Are Neural Topic Models Broken?

Alexander Hoyle Computer Science
Pranav Goel Computer Science
Rupak Sarkar Computer Science
Philip Resnik UMIACS, Linguistics
University of Maryland
{hoyle,pgoel1,rupak,resnik}@cs.umd.edu

Abstract

Recently, the relationship between automated and human evaluation of topic models has been called into question. Method developers have staked the efficacy of new topic model variants on automated measures, and their failure to approximate human preferences places these models on uncertain ground. Moreover, existing evaluation paradigms are often divorced from real-world use.

Motivated by content analysis as a dominant real-world use case for topic modeling, we analyze two related aspects of topic models that affect their effectiveness and trustworthiness in practice for that purpose: the *stability* of their estimates and the extent to which the model's discovered categories *align with* human-determined categories in the data. We find that neural topic models fare worse in both respects compared to an established classical method. We take a step toward addressing both issues in tandem by demonstrating that a straightforward ensembling method can reliably outperform the members of the ensemble.

1 Introduction

Topic models provide an unsupervised way both to discover implicit categories in text corpora, and to estimate the extent to which any given category applies to a specific text item. As such, topic modeling can be viewed as an automated variety of *content analysis* for text: those two capabilities directly correspond to the practice of developing an emergent *coding system* via examination of a text collection, and then *coding* the text units in the collection (Stemler, 2000; Smith, 2000). This form of content analysis is a dominant use case for topic models and therefore it is our focus here.

Explicitly identifying topic models as a tool for content analysis allows us to characterize what makes topic models *good*: we can measure the extent to which a model achieves the goals of content analysis. This careful consideration of the criteria for "good" topic models is essential because recent results have challenged the validity of the prevailing model evaluation paradigm (Hoyle et al., 2021; Doogan and Buntine, 2021; Doogan, 2022; Harrando et al., 2021). In particular, Hoyle et al. identified a *validation gap* in the automatic evaluation of topic coherence: metrics like the widely-used normalized pointwise mutual information (NPMI) were never validated using human experimentation for the newer neural models once they emerged, and the authors demonstrated that such metrics exaggerate differences between models relative to human judgments. Given that the majority of claimed advances in topic modeling are predicated on these metrics (per Hoyle et al.'s meta-analysis), it would appear that much of the topic model development literature now rests on uncertain ground. Doogan and Buntine also challenged the validity of current automated evaluation measures, and highlighted the disconnect between these measures and use of topic models in real-world settings.

In this paper, we begin with the needs of content analysis, and we use those needs to argue for specific choices of how to measure topic model performance. We then report on comprehensive experimentation using two English-language datasets, four neural topic models that are representative of the current state of the art, and classical LDA with Gibbs sampling as implemented in MALLET (McCallum, 2002). The results indicate that MALLET is a more reliable choice than the more recent neural models from a content analysis perspective. Taking a step toward addressing these issues, we use a straightforward ensemble method that combines the output of models across runs, which reliably yields better results than the usual practice of running a single model.

To summarize our argument and contributions:

- Automated comparison of topic models should be grounded in a use case, and content analysis is a dominant use case for topic

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5350 - 5373

models (§2.2).

- Stability and reliability are necessary—although not sufficient—criteria to ensure the value of a content analysis (§2.3).
- Stability and reliability can be directly measured from model outputs, unlike automated coherence, which prior work has shown is an unreliable proxy for human judgment (§2.4).
- On these metrics, we show that LDA with Gibbs sampling (as implemented in Mallet) is significantly more stable and reliable than newer neural models (§4).
- We present a straightforward ensembling method to mitigate the stability problem (§6). We release all code and data.[1]

2 What makes a topic model "good"?

In considering how to characterize a topic model that works well, we focus on text content analysis as a dominant use case for topic modeling.

2.1 Traditional content analysis

Although content analysis is an extremely broad concept (Krippendorff, 2018), a very widely used paradigm across many disciplines is a manual process of inductive discovery of codesets via *emergent coding* (Stemler, 2000), which "allows categories to emerge from the material without the influence of preconceptions" (Smith, 2000). Weber (1990) describes a "data-reduction process by which the many words of texts are classified into much fewer content categories," and this invocation of data reduction in a manual setting provides a sense of why topic modeling, a dimensionality-reduction technique, can be a good fit when considering ways to automate the process.[2]

Typically the inductive process involves multiple researchers independently reading samples of the text units being analyzed, and proposing categories (usually called "codes") that they see as present and relevant; they then reconcile their independent proposed categories to produce a candidate codeset with associated definitions and coding guidelines. The candidate codeset is then used by two or more people to independently code (i.e. manually label) a sample of the data, and inter-coder reliability is measured using a chance-corrected agreement measure like Krippendorff's α (cf. Artstein and Poesio, 2008). If an acceptable level of reliability has not yet been achieved, the codeset and coding guidelines are revisited and revised, and another iteration of independent coding and reliability measurement takes place. Once reasonable reliability has been achieved, the final set of categories is considered to reflect true structure underlying the text collection. Sometimes the texts in the collection are then manually coded using those categories in order to support quantitative analysis — possibly with further inter-coder reliability measurement for quality control — although sometimes the set of categories itself is the intended result, not item-level coding.

2.2 Topic modeling for content analysis

The models of interest in this paper are exemplified by Latent Dirichlet Allocation (LDA, Blei et al., 2003), within which each of N documents d is represented as an admixture $\boldsymbol{\theta}_d$ of K topics, and each topic is itself represented as a distribution $\boldsymbol{\beta}_k$ over the vocabulary V. Topics can thus be viewed in two complementary ways, as ranking either the words in the vocabulary or the documents in the collection. These views can be interpreted as corresponding closely to two central elements of a traditional text content analysis. First, the rows in the topic-word distributions matrix $\mathbf{B} \in \mathbb{R}^{K \times |V|}$ constitute an inductively determined set of categories analogous to a human-determined codeset; for example, the presence of a topic with top (most probable) words `artist, museum, exhibition` might correspond to a human analyst identifying the code ART. Second, the columns of the document-topic matrix $\boldsymbol{\Theta} \in \mathbb{R}^{N \times K}$ constitute a soft coding of documents using the categories in $\mathbf{B}$.[3] To help illustrate the first step, Table 1 shows the top words from inferred topic-word distributions $\boldsymbol{\beta}_k$ for two model types over multiple runs.

Reviewing the use of topic models. Bearing these correspondences in mind, we reviewed the literature to confirm our subjective impression that text content analysis is indeed the dominant use

[1] `github.com/ahoho/topics`

[2] This process of inductive category discovery contrasts with the use of pre-existing categories, e.g. those coming from relevant theory, and from the use of "manifest" or directly observable characteristics of text. Discussions in the literature often distinguish "quantitative" from "qualitative" content analysis, with the inductive process we describe being associated with the latter category. This terminological distinction may be overly sharp, however; see Schreier (2012) for useful discussion of relationships and differences among quantitative content analysis, qualitative content analysis, and other forms of qualitative research.

[3] This could be converted into traditional discrete coding in a number of ways, e.g. assigning the most probable topic for a document as its code.

Model Type	Run	Top Words
MALLET	base	**storm tropical hurricane winds depression mph september damage cyclone system**
	nearest	**storm tropical hurricane winds depression mph september damage cyclone system**
	median	**tropical storm hurricane depression winds september cyclone mph system** august
	farthest	**tropical storm depression hurricane cyclone system season winds september mph**
D-VAE	base	**tropical mph storm hurricane winds cyclone extratropical utc rainfall**
	nearest	**tropical cyclone hurricane storm winds landfall** depression dissipated convection **extratropical**
	median	convection **landfall** shear nhc utc **tropical** mbar northwestward **cyclone extratropical**
	farthest	dvorak southwestward depressions dissipation intensifying conventionally southeastward

Table 1: Sets of WEATHER topics for two model types for different runs with different hyperparameters on a Wikipedia dataset, represented in conventional fashion using the most probable ten words per topic. The table visually illustrates MALLET's dramatically greater stability: the top words from the base topic appear in corresponding topics across the the full range of the other nine runs (overlap with base topic in **orange**), while for D-VAE, a neural topic model, consistency with the base topic begins to show a significant drop-off even with the nearest topic (overlap in **blue**). See §4 and §5 for discussion.

case for topic modeling.[4] Using Semantic Scholar (`semanticscholar.org`), we collected research studies *outside* the field of computer science published in 2019–2022 that cite Blei et al. (2003), and selected 50 at random. We excluded studies that cite Blei et al. but do not actually use any topic model, as well as studies that do not involve language data. We retain those that employ topic model variants, such as STM (Roberts et al., 2013). Using Semantic Scholar's reported field of study, disciplines represented include medicine, sociology, business, political science, psychology, economics, and history. We find that 94% of the papers use a topic model for inductive discovery of categories for human consumption, 68% of which go on to actually assign human-readable code labels to topics; and 64% of papers use document-topic probabilities as a form of coding for individual text units. We interpret these results as strongly suggesting that, outside topic model development, the primary *use* of topic models is an automated form of text content analysis as characterized in §2.1.

2.3 Criteria for good content analysis

Having established text content analysis as a central topic modeling use case, we consider criteria for "good" analysis motivated by that use case. These then inform the selection of topic model evaluation metrics in §2.4, helping to ensure a correspondence between the way topic models are evaluated and the reasons people are using them.

One key issue in content analysis is *stability* or *intra-coder reliability*: if the same coder were to look at the same data again (say, separated by a long interval to achieve some degree of independence), would they produce the same results? When an individual coder cannot produce stable output, this calls into question the quality of the results they have produced any one of those times.

A second central concern in content analysis is *reproducibility* or *inter-coder reliability*: do two or more independent coders looking at the same data agree with each other? In the absence of externally provided coding to compare against, what establishes trust in categories or coding is consensus, what Weber (1990) refers to as "the consistency of shared understandings" between coders.

A third concern that is often discussed is *validity*: do categories or measurements actually correspond to whatever they are intended to measure (Rubio, 2005)? As Weber (1990) notes, in content analysis this often goes only as far as face validity, i.e. a subjective perception that a measure (or category) appears to be valid. In contrast, Shapiro and Markoff (1997) argue that content analysis "is only valid and meaningful to the extent that the results are related to other measures".

Research in content analysis typically focuses on these three issues — stability, reproducibility, and validity — as necessary considerations when considering whether a content analysis should be used as the basis for inferences about a dataset. Validity, however, is challenging to assess outside the context of specific research questions (see Grimmer and Stewart, 2013, for an example in political science). We therefore focus on stability and reproducibility as the basis for developing metrics to assess topic models for the automated content analysis use case.

[4]This review is in the spirit of Liberati et al. (2009), although we are not striving for that level of formality. See Appendix A.2 for more details.

Note that the criteria we emphasize—stability and reproducibility—are necessary to ensure the value of a content analysis, but not sufficient: topic coherence is a complementary and crucial concern (Newman et al., 2010) that requires further investigation, since prior work has shown automated coherence measurements are an unreliable proxy for human judgment (Hoyle et al., 2021).

2.4 Operationalizing the criteria

Because they are generative models, the development community initially evaluated topic models using held-out perplexity, i.e. their ability to predict unseen text. However, focusing on the goal of producing categories that humans can understand, Chang et al. (2009) established that perplexity actually correlated *negatively* with human determinations of coherence as estimated using behavioral measures. Lau et al. (2014) went on to introduce NPMI as an automated coherence metric positively correlated with human preferences. Since then, NPMI has been the most prevalent way to establish that a new topic modeling method is better than the old ones, including the new generation of neural topic models. However, Hoyle et al. (2021) recently identified a *validity gap* for NPMI: its correspondence to human judgments was never validated for neural topic models, and although recent neural topic models can attain relatively high NPMI, human annotators fail to meaningfully distinguish them from a classical LDA baseline.

That result suggests taking a fresh, well motivated look at topic model evaluation. Any model evaluation should be grounded in consideration of the model's intended purpose, which leads us to suggest grounding formal evaluation metrics in the content analysis use case.[5]

2.4.1 Stability

§2.3 notes *stability* as an important criterion in content analysis. Whether codes are being produced by a human coder or a topic model, if there is meaningful latent structure in the text collection, one would expect either humans or models to consistently uncover that structure.

To ground our evaluation in our use case, we measure the stability of models across hyperparameter settings (for a fixed topic number K). In the absence of an unsupervised metric to optimize or reliable "default" values, a practitioner is forced to explore different hyperparameter settings. All else equal, a topic model that is less sensitive to changes in hyperparameter settings is preferable to one that is more sensitive (we also evaluate results for fixed hyperparameters with different random seeds, see Appendix A.1).

Translating these ideas into a formal measurement, we follow Greene et al. (2014) in operationalizing model stability by measuring the total distances between the topic-word estimates for each run, extending their method to measure stability of both the sorted rows of the topic-word estimates $\mathbf{B}$ or the sorted columns of the document-topic estimates $\mathbf{\Theta}$; the smaller these distances, the more stable the estimates.

Without loss of generality, we focus on the topic-word distributions to operationalize stability as *total topic distance*. We collect a set of $\boldsymbol{\beta}_k^{(i)}, i \in 1\ldots, m, k \in 1\ldots K$ estimates from m model runs on the same dataset. For each pair of $\binom{m}{2}$ runs, we compute the pairwise distance d between all K topics in each run. We use the Rank-Biased Overlap distance (RBO, Webber et al., 2010), which is used to measure the distance between two rankings giving more importance to similarity of the top-ranked items, i.e., the measure is *top-weighted*, making it ideal for measuring the distance between topics (Mantyla et al., 2018).[6] Within a pair of runs $\mathbf{B}^{(i)}, \mathbf{B}^{(j)}$, the goal is to find a permutation of rows $\pi(\cdot)$ to minimize

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

This problem is an instance of bipartite matching distance minimization, which we solve with the modified Jonker-Volgenant algorithm of Crouse (2016). If the set of $\binom{m}{2}$ total distances $\mathcal{TD}$ (i.e., the minimized costs) for one model are significantly less than a second model, the first model is more stable.

Prior topic modeling work has identified stability as a crucial concern for robust application to social

[5] It should be noted that we are focusing on only the most central part of the content analysis use case. Smith (2000) situates codeset discovery and coding within a broader process that begins with identifying the research problem, selecting appropriate materials, etc., and ends with actually using the codeset and coding to generate research findings. Bayard de Volo et al. (2020) situate topic model creation within a corresponding end-to-end workflow; see also Boyd-Graber et al. (2014) for practical discussion of topic modeling including discussion of other use cases.

[6] Experiments with distances that used the full distribution, like Jensen-Shannon divergence, led to matched topics that were less interpretable.

sciences (Koltcov et al., 2014; Ballester and Penner, 2022), for better incorporation of topic models in downstream automated NLP tasks (Miller and McCoy, 2017), as a criterion for tuning LDA parameters (Greene et al., 2014), and has offered ways to improve it for LDA estimates (Agrawal et al., 2018; Mantyla et al., 2018). Chuang et al. (2015) introduced an interactive tool to help humans assess a topic model's stability. However, in a meta-analysis of 35 papers proposing new "state-of-the-art" neural topic models over the past three years (2019-2022), we find that *none* of them compared the models on stability.[7]

2.4.2 Inter-coder reliability

§2.3 notes that *reproducibility* or *inter-coder reliability* is also a central consideration in content analysis. Going beyond intra-coder consistency, if a set of codes cannot be applied consistently by multiple coders, this also calls into question whether it is doing a good job capturing meaningful content categories.

We treat a topic model $\langle \mathbf{B}, \mathbf{\Theta} \rangle$ as a coder, and approach inter-coder reliability from the perspective of reproducing categories from other coders who are human, instantiated as a set of human-assigned "ground truth" labels for the documents in the collection. Since what we care about here are the categories discovered by a topic model, not actual labels, we measure the extent to which categories induced by the model *align with* that ground truth. Intuitively, for example, if documents that are assigned to a topic by the model all have the same ground-truth label, the topic is a good fit for human categorization of the data (and this can be determined just using documents assigned to the topic, without any generation or evaluation of labels). Conversely, if documents all assigned to the same topic in the model have a wide variety of ground truth labels, this mismatch suggests that the topic is missing something important relative to the underlying category structure in the collection.

By taking the most probable topic for a document $\hat{\ell_d} = \text{argmax}_{k'}\ \theta_{d,k'}$ as its assigned topic or "code", we can apply standard metrics of cluster quality.[8] We borrow exposition of cluster quality metrics from Poursabzi-Sangdeh et al. (2016), with all metrics using the predicted clustering from a model, $\hat{\mathcal{L}} = \{\ell_d : d = 1, \ldots, n\}$, and a given set of gold labels $\mathcal{L}^*$.

Adjusted Rand Index. The Rand Index compares all pairs of the two labelings over documents, counting the proportion of pairs that have the same (TP) or different (TN) assignments (Rand, 1971).

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

The adjusted rand index further corrects for chance (Steinley, 2004).

Normalized Mutual Information (NMI) measures the mutual information between two clusterings, and is invariant to cluster permutations (Strehl and Ghosh, 2002). Here, $\mathbb{I}$ is the mutual information and $\mathbb{H}$ are the entropies for each clustering.

$$\text{NMI} = \frac{2\mathbb{I}\left(\hat{\mathcal{L}}, \mathcal{L}^*\right)}{\hat{\mathbb{H}} + \mathbb{H}^*} \tag{2}$$

Purity takes all documents contained in a single *predicted* cluster and measures the number of associated gold labels that appear in it — it is roughly akin to precision (Zhao and Karypis, 2002). A small number of gold labels present in a predicted cluster means that there is high alignment between the discovered "concept" and the true one.

$$P(\hat{\mathcal{L}}, \mathcal{L}^*) = \frac{1}{n} \sum_k \text{max}_{k'} |\hat{\mathcal{L}}_k \cap \mathcal{L}^*_{k'}| \tag{3}$$

With $\mathcal{L}_k = \{\ell_d : d = 1, \ldots, n; \ell_d = k\}$. Purity is not symmetrical, so we define *inverse purity* as $P^{-1} = P(\mathcal{L}^*, \hat{\mathcal{L}})$, and P_1 as their harmonic mean (analogous to F_1).

Prior topic modeling work has looked at how well topics discovered by a model align with reference codes (Chuang et al., 2013; Korenčić et al., 2021). However, in the same meta-analysis discussed above, only *six* of the 35 neural topic modeling development papers compared models on a version of alignment. This suggests that even though stability and alignment have been identified as important and useful criteria in topic modeling literature in prior work — especially when using and examining LDA and its variants — they have seen precious little uptake. We hope that our strong use-case motivations and experimental results will change this.

[7] We select publications from the meta-analysis in Hoyle et al. (2021), updated with recent work from the most common venues in that list. Although papers do not measure the stability of the estimates directly as we do here, six papers do report variance over chosen metrics.

[8] It is common in the topic model development literature to evaluate models by learning a mapping $f : \boldsymbol{\theta}_d \rightarrow \ell_d$ from training data and calculating a held-out F_1 score—i.e., to train a classifier—but this process does not correspond to any common real-world use of topic models.

3 Experiments

Having argued that topic models should be subject to evaluations designed with real-world uses in mind, and having motivated specific ways to operationalize evaluative measurements based on criteria that matter for text content analysis, we evaluate nominally "state-of-the-art" topic models to understand how well they perform relative to those criteria.

3.1 Datasets

We use two standard English datasets of varying characteristics: 14,000 "good" articles from Wikipedia (`Wiki`, Merity et al., 2017) and 32,000 bill summaries from the 110-114th U.S. congresses (`Bills`).[9] The documents in both datasets have hierarchical labels, which serve as ground truth when evaluating the quality of the document-topic posteriors (§2.4). The `Wiki` dataset has 45 labeled high-level and 279 low-level labels; the `Bills` dataset has 21 high-level and 114 low-level labels. We process each with the standardized setup of Hoyle et al. (2021), setting the vocabulary size to either 5,000 or 15,000 terms, limiting by term-frequency (Blei and Lafferty, 2006).

Prior evidence suggests that neural topic models may produce topics with narrower scope than classical models (e.g., `agnes_martin, sol_lewitt, minimalism` rather than `art, painting, museum`, cf. Hoyle et al., 2021). We therefore generate held-out sets for both datasets to facilitate exploration of this phenomenon. Namely, we ensure that both the training and held-out sets contain documents from all *high-level* categories, but partition the *low-level* categories into seen and unseen labels. For example, Wikipedia articles about `television` are present in both subsets, but those about `30 Rock` episodes are exclusively in the training set whereas `Simpsons` episodes are unseen. Although not an emphasis of the present work, our high-level conclusions remain the same for the held-out data (i.e., MALLET is better-aligned, Appendix A.1); we leave further analysis to future efforts.

3.2 Models and experimental contexts

Classical topic models use Gibbs sampling or variational inference to infer the posteriors over the latent variables; more recent *neural* topic models use contemporary techniques that involve neural networks, such as variational auto-encoders (Kingma and Welling, 2014).

We evaluate one classical topic model and four neural topic models. Each model is evaluated in one of 16 *experimental contexts*: a tuple of dataset (`Bills`, `Wiki`), vocabulary size (5k, 15k), and number of topics (25, 50, 100, 200).

In light of the finding that automated coherence cannot meaningfully reproduce human judgments (Hoyle et al., 2021), there is no unsupervised metric that we can optimize to avoid the problem of instability, while optimizing for K remains an open research problem. Therefore we vary K and, for all contexts, we train the models ten times using a different set of randomly-selected hyperparameters, where value ranges are based on prior literature (§A.3).

Although "optimal" hyperparameters will often change depending on context, we also report results with fixed hyperparameters and varying seeds in Appendix A.1.

MALLET. Given its prevalence among practitioners and strong qualitative human ratings in prior work (Hoyle et al., 2021), as a classical model we use LDA estimated with Gibbs sampling (Griffiths, 2002), implemented in MALLET (McCallum, 2002). While LDA is a common baseline in the topic model development literature, it is often estimated with variational methods, which anecdotally produce lower-quality topics (Goldberg, 2020).[10]

SCHOLAR. A popular neural alternative to the structural topic model (Roberts et al., 2014), flexibly incorporating supervised signals and external covariates into the model (Card et al., 2018).

SCHOLAR+KD. Hoyle et al. (2020) apply knowledge distillation (KD) to improve on SCHOLAR using a BERT-based autoencoder. Gao et al. (2021) show that domain experts prefer the outputs of an adapted SCHOLAR+KD over other models (MALLET, ETM, Dieng et al., 2020).

Dirichlet-VAE. The Dirichlet-VAE (D-VAE, Burkhardt and Kramer, 2019) is a variant of LDA that (a) uses a VAE to approximate the posterior over the latent document-topic distribution, and

[9] "Featured" Wikipedia articles have an incompatible labeling scheme and are therefore excluded. Raw bill data was extracted from `https://www.govtrack.us/data/us/`.

[10] Mimno (2022) provides discussion of why stochastic variational Bayes, which has seen widespread use in topic modeling using the `gensim` library (`https://radimrehurek.com/gensim/`), may be particularly problematic.

(b) follows PRODLDA by using unnormalized estimates of the topic-word values β, as opposed to a proper distribution. Annotators rate D-VAE's topics similarly to MALLET (Hoyle et al., 2021).

Contextualized Topic Model. Typically, VAE-based neural topic models encode the bag-of-words representation of a document with a neural network to parameterize that document's distribution over topics. The popular model introduced in Bianchi et al. (2021a) extends this representation with a contextualized document embedding from a large pretrained language model.

4 Results

Recall that measuring *stability* is motivated by intra-coder reliability in content analysis: producing the same result every time increases confidence that the analysis reflects actual latent structure in the data. MALLET is significantly more stable than other models across the vast majority of contexts, often by a large margin (Table 2). Most striking are the topic-word distributions $\mathbf{B}$: none of the neural models even approach its consistent level of stability. CTM sometimes achieves comparable stability for $\mathbf{\Theta}$; this may be due to its use of pretrained document embeddings, which are transformed in order to parameterize the estimate.

Recall also that *alignment* is motivated by *inter*-coder reliability in content analysis: is the model, in the role of analyst, agreeing with human-derived categorization for the data? MALLET shows strong consistency in providing the numerically best alignment with human categorization across datasets (Table 3).[11] Among neural models, D-VAE and SCHOLAR sometimes achieve statistically indistinguishable performance, but they do not approach MALLET's consistency across datasets, number of topics, and metrics.

Now, we argued in §2.4.1 that practitioners do not have access to optimal hyperparameters for a given model, because what is optimal will depend on the dataset, number of topics, preprocessing, and other experimental decisions. The above results show that model estimates can be very sensitive to different hyperparameter settings and they clearly favor MALLET on our metrics. However, in many real-world scenarios, a practitioner may simply rely on some "default" settings. We therefore also evaluate models for *fixed* hyperparameters using reasonable default values.[12]

To generate the defaults, for each dataset and model we find the hyperparameter settings that yield the best alignment performance across experimental contexts (vocabulary size, number of topics, alignment metric, and label hierarchy). Specifically, within each context, we first rescale the alignment metric values over the 10 runs for that model to avoid differences in metric values; we then select the hyperparameters which have the largest average values across all contexts, for a given dataset. Finally, to approximate a common use case and to avoid overfitting to the dataset, we use the hyperparameters obtained from one dataset to train models on the *other* dataset (e.g., we select defaults based on the `Bill` alignment metrics and set those for new models run on `Wiki`; defaults in Appendix A.3).

Results are in Appendix A.1. Unsurprisingly, fixing "good default" hyperparameters for the neural models improves their stability and alignment. In particular, D-VAE has competitive alignment metrics in the $|V| = 5k$ case, although it is hampered by its relatively poor stability. MALLET's stability is marginally affected: while it is no longer as consistently dominant, it remains more stable and better-aligned in the majority of contexts.

5 A close reading of model stability

Table 1 illustrates corresponding versions of a topic from different runs of D-VAE and MALLET. For a given context (here, $K = 50, |V| = 15,000$), we collect the topic-word estimates $\mathbf{B}$ across the 10 runs for each of the two models, each run using a different set of randomly-selected hyperparameters. One weather-related topic across runs was chosen manually as the "base" run, and then the corresponding topics in the other nine runs for the same model were ranked by their RBO distance to that topic. The nearest, median, and most-distant topics in that ranking, shown in the table, therefore capture the range of variation across different hyperparameter settings.

It is immediately clear that even the nearest topic for D-VAE has fewer words in common with the base topic, compared to MALLET. And as distance increases, the top words for MALLET stay consistent, whereas those for D-VAE change dramatically, even if they relate to the same weather concept.

[11] The distribution of scores across runs and results for other contexts are shown in Appendix A.4.

[12] We thank an anonymous reviewer for pointing this out and suggesting the additional experimentation.

		$\lvert V\rvert = 5$k								$\lvert V\rvert = 15$k							
		$k = 25$		$k = 50$		$k = 100$		$k = 200$		$k = 25$		$k = 50$		$k = 100$		$k = 200$	
		Θ	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**
Bills	MALLET	**0.78**	**0.27**	**0.74**	**0.28**	**0.74**	**0.32**	0.79	**0.41**	**0.79**	**0.29**	**0.80**	**0.33**	**0.77**	**0.35**	0.76	**0.39**
	SCHOLAR	0 .88	0.63	0.82	0.56	0.76	0.50	0.78	0.57	0.86	0.82	0.85	0.76	0.83	0.71	0.84	0.71
	SCHLR+KD	0.91	0.67	0.89	0.59	0.87	0.64	0.83	0.65	0.90	0.75	0.88	0.69	0.85	0.67	0.88	0.70
	D-VAE	0.97	0.62	0.97	0.77	0.96	0.73	0.96	0.76	0.97	0.75	0.97	0.81	0.97	0.84	0.95	0.83
	CTM	0.79	0.43	0.80	0.51	0.76	0.54	**0.74**	0.58	0.81	0.44	0.83	0.55	0.81	0.60	**0.76**	0.65
Wiki	MALLET	**0.70**	**0.22**	**0.69**	**0.29**	**0.62**	**0.30**	**0.67**	**0.37**	**0.71**	**0.26**	**0.70**	**0.32**	**0.66**	**0.34**	**0.70**	**0.39**
	SCHOLAR	0.82	0.49	0.73	0.38	0.80	0.45	0.83	0.52	0.84	0.66	0.77	0.54	0.77	0.67	0.79	0.61
	SCHLR+KD	0.83	0.47	0.80	0.47	0.86	0.52	0.83	0.42	0.88	0.65	0.84	0.60	0.86	0.64	0.89	0.60
	D-VAE	0.92	0.46	0.92	0.55	0.91	0.54	0.92	0.67	0.92	0.67	0.92	0.63	0.90	0.76	0.87	0.72
	CTM	0.76	0.42	0.73	0.39	0.73	0.46	0.72	0.51	0.76	0.39	0.76	0.43	0.76	0.50	0.73	0.56

Table 2: Stability for topic-word **B** and document-topic **Θ** estimates, over 10 runs. Smallest per-column values are **bolded** and are sig. smaller than unbolded values (2-sided t-test, $p < 0.05$); underlined values have $p > 0.05$.

		$k = 25$			$k = 50$			$k = 100$			$k = 200$		
		ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI	P_1
Bills $\ell = 114$	MALLET	**0.30**	**0.45**	**0.46**	**0.34**	**0.48**	**0.47**	**0.32**	**0.50**	**0.43**	0.22	**0.50**	0.35
	SCHOLAR	0.12	0.28	0.27	0.19	0.40	0.34	0.15	0.40	0.29	0.12	0.39	0.25
	SCHLR+KD	0.11	0.28	0.27	0.16	0.37	0.35	0.14	0.41	0.33	0.11	0.38	0.25
	D-VAE	0.26	0.45	0.44	0.24	0.43	0.40	0.24	0.46	0.40	**0.24**	0.46	**0.38**
	CTM	0.21	0.40	0.38	0.26	0.45	0.41	0.25	0.48	0.39	0.19	0.49	0.34
Wiki $\ell = 279$	MALLET	**0.23**	**0.65**	**0.41**	**0.32**	**0.69**	**0.50**	**0.37**	**0.71**	**0.53**	0.32	**0.70**	0.48
	SCHOLAR	0.21	0.61	0.38	0.31	0.68	0.48	0.34	0.69	0.50	0.29	0.68	0.44
	SCHLR+KD	0.19	0.61	0.37	0.26	0.65	0.43	0.29	0.65	0.44	0.22	0.62	0.36
	D-VAE	0.22	0.64	0.39	0.30	0.68	0.48	0.27	0.65	0.45	0.30	0.68	**0.49**
	CTM	0.21	0.60	0.36	0.27	0.64	0.43	0.31	0.67	0.46	**0.34**	0.69	0.47

Table 3: Average alignment metrics across 10 runs, measured against gold labels at the lowest hierarchy level, $|V| = 15,000$. Largest values in each column are **bolded**, which are significantly greater than unbolded values in a two-sided t-test ($p < 0.05$); underlined values have $p > 0.05$.

Note that in this example, consistent with anecdotal reports from other practitioners and our own experience, the neural model tends toward less frequent or more specific words. The idea that neural models may be capturing topics that are in some sense narrower, with instability leading to different such topics in each run, leads directly to the idea that a cross-run ensemble might be expected perform better than the individual runs—which is important in the absence of a reliable automated method for optimizing hyperparameters.

6 Ensembling estimates

We have highlighted lack of stability as a serious problem for neural topic models, but neural models can also have desirable properties. How can we increase the odds of obtaining a good neural topic model in the face of extreme variation? The distance metrics we use to measure instability offer one solution: clustering to aggregate similar estimates over runs to form an *ensemble*. We adopt an approach similar to prior work (Miller and McCoy, 2017; Mantyla et al., 2018), going further by accounting for the document-word estimates **Θ** and by evaluating ensembles' alignment against human categorization. Specifically, we concatenate run estimates over the m runs $\bar{\mathbf{B}} = \left[\mathbf{B}^{(i)}\right]_i^m$ and $\bar{\mathbf{\Theta}} = \left[\mathbf{\Theta}^{\top(i)}\right]_i^m$, where each row in the concatenated matrix is a topic. We then compute pairwise distances between topics, $\mathcal{D}(\bar{\mathbf{B}})$ and $\mathcal{D}(\bar{\mathbf{\Theta}})$, and cluster based on a linear interpolation of the two distances, $\lambda\mathcal{D}(\bar{\mathbf{B}}) + (1-\lambda)\mathcal{D}(\bar{\mathbf{\Theta}})$, where λ is a hyperparameter. The estimate of each topic k from each run i, $\langle \boldsymbol{\theta}_k^{\top(i)}, \boldsymbol{\beta}_k^{(i)} \rangle$, is assigned to a cluster, and to infer new document-topic or topic-word scores for the ensemble, we take the element-wise mean over the estimates assigned to each cluster.[13]

[13]We experimented with k-medoids (Lloyd, 1957; Kaufman and Rousseeuw, 2008) and hierarchical agglomerative clustering (Day and Edelsbrunner, 1984), and also varied the distance metric — either RBO (§2.4.1) or the Average Jaccard score (Greene et al., 2014). Both metrics are used to measure the distance between topics in prior work (Mantyla et al., 2018; Greene et al., 2014). Clustering algorithms were implemented using scikit-learn (Pedregosa et al., 2011).

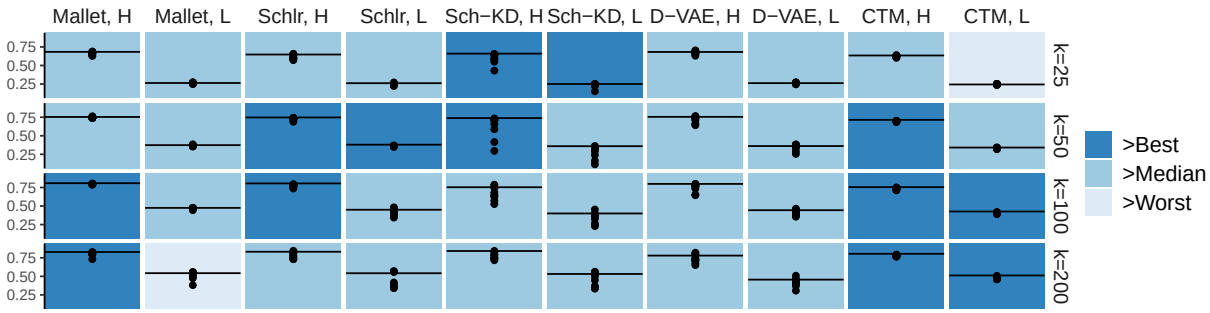


Figure 1: Ensembling performance on alignment (purity metric, § 2.4.2) for the `Wiki` dataset. Each box represents a context: the columns identify model type and the label granularity used in evaluation (e.g. top left is MALLET with **H**igh-level categories), and the rows correspond to different values K for the number of topics. Dots are alignment scores for individual runs; the horizontal line is the alignment score for that *ensemble* of runs using our method. Shading indicates when the ensemble method has beaten the score of the best individual run (darkest), the median (middle), or has outperformed the worst individual run (lightest). The ensemble is typically in the top quartile of the component runs. Ensembles virtually always outdo the median, and frequently outperform the best individual run.

To evaluate this method, we compare the alignment score of the ensemble (§2.4.2) combining the $m = 10$ runs, versus the alignment score of each individual member run. We do so across each of the 400 contexts (model, dataset, K, high versus low label granularity, and metric). Figure 1 illustrates a summary of results for the purity alignment metric on the Wikipedia dataset. Across the full range of our experimentation, the ensemble improves on the median member in 97% of all the contexts, and it is *always* better than the worst member (full results in Appendix A.5).

7 Conclusions

A tool can be considered broken when it doesn't work well for its intended use. In this paper we have focused on a widespread use case for topic models, their application in text content analysis; we have carefully motivated criteria for measuring the extent to which a topic model is serving those needs; and we have demonstrated through comprehensive and replicable experimentation that, when measured on those criteria, recent and representative neural topic models fail to improve on the classical implementation of LDA in MALLET. In particular, MALLET is much more stable, reducing concerns from the content analysis perspective that different runs could yield very different codesets. Equally important, across the vast majority of contexts, its discovered categories are reliable as measured via alignment with ground-truth human categories. For people seeking to use topic modeling in content analysis, therefore, MALLET may still be the best available tool for the job.

That said, there are still good reasons to investigate neural topic models. Foremost among these is the fact that they can benefit from pretraining on vast, general samples of language (e.g. Hoyle et al., 2020; Bianchi et al., 2021a; Feng et al., 2022). Neural realizations of topic models can also be integrated smoothly for joint modeling within larger neural architectures (e.g. Lau et al., 2017; Wang et al., 2019, 2020), and hold the promise of being more straightforward to use multilingually (e.g. Wu et al., 2020; Bianchi et al., 2021b; Mueller and Dredze, 2021) or multimodally (e.g. Zheng et al., 2015).[14] We therefore introduced one possible way to address the shortcomings we identified using a straightforward ensemble technique.

Perhaps the most important take-away we would suggest is that *development* of new topic models—indeed, of all NLP models—should be done with *use cases* firmly in mind. Some models are enabling technologies, without a direct user-facing purpose, and others are intended to produce results directly for human consumption. But whatever the goal, the driving question for methodological development and evaluation should *not* be how to demonstrate an improvement in "state of the art", it should be why the model is being created in the first place and what measurements will demonstrate improved performance for that intended purpose.

Limitations

Our studies used only English datasets, while topic modeling has been used to characterize texts in

[14] See Zhao et al. (2021) for more potential advantages of neural topic modeling.

many languages. While *theoretically* we see no reason why our results and findings should not generalize beyond the English language, *empirical* generalizability across languages remains to be determined.

Our method for measuring alignment of model-induced categories with human-determined categories relies on ground-truth human labels, potentially limiting its broader applicability. In addition, the categories in the Wikipedia data were not, to our knowledge, produced via a traditional human content analysis process. We are currently designing a follow-up study in which human subject matter experts perform traditional content analysis from scratch on the same dataset used for topic modeling, in order to provide a head-to-head comparison between automated and traditional methods and to establish human upper bounds on inter-coder reliability.

Our literature review of topic modeling use cases was not a formal systematic review (Moher et al., 2009). It relied on Semantic Scholar's content and its discipline categorization, and potentially excluded papers in computer science that were about the use of topic models rather than method development. It seems clear that text content analysis *a* dominant use case for topic modeling, if not *the* dominant use case. In the social sciences, we also note frequent use of the Structural Topic Model (Roberts et al., 2014) which, like SCHOLAR, can incorporate metadata into model estimation—we leave an evaluation of this use case to future work.

8 Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grants 2031736 and 2008761 and by Amazon. We thank David Mimno and Xanda Schofield for their input on earlier drafts, as well as our anonymous reviewers for their helpful comments.

References

Amritanshu Agrawal, Wei Fu, and Tim Menzies. 2018. What is wrong with topic modeling? and how to fix it using search-based software engineering. *Information and Software Technology*, 98:74–88.

Zack W. Almquist and Benjamin E. Bagozzi. 2019. Using radical environmentalist texts to uncover network structure and network features. *Sociological Methods & Research*, 48:905 – 960.

Mahdiyeh Amozegar. 2021. *Tweeting in Times of Crisis: Shifting Personal Value Priorities in Corporate Communications and Impact on Consumer Engagement*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Qostal Aniss, Moumen Aniss, and Lakhrissi Younes. 2021. The role of social networks in personality analysis for recruitment of laureates: A systematic review and exploratory study. *SHS Web of Conferences*.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Omar Ballester and Orion Penner. 2022. Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, 16(1):101224.

Tyler Bateman, Shyon Baumann, and Josée Johnston. 2019. Meat as benign, meat as risk: Mapping news discourse of an ambiguous issue. *Poetics*.

Theo Bayard de Volo, Alfredo Gomez, Tatsuki Kuze, and Alexandra Schofield. 2020. LDA in the wild: How practitioners develop topic models. In *West Coast NLP*.

Emily Bell and Tyler A. Scott. 2020. Common institutional design, divergent results: A comparative case study of collaborative governance platforms for regional water planning. *Environmental Science & Policy*, 111:63–73.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. 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, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems*, 18:147.

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

Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 225255.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model.

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, Melbourne, Australia. Association for Computational Linguistics.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc.

Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. 2013. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pages 612–620. PMLR.

Jason Chuang, Margaret E. Roberts, Brandon M. Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. 2015. TopicCheck: Interactive alignment for assessing topic model stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–184, Denver, Colorado. Association for Computational Linguistics.

Alberto Coco, Nicola Viegi, et al. 2020. *The monetary policy of the South African Reserve Bank: Stance, communication and credibility*. Economic Research and Statistics Department, South African Reserve Bank.

David F. Crouse. 2016. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696.

Dipok Chandra Das, Dipa Rani Saha, T Kabir, Prosanto Deb, and Joyjeet Bhowmik. 2020. Analysis of covid-19 coverage in bangladesh news media using topic modelling.

William HE Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24.

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

Caitlin Doogan. 2022. *A Topic is Not a Theme: Towards a Contextualised Approach to Topic Modelling*. Ph.D. thesis, Monash University.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Tiago Ribeiro dos Santos, Jorge Louçã, and Hélder Coelho. 2019. The digital transformation of the public sphere. *Systems Research and Behavioral Science*.

T Philipp Dybowski, Bernd Kempa, et al. 2019. The ecb's monetary pillar after the financial crisis. *Center for Quantitative Economics Working Papers*, (8519).

Han-Sung Eom, Sungyun Choi, and Sang Ok Choi. 2021. Marketable value estimation of patents using ensemble learning methodology: Focusing on u.s. patents for the electricity sector. *PLoS ONE*, 16.

Jiachun Feng, Zusheng Zhang, Cheng Ding, Yanghui Rao, Haoran Xie, and Fu Lee Wang. 2022. Context reinforced neural topic modeling over short texts. *Information Sciences*.

Raul Fernandez, Brenda Palma Guizar, and Caterina Rho. 2021. A sentiment-based risk indicator for the mexican financial sector. *Latin American Journal of Central Banking*, 2(3):100036.

Anthony A. Fung, Andy Zhou, Jennifer K. Vanos, and Geert W. Schmid-Schönbein. 2021. Enhanced intestinal permeability and intestinal co-morbidities in heat strain: A review and case for autodigestion. *Temperature: Multidisciplinary Biomedical Journal*, 8:223 – 244.

Shuang Gao, Shivani Pandya, Smisha Agarwal, and João Sedoc. 2021. Topic modeling for maternal health using Reddit. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 69–76, online. Association for Computational Linguistics.

Yoav Goldberg. 2020. Tweet from @yoavgo on 2020-12-19.

Derek Greene, Derek O?Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 498–513. Springer.

Tom Griffiths. 2002. Gibbs sampling in the generative model of latent dirichlet allocation. 518.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Itika Gupta, Barbara Di Eugenio, Devika Salunke, Andrew Boyd, Paula Allen-Meares, Carolyn Dickens, and Olga Garcia. 2020. Heart failure education of African American and Hispanic/Latino patients: Data collection and analysis. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 41–46, Online. Association for Computational Linguistics.

Ismail Harrando, Pasquale Lisena, and Raphael Troncy. 2021. Apples to apples: A systematic evaluation of topic models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 483–493, Held Online. INCOMA Ltd.

Jinbo He, Jingshi Kang, Shaojing Sun, Marita Cooper, Hana F Zickgraf, and Yujia Zhai. 2022. The landscape of eating disorders research: A 40-year bibliometric analysis. *European eating disorders review : the journal of the Eating Disorders Association*.

Daniel J. Hopkins, Eric Schickler, and David Azizi. 2020. From many divides, one? the polarization and nationalization of american state party platforms, 1918-2017. *Social Science Research Network*.

Keke Hou, Tingting Hou, and Lili Cai. 2021. Public attention about covid-19 on social media: An investigation based on data mining and text analysis. *Personality and Individual Differences*, 175:110701 – 110701.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving Neural Topic Models using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1752–1771, Online. Association for Computational Linguistics.

Tao Hu, Siqin Wang, Wei Luo, Mengxi Zhang, Xiao Huang, Yingwei Yan, Regina Liu, Kelly Ly, Viraj Kacker, Bing She, and Zhenlong Li. 2021. Revealing public opinion towards covid-19 vaccines with twitter data in the united states: Spatiotemporal perspective. *Journal of Medical Internet Research*, 23.

Hyeju Jang, Emily S. Rempel, David Roth, Giuseppe Carenini, and Naveed Zafar Janjua. 2021. Tracking covid-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of Medical Internet Research*, 23.

Min Jing, Raymond R. Bond, Louise J Robertson, Julie Moore, Amanda M Kowalczyk, Ruth K Price, William P Burns, M. Andrew Nesbit, James Mclaughlin, and Tara Moore. 2021. User experience analysis of abc-19 rapid test via lateral flow immunoassays for self-administrated sars-cov-2 antibody testing. *Scientific Reports*, 11.

Amy K Johnson, Runa Bhaumik, Debarghya Nandi, Abhishikta Roy, and Supriya D Mehta. 2022. Is this herpes or syphilis?: Latent dirichlet allocation analysis of sexually transmitted disease-related reddit posts during the covid-19 pandemic. *medRxiv*.

Leonard Kaufman and Peter J. Rousseeuw. 2008. *Partitioning Around Medoids (Program PAM)*, pages 68–125. John Wiley & Sons, Inc.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kilkon Ko, Hyun Hee Park, Dong Chul Shim, and Kyungdong Kim. 2021. The change of administrative capacity in korea: contemporary trends and lessons. *International Review of Administrative Sciences*, 87:238 – 255.

Sergei Koltcov, Olessia Koltsova, and Sergey Nikolenko. 2014. Latent dirichlet allocation: stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on Web science*, pages 161–165.

Damir Korenčić, Strahil Ristov, Jelena Repar, and Jan Šnajder. 2021. A topic coverage approach to evaluation of topic models. *IEEE Access*, 9:123280–123312.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Hyun kyu Park, Robert Phaal, Jae-Yun Ho, and Eoin O'Sullivan. 2020. Twenty years of technology and strategic roadmapping research: A school of thought perspective. *Technological Forecasting and Social Change*, 154:119965.

Tahleen A. Lattimer, Kelly E. Tenzek, Yotam Ophir, and Suzanne S. Sullivan. 2022. Exploring web-based twitter conversations surrounding national healthcare decisions day and advance care planning from a sociocultural perspective: Computational mixed methods analysis. *JMIR Formative Research*, 6.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada. Association for Computational Linguistics.

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.

Junho Lee, Ikjun Kim, Hyomin Kim, and Juyoung Kang. 2021a. Swot-ahp analysis of the korean satellite and space industry: Strategy recommendations for development. *Technological Forecasting and Social Change*.

Sae-Mi Lee and Soongoo Hong. 2020. Policy agenda proposals from text mining analysis of patents and news articles. *Journal of Digital Convergence*, 18:1–12.

Yoon Kyung Lee, Yoonwon Jung, Inju Lee, Jae Eun Park, and Sowon Hahn. 2021b. Building a psychological ground truth dataset with empathy and theory-of-mind during the covid-19 pandemic. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Kai Li, Xing Liu, Feng Mai, and Tengfei Zhang. 2021. The role of corporate culture in bad times: Evidence from the covid-19 pandemic. *Journal of Financial and Quantitative Analysis*, 56:2545 – 2583.

Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos Kleijnen, and David Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10):e1–e34.

SP Lloyd. 1957. Least square quantization in pcm. bell telephone laboratories paper. published in journal much later: Lloyd, sp: Least squares quantization in pcm. *IEEE Trans. Inform. Theor.(1957/1982)*, 18:5.

Peter Madzík and Lukas Falat. 2022. State-of-the-art on analytic hierarchy process in the last 40 years: Literature review based on latent dirichlet allocation topic modelling. *PLoS ONE*, 17.

Harry Mamaysky. 2020. News and markets in the time of covid-19. *Econometric Modeling: Capital Markets - Asset Pricing eJournal*.

Mika V Mantyla, Maelick Claes, and Umar Farooq. 2018. Measuring LDA topic stability from clusters of replicated runs. In *Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement*, pages 1–4.

Brittany R Markides, Rachel Laws, Kylie D. Hesketh, Ralph Maddison, Elizabeth A Denney-Wilson, and Karen Jane Campbell. 2022. A thematic cluster analysis of parents' online discussions about fussy eating. *Maternal & Child Nutrition*, 18.

Pablo Marshall. 2021. Contribution of open-ended questions in student evaluation of teaching. *Higher Education Research & Development*.

Andrew Kachites McCallum. 2002. Machine learning with MALLET. `http://mallet.cs.umass.edu`.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. *ICLR*.

Daniel Taninecz Miller. 2019. *Three International Studies Computational Social Science Inquiries Examining Large Corpora of Natural Data*. Ph.D. thesis.

John Miller and Kathleen McCoy. 2017. Topic model stability for hierarchical summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 64–73, Copenhagen, Denmark. Association for Computational Linguistics.

David Mimno. 2022. Why I don't recommend stochastic variational Bayes for topic models. Blog post. Accessed October 24, 2022.

Andrew Mitchell. 2020. Mode-2 knowledge production within community-based sustainability projects: Applying textual and thematic analytics to action research conversations. *Administrative Sciences*, 10:90.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, 151(4):264–269.

Aaron Mueller and Mark Dredze. 2021. Fine-tuning encoders for improved monolingual and zero-shot polylingual neural topic modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3054–3068, Online. Association for Computational Linguistics.

Federico Neresini, Paolo Giardullo, Emanuele Di Buccio, and Alberto Cammozzo. 2019. Exploring sociotechnical future scenarios in the media: the energy transition case in italian daily newspapers. *Quality & Quantity*, 54:147–168.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

Reuben Ng and Nicole Indran. 2021. Role-based framing of older adults linked to decreased ageism over 210 years: Evidence from a 600-million-word historical corpus. *The Gerontologist*, 62:589 – 597.

Li Ning, Peng Lifang, and He Huixin. 2020. Prediction correction topic evolution research for metabolic pathways of the gut microbiota. *Frontiers in Molecular Biosciences*, 7.

Peter-John Mäntylä Noble, Charlotte Appleton, Alan D. Radford, and Goran Nenadic. 2021. Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in uk dogs. *PLoS ONE*, 16.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Erik L. Peterson and Crystal L. Hall. 2020. "what is dead may not die": Locating marginalized concepts among ordinary biologists. *Journal of the history of biology*.

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. ALTO: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1158–1169, Berlin, Germany. Association for Computational Linguistics.

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.

Doris McGartland Rubio. 2005. *Content validity*. Elsevier.

William J. Scarborough and Rowena Crabbe. 2021. Place brands across u.s. cities and growth in local high-technology sectors. *Journal of Business Research*, 130:70–85.

Margrit Schreier. 2012. *Qualitative content analysis in practice*. Sage publications.

G Shapiro and J Markoff. 1997. *A Matter of Definition*. Mahwah: Lawrence Erlbaum Assoc.

Jinnie Shin, Qi Guo, and Mark Gierl. 2019. Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology*, 10.

Charles P Smith. 2000. Content analysis and narrative analysis. *Handbook of research methods in social and personality psychology*, 2000:313–335.

Allison Patricia Squires, Maya N. Clark-Cutaia, Marcus D. Henderson, Gavin Arneson, and Philip Resnik. 2022. "should i stay or should i go?" nurses' perspectives about working during the covid-19 pandemic's first wave in the united states: A summative content analysis combined with topic modeling. *International Journal of Nursing Studies*, 131:104256 – 104256.

Douglas Steinley. 2004. Properties of the hubert-arabie adjusted rand index. *Psychological methods*, 9:386–96.

Steve Stemler. 2000. An overview of content analysis. *Practical assessment, research, and evaluation*, 7(1):17.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Marlon Santiago Viñán-Ludeña and Luis M de Campos. 2021. Analyzing tourist data on twitter: a case study in the province of granada at spain. *Journal of Hospitality and Tourism Insights*.

Seppo Virtanen. 2021. Uncovering dynamic textual topics that explain crime. *Royal Society Open Science*, 8.

Chihuangji Wang, Edward S. Steinfeld, Jordana L. Maisel, and Bumjoon Kang. 2021. Is your smart city inclusive? evaluating proposals from the u.s. department of transportation's smart city challenge. *Sustainable Cities and Society*, 74:103148.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. 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 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online. Association for Computational Linguistics.

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

Robert Philip Weber. 1990. *Basic content analysis*, volume 49. Sage.

Lino Wehrheim, Tobias Alexander Jopp, and Mark Spoerer. 2021. Turn, turn, turn: A digital history of german historiography, 1950-2019. Technical report,

Working Papers of the Priority Programme 1859" Experience and Expectation ....

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Learning multilingual topics with neural variational inference. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 840–851. Springer.

Taizo Yamada and Satoshi Inoue. 2019. Detection and time series variation of latent topic from diary in northern and southern courts period of japan. *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, pages 1–8.

Tian Yang and Kecheng Fang. 2021. How dark corners collude: a study on an online chinese alt-right community. *Information, Communication & Society*.

Qian Ye, Xiaohong Chen, Kaan Ozbay, and Yonggang Wang. 2020. How people view and respond to special events in shared mobility: Case study of two didi safety incidents via sina weibo. In *20th COTA International Conference of Transportation Professionals: Advanced Transportation Technologies and Development-Enhancing Connections, CICTP 2020*, pages 3229–3240. American Society of Civil Engineers (ASCE).

Faxi Yuan, Min Li, and Rui Liu. 2020. Understanding the evolutions of public responses using social media: Hurricane matthew case study. *International journal of disaster risk reduction*, 51:101798.

Yipeng Zhang, Hanjia Lyu, Yubao Liu, Xiyang Zhang, Yu Wang, and Jiebo Luo. 2021. Monitoring depression trends on twitter during the covid-19 pandemic: Observational study. *JMIR Infodemiology*, 1.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Ying Zhao and George Karypis. 2002. Criterion functions for document clustering: Experiments and analysis.

Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. 2015. A deep and autoregressive approach for topic modeling of multimodal data. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1056–1069.

A Appendix

A.1 Additional Results

Fixed Hyperparameters. In tables 4 and 5, we report the equivalents of tables 2 and 3 when holding hyperparameters *fixed*, rather than letting them vary. We identify the hyperparameters for each model that achieve the highest average alignment metrics on across experimental contexts for one dataset, then use those hyperparameters to estimate models on the other dataset (hyperparameter values in appendix A.3). In this way, we follow a common paradigm in practical application of machine learning models: hyperparameters are determined based on an initial experimental context, then used in another. Broadly, MALLET is more stable and better-aligned than its neural counterparts in this setup, although the difference is not as stark as when hyperparameters are allowed to vary.

Held-out data. In table 6, we report the alignment metrics for unseen category labels. To form the held-out data, we keep all high-level categories consistent between the training and held-out sets, but partition the low-level categories such that some are never seen during training (e.g., although documents from the high-level `architecture` category will be included in both splits, documents on `bridges` are only seen in training while those on `lighthouses` are held-out). Here too, MALLET generally has the highest alignment metrics over experimental contexts.

A.2 Details of LDA applications meta-analysis

Summary statistics of our meta-analysis of studies using LDA outside computer science are shown in Table 7. The major results were discussed in Section 2.2. We find that about half of the papers did not specify the exact LDA implementation they used in their study, which raises larger reproducibility concerns for scientific research. Note that one paper can be assigned multiple subject or fields of study by Semantic Scholar. All the papers used for the meta-analysis are shown in tables 9 and 10.

A.3 Hyperparameters

Hyperparameters are included in the supplementary materials as `<model name>.yml` files. The full range of hyperparameters can also be found in Table 11.

		$\lvert V\rvert = 5$k								$\lvert V\rvert = 15$k							
		$k = 25$		$k = 50$		$k = 100$		$k = 200$		$k = 25$		$k = 50$		$k = 100$		$k = 200$	
		Θ	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**	**Θ**	**B**
Bills	MALLET	0.76	**0.26**	0.75	**0.29**	0.73	**0.30**	0.70	0.36	**0.73**	**0.26**	0.74	**0.30**	0.72	0.34	**0.70**	0.37
	SCHOLAR	**0.72**	0.42	**0.67**	0.39	**0.67**	0.38	**0.68**	0.38	0.78	0.52	**0.73**	0.47	**0.70**	0.44	0.71	0.43
	SCHLR+KD	0.84	0.43	0.78	0.38	0.70	0.31	0.69	**0.29**	0.83	0.46	0.80	0.38	0.74	**0.31**	0.72	**0.29**
	D-VAE	0.91	0.34	0.86	0.54	0.85	0.68	0.74	0.81	0.95	0.40	0.90	0.57	0.90	0.71	0.91	0.79
	CTM	0.78	0.43	0.77	0.47	0.73	0.51	0.70	0.52	0.80	0.44	0.81	0.53	0.79	0.60	0.73	0.62
Wiki	MALLET	**0.70**	**0.22**	**0.60**	0.26	**0.56**	0.28	0.54	0.31	**0.69**	**0.26**	**0.62**	**0.29**	**0.55**	**0.29**	**0.50**	**0.31**
	SCHOLAR	0.77	0.39	0.66	0.31	0.62	0.33	0.59	0.32	0.75	0.45	0.69	0.38	0.65	0.38	0.60	0.37
	SCHLR+KD	0.77	0.41	0.66	0.33	0.62	0.33	**0.53**	**0.28**	0.78	0.52	0.70	0.44	0.62	0.38	0.54	0.34
	D-VAE	0.93	0.24	0.88	**0.25**	0.83	**0.26**	0.82	0.29	0.95	0.28	0.90	0.30	0.83	0.32	0.78	0.33
	CTM	0.75	0.39	0.73	0.39	0.72	0.45	0.70	0.51	0.77	0.42	0.74	0.41	0.75	0.51	0.73	0.57

Table 4: Stability for topic-word **B** and document-topic **Θ** estimates, across 10 seeds, for *fixed hyperparameters*. Smallest per-column values are **bolded** and are sig. smaller than unbolded values (2-sided t-test, $p < 0.05$); underlined values have $p > 0.05$.

		$k = 25$			$k = 50$			$k = 100$			$k = 200$		
		ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI	P_1
Bills $\ell = 114$	MALLET	**0.30**	**0.46**	**0.47**	**0.35**	**0.49**	**0.48**	**0.34**	0.51	**0.45**	**0.22**	0.51	0.37
	SCHOLAR	0.25	0.45	0.42	0.25	0.48	0.43	0.21	**0.51**	0.40	0.15	**0.52**	0.34
	SCHLR+KD	0.24	0.42	0.40	0.23	0.46	0.43	0.20	0.49	0.40	0.13	0.49	0.34
	D-VAE	0.26	0.45	0.45	0.21	0.45	0.45	0.10	0.43	0.42	0.04	0.39	**0.41**
	CTM	0.23	0.40	0.39	0.27	0.46	0.43	0.24	0.48	0.40	0.18	0.49	0.34
Wiki $\ell = 279$	MALLET	0.23	0.65	0.41	0.32	0.70	**0.50**	0.39	**0.73**	**0.56**	**0.39**	0.74	**0.56**
	SCHOLAR	0.22	0.62	0.39	**0.33**	0.68	0.49	**0.39**	0.72	0.54	0.38	**0.74**	0.54
	SCHLR+KD	0.20	0.61	0.37	0.30	0.67	0.47	0.39	0.71	0.53	0.39	0.74	0.54
	D-VAE	**0.24**	**0.66**	**0.41**	0.32	**0.70**	0.50	0.36	0.72	0.54	0.36	0.72	0.54
	CTM	0.21	0.60	0.36	0.28	0.65	0.44	0.32	0.67	0.47	0.35	0.70	0.48

Table 5: Average alignment metrics across 10 seeds, for *fixed hyperparameters*. Measured against gold labels at the lowest hierarchy level, $\lvert V\rvert = 15,000$. Largest values in each column are **bolded**, which are significantly greater than unbolded values in a two-sided t-test ($p < 0.05$); underlined values have $p > 0.05$.

A.4 Additional alignment results

Results summarized in Table 3 are shown in figure 2. Results for rest of the settings for vocabulary and label hierarchy level are shown in figs. 3 to 5.

A.5 Additional ensembling results

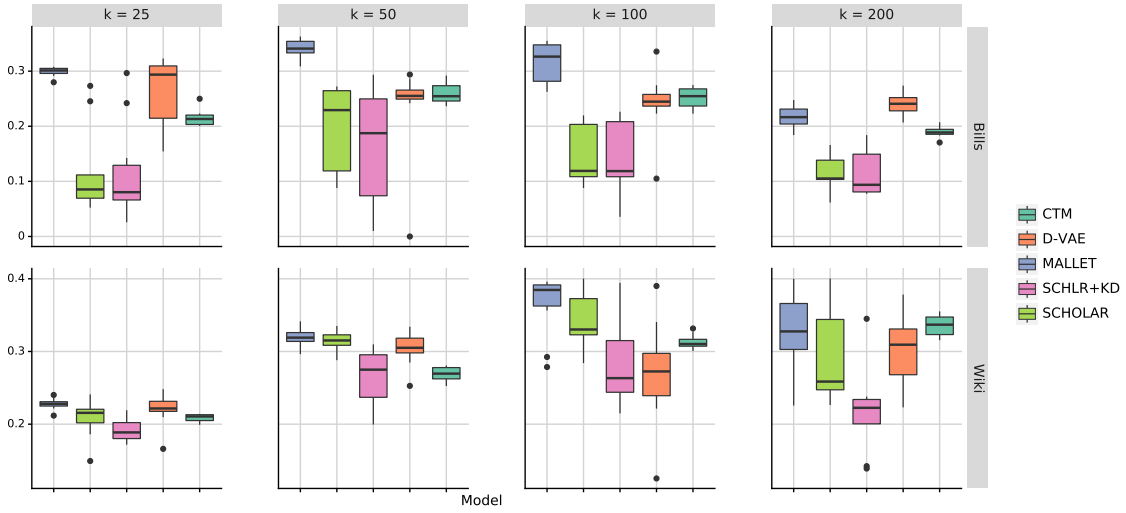
In Table 8, we list the best-performing ensemble per model type, alongside a method that fares well across all models. For two out of the five models, the ensemble outperforms the median member in all 40 settings. Most ensembles improve upon the best member at least half the time. We also identify a set of hyperparameters (distance metric, λ, and clustering algorithm) that can ensemble the results of any of our models (*Overall* row in Table 8).

A.6 Compute infrastructure

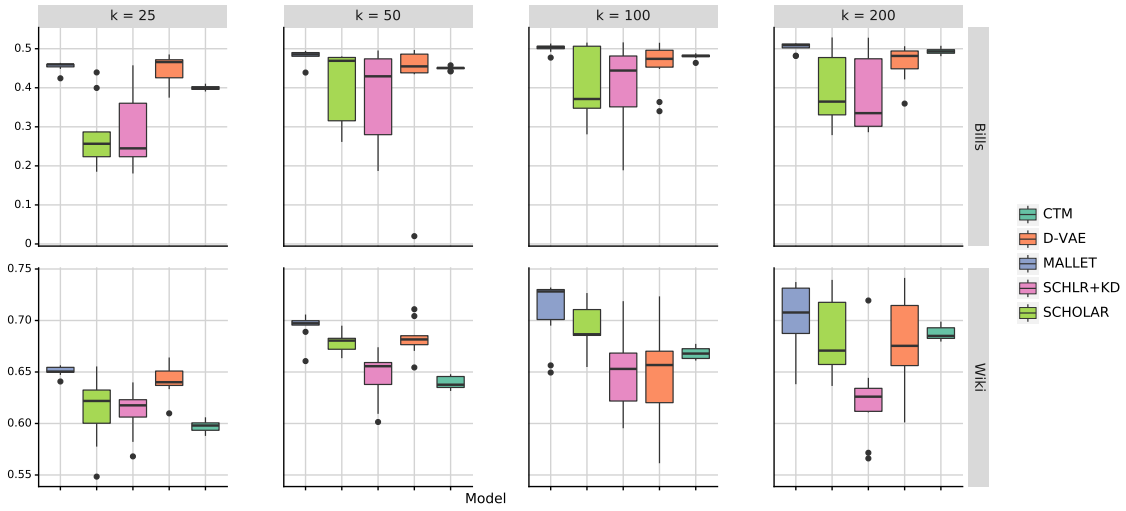
We used AWS ParallelCluster to provide a cloud-computing computing cluster. Neural topic models ran on NVIDIA T4 GPUs using `g4dn.xlarge` instances with 16 GiB memory and 4 CPUs.[15] MALLET ran on CPU only, with `m5d.2xlarge` instances (with 32 GiB memory, 8 CPUs).[16]

[15] https://aws.amazon.com/hpc/parallelcluster/

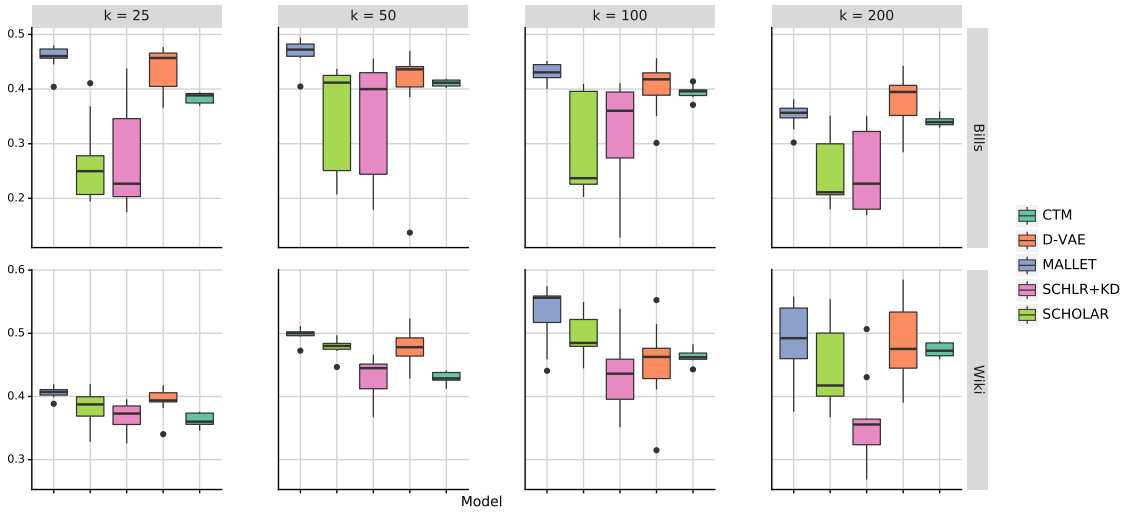
[16] See https://aws.amazon.com/ec2/instance-types/ for further details.



(a) Alignment metric = ARI

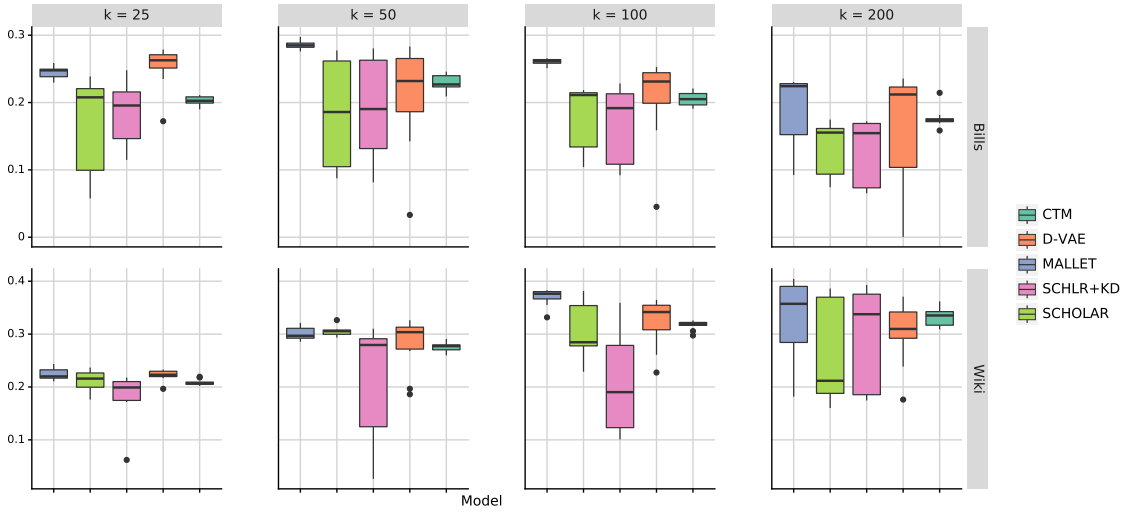


(b) Alignment metric = NMI

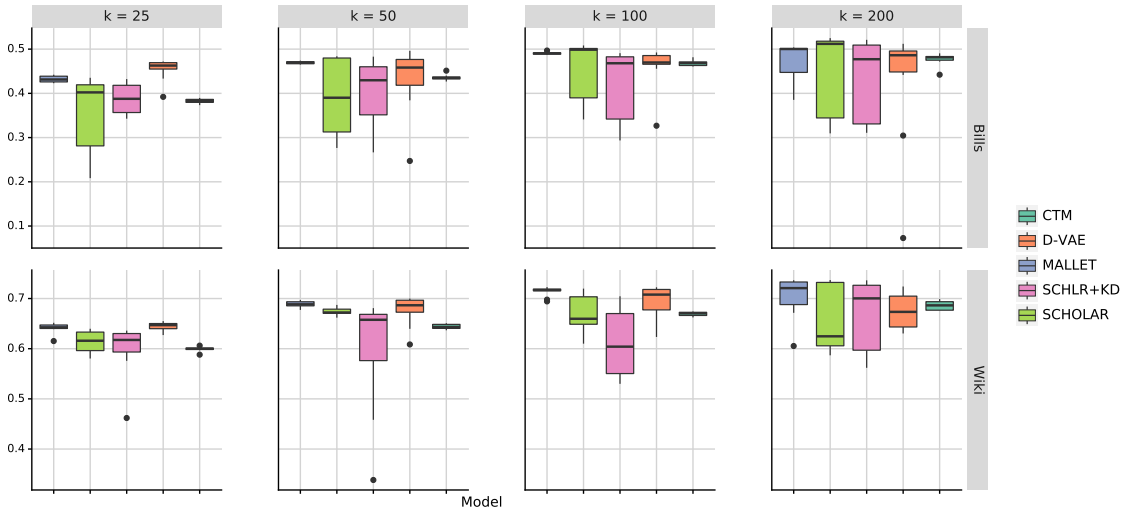


(c) Alignment metric = P_1

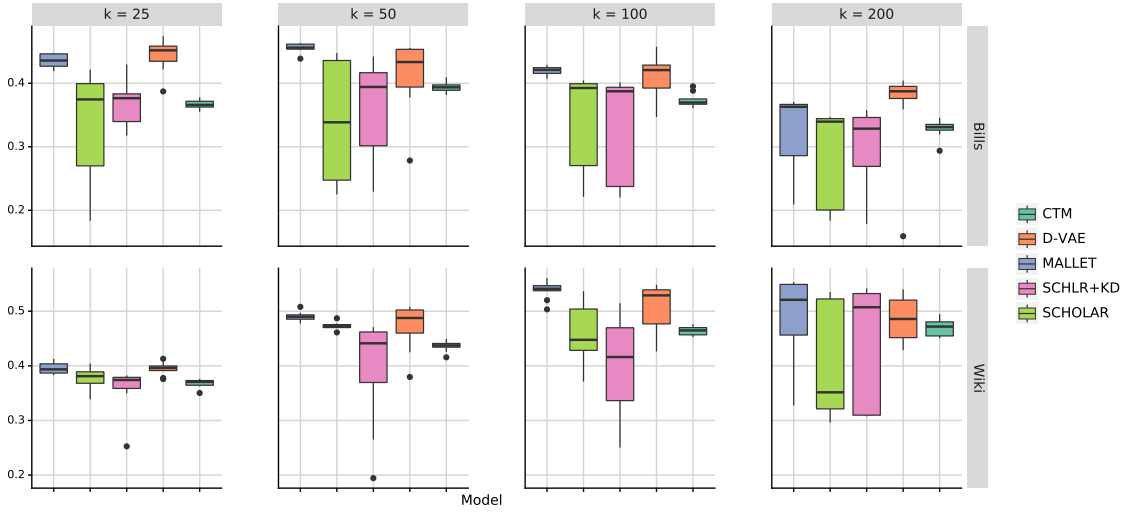
Figure 2: Training set alignment metrics across five models, measured against gold labels at the lowest hierarchy level, $|V| = 15,000$.



(a) Alignment metric = ARI

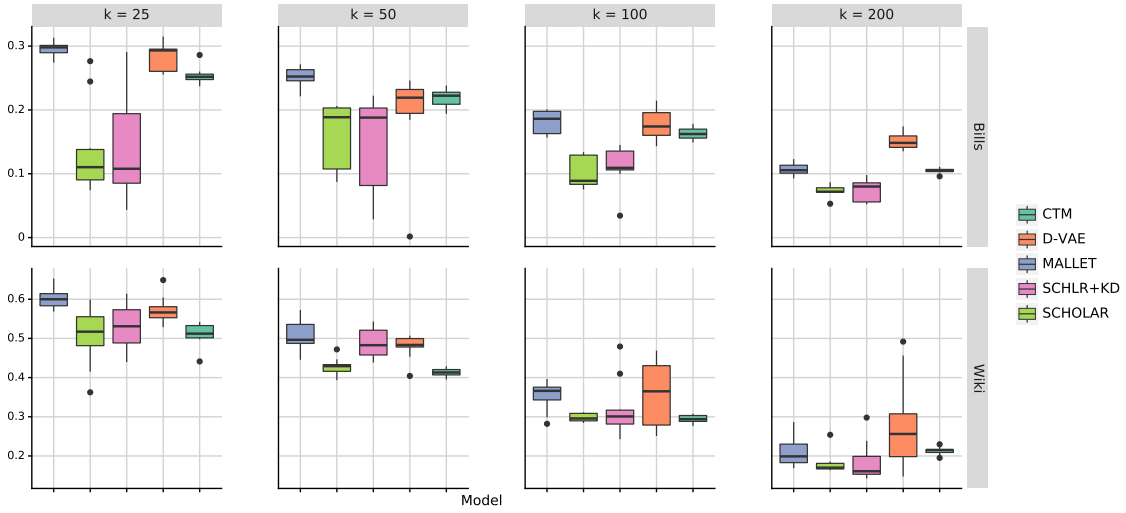


(b) Alignment metric = NMI

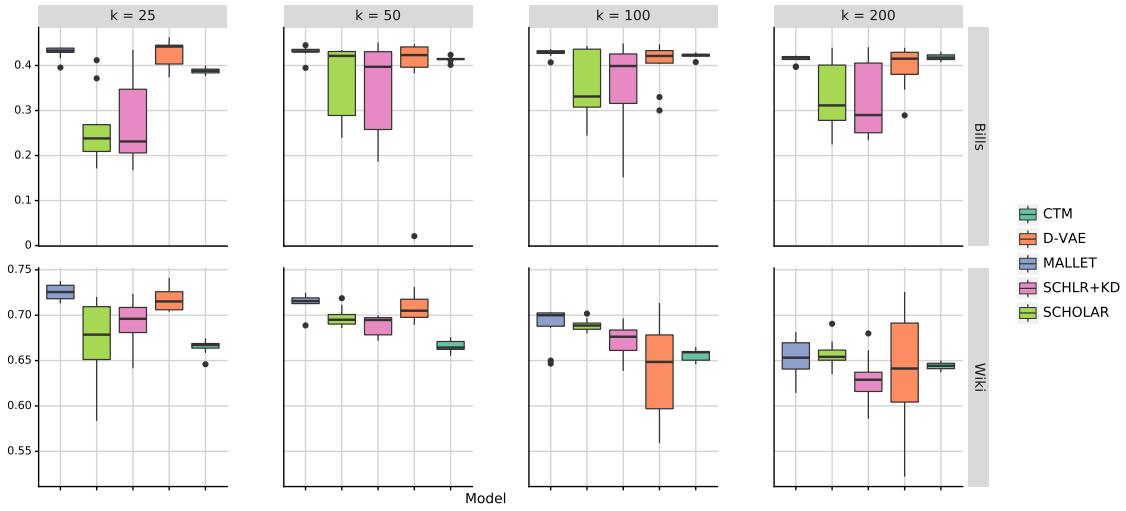


(c) Alignment metric = P_1

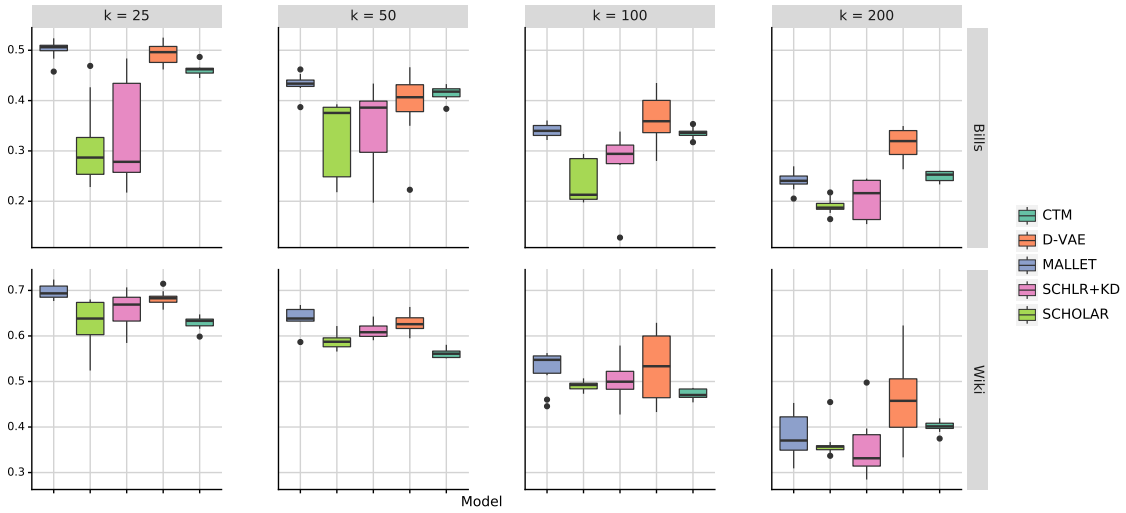
Figure 3: Training set alignment metrics across five models, measured against gold labels at the lowest hierarchy level, $|V| = 5,000$.



(a) Alignment metric = ARI

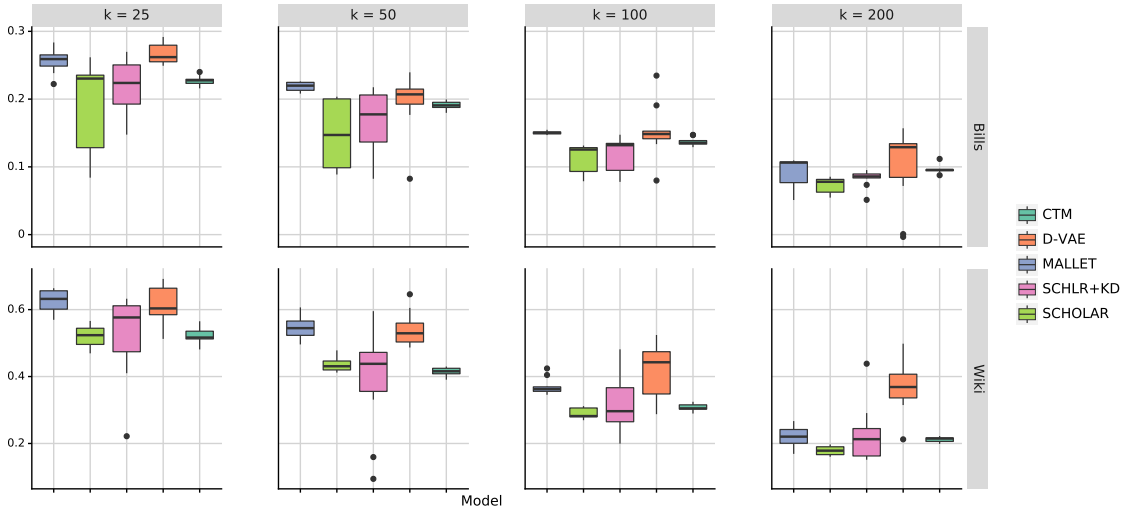


(b) Alignment metric = NMI

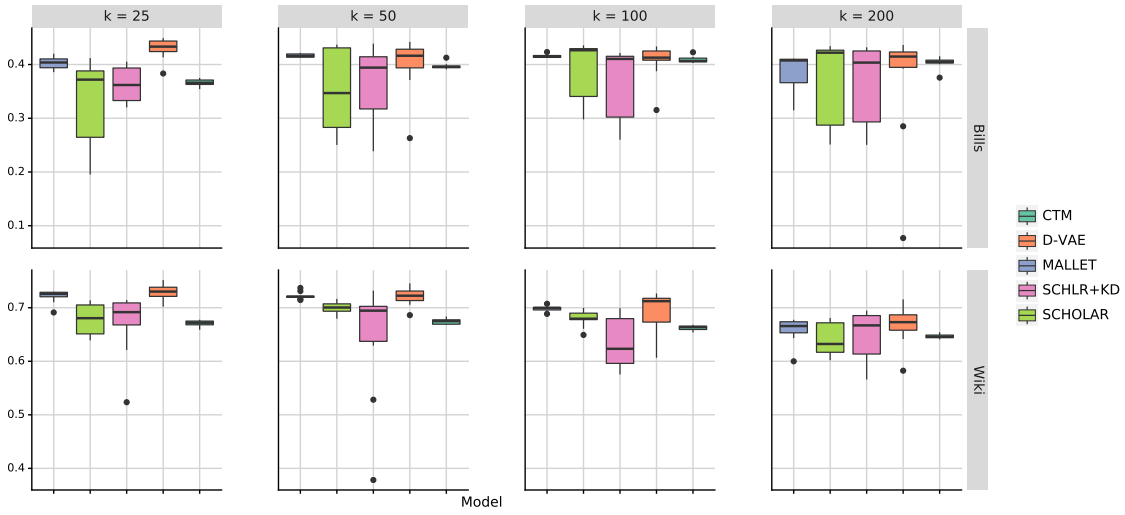


(c) Alignment metric = P_1

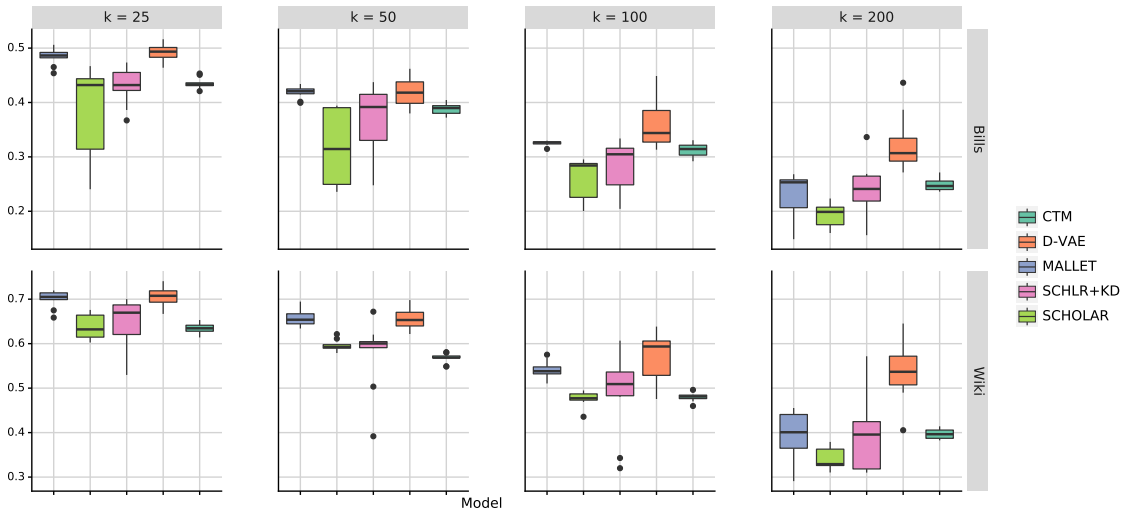
Figure 4: Training set alignment metrics across five models, measured against gold labels at the highest hierarchy level, $|V| = 15,000$.



(a) Alignment metric = ARI



(b) Alignment metric = NMI



(c) Alignment metric = P_1

Figure 5: Training set alignment metrics across five models, measured against gold labels at the highest hierarchy level, $|V| = 5,000$.

		$k = 25$			$k = 50$			$k = 100$			$k = 200$		
		ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI	P_1
Bills $\ell = 114$	MALLET	**0.17**	**0.34**	**0.37**	**0.17**	**0.37**	**0.37**	**0.17**	**0.40**	**0.37**	**0.15**	**0.41**	**0.35**
	SCHOLAR	0.06	0.18	0.22	0.09	0.27	0.26	0.07	0.27	0.23	0.05	0.26	0.19
	SCHLR+KD	0.06	0.19	0.24	0.07	0.25	0.27	0.06	0.27	0.26	0.04	0.26	0.19
	D-VAE	0.13	0.30	0.34	0.11	0.29	0.32	0.11	0.32	0.32	0.10	0.34	0.31
	CTM	0.11	0.28	0.31	0.12	0.32	0.31	0.10	0.35	0.29	0.08	0.37	0.26
Wiki $\ell = 279$	MALLET	**0.34**	**0.65**	**0.53**	**0.37**	**0.66**	**0.53**	**0.37**	**0.68**	**0.53**	**0.33**	**0.68**	**0.51**
	SCHOLAR	0.29	0.61	0.48	0.32	0.64	0.51	0.30	0.63	0.47	0.24	0.61	0.41
	SCHLR+KD	0.28	0.61	0.48	0.31	0.63	0.49	0.28	0.63	0.46	0.19	0.59	0.35
	D-VAE	0.32	0.64	0.51	0.35	0.66	0.52	0.28	0.62	0.48	0.31	0.64	0.48
	CTM	0.32	0.61	0.49	0.32	0.63	0.48	0.29	0.64	0.46	0.27	0.65	0.45

(a) Random hyperparameters

		$k = 25$			$k = 50$			$k = 100$			$k = 200$		
		ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI	P_1
Bills $\ell = 114$	MALLET	**0.17**	**0.34**	**0.37**	**0.17**	**0.37**	**0.38**	**0.17**	**0.40**	**0.38**	**0.15**	**0.42**	**0.36**
	SCHOLAR	0.14	0.32	0.34	0.13	0.34	0.34	0.14	0.38	0.34	0.12	0.40	0.32
	SCHLR+KD	0.10	0.27	0.31	0.09	0.29	0.32	0.08	0.33	0.30	0.06	0.32	0.31
	D-VAE	0.11	0.27	0.32	0.05	0.26	0.31	0.03	0.25	0.32	0.01	0.21	0.34
	CTM	0.11	0.28	0.31	0.12	0.32	0.31	0.10	0.34	0.29	0.07	0.36	0.26
Wiki $\ell = 279$	MALLET	0.33	0.65	0.52	0.37	0.66	0.54	**0.38**	0.68	**0.54**	**0.34**	0.69	**0.52**
	SCHOLAR	0.31	0.62	0.50	0.33	0.64	0.50	0.34	0.67	0.51	0.32	0.68	0.49
	SCHLR+KD	0.30	0.63	0.49	**0.38**	**0.67**	**0.54**	0.37	**0.69**	0.52	0.34	**0.70**	0.51
	D-VAE	**0.34**	**0.65**	**0.52**	0.37	0.67	0.54	0.36	0.67	0.52	0.32	0.66	0.49
	CTM	0.31	0.61	0.48	0.32	0.63	0.49	0.30	0.65	0.47	0.28	0.66	0.46

(b) Fixed hyperparameters

Table 6: Average alignment metrics across 10 runs *for unseen category labels*. Measured against gold labels at the lowest hierarchy level, $|V| = 15,000$. Largest values in each column are **bolded**, which are significantly greater than unbolded values in a two-sided t-test ($p < 0.05$); underlined values have $p > 0.05$.

Coding Question	Answer	Count	
Is LDA used for inductive discovery of categories for human consumption?	Yes	47	(94%)
Which LDA estimates were used for the above discovery?	B only	33	(70%)
	Both B and Θ	11	(23%)
Is LDA used to categorize or represent individual units of text (using Θ estimate)?	Yes	32	(64%)
Are human-readable code labels assigned to topics (formal *content analysis*)?	Yes	34	(68%)
Is the exact LDA implementation specified?	Yes	27	(54%)
Field of study	Medicine	21	(42%)
	Sociology	9	(18%)
	Business	7	(14%)
	Political Science	7	(14%)
	Psychology	4	(8%)
	Economics	2	(4%)
	History	2	(4%)

Table 7: Meta-analysis of fifty topic modeling papers outside the field of computer science (denominator may change, as not all conditions are always applicable). Content analysis is the dominant use case for topic models. The reliability, validity, and reproducibility of LDA estimates is critical to this use-case.

	Algo.	d	λ	> Worst	Med.	Best
Overall	k-med.	RBO	1.00	100%	97%	52%
MALLET	k-med.	RBO	1.00	100%	98%	55%
SCHOLAR	k-med.	Jcd.	0.25	100%	100%	66%
SCHLR+KD	Aggl.	RBO	0.75	100%	99%	60%
D-VAE	Aggl.	Jcd.	0.25	100%	92%	29%
CTM	Aggl.	Jcd.	0.25	100%	100%	90%

Table 8: Alignment metrics for ensembles of each model, and how often they improve over the worst, median, and best member of the ensemble across 80 evaluation settings.

Table 9: Part one of all papers and their assessment for the meta-analysis of LDA use (Section 2.2).

Source	Field of Study	Is LDA used for inductive discovery of categories for human consumption?	Which LDA estimates were used for the inductive discovery?	Is LDA used to categorize or represent individual units of text (using Θ estimate)?	Are human-readable code labels assigned to topics (formal content analysis)?	Is the exact LDA implementation specified?
(dos Santos et al., 2019)	Political Science	N	N/A	Y	N	N
(Markides et al., 2022)	Medicine	Y	Both	Y	Y	Y
(Miller, 2019)	Political Science	Y	Both	N	N	Y
(Scarborough and Crabbe, 2021)	Business	Y	Topic-Word Only	Y	N	N
(Jang et al., 2021)	Medicine	Y	Topic-Word Only	Y	Y	Y
(Noble et al., 2021)	Medicine	Y	Topic-Word Only	Y	Y	Y
(Yamada and Inoue, 2019)	History	Y	Both	Y	N	Y
(Zhang et al., 2021)	Medicine	Y	Topic-Word Only	Y	Y	N
(Lee et al., 2021b)	Medicine	Y	Topic-Word Only	N	Y	N
(Lattimer et al., 2022)	Medicine	Y	Both	Y	Y	N
(Madzík and Falat, 2022)	Medicine	Y	Topic-Word Only	Y	Y	Y
(Wehrheim et al., 2021)	History	Y	Topic-Word Only	Y	Y	Y
(Marshall, 2021)	Psychology	Y	Topic-Word Only	N	Y	N
(Hopkins et al., 2020)	Political Science	Y	Topic-Word Only	Y	Y	N
(Ning et al., 2020)	Medicine	Y	Topic-Word Only	Y	Y	N
(Aniss et al., 2021)	Psychology	Y	Topic-Word Only	N	Y	Y
(Neresini et al., 2019)	Sociology	Y	Topic-Word Only	N	Y	N
(Coco et al., 2020)	Economics	Y	Topic-Word Only	Y	N	N
(Almquist and Bagozzi, 2019)	Sociology	Y	Both	Y	Y	Y
(Lee and Hong, 2020)	Political Science	Y	Topic-Word Only	N	Y	Y
(Li et al., 2021)	Business	Y	Both	Y	Y	Y
(Hou et al., 2021)	Medicine	Y	Topic-Word Only	Y	Y	N
(He et al., 2022)	Medicine	Y	Topic-Word Only	Y	Y	Y
(Wang et al., 2021)	Business	Y	Topic-Word Only	N	N	Y
(Dybowski et al., 2019)	Economics	Y	Both	Y	N	Y

Table 10: Part two of all papers and their assessment for the meta-analysis of LDA use (Section 2.2).

Source	Field of Study	Is LDA used for inductive discovery of categories for human consumption?	Which LDA estimates were used for the inductive discovery?	Is LDA used to categorize or represent individual units of text (using Θ estimate)?	Are human-readable code labels assigned to topics (formal content analysis)?	Is the exact LDA implementation specified?
(Gupta et al., 2020)	Psychology	Y	Topic-Word Only	N	Y	Y
(Amozegar, 2021)	Business	Y	Topic-Word Only	Y	Y	Y
(Bell and Scott, 2020)	Business	Y	Topic-Word Only	Y	N	N
(Viñán-Ludeña and de Campos, 2021)	Sociology	Y	Topic-Word Only	N	N	Y
(Yang and Fang, 2021)	Sociology	Y	Both	N	Y	N
(Das et al., 2020)	Political Science	Y	Both	Y	Y	Y
(Virtanen, 2021)	Medicine	Y	Topic-Word Only	Y	Y	Y
(Ye et al., 2020)	Sociology	Y	Topic-Word Only	Y	Y	N
(Jing et al., 2021)	Medicine	Y	Topic-Word Only	N	N	N
(Peterson and Hall, 2020)	Medicine,Sociology	N	N/A	Y	N	Y
(Ko et al., 2021)	Political Science	Y	Doc-Topic Only	Y	Y	N
(Mitchell, 2020)	Sociology	Y	Topic-Word Only	N	N	N
(Fernandez et al., 2021)	Business	Y	Both	N	Y	N
(Bateman et al., 2019)	Political Science	Y	Doc-Topic Only	Y	Y	Y
(Fung et al., 2021)	Medicine	Y	Topic-Word Only	Y	Y	Y
(Yuan et al., 2020)	Sociology	Y	Topic-Word Only	Y	Y	N
(Eom et al., 2021)	Medicine	Y	Topic-Word Only	N	N	Y
(Johnson et al., 2022)	Medicine	Y	Topic-Word Only	Y	Y	N
(Squires et al., 2022)	Medicine	N	N/A	N	N	N
(Hu et al., 2021)	Medicine	Y	Topic-Word Only	Y	Y	N
(Ng and Indran, 2021)	Medicine	Y	Topic-Word Only	N	N	N
(Lee et al., 2021a)	Business	Y	Topic-Word Only	N	Y	Y
(Mamaysky, 2020)	Medicine	Y	Topic-Word Only	Y	Y	Y
(Shin et al., 2019)	Medicine, Psychology	Y	Both	N	N	Y
(kyu Park et al., 2020)	Sociology	Y	Doc-Topic Only	Y	Y	Y

Model: MALLET			
α	β	Optim. Interval	#Steps
$\{0.01, 0.05, 0.1, 0.25, 1.0^{*\dagger}\}$	$\{0.01, 0.05^{*}, 0.1^{\dagger}\}$	$\{0, 10^{*\dagger}, 500\}$	$\{2000\}$

(a) Hyperparameter ranges for MALLET. α is the topic density parameter. β is the word density parameter. Optim. Interval sets the number of iterations between Mallet's own internal hyperparameter updates. #Steps are training iterations.

Model: SCHOLAR			
α	η	η_{BN}	#Steps
$\{0.001, 0.005, 0.01, 0.5^{\dagger}, 1.0^{*}\}$	$\{0.001^{*\dagger}, 0.002\}$	$\{0.25^{*}, 0.5^{\dagger}, 0.75\}$	$\{200^{\dagger}, 500^{*}\}$

(b) Hyperparameter ranges for SCHOLAR. α is the Dirichlet prior. η is the learning rate. η_{BN} is the epoch when batch-norm annealing ends (i.e., $\eta \times$ Steps). #Steps are training epochs.

Model: SCHLR+KD						
α	η	η_{BN}	clipping	T	λ	#Steps
$\{0.001, 0.005, 0.01, 0.5^{\dagger}, 1.0^{*}\}$	$\{0.001^{*}, 0.002^{\dagger}\}$	$\{0.25, 0.5^{*}, 0.75^{\dagger}\}$	$\{0.0^{*}, 1.0^{\dagger}, 2.0\}$	$\{1.0^{\dagger}, 2.0^{*}\}$	$\{0.5^{*}, 0.75, 0.99^{\dagger}\}$	$\{200^{\dagger}, 500^{*}\}$

(c) Hyperparameter ranges for SCHLR+KD. α is the Dirichlet prior. η is the learning rate. η_{BN} is the epoch when batch-norm annealing ends. λ is weight on the teacher model logits, T is the softmax temperature, and clipping controls how much of the logit distribution to clip. #Steps are training epochs.

Model: D-VAE					
α	η	$\beta_{reg.}$	γ_{BN}	γ_{KL}	#Steps
$\{0.001^{\dagger}, 0.01^{*}, 0.1\}$	$\{0.001^{\dagger}, 0.01^{*}\}$	$\{0.0, 0.01, 0.1^{*}, 1.0^{\dagger}\}$	$\{0, 1, 100^{*}, 200^{\dagger}\}$	$\{100^{*\dagger}, 200\}$	$\{500\}$

(d) Hyperparameter ranges for D-VAE. α is the Dirichlet prior. η is the learning rate. $\beta_{reg.}$ is the L_1-regularization of the topic-word distribution. γ_{BN} and γ_{KL} are the number of epochs to anneal the batch normalization constant and KL divergence term in the loss, respectively. #Steps are training epochs.

Model: CTM			
$e(\cdot)$	Learn Priors?	γ_{η}	#Steps
{ `paraphrase-distilroberta-base-v2`, `multi-qa-mpnet-base-dot-v1`*†, `all-mpnet-base-v2` }	{False†, True*}	$\{0.001^{*}, 0.002\}$	$\{100, 200^{*\dagger}\}$

(e) Hyperparameter ranges for CTM. $e(\cdot)$ is the Sentence Transformers document-embedding model(Reimers and Gurevych, 2019). η is the learning rate. W_{decay} is the L_2 regularization constant. γ_{η} is an indicator of whether learning rate is annealed. #Steps are training epochs.

Table 11: Hyperparameter settings for MALLET, D-VAE, CTM, SCHLR+KD and SCHOLAR. $*$: Best setting for `Bills`, †: best setting for `Wiki`; based on the best average alignment metrics across experimental contexts.