Are Neural Topic Models Broken?

Alexander Hoyle, Pranav Goel, Rupak Sarkar, Philip Resnik

Background

Topic models are widely used

Practitioners want to make valid inferences from model outputs

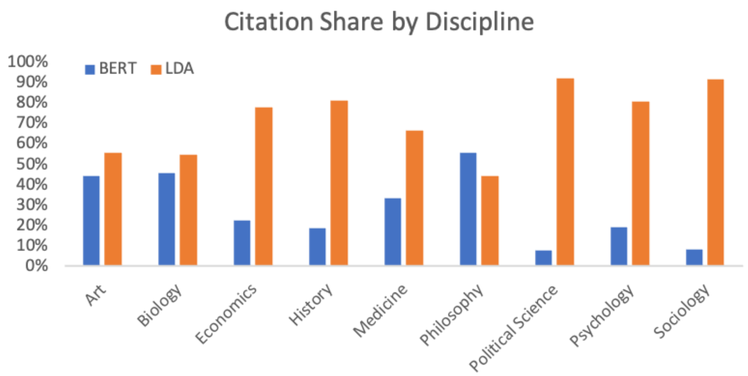

Fig 1: Citation share since 2019, and for just one popular topic model: LDA! (Source: Semantic Scholar)

And new topic model variants are getting introduced in top ML/NLP conferences all the time!

Prior work: coherence metrics are flawed, differentiating models when humans do not

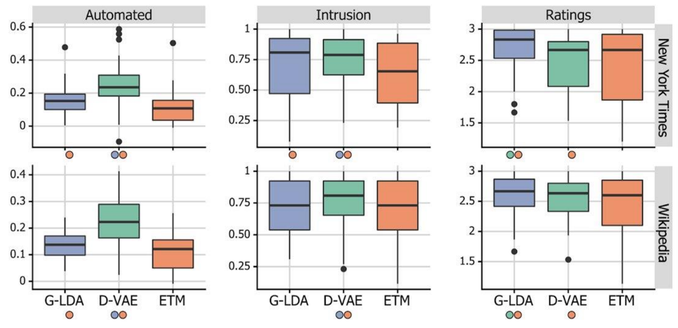

Figure 2: While automated evaluations (here, NPMI) suggest a clear winner between models, human evaluation is more nuanced. Human judgments exhibit greater variability over a smaller range of values. Colored circles correspond to pairwise one-tailed significance tests between model scores at $\alpha = 0.05$; for example, the rightmost orange circle at bottom right shows that human intrusion ratings for D-VAE are significantly higher than ETM for topics derived from Wikipedia.

Lau et al. (2014). "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality." *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*

Hoyle et al. (2021). "Is automated topic model evaluation broken?". NeurIPS 2021.

Doogan & Buntine (2021). "Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Method

In the absence of unsupervised metrics that predict human interpretability, *how can we evaluate topic models?*

Idea: Use metrics that correspond to *content analysis*, a dominant use case

Alignment: do I match human labels?

Using standard clustering metrics, we measure the extent to which assigned labels for documents (that is, the most probable topic per document) agree with human-provided labels.

Shown here: ***Adjusted Rand Index.*** The rand index compares all pairs of the gold labels and assigned labels over documents, counting the proportion of pairs that have the same (TP) or different (TN) assignments. ARI corrects for chance. Other metrics in paper

$$\text{RI} = \frac{TP + TN}{TP + FP + TN + FN}$$

Stability: are my estimates reliable?

Intuitively, estimates from a model on the same set of data should be similar over multiple runs. That is, we want the distance between the topic-word estimates of each run to be small.

Collect the topic-word estimates for m runs and K topics $\beta_k^{(i)}, i \in 1, \ldots, m; k \in 1, \ldots, K$. For each run pair, compute pairwise RBO distance d between all K topics in each run. For a pair of runs i, j, we want to find a permutation of rows $\pi(\cdot)$ that minimizes the total distance

$$\mathcal{TD}(\mathbf{B}^{(i)}, \mathbf{B}^{(j)}) = \frac{1}{K} \sum_k d\left(\boldsymbol{\beta}_k^{(i)}, \boldsymbol{\beta}_{\pi(k)}^{(j)}\right)$$

If the total distances for the set of $\binom{m}{2}$ runs for one model is smaller than a second model, then the first model is more stable.

Results

Alignment

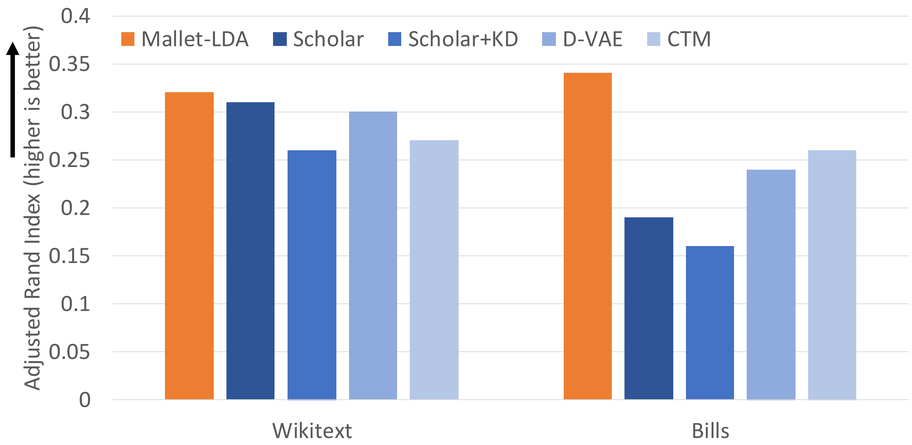

Fig 3: Classical models are better-aligned: this graph shows the adjusted Rand Index for Wikipedia and Bills dataset, K=50 topics, averaged over 10 runs (with randomly varied hyperparameters and seeds). Orange is a classical model with Gibbs-sampling; blue are neural models.

Stability

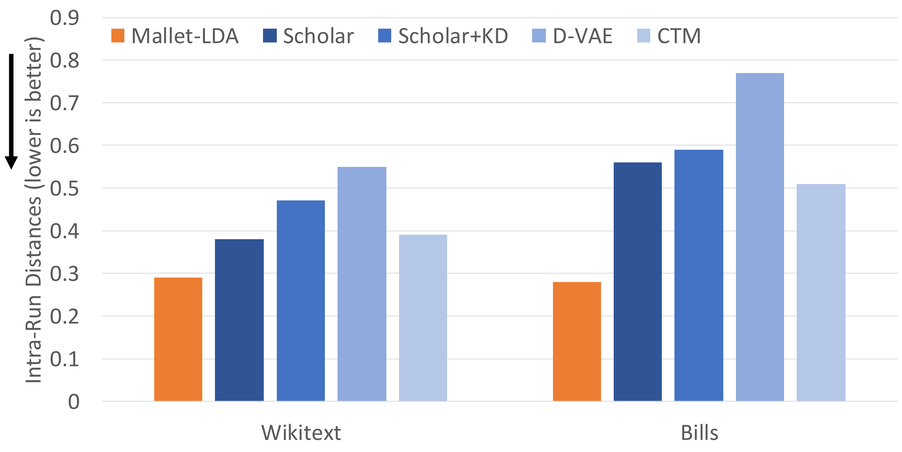

Fig 4: Classical models are more stable: this graph shows the total distance between topic-word estimates over 10 runs (randomly varied hyperparameters and seeds).

Possible solution and takeaways

- In paper, we introduce a simple ensembling method that improves both alignment and stability
- Evaluation should be aligned with use case!