Benchmarking Neural Topic Models: An Empirical Study

Thanh-Nam Doan
University of Tennessee at Chattanooga
Chattanooga, TN, USA
thanh-nam-doan@utc.edu

Tuan-Anh Hoang
VNU University of Science
334 Nguyen Trai, Hanoi, Vietnam
hoangtuananh@hus.edu.vn

Abstract

Neural topic modeling approach has been attracting much attention recently as it is able to leverage the advantages of both neural networks and probabilistic topic models. Previous works have proposed several models that are based on this framework and obtained impressive experimental results compared to traditional probabilistic models. However, the reported result is not consistent across the works, making them hard for gaining a rigorous assessment of these approaches. This work aims to address this issue by offering an extensive empirical evaluation of typical neural topic models in different aspects using large, diverse datasets as well as a thorough set of metrics. Precisely, we examine the performance of these models in three tasks, namely *uncovering cohesive topics*, *modeling the input documents*, and *representing them for downstream classification*. Our results show that while the neural topic models are better in the first and the third tasks, the traditional probabilistic models are still a strong baseline and are better in the second task in many cases. These findings give us more insights for choosing off-the-shelf topic modeling toolboxes in different contexts, as well as for designing more comprehensive evaluation for neural topic models.

1 Introduction

Classical topic modeling approach consists of statistical learning methods for uncovering the latent topics from a corpus and the semantic meaning of each document in the corpus. Notable works include the pioneering ones by (Hofmann, 1999; Blei et al., 2003). During its more than 20 years of research, topic modeling has also been applied to other fields beyond its original scope such as image analysis (Fei-Fei and Perona, 2005), and recommender systems (McAuley and Leskovec, 2013).

Recently, neural topic modeling approach has been attracting much research attention for understanding topic models of corpus due to its ability to leverage the advantages of both neural networks and probabilistic generative models. Compared to classical ones, this approach has three main advantages. First, its inference is amortized (Miao et al., 2016) and hence is much computationally simpler than that of the classical approach which requires to deal with a complicated optimization problem (Blei et al., 2003). Second, the gap between prototype and deployment processes become closer thanks to the power of some deep learning frameworks such as Pytorch (Paszke et al., 2019), Tensorflow (Abadi et al., 2016), and Flux.jl (Innes, 2018). Third, neural topic models are easy to be integrated with prior knowledge such as pre-trained word and text embeddings, which are prevalent and have shown the tremendous usefulness, e.g. GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019).

Several neural topic models are proposed recently to investigate the above advantages. Their reported experimental results are promising compared to the classical topic models. However, the reported results are not consistent or even contradict across these works. For example, for the same 20News dataset, (Miao et al., 2016) reports that their model i.e. NVDM obtains much better performance (measured by perplexity) than the classical LDA does, while the same experiments of (Srivastava and Sutton, 2017) show that LDA outperforms NVDM significantly. All these inconsistency and contradiction make it hard to assess the neural topic modeling approach comprehensively.

In this work, we would like to address this gap by conducting an extensive empirical study to accurately evaluate the typical existing neural topic models. Precisely, using a diverse set of large datasets, we examine the performance of the models in several tasks, including document modeling, topic discovery, and document representation for downstream classification. Our contributions are:

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4363–4368

- To the best of our knowledge, our work is the first one which examines state-of-the-art neural topic models in a systematical mechanism.
- We also public our implementation [1] as an additional resource for better reproducibility and further improvement in topic modeling research.
- Based on the results of our extensive experiments, we provide some practical guidelines of using neural topic models for specific tasks.

2 Experiment Settings

In this section, we describe settings used in our experiments. We start by briefing the models chosen to examined. We then introduce the datasets and define the metrics used to examine the models.

2.1 Evaluated Models

Neural topic models. Existing neural topic models are generally based on encoder-decoder architectures. In those models, the encoders are some variants of variational auto-encoder (VAE) (Kingma and Welling, 2013) whose input is the bag-of-word vector of an input document. The decoders, on the other hand, are designed to recover the input document from the encoded vector. For a fair comparison with the classical topic models that work purely on bag-of-words representation of documents, we surveyed the state-of-the-art works on neural topic modeling and chose to examine the following models whose decoders also work with the same representation of the documents.

- **NVDM** (Miao et al., 2016): A pioneering work that applies the VAE architecture for topic modeling with the encoder is implemented by multilayer perceptron, the variational distribution is a Gaussian distribution, and the variational inference is based on minimizing the KL-divergence (Blei et al., 2017).
- **GSM** (Miao et al., 2017): A variant of **NVDM** whose variational distribution is softmax of a Gaussian distribution.
- **NVLDA** (Srivastava and Sutton, 2017): Another variant of **NVDM** in which the variational distribution is Dirichlet distribution and approximated by a Laplace approximation based on Gaussian distribution.
- **ProdLDA** (Srivastava and Sutton, 2017): A variant of **NVLDA** in which the decoder is designed by following the product of expert model and the topics are unnormalized.
- **Scholar** (Card et al., 2018): This model is designed to incorporate the metadata and labels associated with documents into the modeling of topics. If no metadata and no labels are available, this is similar to **ProdLDA** except that the encoder has only a single linear layer.
- **NSMDM** (Lin et al., 2019): The sparsity of documents' topics is modeled by Gaussian sparsemax distribution, and the variational inference is based on the Wasserstein distance.
- **NSMTM** (Lin et al., 2019): This model is a variant of **NSMDM** in which the Gaussian sparsemax distribution is also used for modeling the sparsity of the topics .
- **NVCTM** (Liu et al., 2019): Again, this model is much similar to **NVDM** except that the decoder make uses of a complicated transformation sequence for modeling the correlation of topics within documents.

The above list is certainly not exhaustive. There are several other notable models proposed recently. We however do not examine them in this work as they either do not model documents as bags but sequences of words (e.g., (Dieng et al., 2016)) or are not generative models (e.g., (Yang et al., 2020)).

In our experiments, for each of the aforementioned models, we make use of the implementations provided by the authors if these are any. We re-implement the provided implementation if it is not written in Pytorch, strictly follow the original one. If the implementation is not provided, we implement the models ourselves, strictly follow the description and settings in the published papers.

Classical topic models. We compare these methods above with two classical topic models, namely Non-negative Matrix Factorization (**NMF**) (Zhao et al., 2017) and Latent Dirichlet Allocation (**LDA**) (Blei et al., 2003). We examine both two widely used learning methods for LDA: online variational inference (**o_LDA**) and Gibbs sampling algorithm (**g_LDA**).

2.2 Datasets

We uses two types of corpus: long text corpus containing *W2E-content* and *20News* (Lang, 2008);

[1] https://github.com/smutahoang/ntm

Table 1: Dataset statistic

Dataset	#documents	#words	Avg. length	#labels
20News	15,465	4,159	73.52	20
W2E-content	84,017	11,123	256.62	30
W2E-title	105,522	4,051	6.90	30
Web Snippets	12,295	4,722	14.42	8

and short text corpus including *W2E-title* and *Web Snippets* (Ueda and Saito, 2003). Both W2E-content and W2E-title are derived from news articles in W2E dataset (Hoang et al., 2018) by using the whole content or title respectively of the articles in the top 30 topics (by the number of news articles). Documents in these datasets are labeled by their topics (*W2E-content* and *W2E-title*), category of the discussion they belong to (*20News*), or the category of the webpage they were collected from (*Web Snippets*).

For each dataset, we preprocess the datasets by: removing stopwords, then iteratively removing infrequent words and too short documents. These are conventional in prior works on topic modeling, and is to make sure that we have sufficient data for learning meaningful topics. The basic statistics of the preprocessed datasets are shown in Table 1. The diversity and the large sizes of these datasets allow us to evaluate the models accurately.

2.3 Evaluation Metrics

For a comprehensive evaluation, we examine the performance of the models in three aspects: (i) *document modeling* – measured by the perplexity of unseen documents or held-out words (Blei et al., 2003), (ii) *topic discovery* – evaluated by topic coherence, and (iii) *document representation* – quantified through the performance of the obtained documents' topic vectors in downstream tasks.

For measuring the perplexities, we train the model using 90% of documents (or 90% of words in each document, respectively) in the dataset, and compute the perplexity on the remaining 10% of documents (or words, respectively). The coherence of the learned topics are measured by *normalized point-wise mutual information (NPMI)* (Lau et al., 2014) of their top words. Lastly, we use classification task to examine the obtained documents' topic vectors. That is, we use these vectors to train a logistic regression model to classify the documents in each dataset by their labels. For this experiment, we perform 10-fold cross validation, and report the average micro F1 scores across the folds. To obtain robust findings, and also to examine the consistency of the models' performance, for each metric, each model, each number of topics, and each dataset, we run the experiments 10 times, and report the mean and variance of the performance across the runs. The number of topics is varied from half to three times the number of labels of documents in the corresponding dataset.

3 Experiment Results & Findings

Figure 1 shows the mean and variance of the models' unseen-document perplexity on different datasets and with different number of topics. Similarly, Figures 2, 3, and 4 show the means and variances of the models' held-out word perplexity, average F1 scores, and topic coherence respectively on the same datasets and with different number of topics. Note that the **o_LDA** model has no unseen-document perplexity as it does not has a straightforward method for computing the perplexity, while **NMF** has no perplexities as it is not a generative model. From the figures, we can observe that:

- **NVDM** and **NVCTM** are generally better than other models in modeling and representing documents as their perplexities are significantly lower and their average F1 scores are significantly higher. This is reasonable and expected as these two models learn the unnormalized topic vectors for documents while the others learn the normalized ones[2]. However, they underperform other models in discovery cohesive topics as their coherence scores are much lower than others' in most cases.
- None of the neural topic models in our study outperforms the classical **LDA** on all the datasets and in all the metrics.
- The classical models (i.e., **o_LDA**, **g_LDA**, and **NMF**) are generally more stable than the neural ones as their variances are generally much smaller than those the neural models.

From the above observations, we make the following implications, which can be served as some guidelines for evaluation and practical usage of neural topic models.

- **LDA**, especially **g_LDA**, is a strong baseline for topic modeling. It should be used to evaluate neural topic models comprehensively.
- The performance of neural models may not be stable and may vary significantly from run

[2]Here the normalized vectors are probability distribution vectors: their elements are non-negative and summing up to 1

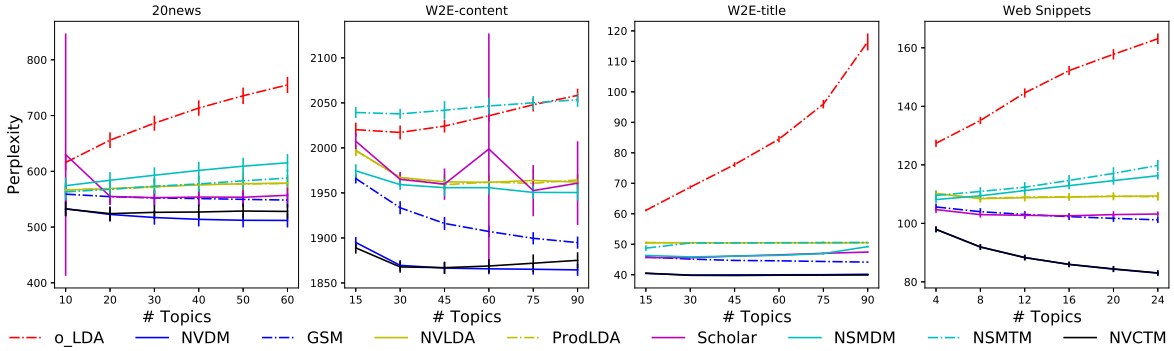

Figure 1: Unseen-document perplexity of the examined models: the lower is better

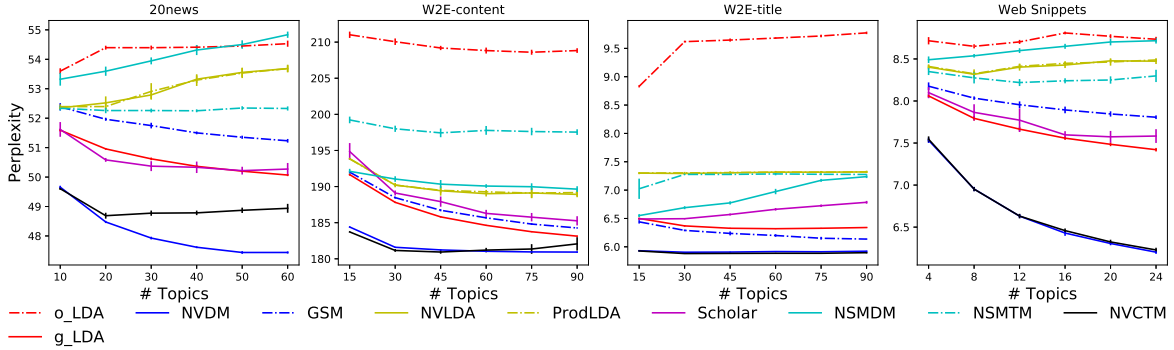

Figure 2: Held-out word perplexity of the examined models: the lower is better

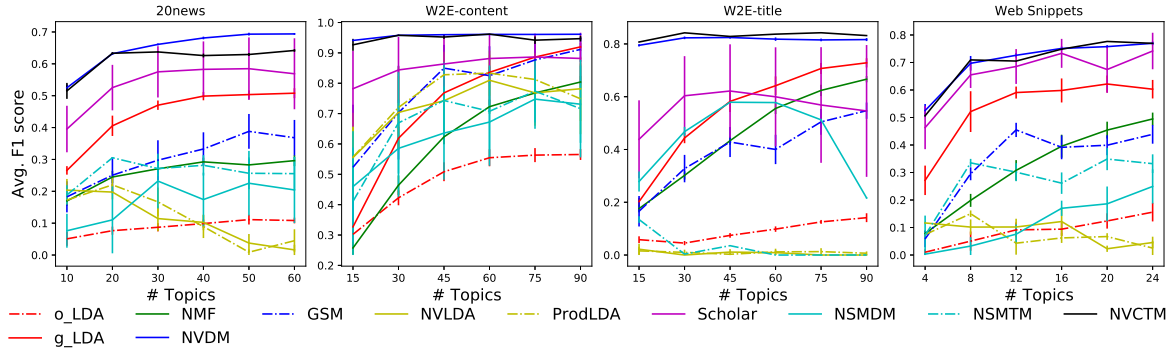

Figure 3: Average F1 scores of the examined models in downstream classification tasks: the higher is better

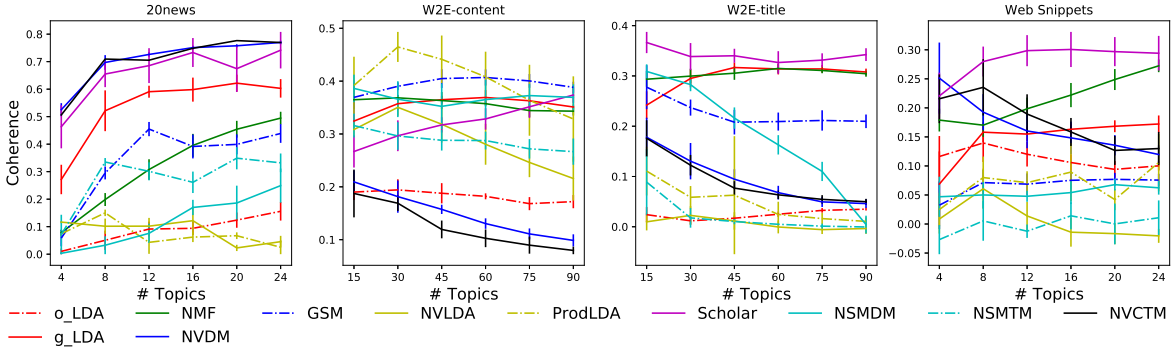

Figure 4: Coherence of topics learned by the examined models: the higher is better

to run. One should perform multiple runs on the same dataset to have accurate evaluation.

- Among the examined neural models, **Scholar** is more suitable for uncovering cohesive, meaningful topics from the corpus, while the unnormalized models (i.e., **NVDM** and **NVCTM**) are more suitable for modeling and representing documents in the corpus.

4 Conclusion

We have examined the performance state-of-the-art neural topic models in a systematic mechanism. From our extensive experiments, we found that classical methods, e.g. **LDA**, still have comparable expressive ability. We have also suggested some considerations for a comprehensive evaluation and practical usage of those models.

In the future, we would like to extend further by designing a fair comparison framework for neural methods that use other representation of documents, e.g. sequences of words, and from other approaches. We would also want to benchmark other issues in neural topic models, e.g., the posterior collapsing problem (He et al., 2019).

Acknowledgments

This work is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2020.DA14

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

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

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. 2016. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*.

Li Fei-Fei and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531. IEEE.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*.

Tuan-Anh Hoang, Khoi Duy Vo, and Wolfgang Nejdl. 2018. W2e: A worldwide-event benchmark dataset for topic detection and tracking. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1847–1850, New York, NY, USA. Association for Computing Machinery.

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.

Mike Innes. 2018. Flux: Elegant machine learning with julia. *Journal of Open Source Software*, 3(25):602.

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

Ken Lang. 2008. The 20 news groups data set. *http://people. csail. mit. edu/jrennie/20Newsgroups/*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Tianyi Lin, Zhiyue Hu, and Xin Guo. 2019. Sparsemax and relaxed wasserstein for topic sparsity. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 141–149.

Luyang Liu, Heyan Huang, Yang Gao, Yongfeng Zhang, and Xiaochi Wei. 2019. Neural variational correlated topic modeling. In *The World Wide Web Conference*, pages 1142–1152.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2410–2419.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

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

Naonori Ueda and Kazumi Saito. 2003. Parametric mixture models for multi-labeled text. In *Advances in neural information processing systems*, pages 737–744.

Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. 2020. Graph attention topic modeling network. In *Proceedings of The Web Conference 2020*, pages 144–154.

Renbo Zhao, Vincent Tan, and Huan Xu. 2017. Online nonnegative matrix factorization with general divergences. In *Artificial Intelligence and Statistics*, pages 37–45. PMLR.