Expanding Horizons in Short Text Analysis: Integrating LLMs and VAEs for Enhanced Topic Modeling

Anonymous ACL submission

Abstract

Topic models are one of the compelling methods for discovering latent semantics in a document collection. However, it assumes that a document has sufficient co-occurrence information to be effective. However, in short texts, co-occurrence information is minimal, which results in feature sparsity in document representation. Therefore, existing topic models (probabilistic or neural) mostly fail to mine patterns from them to generate coherent topics. In this paper, we take a new approach to short-text topic modeling to address the data-sparsity issue by extending short text into longer sequences using large language models (LLMs) and decoding topics using a variational autoencoder (VAE). We observe that our model can substantially improve the performance of short-text topic modeling. Extensive experiments on multiple real-world datasets under extreme data sparsity scenarios show that our models can generate high-quality topics that outperform state-of-the-art models. [1]

1 Introduction

In the digital era, short texts dominate the Web, such as tweets, web page titles, news headlines, image captions, product reviews, etc. These short texts are one of the most effective mediums for sharing knowledge. However, the volume of short texts is also huge because of the information explosion, which demands an external mechanism for extracting key information from them. Topic modeling is one such mechanism for uncovering latent topics from short texts, which has a wide range of applications, such as comment summarization (Ma et al., 2012), content characterization (Ramage et al., 2010; Zhao et al., 2011), emergent topic detection (Lin et al., 2010), document classification (Sriram et al., 2010), user interest profiling (Weng et al., 2010), and so on.

Traditional topic models (e.g., LDA, PLSA) (Blei et al., 2003; Hofmann, 1999) are primarily used to discover latent topics from text corpora. However, these models largely assume that each given text document has rich context information to infer topic structures from the corpus. Therefore, the lack of ample context information in short texts makes topic modeling a challenging task. This issue is also called the data sparsity problem, where the co-occurrence information in short texts is minimal, making traditional models less effective in high-quality topic mining.

While various strategies have been developed for modeling topics in short texts, each has its limitations. E.g., aggregating short texts into longer pseudo-documents based on metadata like user information, hashtags, or external corpora is a common approach Weng et al. (2010); Mehrotra et al. (2013); Zuo et al. (2016); however, the availability of such metadata can be inconsistent. To overcome this, some methods rely on structural or semantic information within the texts themselves, such as the Biterm Topic Model (Yan et al., 2013) and its extensions (Zhu et al., 2018), which focus on word pairs but often cannot provide individual document topic distributions. Another method Yin and Wang (2014) limits texts to a single topic, simplifying the model but potentially overlooking texts that span multiple topics.

Considering the above limitations, in this paper, we first try to understand the characteristics of short texts and how humans process short texts while mining topics. A short text (e.g., title, caption) is usually a summarized version of an existent longer text, providing an excellent hint to readers about the longer text. To judge the topics of a short text, humans usually "*imagine*" the context of the short text. From the headline: "No tsunami but FIFA's corruption storm rages on", humans may guess its content and gather context about "FIFA" through imagination; based on this, they can understand the

[1] Code and data will be released after the review process.

headline is about the topic “sports”.

Now, can machines also “*imagine*” the context to better understand the topics of a short text? Recently, large language models (LLMs) such as GPT-3 (Brown et al., 2020), LLAMA2 (Touvron et al., 2023) and T5 (Raffel et al., 2020; Chung et al., 2022) have appeared as amazing open-ended text generator capable of rendering surprisingly fluent text from a limited preceding context. E.g., from the previously specified news headline, a relatively smaller language model T5 generates an extended sequence (as shown in the second column of Table 1) with tokens like “Sepp Blatter”, “Fernando Torres”, and “kicking” that are strongly related to sports soccer.

Considering the above scenario, one potential solution for short text topic modeling can be to leverage the extensive knowledge encoded within LLMs. LLMs, trained on diverse and voluminous datasets, possess a broad understanding of various subjects, enabling them to generate coherent, humanlike texts from minimal input. By using LLMs to expand short texts into longer, context-enriched narratives, we can create a proxy for the detailed context that traditional topic modeling techniques lack when dealing with short texts. This approach harnesses the LLMs’ ability to synthesize information from their training data, effectively ’imagining’ the broader context that surrounds a given piece of short text.

Now, the question is, how can we use the longer texts generated by LLMs for topic modeling? To answer this question, we delve deeper into the relationship between LLMs’ text-generation capabilities and traditional topic-modeling techniques. LLMs, by design, engage in an implicit form of topic modeling, as outlined in the research by (Wang et al., 2023). They navigate a latent conceptual space to generate text, making each token generation a decision influenced by an underlying topic variable. This implies that LLMs, despite not learning discrete topic variables explicitly like LDA, can infer and engage with these variables implicitly through their generative process. The challenge then becomes how to effectively extract or infer these latent topic representations from LLMs outputs, bridging the gap between the continuous, nuanced understanding of LLMs and the discrete topic models traditionally used in text analysis.

To bridge the gap between LLMs’ continuous generation process and discrete topic modeling, in

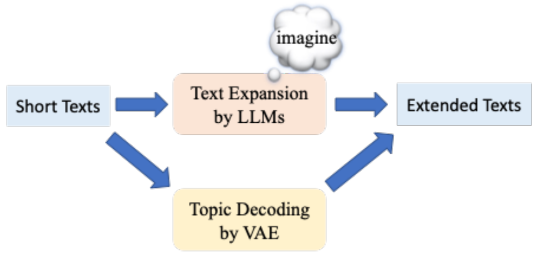

Figure 1: Overview of the proposed architecture.

this paper, we leverage variational autoencoders (VAEs) (Kingma and Welling, 2013). VAEs are a class of machine learning models designed to learn compressed, latent representations of data. Given a short text x, an LLM is used to generate an augmented, contextually enriched version x'. The VAE aims to learn a latent representation z that captures the underlying topic distribution of x, with the generative process designed to reconstruct x' from z. In other words, VAE works as a proxy to infer discrete topic representation (z) that is continuously inherent (θ) in LLMs while reconstructing extended texts from z.

To summarize, our **contributions** in this paper are the following. Firstly, we propose to leverage LLMs for its inherent topic modeling capability. Specifically, we extend a short text into a long sequence using LLMs (i.e., LLAMA2(Touvron et al., 2023)). Secondly, to decode the discrete topic representations from the continuous domain of text generation of LLMs, we use VAE. In other words, we use the VAE to learn topic representations of short texts by having the capability of regenerating the extend texts from LLMs. Finally, we conduct a comprehensive set of experiments on multiple datasets over different tasks, demonstrating our models’ superiority against existing baselines.

2 Proposed Methodology

Our proposed framework consists of two components. The first component generates longer text given a short text. The second one utilizes the generated longer texts for topic modeling. The overall framework is shown in Figure 1.

2.1 Short Text Extension

As specified before, according to (Wang et al., 2023), LLMs inherently perform topic modeling. This is achieved by treating each token generation as a decision informed by a latent topic or concept variable θ, suggesting that LLMs understand and generate text by navigating a latent conceptual space. More specifically, LLMs generate new tokens based on all previous tokens $P(w_{1:T}) = \prod_{i=1}^{T} P(w_i|w_{i-1}, \ldots, w_1)$ and it can be decomposed as below:

$$P_M(w_{t+1:T}|w_{1:t}) = \int_{\Theta} P_M(w_{t+1:T}|\theta) P_M(\theta|w_{1:t}) d\theta$$

where M is a specific LLM. This illustrates the LLM's process of generating text conditioned on previous tokens and a latent topic variable, integrating over all possible conceptual themes Θ that could inform the generation. However, we can not explicitly obtain the latent concept variable to understand the topic. Therefore, we formulate the short text extension as a conventional conditional sentence generation task, i.e., generating longer text sequences given a short text. Formally, we use the standard sequence-to-sequence generation formulation with a PLM $\mathcal{M}$: given input a short text sequence x, the probability of the generated long sequence $y = [y_1, \ldots, y_m]$ is calculated as:

$$\mathbf{Pr}_{\mathcal{M}}(y|x) = \sum_{i=1}^{m} \mathbf{Pr}_{\mathcal{M}}(y_i|y_{<i}, x),$$

where $y_{<i}$ denotes the previous tokens $y_1, \ldots, y_{i-1}$. The LLM $\mathcal{M}$ specific text generation function $f_{\mathcal{M}}$ is used for sampling tokens and the sequence with the largest $\mathbf{Pr}_{\mathcal{M}}(y|x)$ probability is chosen. Later, we use the extended text to decode the inferent topic in LLMs.

2.2 Topic Model on Generated Long Text

Upon optioning the longer text sequences from the previous step, one possible straightforward way can be using existing topic models that work better for long text documents. As the longer texts have better co-occurrence context than the original short texts, it is expected to reduce the data sparsity problem of short-text topic modeling. Therefore, exploring existing probabilistic and neural topic models is intuitive on top of the generated longer text sequences. Therefore, we directly utilize different existing topic models on generated texts as one solution.

However, as the pre-trained knowledge is directly used for text generation without finetuning on the target dataset, one possible issue with this straightforward approach is that the generated text may shift from the original domain or topic of the given short text (or partially cover the topics). One such inconsistency is shown in the third column of Table 1 where we see a longer sequence generated from a given short text using a LM GPT-2. We observe that the generated sequence is coherent and easily readable sentences with many related words to the given short text. E.g., as the short text has content about the court proceeding, the generated long text has many such related words like "judgment", "plaintiffs" and so on. However, the generated text has partially shifted from the original topic of the text. More specifically, the "sports" aspect of the given short text is entirely missing in the generated longer text. Therefore, only relying on this generated text for topic modeling will likely miss the expected topics distribution in the result.

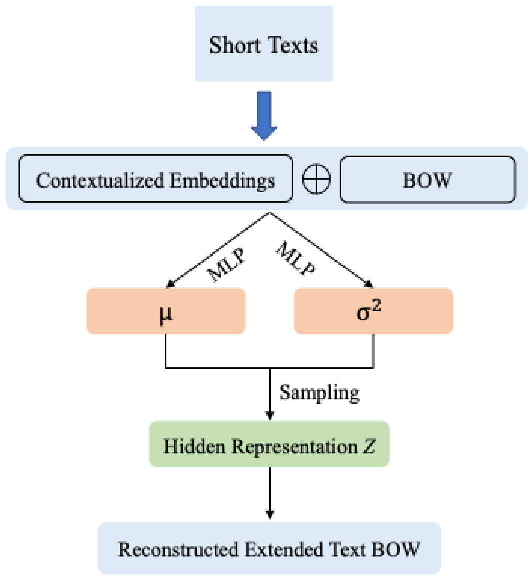

Figure 2: VAELink

To solve this issue, we propose a simple yet very effective solution by using variation autoencoders (VAEs), which we call VAELink: linking LLM's generated text with VAE, as shown in Figure 2.

VAELink: As solely relying on generated long texts creates the problem of topic shift or incomplete topic coverage of a document, we use the generated sequence only as output to be reconstructed from short text. Inspired by a previous work (Bianchi et al., 2020), we incorporate the contextualized representation of short text along with the given short text bow as input of the topic model. This will enrich the context information of the given short text without much deviation from the original topics of the text.

Formally, the model extends an existing topic model called ProdLDA (Srivastava and Sutton,

Short Texts	no tsunami but fifa's corruption storm rages on	court agrees to expedite n.f.l.'s appeal
Extended Texts	no tsunami but fifa's corruption storm rages on. fifa president sepp blatter speaks out about corruption scandals . but fifa's stewardship is far from over and fifa are not at fault . Fernando torres, fifa's head of integrity, is still alive and kicking . fa and fifa must stop corruption before fifa takes over . fifa fans are not safe when it comes to their vote, this is not the place..	court agrees to expedite N.F.L. appeal.May 5, 1987. The Third United States Circuit Court of Appeals issues an order denying Enron's request for summary judgment in his suit seeking summary judgment from Enron in his suit for injunctive relief to prevent Enron from misusing the trademark ""energy"" in commerce. Judge Joseph S.Tumlinson's order states that both plaintiffs..

Table 1: Example short texts and corresponding extended texts using PLMs.

2017). ProdLDA is a neural topic model based on the Variational AutoEncoder (VAE) mechanism (Kingma and Welling, 2013). The encoder part of this model maps the BOW representation of a document to a continuous latent representation by training a neural variational inference network. More specifically, the model first generates mean vector μ and variance vector σ^2 by two separate MLPs from a document. The μ and σ^2 are then used to sample a latent representation Z assuming Gaussian distribution. Then, a decoder network reconstructs the BOW representation of the extended long texts by LLMs by generating its words from Z. In our model, instead of using only the short text BOW as input, we concatenate it with the contextualized representation of generated long text using an embedding representation (i.e., SBERT (Reimers and Gurevych, 2019)). The model is trained with the original objective function (Srivastava and Sutton, 2017) called the evidence lower bound (ELBO) as follows:

$$\mathcal{L}(\Theta) = \sum_{d \in \mathcal{D}} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{dn} \mid Z_d)] - \sum_{d \in \mathcal{D}} KL(q(Z_d; w_d, \Theta) \mid\mid p(Z_d)), \quad (1)$$

where w_{dn} is the n-th token in a document d with length N_d from the corpus $\mathcal{D}$. Θ represents learnable parameters in the model. $q(\cdot)$ is a Gaussian whose mean and variance are estimated from two separate MLPs.

3 Experiments

In this section, we employ empirical evaluations, which are designed mainly to answer the following research questions (RQs):

- **RQ1.** How effectively does the proposed VAELink improve the performance of topic modeling for short texts?
- **RQ2.** Does the LLMs grounded text extension improve the performance of existing topic models?
- **RQ3.** How qualitatively different are the topics discovered by the proposed architecture from existing baselines?
- **RQ4.** How does the size of LLMs affect the performance of topic modeling?

Datasets	# of docs	Average length	# of class labels	Vocabulary size
StackOverflow	19899	4.49	20	2013
TagMyNews	4918	3.88	7	1410
WebSnippets	4067	14.52	8	12329

Table 2: Statistics of datasets after preprocessing.

3.1 Experiment Setup

Datasets. We use the following datasets to evaluate our proposed architecture. The detailed statistics of these datasets are shown in Table 2.

- **TagMyNews:** Titles and contents of English news articles published by Vitale et al. (2012) are included in this dataset . In our experiment, we use the headlines from the news as brief paragraphs. Every news item is given a ground-truth name, such as "sci-tech", "business", etc.
- **Google News:** The web content from Google search snippets makes up the dataset provided by Yin and Wang (2014). It is a snapshot of Google News on November 27, 2013. It includes the titles and brief descriptions of 11,109 news articles, which are organized into 152 distinct categories or clusters.
- **StackOverflow:** This dataset was created using the challenge information that was provided in Kaggle[2]. We make use of the dataset which contains 20,000 randomly chosen question titles. Information technology terms like "matlab", "osx", and "visual studio" are labeled next to each question title.

Baselines. We compare our models with the following baselines.

- **LDA:** We used one of the widely used proba-

[2] https://www.kaggle.com/datasets/stackoverflow/stackoverflow

bilistic topic models, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as a baseline for this work.

- **BTM:** Biterm Topic Model (Yan et al., 2013) uses extra structural information by directly constructing the topic distributions over unordered word pairs (biterms). This model is specialized for short text topic modeling.
- **NQTM:** A state-of-the-art neural short text topic model with vector quantization. (Wu et al., 2020)
- **CLNTM**: Contrastive Learning for Neural Topic Model combines contrastive learning paradigm with neural topic models by considering both effects of positive and negative pairs (Nguyen and Luu, 2021).
- **TSCTM**: It is another constrastive learning-based approach that uses quantization for better positive and negative sampling. (Nguyen and Luu, 2021).
- **CTM:** Contextualized Topic Model combines contextualized representations of documents with neural topic models (Bianchi et al., 2020).

We mainly use llama2 (Touvron et al., 2023) for extending short texts into longer texts. The implementation details are shown in Appendix A.

3.2 Topic Quality Evaluation

Evaluation Metrics. We evaluate each model using two different metrics: two for topic coherence (i.e., NPMI and CWE) and one for topic solution diversity (i.e., IRBO).

- C_V: We use the widely used coherence score for topic modeling named C_V. It is a standard measure of the interpretability of topics (Wu et al., 2020).
- $IRBO$: Inverted Rank-Biased Overlap (IRBO) evaluates the topic diversity by calculating rank-biased overlap over the generated topics introduced in (Webber et al., 2010).

Results and Discussions. We first analyze the quality of the topics from VAELink compared to state-of-the-art methods (described in Section 2.1). The topic quality scores (C_V, and $IRBO$) in Table 3 show the apparent dominance of VAELink. The best C_V and $IRBO$ scores for all three datasets are from VAELink with significant improvement in topic coherency and comparable diversity. This clearly shows that extension of short text using LLMs and encoding it through VAEs help discover higher-quality topics that are more coherent and diverse.

Now, considering the topic quality performance of the proposed VAELink, we identify some interesting findings. In almost all cases, we get an improvement in topic quality scores compared to the short-text counterparts. More specifically, in TagMyNews and StackOverflow datasets, we obtained a significant performance boost in terms of coherence and diversity scores compared to all other baselines. E.g., in the TagMyNews dataset, compared to the most similar model CTM, the C_V score for VAELink increases from 0.595 to 0.722 (for K=20 topics).

However, in the GoogleNews dataset, the improvement in topic quality is not as promising as baselines on extended texts. One possible reason for this is that this dataset's average document text length is extremely short (i.e., as shown in Table 2). And each of these short texts carries very limited (or absent) topic-indicative words. Therefore, while the VAE reconstructs this short text during training, the generated topics may become less coherent. On the other hand, for the baselines that solely use the generated long texts, this problem is resolved by coherent tokens from the extended texts.

3.3 Text Classification Evaluation

Although text classification is not the main purpose of topic models, the generated document topic distribution can be used as the document feature for learning text classifiers. Therefore, we evaluate how learned document topic distribution is distinctive and informative enough to represent a document to be used for classifying a document correctly. We employ four different classification models on top of document topic distribution learned by different models. The classification models are Support Vector Machine (SVM) (Cortes and Vapnik, 1995) and Logistic Regression (LR) (Wright, 1995). We use classification accuracy over 5-fold cross-validation to compare the performance of multiple classifiers.

Results and Discussions. The classification result is presented in Table 4. Overall, the proposed VAELink is the best-performing model regarding classification accuracy, leveraging both the generated text and considering the topics shift (or incomplete coverage of topics) problem. As specified before, when using LLMs without finetuning on the target corpus, the generated text may not cover the original topics of the document or shift from them.

Method	TagMyNews Titles				Google News				StackOverflow			
	K=20		K=50		K=20		K=50		K=20		K=50	
	C_V	$IRBO$	C_V	$IRBO$	C_V	$IRBO$	C_V	$IRBO$	C_V	$IRBO$	C_V	$IRBO$
LDA	0.399	0.981	0.369	0.983	0.326	0.996	0.347	0.998	0.453	0.980	0.396	0.991
BTM	0.399	0.959	0.401	0.974	0.341	0.995	0.383	0.995	0.003	0.893	0.447	0.919
NQTM	0.322	0.941	0.345	0.937	0.258	0.973	0.289	0.942	0.291	0.993	0.327	0.991
CLNTM	0.311	0.972	0.356	0.942	0.324	0.995	0.356	0.942	0.324	0.995	0.296	0.845
TSCTM	0.363	1.000	0.304	1.000	0.284	1.000	0.298	1.000	0.124	1.000	0.121	0.997
CTM	0.481	1.000	0.531	0.991	0.351	1.000	0.393	0.994	0.410	1.000	0.392	0.986
VAELink	**0.598**	**1.000**	**0.559**	**1.000**	**0.440**	**1.000**	**0.441**	**1.000**	**0.437**	**1.000**	**0.402**	**0.997**

Table 3: Topic coherences (CV) and diversity (IRBO) scores of topic words. K is the topic number. The best in each case is shown in **bold**.

	TagMyNews Titles				Google News				StackOverflow			
	K=50		K=100		K=50		K=100		K=50		K=100	
	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR
LDA	0.247	0.317	0.259	0.303	0.235	0.354	0.563	0.645	0.381	0.431	0.561	0.605
BTM	0.441	0.484	0.457	0.479	0.287	0.475	0.567	0.695	0.462	0.555	0.651	0.683
NQTM	0.123	0.254	0.123	0.254	0.023	0.0387	0.114	0.309	0.05	0.05	0.05	0.05
CLNTM	0.123	0.254	0.123	0.254	0.023	0.038	0.114	0.309	0.050	0.051	0.066	0.057
TSCTM	0.423	0.473	0.485	0.527	0.337	0.518	0.498	0.765	0.665	0.736	0.774	0.784
CTM	0.595	0.619	0.668	0.694	0.283	0.512	0.514	0.679	0.581	0.739	0.814	**0.817**
VAELink	**0.722**	**0.744**	**0.755**	**0.765**	**0.326**	**0.569**	**0.585**	**0.766**	**0.583**	**0.787**	**0.825**	**0.817**

Table 4: Text classification accuracy over 5-fold cross validation. The best results in each case are shown in **bold**.

Even if the StackOverflow dataset is about a particular technical domain, the LLMs are more likely to generate tokens from general domains. That is why the learned topics from the extended texts may not represent the original documents, resulting in poor classification performance. This effect is comparatively less in the other two datasets, as those are about more general topics like "politics", "sports", etc. On the other hand, the VAELink reduces this effect by using the original short texts as input during training, which is also visible in the classification result.

From the above results, it is evident that VAELink makes a tradeoff between topic quality and classification performance, while others improve in one direction only.

We have also shown the effect of the different generated text sizes on the topic quality in Appendix B.

3.4 Topic Examples Evaluation

To evaluate the proposed models qualitatively, we show the top five words for each of the three topics generated by different models in Table 5. We observe that some models on short texts generate topics with repetitive words (e.g., CLNTMa and

Models	Topic Words (on Short Text)	Topic Words (on LLAMA2 Long Text)
LDA	application window load open test linq oracle sql query table matlab update image value field	application spring api java library database query table sql oracle matlab image number size color
CLNTM	pl sql outer procedure join pl sql script mark os script pl sql linqtosql not	clause join query hql desc ipad usb iphone icloud player maven tomcat npm gradle restful
CTM	good best framework way web scala class method java object mac os osx run application	bash script shell command path svn repository git subversion branch sql database query oracle statement
BTM	use file visual excel studio use file magento drupal hibernate use magento file oracle way	-
GraphBTM	example axis applescript log properly derive hold partition line spreadsheet applescript parent hold example axis	-
NQTM	custom bit lambda depth map specific crash dead svn handling use file excel wordpress magento	-
VAELink	-	oracle database sql store procedure bash script command line shell ajax apache request rewrite jquery

Table 5: Topic words examples under k = 20.

BTM). Although the CTM on short texts generates diverse topics, they are less informative (i.e., with words like "best", "good", etc.). On the other hand, topics in generated long texts are less repetitive with much more coherency, although some also tend to generate topics with general words like "number" and "size". Finally, the VAELink generates both non-repetitive and informative topics. E.g., it is easy to detect that the three discovered topics are database, shell, and web programming.

4 Related Work

4.1 Traditional Topic Models

The widely used traditional topic models, also known as probabilistic topic models, such as Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), performs well when the given corpus consists of large-sized documents. These models assume that the documents have sufficient co-occurrence information to capture latent topic structures from the corpus. Thus, these models typically fail to infer high-quality topics from short texts such as news titles and image captions. To solve this issue, one strategy in existing probabilistic topic models uses structural and semantic information from texts such as Biterm Topic Model (BTM) (Yan et al., 2013). Another strategy aggregates a subset of short texts into a longer pseudo document using various metadata (e.g., hashtags, external corpora) before applying conventional topic models (Mehrotra et al., 2013; Zuo et al., 2016). Another line of short-text topic modeling restricts the document-topic distribution by assuming each document is sampled from a single topic such as Dirichlet Multinomial Mixture (DMM) model (Yin and Wang, 2014; Nigam et al., 2000). Although this is intuitive considering the limited context in shorts, this simplification may be too strict in practice as many short texts could cover more than one topic.

4.2 Neural Topic Models

With the recent developments in deep neural networks (DNNs) and deep generative models, there has been an active research direction in leveraging DNNs for inferring topics from corpus, also called neural topic modeling. The recent success of variational autoencoders (VAE) (Kingma and Welling, 2013) has opened a new research direction for neural topic modeling (Nan et al., 2019). The first work that uses VAE for topic modeling is called the Neural Variational Document Model (NVDM) (Miao et al., 2016), which leverages the reparameterization trick of Gaussian distributions and achieves a fantastic performance boost. Another related work called ProdLDA (Srivastava and Sutton, 2017) uses Logistic Normal distribution to handle the difficulty of the reparameterization trick for Dirichlet distribution.

There also have been several works in neural topic modeling (NTM) for short texts. E.g., (Zeng et al., 2018) combines NTM with a memory network for short text classification. (**?**) takes the idea of the probabilistic biterm topic model to NTM where the encoder is a graph neural network (GNN) of sampled biterms. However, this model is not generally able to generate the topic distribution of an individual document. (Lin et al., 2020) introduce the Archimedean copulas idea in the neural topic model to regularise the discreteness of topic distributions for short texts, which restricts the document from some salient topics. From a similar intuition, (Feng et al., 2022) proposes an NTM by limiting the number of active topics for each short document and also incorporating the word distributions of the topics from pre-trained word embeddings. Another neural topic model (Wu et al., 2020) employs a topic distribution quantization approach to generate peakier distributions that are better suited to modeling short texts.

4.3 PLMs in Topic Models

Previously, some neural topic models attempted to use PLMs as input representations of given documents. E.g., a model called the contextualized topic model (CTM) (Bianchi et al., 2020) complements the Bag of Words (BOW) representation of a document with its contextualized vector representation from PLMs like BERT (Devlin et al., 2018). As PLMs are pre-trained on large-scale text corpora such as Wikipedia and hold rich linguistic features, they are supposed to capture the context and order information in a text ignored in BOW representation. Similarly, BERTopic (Grootendorst, 2022) also uses PLM-based document embedding to cluster them and TF-IDF to find representative words from each cluster as topics. However, as it uses TF-IDF metrics, it fails to take benefit of the distributed representations of PLMs, which are better at capturing word semantics than frequency-based statistics. Moreover, the above approaches do not solve the data sparsity problem in short text topic modeling but rather use PLMs only for better representation of input documents for general-purpose topic modeling. Unlike these neural topic models, the proposed framework in this paper uses PLMs to enrich contextual information of short documents by conditional text generation.

5 Conclusion

In this paper, we proposed a simple yet effective approach for short-text topic modeling leveraging the "imagination" capability of PLMs. To solve the data-sparsity problem of short texts, we first extend them into longer sequences using a PLM. These longer sequences are then used to mine topics by existing topic models. To further reduce the effect of the domain-shift problem of a pre-trained model, we propose a solution extending a neural topic model. A set of empirical evaluations demonstrate the effectiveness of the proposed framework over the state-of-the-art.

Limitations

The proposed framework directly utilize PLMs for text generation conditioned on the given short texts. As we have specified before, this may result in noisy out-of-domain text generation, which hurts the document representativeness of the generated topics. This problem may worsen when the target domain is very specific. Although the proposed LCSNTM tries to solve this problem by a simple mechanism of short text reconstruction, it does not work in extreme sparsity scenarios, as we observed in the TagMyNews dataset. Therefore, controlling the generation process such that it outputs more relevant text in the target domain is a possible future research direction in this line.

References

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jiachun Feng, Zusheng Zhang, Cheng Ding, Yanghui Rao, Haoran Xie, and Fu Lee Wang. 2022. Context reinforced neural topic modeling over short texts. *Information Sciences*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, and Jiawei Han. 2010. Pet: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 929–938.

Lihui Lin, Hongyu Jiang, and Yanghui Rao. 2020. Copula guided neural topic modelling for short texts. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1773–1776.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2012. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 265–274.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. *arXiv preprint arXiv:1907.12374*.

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

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. In *Fourth international AAAI conference on weblogs and social media*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Daniele Vitale, Paolo Ferragina, and Ugo Scaiella. 2012. Classification of short texts by deploying topical annotations. In *European Conference on Information Retrieval*, pages 376–387. Springer.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*, page 3.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270.

Raymond E Wright. 1995. Logistic regression.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. 2018. Topic memory networks for short text classification. *arXiv preprint arXiv:1809.03664*.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338–349. Springer.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2105–2114.

A Implementation Details.

There are some parameters for both the proposed architecture and baselines we need to set. For text generation from LLMs. we use the maximum new tokens length as 500. We find that using beam-search decoding with a beam size of 2 generates more coherent text for BART, while multinomial sampling works better in GPT-2 and T5 for all three datasets. The number of iterations for all the topic models is set to 100, except LDA uses 200 as the maximum number of iterations. For the contextualized representation of input documents in CTM and LCSNTM, we use pre-trained SBERT[3] with a maximum sequence length of 512. All parameters during calculating evaluation metrics are set to the same value across all the models. E.g., the number of top words for each topic for calculating C_V and IRBO is set to 10. In text classification experiments, we use the default parameters for MNB from scikit-learn[4]. For SVM, we use the hinge loss with the maximum iteration of 5. For logistic regression, the maximum iteration is set to 1000, and the tree depth for RF is set to 3 with the number of trees as 200.

B Effect of extended text lengths

In this section, we analyzed the effect of generated text length on the topic quality (shown in 6). Here, we use GPT2 on CTM (as it purely uses extended texts, the effects will be easily analyzed). We use different generated text sizes of 10, 20, 50, and 100. Here, for almost all the cases, we can see improvement in topic quality in coherence (NPMI, CWE) when we increase the minimum generated sequence length with stable diversity scores (IRBO). This shows that when we have more context in the generated text, the learned topics are more coherent (interpretable) without hampering diversity.

Text-Length	20	30	50	100
Stack Overflow				
NPMI	0.072	0.077	0.082	0.083
CWE	0.157	0.158	0.159	0.153
IRBO	0.992	0.992	0.992	0.994
TagMyNews				
NPMI	0.032	0.037	0.044	0.045
CWE	0.189	0.201	0.199	0.201
IRBO	0.991	0.992	0.992	0.990
WebSnippets				
NPMI	-0.015	-0.028	-0.008	0.008
CWE	0.227	0.212	0.237	0.234
IRBO	0.992	0.990	0.992	0.996

Table 6: Effect of generated text length on Topic quality

[3] https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v2

[4] https://scikit-learn.org