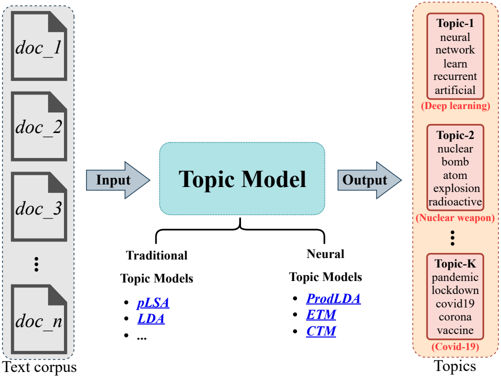
Improving Contextualized Topic Models with Negative Sampling

Suman Adhya[†], Avishek Lahiri[†], Debarshi Kumar Sanyal[†], Partha Pratim Das[§]

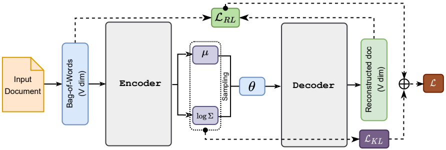
[†] Indian Association for the Cultivation of Science, Kolkata-700032, India
[§] Ashoka University, Sonipat, Haryana-131029, India &
Indian Institute of Technology Kharagpur, West Bengal-721302, India

What is a Topic Model?

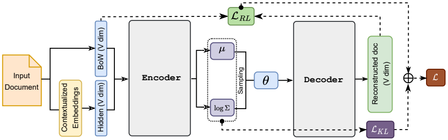


- **Type**: Unsupervised learning;
- **Input**: Set of documents;
  Output: Set of topics;
- **Topic**: Distribution over the words;
- **Topic label**: Manually, by looking at the top words.

Framework of VAE-based Neural Topic Models



ProdLDA



Contextualized Topic Models (CTM)

Negative Sampling

- **Sampling:** $-$ve examples are sampled from a noise distribution.
- **Training objective:** To distinguish between the $+$ve and $-$ve samples.

Used to:

1. Reduce the computational cost of training.
2. Identify out-of-distribution examples.
3. Make the model more robust to adversarial attacks.

Negative Sampling for Topic Models

1. **NQTM** [9] ⇒ (A) Topic distribution **quantization mechanism** and (B) **negative sampling decoder** to produce peakier topic distributions for short texts.
2. **ToMCAT** [4] ⇒ **CycleGAN**-based model.
3. **ATM** [7] ⇒ **GAN**-based model.
4. **BAT** [8] ⇒ **GAN**-based model.

Issues:

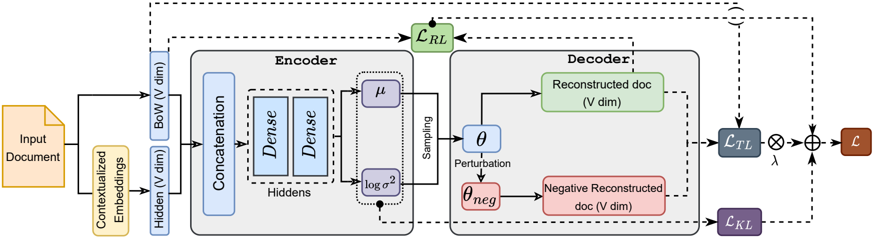
1. Task specific;
2. Negative sample generation technique is not very convenient;

Our Contributions

Primary contributions are:

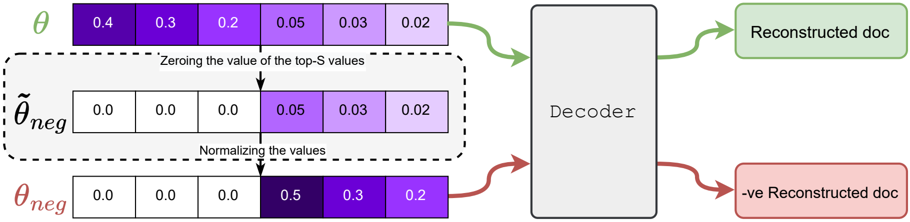
1. Proposed a simple but effective unsupervised negative sampling generation technique.
2. Increase in topic quality due to the proposed methodology.
3. In downstream task (document classification) proposed model outperforms the existing models.

Framework of the Proposed Model



Framework for CTM with negative sampling (CTM-Neg).

Negative Sample Generation



$$\boldsymbol{\theta}_{neg} = \frac{\tilde{\boldsymbol{\theta}}_{neg}}{\sum_{i=1}^{T} \tilde{\theta}_{neg}[i]}$$

$$\text{where, } \tilde{\theta}_{neg}[i] = \begin{cases} 0, & \text{if } i \in \operatorname{argmax}(\boldsymbol{\theta}, S) \\ \theta[i], & \text{otherwise} \end{cases}$$

Triplet Loss Term

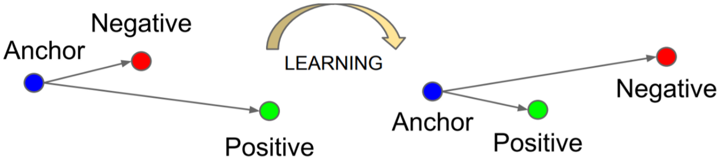


Illustration of triplet loss given one positive and one negative per anchor. Image source: [5]

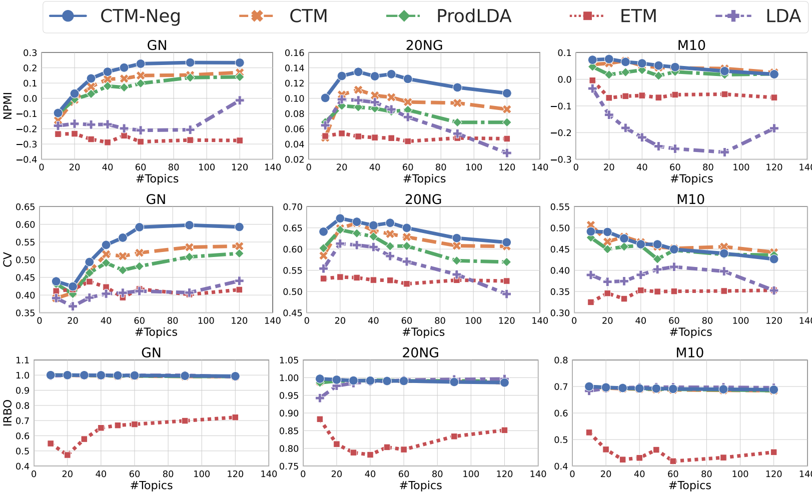
$$\mathcal{L}_{\text{TL}} = \max(||\hat{\mathbf{x}} - \mathbf{x}_{\text{BoW}}||_2 - ||\hat{\mathbf{x}} - \hat{\mathbf{x}}_{neg}||_2 + m, 0)$$

Experimental Setup

Dataset	Type	#Documents
GoogleNews (GN)	News articles	11,109
20NewsGroups (20NG)	Newsgroups posts on 20 topics	16,309
M10	Scientific publications	8,355

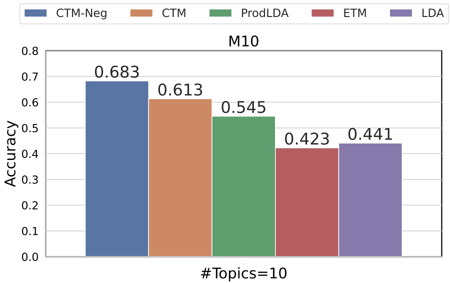
- **Datasets**:
  - **GN**, **20NG**, **M10**
- **Baselines**:
  - **CTM** [1], **ProdLDA** [6], **ETM** [3], **LDA** [2]
- **Evaluation metrics**:
  - Topic coherence (intra topic words relevancy): **NPMI**, **CV**.
  - Topic diversity (topics distinction): **IRBO**.

Quantitative Evaluation



Document Classification

No.	Label	#Documents
1	Agriculture	643
2	Archaeology	131
3	Biology	1059
4	Computer Science	1127
5	Financial Economics	978
6	Industrial Engineering	944
7	Material Science	873
8	Petroleum Chemistry	886
9	Physics	717
10	Social Science	997



- **Dataset**: M10, **Classes**: 10, **Train:Test:Valid** = 70:15:15, #**Topics** = 10
- **Document representation**: T-dim document-topic (θ) vector
- Linear SVM is trained with θ of the training subset and the performance on the test subset is recorded.

Conclusion & Future Directions

Conclusion:

- Proposed a negative sampling strategy for a neural contextualized topic model.
- Experimental results on three publicly available datasets validate the effectiveness of the proposed methodology.

Future work:

- Comparison with other adversarial topic models.
- Integrate with other neural topic models to judge their performance.

References

[1] Federico Bianchi, Silvia Terragni, and Dirk Hovy. "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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 759–766. DOI: 10.18653/v1/2021.acl-short.96. URL: https://aclanthology.org/2021.acl-short.96.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet allocation". In: *Journal of Machine Learning Research* 3.Jan (2003), pp. 993–1022. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[3] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. "Topic Modeling in Embedding Spaces". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453. DOI: 10.1162/tacl_a_00325. URL: https://aclanthology.org/2020.tacl-1.29.

[4] Xuemeng Hu et al. "Neural Topic Modeling with Cycle-Consistent Adversarial Training". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 9018–9030. URL: https://aclanthology.org/2020.emnlp-main.725/.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *CoRR* abs/1503.03832 (2015). arXiv: 1503.03832. URL: http://arxiv.org/abs/1503.03832.

[6] Akash Srivastava and Charles Sutton. "Autoencoding Variational Inference for Topic Models". In: *Proceedings of the 5th International Conference on Learning Representations*. 2017. URL: https://arxiv.org/abs/1703.01488.

[7] Rui Wang, Deyu Zhou, and Yulan He. "ATM: Adversarial-neural topic model". In: *Information Processing & Management* 56.6 (2019), p. 102098. URL: https://www.sciencedirect.com/science/article/abs/pii/S0306457319300500.

[8] Rui Wang et al. "Neural Topic Modeling with Bidirectional Adversarial Training". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 340–350. DOI: 10.18653/v1/2020.acl-main.32. URL: https://aclanthology.org/2020.acl-main.32.

[9] Xiaobao Wu et al. "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)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1772–1782. DOI: 10.18653/v1/2020.emnlp-main.138. URL: https://www.aclweb.org/anthology/2020.emnlp-main.138.

Thank you all for listening...

Code: `https://github.com/AdhyaSuman/CTMNeg`