of the learned topics.

5.2 Ablation Study

5.2.1 Effects of VIC Terms

We evaluated the effects of each VIC Regularization term on NPMI and perplexity using the 20NG dataset with $k = 50$. The evaluation followed the same experimental settings and hyperparameters as in Table 1. Table 2 presents the NPMI and perplexity results for each VIC Regularization term. The results indicate that when only one term is present, the variance term (V) alone shows the best performance. However, when considering any combination of two terms, the combination of invariance (I) and covariance (C) yields better results. Nevertheless, the best performance is ultimately achieved when all three terms—variance (V), invariance (I), and covariance (C)—are applied together.

Table 2: For 20NG dataset, we present the NPMI and perplexity values along with their sample standard deviations for topic numbers $k = 50$. A higher NPMI indicates better topic coherence, while a lower perplexity indicates better predictive performance of the model.

Used Reg Terms	NPMI	Perplexity
VIC	**0.362** (±0.007)	1675 (±44)
VI	0.352 (±0.005)	1681 (±43)
VC	0.354 (±0.008)	1675 (±30)
IC	0.355 (±0.007)	1679 (±39)
V	0.354 (±0.006)	1665 (±30)
I	0.351 (±0.011)	1697 (±23)
C	0.352 (±0.010)	1688 (±36)

5.2.2 Effects of Combining Context

To examine the effect of contextual embeddings, we evaluated the model using the 20NG dataset with $k = 50$. Table 3 presents the results for four cases: using only BoW embeddings (BoW), using only contextual embeddings (SBERT), using only BoW with VIC Regularization (VICNTM), and using a combination of both embeddings (CTM). The results indicate that using only contextual embeddings significantly degrades the performance across the evaluation metrics. In contrast, the model combining BoW and contextual embeddings (CTM) shows improved topic coherence compared to using BoW embeddings alone. However, in all cases, the

Table 3: For 20NG dataset, we present the NPMI and perplexity values along with their sample standard deviations for topic numbers $k = 50$. A higher NPMI indicates better topic coherence, while a lower perplexity indicates better predictive performance of the model.

Embedding Representations	NPMI	Perplexity
BoW+SBERT+VICReg(VICCTM)	**0.362** (±0.007)	1675 (±44)
BoW	0.353 (±0.006)	1655 (±31)
SBERT	0.163 (±0.032)	2610 (±408)
BoW+VICReg(VICNTM)	0.352 (±0.014)	1682 (±38)
BoW+SBERT(CTM)	0.355 (±0.007)	1656 (±25)

model utilizing VIC Regularization outperforms the others, demonstrating the effectiveness of the proposed approach.

6 Conclusion

In this study, we proposed a neural topic model incorporating VIC Regularization, which is commonly used in multimodal analysis, with the expectation that it would also be effective in a contextualized neural topic model. The proposed model combines traditional BoW embeddings with contextual embeddings and applies VIC Regularization to the loss function in contrastive learning, where positive samples are generated using a tf-idf-based measure. Instead of using a contrastive loss, the model applies the three regularization terms: Variance, Invariance, and Covariance. The performance of the proposed model was evaluated through experiments, which demonstrated that it improves topic coherence while maintaining predictive performance compared to conventional neural topic models. Additionally, the results showed that setting an appropriate number of topics further improves topic coherence. This improvement is attributed to the constraints imposed by VIC Regularization, which reduce redundancy and dispersion in topic representations during model training.

7 Limitation

Our experiments revealed that the proposed method maintains predictive performance while effectively capturing semantic relationships. However, several limitations remain. The model's performance heavily depends on the selection of the topic number; while appropriate topic numbers yield better results, excessively low or high values degrade topic coherence. This finding highlights the need for automated or dynamic optimization techniques for topic number selection. Additionally, the VICReg parameters—variance, invariance, and covariance—significantly impact model performance. Tuning these parameters requires substantial computational resources, underscoring the need for more efficient optimization strategies.

References

Adrien Bardes, Jean Ponce, and Yann LeCun. 2022. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):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.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the Workshop on NLP for Conversational AI*. Data available at github.com/PolyAI-LDN/conversational-datasets.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrew L. 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, Portland, Oregon, USA. Association for Computational Linguistics.

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

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.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

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

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

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, D. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, A. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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