Do Neural Topic Models Really Need Dropout? Analysis of the Effect of Dropout in Topic Modeling

Suman Adhya, Avishek Lahiri, Debarshi Kumar Sanyal
Indian Association for the Cultivation of Science, Jadavpur, Kolkata-700032, India
{ adhyasuman30, avisheklahiri2014, debarshisanyal }@gmail.com

Abstract

Dropout is a widely used regularization trick to resolve the overfitting issue in large feedforward neural networks trained on a small dataset, which performs poorly on the held-out test subset. Although the effectiveness of this regularization trick has been extensively studied for convolutional neural networks, there is a lack of analysis of it for unsupervised models and in particular, VAE-based neural topic models. In this paper, we have analyzed the consequences of dropout in the encoder as well as in the decoder of the VAE architecture in three widely used neural topic models, namely, contextualized topic model (CTM), ProdLDA, and embedded topic model (ETM) using four publicly available datasets. We characterize the dropout effect on these models in terms of the quality and predictive performance of the generated topics.

1 Introduction

Dropout (Hinton et al., 2012) is used while training neural networks, by stochastically dropping out the activation of neurons to prevent complex co-adaptations of feature vectors (Baldi and Sadowski, 2013). The working of dropout is attributed to the implicit averaging over an ensemble of neural networks (Labach et al., 2019; Warde-Farley et al., 2014). It has been shown to be effective on supervised learning tasks to prevent overfitting (Srivastava et al., 2014).

As the volume of digital documents significantly increases with time, organizing them manually is becoming quite an inconvenient task. Because of the ability of topic models to learn a thematic structure from a set of documents in an unsupervised manner and label the documents with their corresponding dominant topics, the significance of topic models is enormous in this area (Hall et al., 2008; Adhya and Sanyal, 2022). But in the traditional topic models, not only the computation cost of the approximate posterior is very high but also for a small change in the modeling assumption, re-derivation of the inference method is needed. With greater flexibility and scalability than traditional topic models, a class of Neural Topic Models (NTMs) aim to leverage the potential of neural networks using the AEVB (Kingma and Welling, 2014) based inference technique. Following (Zhao et al., 2021), we refer to this class of models as VAE-NTMs where the training objective is to maximize the log-likelihood of the reconstruction of the input document while minimizing the KL-divergence of the learned posterior distribution of the latent space from a known prior distribution.

An earlier study by (Ha et al., 2019) of the dropout effect on two traditional topic models LDA (Blei et al., 2003) and BTM (Yan et al., 2013) shows that the correct choice of the dropout rate not only decreases the learning time of the models but also significantly improves the predictive performance and generalization for short texts. However, the study does not consider neural topic models.

In this work, we propose the use of dropout on VAE-NTMs as a hyperparameter in order to achieve much better performance in terms of topic coherence, topic diversity, and topic quality. We test this proposition on a range of standard VAE-NTM architectures. To the best of our knowledge, there has been no other study focusing specifically on the use of dropout in neural topic models. We have made our analysis publicly available[1].

In summary, our contributions are as follows:

1. We comprehensively show both quantitatively and qualitatively that topic quality undergoes a massive improvement with either very low or zero dropout settings in both the encoder and the decoder of a VAE-NTM.

2. We show that for VAE-NTMs the systematic choice of low dropout rates can lead to a sig-

[1] https://github.com/AdhyaSuman/NTMs_Dropout_Analysis

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2220–2229

nificant improvement in downstream tasks like document classification.

3. We study the dependence of dropout on the length of the input documents.

4. We present an empirical analysis for the increase in performance of VAE-NTMs with a decrease in dropout.

2 Task Formulation

Given a corpus $\{D_1, D_2, \ldots, D_N\}$ of N documents with vocabulary $\{w_1, w_2, ..., w_V\}$ of V words, topic models describe a document D_i as a distribution over K topics $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_K\}$, where an individual topic $\boldsymbol{\beta}_k$ is a distribution over V-words.

2.1 VAE Framework in Neural Topic Models

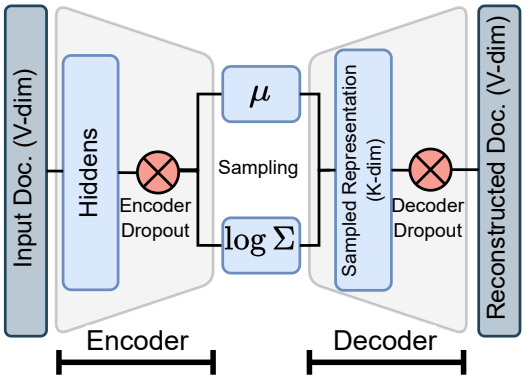


Figure 1: VAE framework in neural topic models.

Given an input sample $\mathbf{x}$, a VAE encoder learns the approximate posterior distribution $q_W(\mathbf{z}|\mathbf{x})$ where W is the encoder's weights that are to be learned and $\mathbf{z}$ is a latent variable. Given a sample $\mathbf{z} \sim q_W(\mathbf{z}|\mathbf{x})$, the VAE decoder learns the likelihood $p_{W'}(\mathbf{x}|\mathbf{z})$ where W' is the learnable decoder's weights.

In VAE-NTMs the input to the encoder is a document representation (e.g., bag-of-words) $\mathbf{x}_{V\times1}$. The encoder then returns the Gaussian parameters $(\boldsymbol{\mu}_{K\times1}, \boldsymbol{\Sigma}_{K\times1})$ that approximate the true posterior where K is the dimension of latent (topic) space, $\boldsymbol{\mu}_{K\times1}$ is the mean, and $\boldsymbol{\Sigma}_{K\times1}$ is the diagonal covariance matrix. Upon taking these Gaussian parameters as input, the decoder samples a latent representation $\mathbf{z}_{K\times1}$ from $\mathcal{N}(\boldsymbol{\mu}_{K\times1}, \boldsymbol{\Sigma}_{K\times1})$ using the reparametrization trick as follows:

$$\mathbf{z}_{K\times1} = \boldsymbol{\mu}_{K\times1} + \boldsymbol{\Sigma}_{K\times1}^{\frac{1}{2}} \odot \boldsymbol{\epsilon}_{K\times1}$$

where $\boldsymbol{\epsilon}_{K\times1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ represents the element-wise product. Then the document-topic distribution vector $(\boldsymbol{\theta}_{K\times1})$ is generated such that $\boldsymbol{\theta}_{K\times1} = \sigma(\mathbf{z}_{K\times1})$ where $\sigma(\cdot)$ is a softmax function. The input document-term distribution vector is reconstructed with the product of $\boldsymbol{\theta}_{K\times1}$ and $\boldsymbol{\beta}_{K\times V}$, the topic-word matrix, in the following manner:

$$\tilde{\mathbf{x}}_{V\times1} = \begin{cases} \boldsymbol{\beta}^T\boldsymbol{\theta} & \text{if } \boldsymbol{\beta} \text{ is normalized.} \\ \sigma\left(\boldsymbol{\beta}^T\boldsymbol{\theta}\right) & \text{if } \boldsymbol{\beta} \text{ is unnormalized.} \end{cases}$$

As shown in Figure 1, in the encoder, dropout is applied with probability E_p on the output of the hidden layer(s) of the multi-layer feed-forward neural network (FFNN). This output is then fed to two separate layers to get the approximate posterior $q_W(z|x)$. In the decoder, dropout is applied with probability D_p on the document-topic distribution vector $(\boldsymbol{\theta}_{K\times1})$, just before the reconstruction process.

2.2 Task Description

The goal is to measure the effect that dropout has on the *performance* of VAE-NTMs by varying the dropout rates from 0.0 to 0.6 in steps of 0.1, in both the encoder and the decoder. We have chosen 0.6 as the upper bound of the dropout rates for our experiments because it is the highest dropout rate used in any VAE-NTMs that we have considered as a baseline in this work. We measure performance using: *topic coherence*, *topic diversity*, and *topic quality*. We use NPMI (Lau et al., 2014; Röder et al., 2015) to measure topic coherence. Topic diversity (Dieng et al., 2020) shows the uniqueness of topics. Topic quality is the product of coherence and diversity (Dieng et al., 2020). As the automated topic model measures do not always accurately capture the quality of the topics (Hoyle et al., 2021), we also perform a manual evaluation of the topics and study their predictive performance on the document classification task.

3 Empirical Study

We perform all experiments in OCTIS (Terragni et al., 2021), which is an integrated framework for topic modeling.

3.1 Datasets

We have used four publicly available datasets in our experiments. Among them, **20NG**[2] and **BBC**

[2] http://qwone.com/~jason/20Newsgroups/

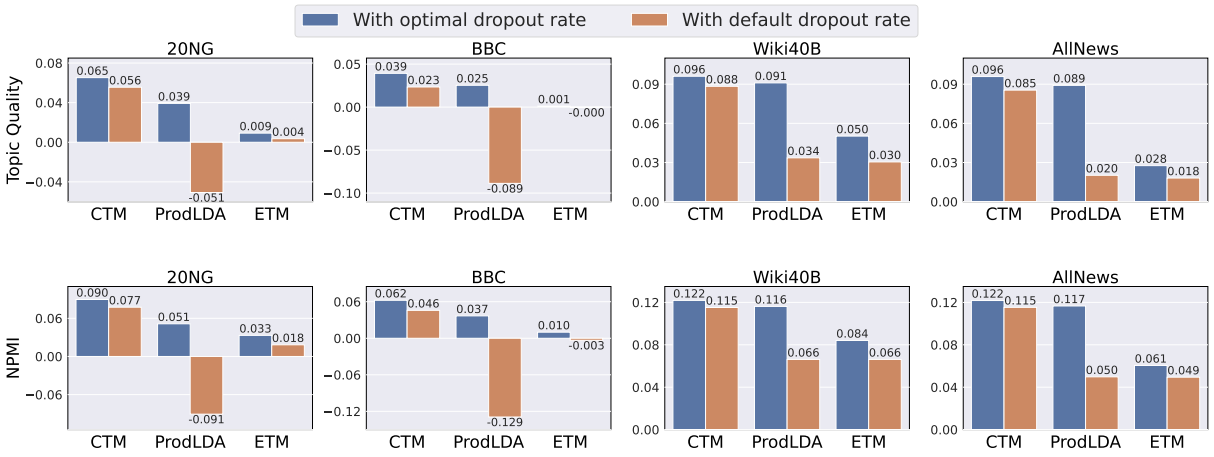


Figure 2: Topic quality and NPMI for different topic models with optimal dropout rate and default dropout rate.

(Greene and Cunningham, 2006) are already available in OCTIS in the pre-processed format while we added **Wiki40B** (Guo et al., 2020) and **AllNews** (Zhu et al., 2018) datasets further. The statistical descriptions of these datasets are mentioned in Table 1. Each corpus is split into train/valid/test sets in the ratio 70 : 15 : 15. The validation set is used for early stopping.

Dataset	#Docs	Avg. #words	|Vocab|
20NG	16309	48.02	1612
BBC	2225	120.12	2949
Wiki40B	24774	541.08	2000
AllNews	49754	229.53	2000

Table 1: Statistics of the used datasets.

3.2 Models

We use the following three VAE-NTMs: **CTM** (Bianchi et al., 2021) which incorporates the contextualized documents embeddings with the neural topic models; **ProdLDA** (Srivastava and Sutton, 2017) which, unlike LDA, relaxes the simplex constraint over the topic-word matrix; (**ETM**) (Dieng et al., 2020) which incorporates word-embeddings in topic modeling to increase robustness in presence of stopwords.

For each of the four datasets, we compute the dropout rate that optimizes the topic quality of each model on that dataset. We train all three topic models for topic-count $K \in \{20, 50, 100\}$ with 30 epochs while keeping all hyperparameter values, except dropout, the same as in their original implementations. To ensure robustness, we average scores over 10 independent runs of each model. For comparison, we use the default dropout rates for each model as mentioned in the original papers that proposed the corresponding model. In Table 2, we show the default and the optimal dropout rates.

Model	20NG	BBC	Wiki40B	AllNews
CTM (0.2, 0.2)	(0.0, 0.0)	(0.0, 0.0)	(0.2, 0.1)	(0.0, 0.1)
ProdLDA (0.6, 0.6)	(0.1, 0.1)	(0.0, 0.0)	(0.1, 0.1)	(0.1, 0.1)
ETM (0.5, 0.0)	(0.0, 0.0)	(0.1, 0.0)	(0.0, 0.0)	(0.1, 0.0)

Table 2: For each of the datasets, the optimal dropout rates of all the models considering the highest topic quality are mentioned in the (E_p, D_p) format in the second through last columns. The default dropout rate is also specified for each model in the first column.

3.3 Results and Analysis

3.3.1 Quantitative Evaluation of Topic Quality

In Figure 2, we compare, for each dataset and each model, the topic quality and the NPMI respectively between the dropout-optimized model that gives the highest topic quality and the model with default dropout rates as mentioned in Table 2.

On **20NG**, the topic quality score for (CTM, ProdLDA, ETM) is improved from $(0.056, -0.051, 0.004)$ to $(0.065, 0.039, 0.009)$ by optimizing the dropout rate. For CTM, the increase in performance is around 16.07% whereas for the other two models it is over 100%. This is because the original implementation of CTM already uses a relatively low dropout rate, i.e.,

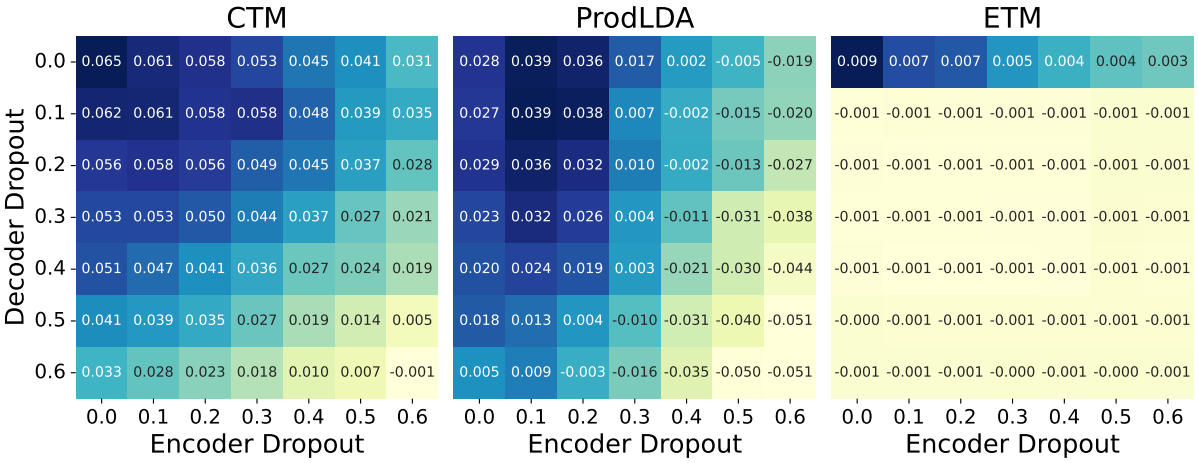


Figure 3: Topic qualities on **20NG** for $(E_p, D_p) \in [0.0,\ 0.6] \times [0.0,\ 0.6]$ with a increment of 0.1.

0.2, for both the encoder and the decoder. The other two models show a significant increase in performance due to their large dropout in the baseline models.

Figure 3 shows that the topic quality on the **20NG** dataset for the VAE-NTMs generally produces better results on keeping the dropout rate for both the encoder and the decoder either to be zero or close to it, especially values like $\{0.0, 0.1\}$. Similar results have been found for the other datasets. Based on these observations, the topic quality is found to reduce with an increase in dropout rates in the encoder and decoder.

3.3.2 Qualitative Evaluation of Topic Quality

To qualitatively evaluate the models, we trained all of them for a topic count of 100 on the **20NG** dataset. We then aligned the topics for each pair of *(optimal-dropout model, default-dropout model)* for all three different models. We followed a two-step strategy for topic alignment. For a given pair of models, namely, one with optimal dropout and another with default dropout, with topics lists P and Q, respectively, we first construct a similarity matrix of the topic lists using Rank-biased Overlap (Webber et al., 2010) (RBO) which computes the similarity between two ordered lists by taking into consideration the rank of the individual elements. For example, for 100 topics, we get a matrix, $\mathbf{A} = (a_{ij})_{1 \leq i,j \leq 100}$ such that, $a_{i,j} = \text{RBO}\left(P[i], Q[j]\right)$. The RBO score lies in $[0, 1]$, where 0 represents no overlap and 1 implies exact overlap. In the final step, we iteratively select the pair of topics for which the similarity score is maximum and simultaneously exclude these two topics from further consideration, i.e. if $\left(P[i_1], Q[j_1]\right)$ and $\left(P[i_2], Q[j_2]\right)$ are two selected pairs then $(i_1 \neq i_2 \wedge j_1 \neq j_2)$.

In Table 3 we show the top words from *aligned* topics of all the models. '$*$' marked models have dropout optimized to give the highest topic quality while others use the default dropout rates as mentioned in Table 2. We see that dropout-optimized models output more interpretable topics.

3.3.3 Effect of Dataset Length

Among the input datasets on which we have experimented, the **20NG** dataset contains relatively short texts, while the others contain longer texts. (Ha et al., 2019) find that their dropout methods are not effective on long texts. But here we see that the performance of all VAE-NTMs decreases uniformly with the increase in the dropout rate, irrespective of the length of the dataset.

3.3.4 Document Classification

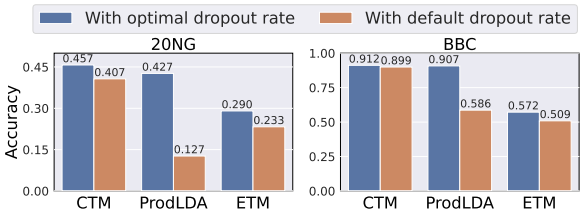


Figure 4: Accuracy for different topic models with optimal dropout and default dropout from Table 2.

We test the predictive performance of the topics produced by the models on a document classification task. We train the models on **20NG** and **BBC** corpora for K topics using the training subset. We represent each document as a K-dimensional document-topic vector and train an SVM, which is then tested on the test subset. We average the

Model	Topics
CTM* $(0.0, 0.0)$	**monitor**, **card**, **video**, **port**, **vga**, **apple**, **connector**, **serial**, **slot**, **output** **firearm**, **weapon**, **dangerous**, **military**, **license**, *file*, *state*, **gun**, **police**, *issue* **christian**, **truth**, **scripture**, *exist*, **belief**, **accept**, *understand*, *word*, **human**, **doctrine**
CTM $(0.2, 0.2)$	**card**, **monitor**, **video**, *offer*, *sale*, *upgrade*, **mouse**, **vga**, **port**, **parallel** **firearm**, **dangerous**, **license**, **weapon**, *section*, *file*, *division*, *device*, *manufacture*, *carry* *interpretation*, **truth**, **scripture**, **christian**, **agree**, **moral**, *understand*, **human**, **faith**, *claim*
ProdLDA* $(0.1, 0.1)$	**window**, **driver**, *mode*, **run**, **mouse**, **session**, **server**, **program**, **manager**, **install** **car**, **engine**, **buy**, **company**, **vehicle**, *make*, **brake**, **tire**, **dealer**, **road** **signal**, **voltage**, **output**, **circuit**, **noise**, **power**, **switch**, **wire**, **connector**, *degree*
ProdLDA $(0.6, 0.6)$	*line*, **window**, *gun*, **read**, **space**, **run**, *statement*, **datum**, **drive**, *make* *make*, **battery**, **engine**, *homosexual*, *assault*, *reason*, *place*, *single*, *large*, *attempt* **voltage**, *damn*, **signal**, *usual*, *label*, *hour*, *bio*, *leg*, *bullet*, *hundred*
ETM* $(0.0, 0.0)$	**version**, **software**, **program**, **file**, *include*, **image**, **application**, *set*, **server**, *support* **armenian**, **turkish**, **village**, **people**, **israeli**, **population**, **muslim**, **kill**, **russian**, **genocide** **system**, **run**, **work**, **window**, *problem*, *include*, *set*, *good*, *support*, **information**
ETM $(0.5, 0.0)$	**file**, **application**, *set*, **program**, *support*, **image**, **display**, *list*, **version**, **bit** **armenian**, **turkish**, **village**, **israeli**, **population**, **muslim**, **genocide**, *son*, *land*, **jewish** **work**, *call*, **system**, **window**, *problem*, **bit**, *set*, **run**, *support*, *good*

Table 3: Some selected topics among 100 topics from **20NG**. '*' indicates models with optimal dropout. The dropout rate is mentioned in the (E_p, D_p) format. The more related words in a topic are highlighted in bold while less related ones are italicized.

accuracy scores over $K \in \{20, 50, 100\}$. Figure 4 shows that accuracy increases when we use the optimized dropout rates.

4 Theoretical Understanding of Results

Our experiments show that by tuning the dropout carefully, we can achieve a significant improvement in the performance of VAE-NTMs. Therefore, we argue that the dropout rate should be treated as an important hyperparameter and carefully selected based on the choice of the model as well as the dataset, especially in the case of VAE-NTMs. More precisely, in most cases, low dropout rates in the encoder and the decoder lead to higher performance than that achieved for higher dropout rates.

Standard dropout and other types of dropout have been extensively used in supervised learning techniques (Srivastava et al., 2014; Wu and Gu, 2015; Tompson et al., 2015; Devries and Taylor, 2017; Cai et al., 2019). The main prerogative of using dropout in the supervised scenario is to introduce noise while training so that the model can recognize the outliers in the testing phase. The drop in performance with high dropout that we see in our experiments is perhaps due to the fact that we are trying to learn a generative model of the data. Dropout makes the model robust against perturbations in the input data and thereby also prevents it from learning the characteristics of the input distribution accurately. This is probably why we see a drop in topic coherence and quality. In the case of document classification, if the topic model is trained with a high dropout, the document-topic vectors are of poor quality and the classifier gets trained on these vectors; this results in poor accuracy on the test documents. This setting is different from the usual supervised learning of neural classifiers where dropout is introduced directly in the classifier to prevent overfitting. We intend to analyze these aspects in more depth in the future.

5 Conclusion

We present a detailed study of the effect of the dropout rate on VAE-NTMs. We find that the model performance generally reduces with the increase in dropout rate in the encoder as well as the decoder.

Limitations

The following limitations are known and should be considered when applying the results of this work or relying on them in future studies: (1) Other variants of dropout can be applied to the VAE-NTMs. (2) Analysis of the dropout effect may be done for other VAE-NTMs as well. (3) Other downstream tasks may be formulated for further analysis.

Acknowledgments

This work is partially supported by the SERB-DST Project CRG/2021/000803 sponsored by the Department of Science and Technology, Government of India at Indian Association for the Cultivation of Science, Kolkata.

References

Suman Adhya and Debarshi Kumar Sanyal. 2022. What does the Indian Parliament discuss? an exploratory analysis of the question hour in the Lok Sabha. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 72–78, Marseille, France. European Language Resources Association.

Pierre Baldi and Peter J Sadowski. 2013. Understanding dropout. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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

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

Shaofeng Cai, Jinyang Gao, Meihui Zhang, Wei Wang, Gang Chen, and Beng Chin Ooi. 2019. Effective and efficient dropout for deep convolutional neural networks. *CoRR*, abs/1904.03392.

Terrance Devries and Graham W. Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 377–384, New York, NY, USA. Association for Computing Machinery.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Cuong Ha, Van-Dang Tran, Linh Ngo Van, and Khoat Than. 2019. Eliminating overfitting of probabilistic topic models on short and noisy text: The role of dropout. *International Journal of Approximate Reasoning*, 112:85–104.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Honolulu, Hawaii. Association for Computational Linguistics.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Lee Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014*.

Alex Labach, Hojjat Salehinejad, and Shahrokh Valaee. 2019. Survey of dropout methods for deep neural networks. *CoRR*, abs/1904.13310.

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, Gothenburg, Sweden. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics.

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. 2015. Efficient object localization using convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 648–656. IEEE Computer Society.

David Warde-Farley, Ian J. Goodfellow, Aaron C. Courville, and Yoshua Bengio. 2014. An empirical analysis of dropout in piecewise linear networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

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

Haibing Wu and Xiaodong Gu. 2015. Towards dropout training for convolutional neural networks. *Neural Networks*, 71:1–10.

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*, WWW '13, page 1445–1456, New York, NY, USA. Association for Computing Machinery.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray L. Buntine. 2021. Topic modelling meets deep neural networks: A survey. *CoRR*, abs/2103.00498.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4663–4672, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Datasets

We run our experiments on the following datasets:

- **20NewsGroups (20NG)**[3] is a dataset of 18,846 documents from 20 different newsgroups posts. The **20NG** dataset is present in OCTIS, so it is already in pre-processed form. All the documents of this dataset have their corresponding category type as the document labels. The details about these categories are mentioned in Table 4.
- **BBC News (BBC)** (Greene and Cunningham, 2006) is a dataset of news articles from BBC. It is also accessible from OCTIS in pre-processed form. The documents of this dataset are categorized into 5 different categories which are *tech*, *business*, *entertainment*, *sports*, and *politics*. The details of these categories are mentioned in Table 5.
- **Wiki40B**(Guo et al., 2020) is a Wikipedia text dataset in 40+ languages, available in TensorFlow dataset format. In our experiment, we take a sample of 24,774 English documents from this dataset.
- **All the News (AllNews)**(Zhu et al., 2018) dataset consists of 50,001 news articles from 15 news publishers.

[3] http://qwone.com/~jason/20Newsgroups/

#No.	Label	#Docs	%Docs
1.	misc.forsale	861	5.28
2.	comp.windows.x	883	5.41
3.	soc.religion.christian	920	5.64
4.	talk.religion.misc	521	3.19
5.	rec.autos	822	5.04
6.	sci.med	866	5.31
7.	talk.politics.misc	689	4.22
8.	talk.politics.mideast	828	5.08
9.	sci.electronics	867	5.32
10.	rec.sport.hockey	843	5.17
11.	rec.sport.baseball	787	4.83
12.	talk.politics.guns	808	4.95
13.	sci.crypt	883	5.41
14.	comp.sys.mac.hardware	838	5.14
15.	comp.sys.ibm.pc.hardware	891	5.46
16.	comp.graphics	836	5.13
17.	comp.os.ms-windows.misc	828	5.08
18.	alt.atheism	689	4.22
19.	sci.space	856	5.25
20.	rec.motorcycles	793	4.86

Table 4: **20NG** labels with corresponding document counts and percentage of documents.

#No.	Label	#Docs	%Docs
1.	tech	401	18.02
2.	business	510	22.92
3.	entertainment	386	17.35
4.	sport	511	22.97
5.	politics	417	18.74

Table 5: **BBC** labels with corresponding document counts and percentage of documents.

A.2 Pre-processing Steps

Using OCTIS, we convert each document to lowercase, remove the punctuations, lemmatize it, filter

the vocabulary with the most frequent 2000 terms, filter words with less than 3 characters, and filter documents with less than 3 words.

A.3 Topic Evaluation Metrics

1. **Coherence metric**: This measures how much the top words of the topics are relevant. Topic coherence (TC) for K topics each of which contains n top words can be calculated as:

$$\mathrm{TC} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{{}^{n}C_{2}}\sum_{i=1}^{n}\sum_{j=i+1}^{n} f\left(w_i^{(k)}, w_j^{(k)}\right)$$

Here, $f(\cdot,\cdot)$ is the Normalized Pointwise Mutual Information or NPMI (Lau et al., 2014) of the words $w_i^{(k)}$ and $w_j^{(k)}$ appearing in topic k:

$$f(w_i^{(k)}, w_j^{(k)}) = \frac{\log \frac{p(w_i^{(k)}, w_j^{(k)})+\epsilon}{p(w_i^{(k)})p(w_j^{(k)})}}{-\log\left(p(w_i^{(k)}, w_j^{(k)})+\epsilon\right)}$$

where, $p(w_i^{(k)}, w_j^{(k)})$ is the probability of the co-occurrence of the words $w_i^{(k)}$ and $w_j^{(k)}$ in a boolean sliding window in topic k and $p(w_i^{(k)})$ and $p(w_j^{(k)})$ represents the probability of the occurrence of the individual words in topic k. ϵ is a small positive constant that is used to avoid zero in the $\log(\cdot)$ function.

2. **Diversity metric**: This measures how much the generated topics are different from each other. To measure the diversity score we have used the metric Topic Diversity (TD) (Dieng et al., 2020) which is defined as the proportion of the number of unique words appearing across all topics. It ranges between $[0, 1]$ where a value close to 0 implies repetitive topics and a value near 1 represents more diversification in the topics.

3. **Topic quality**: This is an overall metric that is defined as the product of the two metrics NPMI and TD.

In our experiments, we take the top 10 words for each topic (i.e., $n = 10$) to compute NPMI and TD scores.

A.4 Computing Infrastructure

Our experiments were run on a workstation with Intel® Xeon™ Gold 6326 CPU @ 2.90GHz, 256.0 GB RAM, NVIDIA A100 80GB PCIe, CUDA Version: 11.7 and Ubuntu 22.04 operating system.

A.5 Detailed Results

The detailed results of our experiments are given in Tables 6, 7, and 8. An asterisk ($*$) against a model in the above tables indicates that it is trained with the optimal dropout rate, and the absence of an asterisk indicates that the default dropout rate is used. The default dropout rate for **CTM** is taken from (Bianchi et al., 2021), for **ProdLDA** from (Srivastava and Sutton, 2017), and for **ETM** from (Dieng et al., 2020).

Model	Topic quality for each dataset			
	20NG	BBC	Wiki40B	AllNews
CTM*	0.0652	0.0392	0.0961	0.0958
CTM	0.0556	0.0234	0.0884	0.0854
ProdLDA*	0.0392	0.0254	0.0910	0.0891
ProdLDA	-0.0507	-0.0887	0.0336	0.0202
ETM*	0.0092	0.0011	0.0502	0.0276
ETM	0.0036	-0.0003	0.0304	0.0181

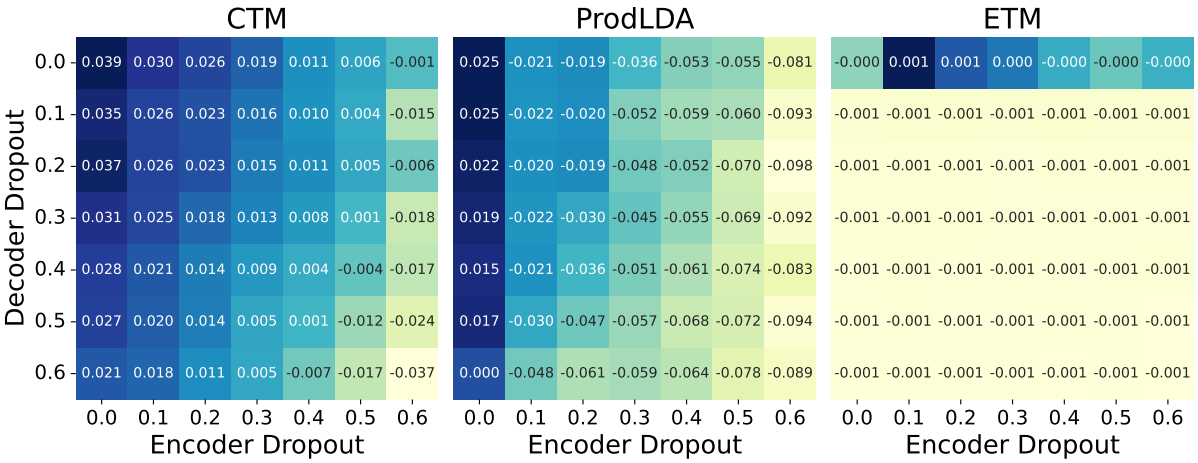
Table 6: Topic Quality values for different VAE-NTMs with optimal dropout rate and default dropout rate (see Table 2). '*' indicates models with optimal dropout.

Model	NPMI for each dataset			
	20NG	BBC	Wiki40B	AllNews
CTM*	0.0896	0.0623	0.1219	0.1218
CTM	0.0774	0.0458	0.1152	0.1153
ProdLDA*	0.0513	0.0367	0.1162	0.1166
ProdLDA	-0.0907	-0.1293	0.0662	0.0498
ETM*	0.0331	0.0100	0.0841	0.0605
ETM	0.0183	-0.0033	0.0662	0.0494

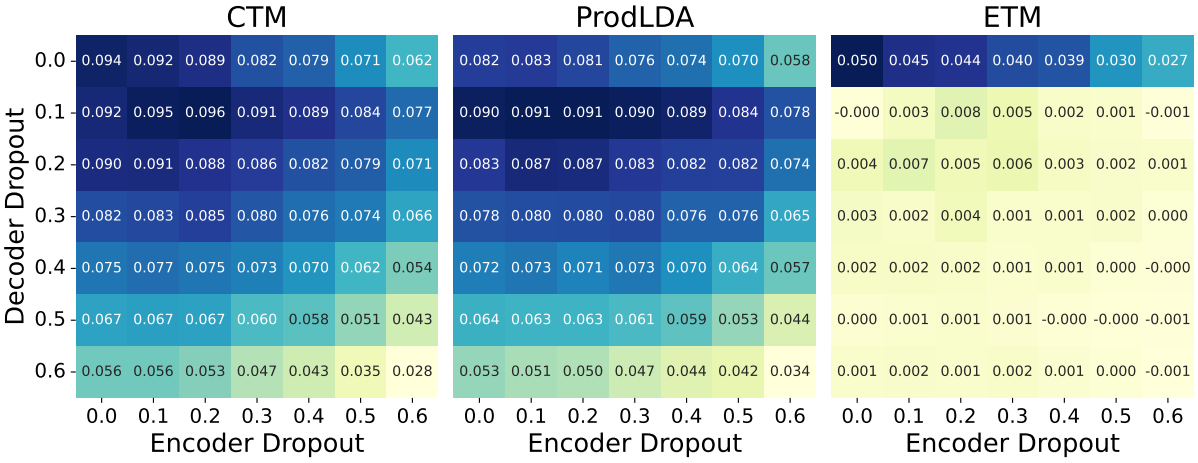
Table 7: NPMI values for different VAE-NTMs with optimal dropout rate and default dropout rate (see Table 2). '*' indicates models with optimal dropout.

Model	TD for each dataset			
	20NG	BBC	Wiki40B	AllNews
CTM*	0.7283	0.6295	0.7883	0.7871
CTM	0.7175	0.51	0.7671	0.7409
ProdLDA*	0.7644	0.6902	0.7829	0.7640
ProdLDA	0.5594	0.6861	0.5071	0.4061
ETM*	0.2776	0.1108	0.5973	0.4561
ETM	0.1949	0.0902	0.4599	0.3659

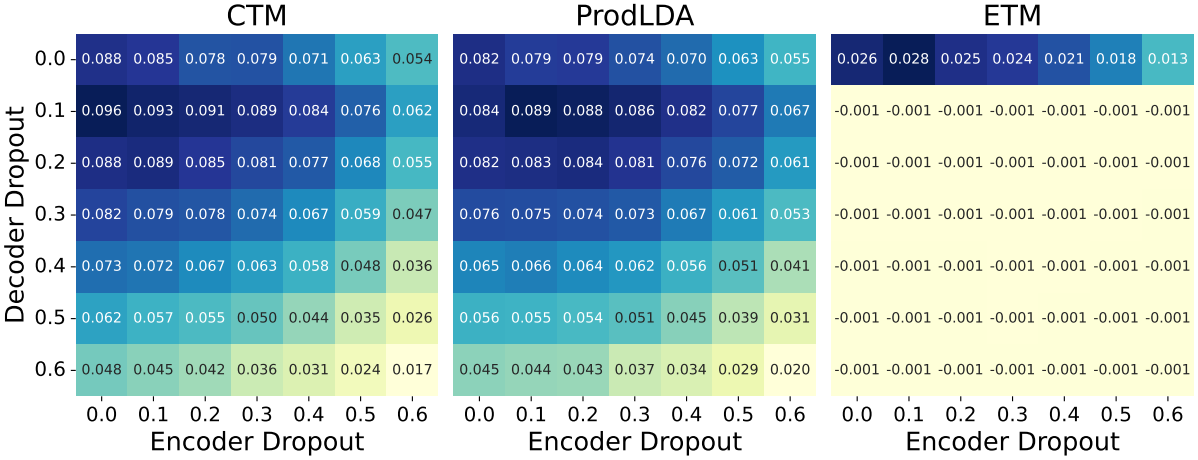
Table 8: Topic Diversity values for different VAE-NTMs with optimal dropout rate and default dropout rate (see Table 2). '*' indicates models with optimal dropout.



(a) BBC

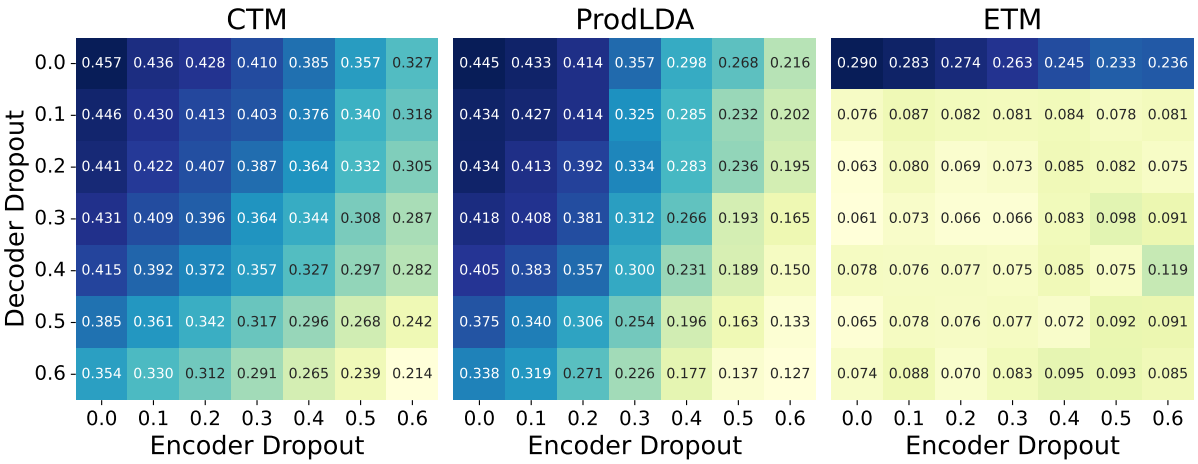


(b) Wiki40B

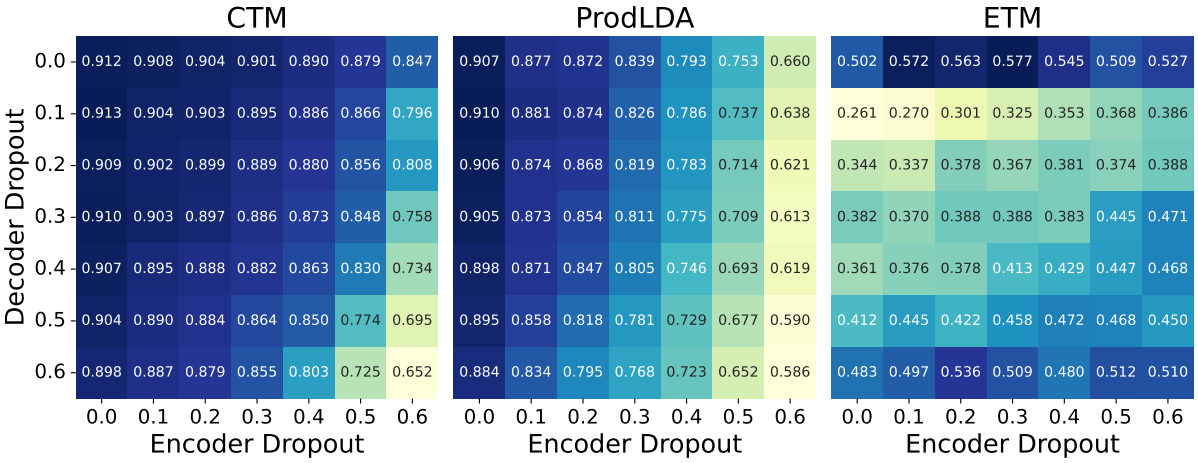


(c) AllNews

Figure 5: Change in **topic quality** for $(E_p, D_p) \in [0.0,\ 0.6] \times [0.0,\ 0.6]$ with a increment of 0.1.



(a) 20NG



(b) BBC

Figure 6: Accuracy scores for $(E_p, D_p) \in [0.0,\ 0.6] \times [0.0,\ 0.6]$ (step $= 0.1$) in document classification task.