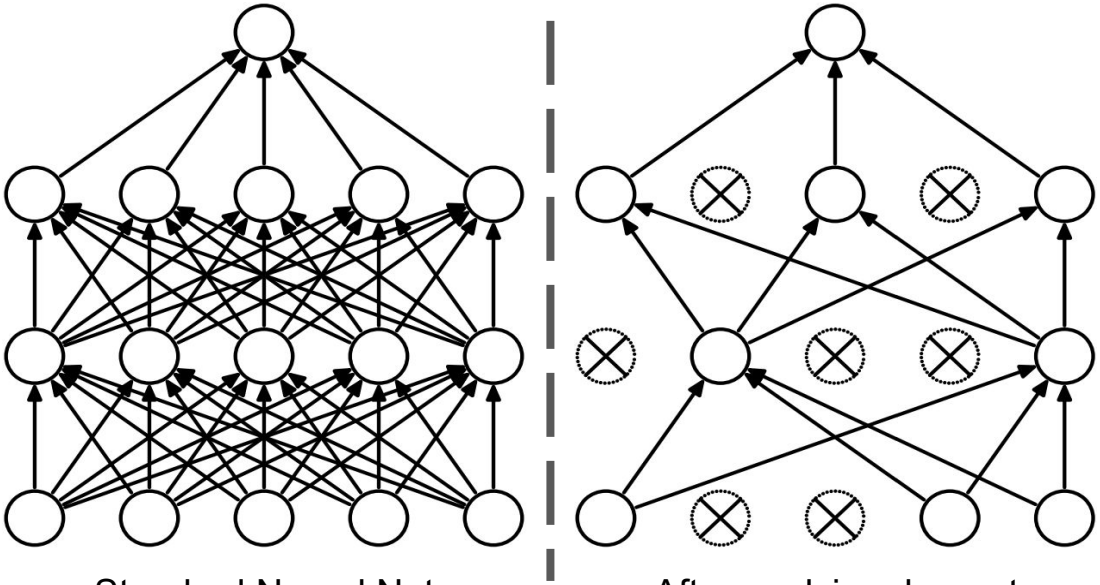
Do Neural Topic Models Really Need Dropout? Analysis of the Effect of Dropout in Topic Modeling

Suman Adhya, Avishek Lahiri, Debarshi Kumar Sanyal

Indian Association for the Cultivation of Science
Jadavpur, Kolkata 700032, INDIA

EACL 2023

Overview of Dropout



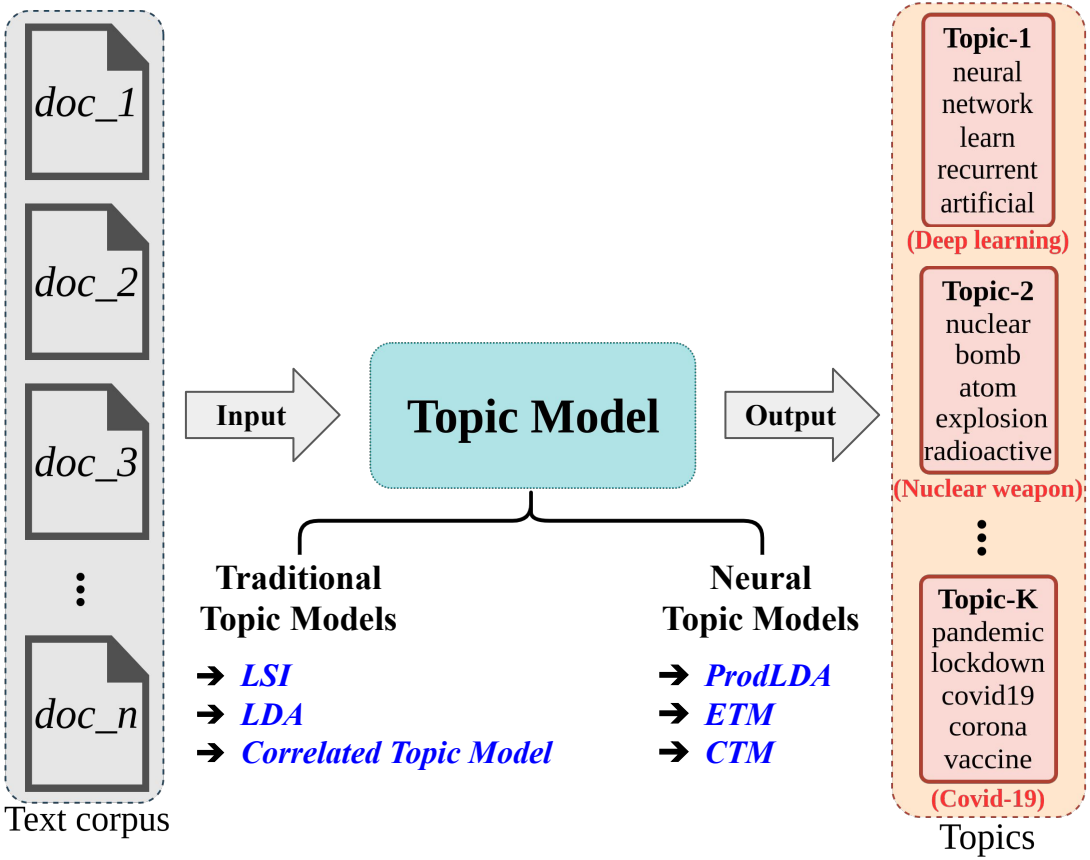
Standard Neural Net

After applying dropout

- **Type:** Regularizer;
- **Usage:** Resolve the overfitting;
- **Method:** Stochastically drops the activation of neurons;
- **Benefit:** Average over an ensemble of neural networks;

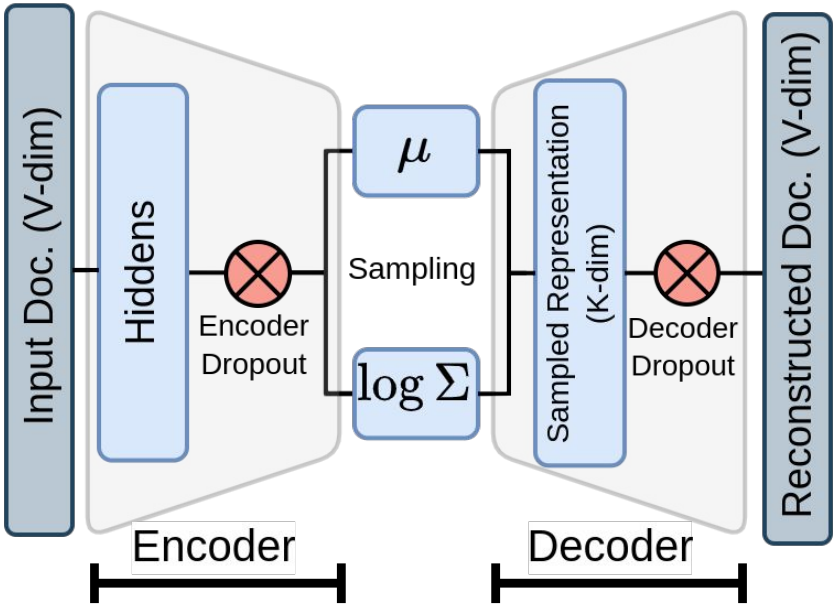
Dropout: A Simple Way to Prevent Neural Networks from Overfitting (Srivastava et al., JMLR 2014)

What is a Topic Model?



- **Type:** Unsupervised learning;
- **Input:** Set of documents;
- **Output:** Set of topics;
- **Topic:** Distribution over the words;
- **Use cases:**
  - Document clustering;
  - Text classification;
  - Information retrieval;

VAE Framework in Neural Topic Models



Encoder:

Input: Document representation;

Dropout on: Output of the hidden layer(s);

Returns: Posterior distribution;

Decoder:

Input: Document-topic distribution vector;

Dropout on: Document-topic distribution;

Returns: Reconstructed document;

Experimental Setup

All experiments are performed using **OCTIS** [1].

Datasets:

Name	Type	#Docs
20NG	Newsgroups posts	16309
BBC	News articles from BBC	2225
Wiki40B	Wikipedia text dataset	24774
AllNews	News articles	49754

Train:valid:test = 70 : 15 : 15. The validation set is used for early stopping.

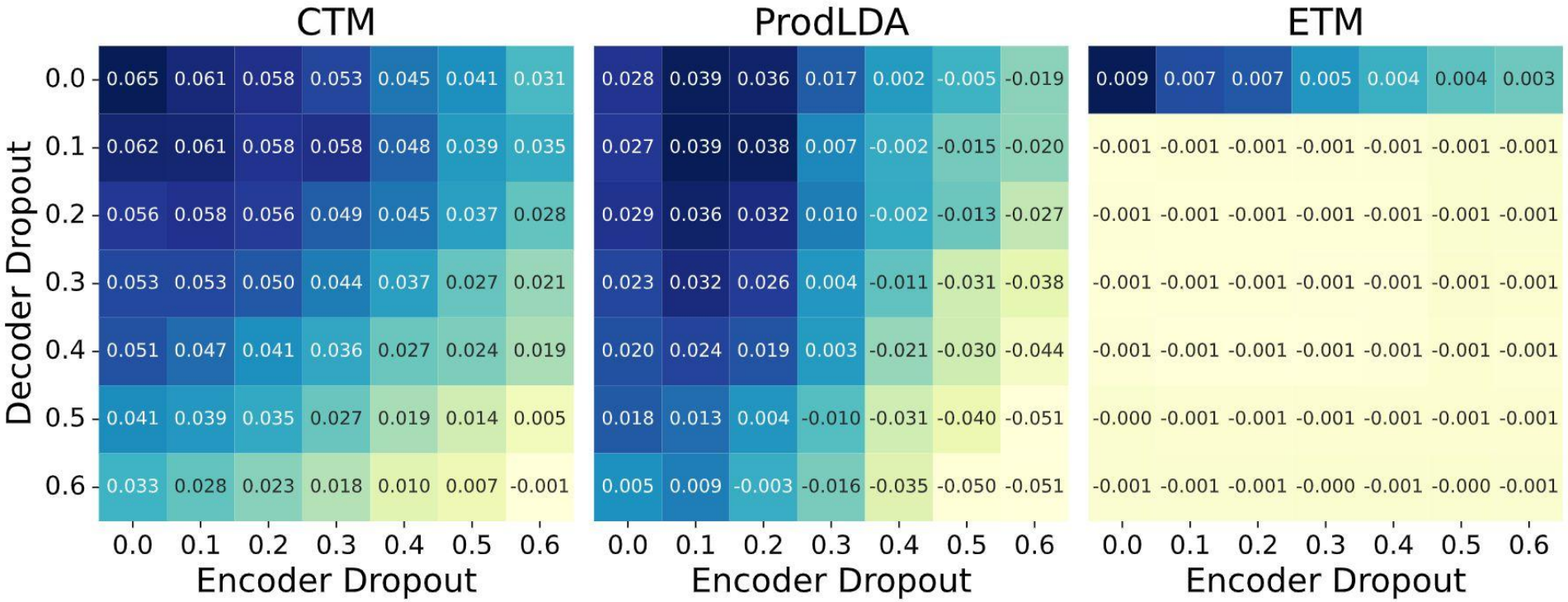
Baselines: **CTM**[2], **ProdLDA**[3], **ETM**[4]

Evaluation:

- **NPMI**: Topic-words relevancy
- **TD**: Topic distinction
- **Topic Quality = NPMI × TD**

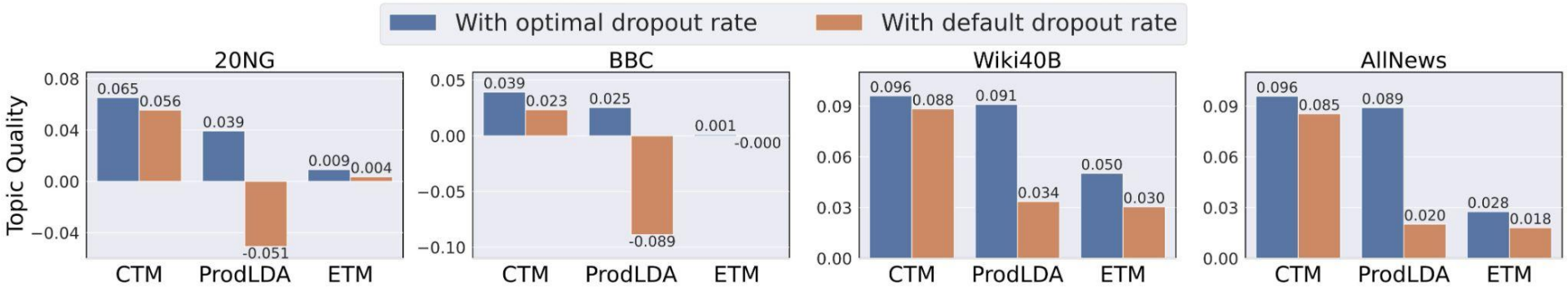
[1] *OCTIS: Comparing and Optimizing Topic models is Simple! (Terragni et al., EACL 2021)*
[2] *Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence (Bianchi et al., ACL-IJCNLP 2021)*
[3] *Autoencoding Variational Inference For Topic Models (Srivastava and Sutton, ICLR 2017)*
[4] *Topic Modeling in Embedding Spaces (Dieng et al., TACL 2020)*

Topic Quality wrt Change in Dropout Rate



Topic qualities on **20NG** for dropout rate [0.0, 0.6] with a increment of 0.1.

Quantitative Evaluation of Topic Quality



Topic quality for the models with **optimal** and **default** dropout rate.

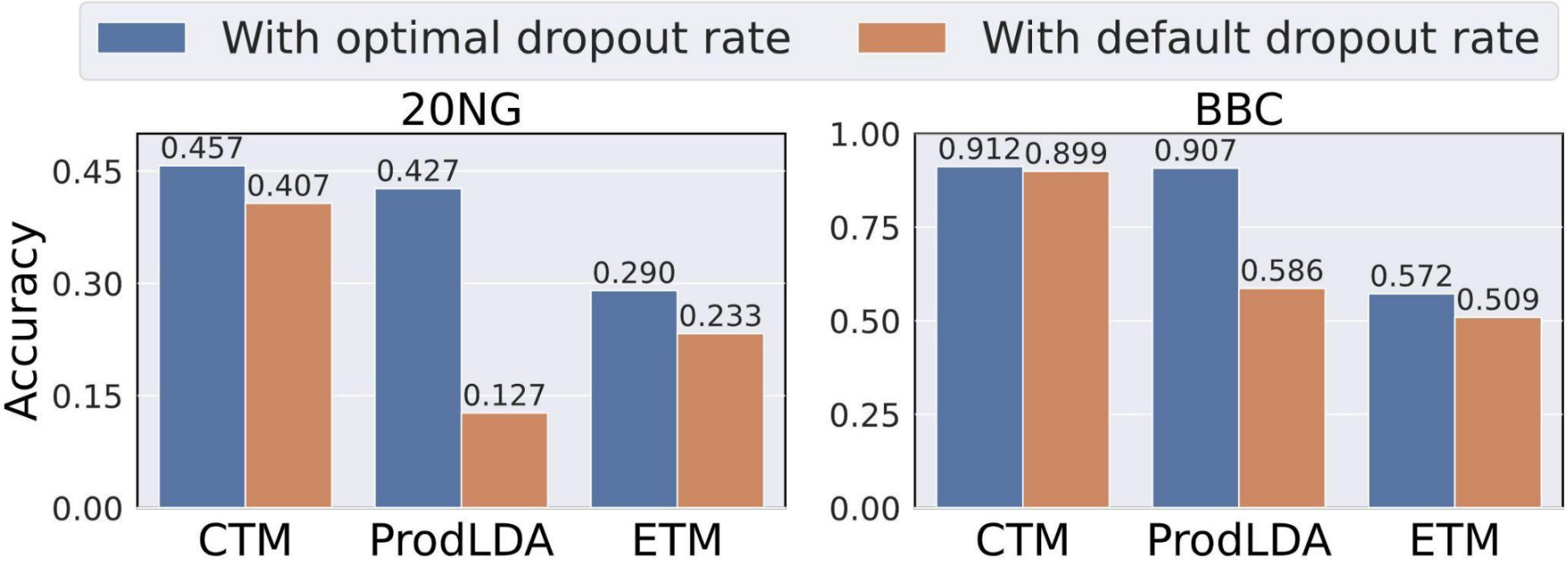
Qualitative Evaluation of Topic Quality

Model	Topics
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*

Some selected topics among **100 topics** from **20NG**. '***' indicates models with optimal dropout**. The more related words in a topic are highlighted in **bold** while less related ones are *italicized*.

Extrinsic Evaluation

Document Classification



Accuracy for different topic models with **optimal** and **default** dropout rate.

Conclusion

A. We found that adjusting dropout can greatly enhance VAE-NTM's performance. Thus, it's a crucial hyperparameter to consider. Typically, lower dropout rates in the encoder and decoder result in better performance.

B. Performance falls at high dropout because we're trying to learn a generative model of the data. Dropout makes the model robust against input data changes, but it also prevents accurate learning of the input distribution characteristics. This is likely why we see a drop in topic coherence and quality.

Future Work

A. General study of the dropout effect on VAE can be done, which may be extended to other generative models, too.

B. Strong theoretical explanation of the inefficiency of dropout for VAE-NTMs is also needed.

Thanks for listening...!!

Code: https://github.com/AdhyaSuman/NTMs_Dropout_Analysis

Indian Association for the Cultivation of Science
Jadavpur, Kolkata 700032, INDIA

EACL 2023