Prefix-VAE: Efficient and Consistent Short-Text Topic Modeling with LLMs

Anonymous EMNLP submission

Abstract

Topic models are 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- whether probabilistic or neural- mostly struggle to mine patterns from them to generate coherent topics. In this paper, we first explore the capability of large language models (LLMs) to generate longer texts from shorter ones before applying them to traditional topic modeling. To further improve the efficiency and solve the problem of the semantic inconsistency from LLM-generated texts, we propose to use prefix tuning to train a smaller language model coupled with a variational autoencoder for 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

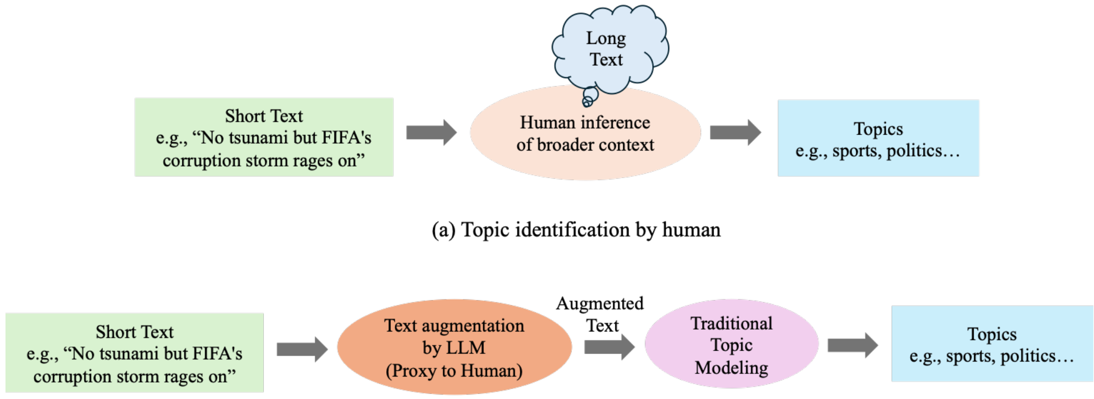
In the digital era, short texts like tweets, web page titles, news headlines, image captions, and product reviews are prevalent for sharing knowledge. However, the sheer volume of these texts necessitates efficient information extraction mechanisms. Topic modeling is a key method for uncovering latent topics in short texts, with applications including 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, such as LDA and PLSA, are designed to uncover latent topics given a corpus of documents by analyzing word co-occurrences within the texts (Blei et al., 2003; Hofmann, 1999). These models assume that each document contains enough text to provide meaningful co-occurrence information. However, in the case of short-text documents such as titles, captions, and headlines, this assumption does not hold due to the limited text available in each document. This scarcity of text per document leads to a data sparsity problem, where the limited word co-occurrences make it difficult for traditional models to effectively mine high-quality topics. In this context, the primary challenge is that each individual document is short, rather than the corpus itself being insufficient in size.

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 limitations mentioned above, in this paper, we first try to understand the characteristics of short texts and how humans process these texts when detecting topics. A short text, such as a title or caption, typically serves as a summarized version of a longer text, providing readers with essential hints about the full content. When judging the topics of short texts, humans often infer the broader context based on their background knowledge and the cues provided in the text. For example, given the headline: "No tsunami but FIFA's

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



(a) Topic identification by human

(b) LLM-based Text Expansion for Short-text Topic Modeling

Figure 1: LLM for short-text topic modeling

corruption storm rages on," readers might use their understanding of "FIFA" to infer that the headline pertains to the topic of "sports."

This leads us to the question: Can a model similarly infer the broader 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 demonstrated remarkable capabilities as open-ended text generators, capable of producing surprisingly fluent text from a limited preceding context. For example, given the abovementioned news headline, LLMs can generate extended sequences (as shown in the third and fourth columns of Table 1 with tokens such as "FIFA World Cup" and "Soccer," which are strongly related to the sport of soccer. This ability to generate contextually relevant information suggests that LLMs can be leveraged to enrich the contextual information of short texts, thereby improving topic modeling.

Considering these capabilities, we first explore the potential solution for short-text topic modeling: leveraging large language models (LLMs) to generate a longer text from each short text in a corpus before applying traditional topic modeling techniques. By expanding short texts into more detailed, context-rich narratives, LLMs can create a proxy for the detailed context that traditional topic modeling techniques often lack when dealing with short texts. In other words, it is a proxy of human-like inference of the broader context surrounding a given short text before mining the topics, as shown in Figure 1.

While leveraging LLMs to expand short texts offers a promising solution, this approach faces several significant **challenges**. First, there is the challenge of *semantic consistency*: ensuring that the generated longer texts accurately reflect the original short texts without introducing irrelevant or inaccurate information is difficult, as LLMs are not always fine-tuned for specific tasks or domains. This can lead to a shift in meaning, distorting the topic modeling results. Second, the issue of *scalability* presents a challenge: generating extended texts for a large corpus of short texts is computationally expensive and time-consuming, making it impractical for real-time applications and large-scale datasets. Although generating texts offline during training might be permissible, the inference time required for real-time topic detection can be impractical.

To tackle these challenges, we aim to avoid directly using LLM-generated longer texts as input. Instead, we train a model to learn topics from short texts and reconstruct longer texts previously generated by an LLM. This minimizes the effects of any shift in meaning in the generated texts. By decoding topics from short texts before generating longer texts, we align with one of the LLM's inherent characteristics. As noted by (Wang et al., 2023), LLMs implicitly engage in topic modeling by navigating a latent conceptual space to generate text, with each token generation influenced by an underlying topic variable. However, directly inferring these latent concepts into discrete topics like Latent Dirichlet Allocation (LDA) is not straightforward.

To bridge this gap, we introduce the *Prefix-VAE Topic Model* (P-VTM), which combines a smaller language model (LM) with a variational autoencoder (VAE) for topic inference. Instead of tuning the entire LM, we employ prefix tuning (Li and Liang, 2021), which fine-tunes only a small set of parameters, effectively capturing domain-specific features from short texts. This reduces the risk of meaning shift associated with larger, general LLMs. The extracted features serve as input for a VAE to decode discrete topics. Both LM and VAE are trained end-to-end on a topic modeling objective.

The key insights of our solution include – (1) Semantic Consistency: By training on short texts and using generated longer texts only as output, we ensure the integrity of the original data and mitigate the risk of introducing irrelevant information. (2) Efficiency: The reduced inference time of smaller LMs and the efficiency of VAEs in learning discrete topics make this method suitable for real-time topic detection applications. (3) Prefix Tuning: This fine-tuning method allows us to capture domain-specific features without the computational overhead of tuning large LLMs, ensuring scalability.

To summarize, our **contributions** in this paper are the following. Firstly, we explore LLMs for extending short texts into longer ones and then apply traditional topic models to the longer texts. Secondly, to improve efficiency and solve the meaning shift problem, we propose a new framework consisting of a jointly trained smaller LM and VAE. Finally, we conduct a comprehensive set of experiments on multiple datasets over different tasks, demonstrating our models' superiority against existing baselines.

2 Related Work

2.1 Traditional Topic Models

Traditional probabilistic topic models like Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) work well with large-sized documents, relying on ample co-occurrence information to capture latent topic structures. However, these models often struggle with short texts such as news titles and image captions. To address this, the Biterm Topic Model (BTM) (Yan et al., 2013) utilizes structural and semantic information, while another strategy aggregates short texts into longer pseudo-documents using metadata (e.g., hashtags, external corpora) before applying conventional topic models (Mehrotra et al., 2013; Zuo et al., 2016). Another approach, the Dirichlet Multinomial Mixture (DMM) model (Yin and Wang, 2014; Nigam et al., 2000), assumes each document is sampled from a single topic. Although intuitive, this assumption can be overly restrictive as many short texts may cover multiple topics.

2.2 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. (Zhu et al., 2018) 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.

2.3 LMs in Topic Models

Previous neural topic models have used language models (LMs) to represent documents. For example, the contextualized topic model (CTM) (Bianchi et al., 2020) combines a document's Bag of Words (BOW) representation with its contextu-

Prompt	Given the short text: "{Short Text}", expand it into a detailed paragraph that provides background and elaborates on the key points to enrich its context. Try to make it as detailed as possible.
Short Text	No tsunami but FIFA's corruption storm rages on
Extended Text (T5-XXL)	After years of corruption scandals, FIFA is in the grip of an even greater one. A year after FIFA's biggest corruption scandal in its history, the world soccer governing body continues to roil with revelations, and the WTA has called for the re-establishment of a special committee to investigate the scandal. A year after the FIFA Ethics Committee launched its investigation into allegations of corruption in the sport, the investigation is still going strong, with new allegations surfacing on a daily basis. After the FIFA World Cup corruption scandal, a wave of resignations has swept across the organization. A tsunami of scandals engulfed FIFA as it awaited a decision on the replacement of the new president. A year after FIFA's most serious corruption scandal broke, the corruption storm continues to rage with FIFA chiefs.
Extended Text (LLAMA2)	Despite the absence of a tsunami, the corruption storm surrounding FIFA, the governing body of international soccer, continues to rage on. The organization has been embroiled in a series of scandals over the past few years, with allegations of bribery, money laundering, and other forms of corruption among top officials and sponsors. The most recent controversy involves the awarding of the 2022 World Cup hosting rights to Qatar, which has been marred by allegations of vote-buying and other unethical practices. The fallout from these scandals has been significant, with several top FIFA officials arrested and charged with criminal offenses, and the organization's reputation taking a hit as a result. Despite efforts to clean up its act, the corruption storm continues to rage on, with many calling for greater transparency and accountability within the organization.

Table 1: Example short text and corresponding extended texts using two different LLMs.

alized vector from LMs like BERT (Devlin et al., 2018), capturing context and order information that BOW misses. Similarly, BERTopic (Grootendorst, 2022) uses LM-based document embeddings for clustering and TF-IDF to identify representative words as topics. However, BERTopic's reliance on TF-IDF doesn't fully utilize LMs' ability to capture word semantics. DeTime (Xu et al., 2023a) improves clusterability and semantic coherence by using Encoder-Decoder-based LLMs for embeddings. Despite these advances, these models don't address the data sparsity issue in short text topic modeling; they only improve document representation for general-purpose topic modeling. In contrast, our proposed framework leverages LMs for conditional text generation to enrich the contextual information of short documents.

3 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.

3.1 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 inherent topic in LLMs.

3.2 Topic Model on Generated Long Text

Upon obtaining the longer text sequences from the previous step, one straightforward approach is to use existing topic models that perform better with 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. Thus, exploring existing probabilistic and neural topic models on the generated longer text sequences is intuitive. Therefore, we directly utilize different existing topic models on generated texts as one solution, as shown in Figure 1.

However, directly using LLMs generated text for topic modeling may pose a risk. The generated text might shift from the original domain or only partially cover the intended topics. For example, consider a short text about "renewable energy sources":

- *Original short text:* "Renewable energy sources like solar and wind power are essential for reducing carbon emissions and combating climate change."
- *ChatGPT-generated longer text (OpenAI, 2023):* "Renewable energy sources, such as solar power and wind turbines, are becoming increasingly popular worldwide. These sources harness natural elements to generate electricity, contributing to the reduction of greenhouse gases. Solar panels capture sunlight and convert it into energy, while wind turbines use the wind's kinetic energy. Additionally, hydroelectric power, geothermal energy, and biomass are also crucial renewable sources. Countries are investing heavily in these technologies to transition from fossil fuels to cleaner energy solutions."

While the generated text provides a detailed overview of various renewable energy sources, it introduces new topics like hydroelectric power, geothermal energy, and biomass. This expansion can be beneficial for providing a broader context but may deviate from the original focus on solar and wind power. The opposite scenario is also possible, where the original short text is about multiple topics, and the generated long text is missing some of these topics, leading to incomplete topic coverage in a document.

To solve this issue, we propose a solution called Prefix-VAE Topic model (P-VTM), as shown in Figure 2.

P-VTM: To address the issues of deviations from the original focus or incomplete topic coverage in generated long texts, we employ the generated

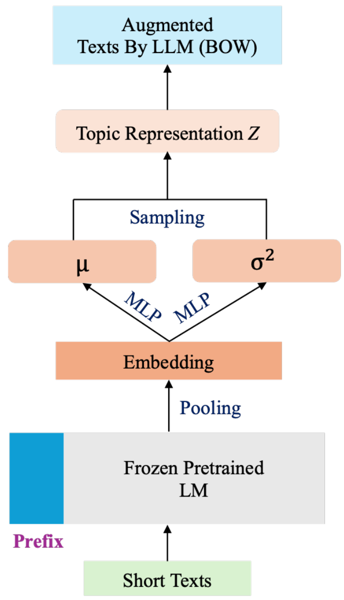


Figure 2: Proposed Architecture of P-VTM

sequence solely as an output to be reconstructed from short text. Formally, our model builds upon an existing topic model known as ProdLDA (Srivastava and Sutton, 2017). ProdLDA is a neural topic model based on the Variational AutoEncoder (VAE) mechanism (Kingma and Welling, 2013). The encoder component of this model maps the BOW representation of a document to a continuous latent representation by training a neural variational inference network. Instead of using BOW input, we employ a smaller language model to encode input short texts for learning features specific to the topic modeling task. However, training the entire LM on this task might be computationally intensive, and we may not need to train the entire set of parameters of the LM. Therefore, we use a parameter-efficient tuning method called Prefix tuning. Prefix-tuning trains a much smaller set of parameters to adjust the model towards a specific task.

We then use the output of the LM as the input for the VAE to perform topic inference. Specifically, the model first generates a mean vector μ and a variance vector σ^2 through two separate MLPs from a document. The μ and σ^2 are then used to sample a latent representation Z assuming a Gaussian distribution. Subsequently, a decoder network reconstructs the BOW representation of the extended long texts generated by LLMs by generating words from Z. The model is trained with the original objective function (Srivastava and Sutton, 2017) called the evidence lower bound (ELBO), defined as follows:

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

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.

4 Experiments

In this section, we employ empirical evaluations, which are designed mainly to fulfill the following objectives:

- How effectively does the proposed P-VTM improve the performance of topic modeling for short texts?
- Does the LLMs grounded text extension improve the performance of existing topic models?
- How qualitatively different are the topics discovered by the proposed architecture from existing baselines?

4.1 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 11,108 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.

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

Datasets	# of docs	Average length	# of class labels	Vocabulary size
TagMyNews Titles	5000	5.78	7	7111
Google News	11108	6.11	152	7187
StackOverflow	19899	4.49	20	8556

Table 2: Statistics of datasets after preprocessing.

- **LDA:** We used one of the widely used probabilistic topic models, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as a baseline for this work.
- **NQTM:** A state-of-the-art neural short text topic model with vector quantization. (Wu et al., 2020)
- **CTM:** Contextualized Topic Model combines contextualized representations of documents with neural topic models (Bianchi 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).
- **vONTSS**: This method (Xu et al., 2023b) presents a semi-supervised neural topic modeling method that leverages von Mises-Fisher (vMF) based variational autoencoders and optimal transport. This approach optimizes topic-keyword quality and topic classification by using a small set of keywords per topic.
- **DeTime**: DeTime (Xu et al., 2023a) leverages encoder-decoder-based large language models (LLMs) to produce highly clusterable embeddings that generate topics with superior clusterability and enhanced semantic coherence. another constrastive learning-based approach that uses quantization for better positive and negative sampling. (Nguyen and Luu, 2021).

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

4.2 Topic Quality Evaluation

Evaluation Metrics. For evaluating the topic quality of each model, we use following two different metrics:

- 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.,

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	C_V c	$IRBO$	C_V	$IRBO$	C_V	$IRBO$	C_V	$IRBO$
LDA	*ST*	0.399	0.981	0.369	0.983	0.326	0.996	0.347	0.998	0.413	0.980	0.396	0.991
	ET	0.523	0.979	0.498	0.989	0.414	0.99	0.433	0.991	0.501	0.638	**0.492**	0.935
NQTM	*ST*	0.322	0.941	0.345	0.937	0.258	0.973	0.289	0.942	0.291	0.993	0.327	0.991
	ET	0.542	1	0.551	0.999	0.405	1	0.468	1	0.301	1	0.218	1
CTM	*ST*	0.481	1.000	0.531	0.991	0.351	1.000	0.393	0.994	0.410	1.000	0.392	0.986
	ET	0.618	0.997	0.566	0.991	0.421	0.988	0.472	0.995	0.411	0.994	0.437	0.99
CLNTM	*ST*	0.311	0.972	0.356	0.942	0.324	0.995	0.356	0.942	0.324	0.995	0.296	0.845
	ET	0.613	0.988	0.541	0.979	**0.503**	0.999	0.513	0.994	0.412	0.998	0.438	0.99
TSCTM	*ST*	0.363	1.000	0.304	1.000	0.284	1.000	0.298	1.000	0.124	1.000	0.121	0.997
	ET	0.585	1	0.391	1	0.35	1	0.338	1	0.151	1	0.108	1
vONT	*ST*	0.409	0.788	0.397	0.93	0.349	0.981	0.348	0.933	0.281	0.723	0.358	0.868
	ET	0.536	0.994	0.457	0.983	0.418	0.999	0.404	0.991	0.413	0.998	0.392	0.982
DeTime	*ST*	0.398	0.779	0.403	0.922	0.288	0.719	0.326	0.903	0.279	0.664	0.361	0.849
	ET	0.427	0.976	0.37	0.963	0.371	0.954	0.32	0.938	0.3812	0.797	0.36	0.907
P-VTM		**0.632**	**1.000**	**0.585**	**1**	0.445	**1**	**0.452**	**1**	**0.558**	**1**	0.462	**1**

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**. *ST:* Short Texts, *ET:* Extended Texts (by LLAMA2)

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 result of existing topic models on the generated text from an LLM (described in Section 3). The topic quality scores (C_V, and IRBO) in Table 3 show the apparent dominance of topic models on extended text compared to short texts. The best NPMI and IRBO scores for all three datasets are from extended texts with significant improvement in topic coherency and comparable diversity. This clearly shows that the extension of short text using LLMs helps discover higher-quality topics that are more coherent and diverse. For example, in LDA, while using extended texts, the coherence score C_V improves from 0.399 to 0.523 compared to short texts.

However, these topic quality results do not always show that the mined topics correctly represent the target dataset. As specified in Section 3.2, the topics may shift because of the LLM-generated texts. We further discuss this through classification results in the next section. Now, considering the topic quality performance of the proposed P-VTM, we identify some interesting findings. In almost all cases, we get an improvement in topic quality scores compared to both the short-texts and extendeded texts counterparts . More specifically, 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 P-VTM increases from 0.618 to 0.632 (for K=20 topics).

4.3 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 two 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 P-VTM 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. Even if the StackOverflow dataset is about a partic-

Method		TagMyNews Titles				Google News				StackOverflow			
		K=20		K=50		K=20		K=50		K=20		K=50	
		SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR
LDA	*ST*	0.247	0.317	0.259	0.303	0.235	0.354	0.432	0.535	0.381	0.431	0.561	0.605
	ET	0.695	0.718	0.725	0.737	0.292	0.531	0.529	0.737	0.522	0.588	0.658	0.707
NQTM	*ST*	0.123	0.254	0.123	0.254	0.023	0.038	0.114	0.309	0.05	0.05	0.05	0.05
	ET	0.172	0.249	0.188	0.241	0.013	0.037	0.011	0.028	0.049	0.054	0.048	0.055
CTM	*ST*	0.595	0.619	0.668	0.694	0.283	0.512	0.514	0.679	**0.705**	0.739	0.814	**0.817**
	ET	0.686	0.721	0.736	0.777	0.339	0.547	0.592	0.762	0.462	0.58	0.656	0.719
CLNTM	*ST*	0.165	0.26	0.165	0.251	0.02	0.066	0.05	0.095	0.065	0.121	0.05	0.1
	ET	0.703	0.718	0.72	0.736	0.343	0.619	0.565	0.782	0.522	0.659	0.624	0.67
TSCTM	*ST*	0.423	0.473	0.485	0.527	0.337	0.518	0.498	0.685	0.565	0.736	0.774	0.784
	ET	0.721	0.751	0.755	0.773	0.314	**0.699**	0.594	0.63	0.557	0.657	0.687	0.726
vONT	*ST*	0.316	0.447	0.166	0.459	0.217	0.474	0.125	0.545	0.412	0.605	0.366	0.662
	ET	0.562	0.721	0.305	0.72	0.15	0.473	0.093	0.45	0.188	0.312	0.167	0.331
DeTime	*ST*	0.145	0.254	0.123	0.254	0.038	0.028	0.031	0.038	0.05	0.1	0.05	0.1
	ET	0.511	0.602	0.176	0.274	0.054	0.142	0.029	0.038	0.059	0.088	0.051	0.075
P-VTM		**0.722**	**0.744**	**0.755**	**0.765**	**0.366**	0.569	**0.595**	**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**.

Models	Topic Words (on Short Text)	Topic Words (on LLAMA2 Long Text)
LDA	application,different,session,edit, use,install,compile,long,design,setup	app,library,use,build,cocoa,project, application,dependency,framework,include
NQTM	image,come,null,application,pdf, hard,qstring,behave,repo,dynamically	spring,application,development,framework,web security,developer,platform,integrate,scalable
CTM	cocoa,mac,app,os,application, osx,iphone,detect,development,audio	spring,application,hibernate,configure,transaction, configuration,session,database,security boot
CLNTM	mac,os,matlab,bash,command, qt,osx,context, url,rewrite	mac,app,os, apple,device, audio,video,cocoa,screen,quality
TSCTM	eexample,axis,applescript,log,properly, derive,hold,partition,line,spreadsheet	studio,fxcop,visual,oslo,projects, awesome,editions,addon,eee,sharp
vONT	oracle,cocoa,sql,datum,application, subversion,convert,different,select,xml	branch,tuple,relational,orm,right, operator,standard,tree,trunk,left
DeTime	bash,sharepoint,page,class,table, string,load,line,variable,item	shell,operator,icon,question,review, second,optimization,word,account,editor
P-VTM	-	oracle database sql store procedure bash script command line shell

Table 5: Topic words examples under k = 10.

ular 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 P-VTM reduces this effect by using the original short texts as input during training, which is also visible in the classification result.

4.4 Topic Examples Evaluation

To evaluate the proposed models qualitatively, we show the top 10 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., CLNTM). 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 P-VTM 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.

5 Conclusion

In this paper, we address the issue of topic modeling for short texts. Our approach focuses on improving the input representation of short texts and enhancing the model's ability to capture latent topics despite the limited contextual information. The input to our method consists of individual short texts, such as a collection of tweets or headlines, and the output is a set of coherent topics that summarize the main themes present in the corpus. By tackling the data sparsity problem, we aim to develop a more effective topic modeling framework for short texts. A set of empirical evaluations demonstrate the effectiveness of the proposed framework over the state-of-the-art.

Limitations

The proposed framework directly utilize LLMs 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 P-VTM tries to solve this problem, 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*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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.

OpenAI. 2023. Chatgpt: A large language model trained by openai. https://www.openai.com/chatgpt. Accessed: 2024-06-15.

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*.

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.

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

Weijie Xu, Xiaoyu Jiang, Srinivasan Sengamedu Hanumantha Rao, Francis Iannacci, and Jinjin Zhao. 2023b. vONTSS: vMF based semi-supervised neural topic modeling with optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4433–4457, Toronto, Canada. Association for Computational Linguistics.

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 5 generates more coherent text. The number of iterations for all the topic models is set to 100. For the smaller pretrained language model we use 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.

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

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