

# Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures

**Caitlin Doogan**

Faculty of Information Technology  
Monash University, Australia  
caitlin.doogan@monash.edu

**Wray Buntine**

Faculty of Information Technology  
Monash University, Australia  
wray.buntine@monash.edu

## Abstract

When developing topic models, a critical question that should be asked is: *How well will this model work in an applied setting?* Because standard performance evaluation of topic interpretability uses automated measures modeled on human evaluation tests that are dissimilar to applied usage, these models' generalizability remains in question. In this paper, we probe the issue of validity in topic model evaluation and assess how informative coherence measures are for specialized collections used in an applied setting. Informed by the literature, we propose four understandings of interpretability. We evaluate these using a novel experimental framework reflective of varied applied settings, including human evaluations using open labeling, typical of applied research. These evaluations show that for some specialized collections, standard coherence measures may not inform the most appropriate topic model or the optimal number of topics, and current interpretability performance validation methods are challenged as a means to confirm model quality in the absence of ground truth data.

## 1 Introduction

Topic modeling has become a popular tool for applied research such as social media analysis, as it facilitates the exploration of large document-collections and yields insights that would not be accessible by manual methods (Sinnenberg et al., 2017; Karami et al., 2020). However, social media data can be challenging to model as it is both sparse and noisy (Zhao et al., 2011). This has resulted in increased demand for short-text topic models that can handle these challenges (Lim et al., 2013; Zuo et al., 2016; Chen et al., 2015).

Topic word-sets, denoted  $T_{ws}$ , are considered to be semantically related words that represent the latent component of the underlying topic's document-collection, denoted  $T_{dc}$ . Meaning is derived from these topics through the interpretation of either the

$T_{ws}$  (Nerghes and Lee, 2019), the corresponding  $T_{dc}$  (Maier et al., 2018), or both (Törnberg and Törnberg, 2016). Since meaning requires topics to be interpretable to humans, empirical assurance is needed to confirm a novel topic models' capacity to generate "semantically interpretable" topics, as well as a method to guide model selection and other parameters such as the number of topics,  $K$ . This is often achieved by calculating the coherence scores for  $T_{ws}$  (Lau and Baldwin, 2016)

Recent literature contradicts previous evaluations of some short-text topic models that claim superior interpretability (Li et al., 2018; Eickhoff and Wieneke, 2018; Bhatia et al., 2017). Such rethinking flows from the fact there is no agreement on the best measure of interpretability (Lau et al., 2014b; Morstatter and Liu, 2017) and is compounded by the unclear relationship between human evaluation methodologies and automated coherence scores (Lau et al., 2014b). Finally, despite assurances of generalizability and applicability, topic model evaluations in machine learning are conducted in experimental settings that are not representative of typical applied use. This raises questions of whether coherence measures are suitably robust to measure topic interpretability and inform model selection in applied settings, particularly with challenging datasets like that of social media.

Advances in topic modeling for static document-collections have produced non-parametric approaches such as HDP-LDA, which employ sophisticated hierarchical priors that allow for different prior proportions (Teh et al., 2006). Non-negative matrix factorization (Zhou and Carin, 2015), the use of word embeddings, and neural network methods (Zhao et al., 2021) are a few of these other innovations.

To support these advances, it is crucial to establish the robustness of topic modeling interpretability measures, especially given the growing trend towards evaluating topic models using coherence

measures, often in the absence of perplexity or other predictive scores (?). Additionally, increasingly sophisticated methods for automatic topic labeling have been developed. Beginning with [Lau et al. \(2011\)](#), this research relies on models which generate interpretable topics. While these advances enhance the technologies available to conduct applied research, they do not address the underlying question of whether topic interpretability can be adequately assessed using coherence measures.

In this paper, we demonstrate a research gap in topic model evaluation methods in light of their growing use in specialized settings. Previously declared state-of-the-art models are under-performing in applied settings ([Li et al., 2018](#); [Arnold et al., 2016](#)), and little work has been done to improve application relevance ([Hecking and Leydesdorff, 2019](#)). Following the work of ([Lau and Baldwin, 2016](#); [Bhatia et al., 2017](#); [Hecking and Leydesdorff, 2019](#)), this study examines whether coherence is a valid predictor of topic model interpretability when interpretability is defined as more than just the ability to label a  $T_{ws}$ , and as the diversity of topic models, datasets and application tasks increases.

Earlier research has established a correlation between novel coherence measures and human ranking of interpretability, as measured by qualitative tests ([Cheng et al., 2014](#); [Newman et al., 2010a](#)). However, since bounded experimental settings constrain these tests, they are unlikely to reliably and consistently indicate topic quality in applied research settings. As a result, we ask the following question: To what extent can we rely on current coherence measures as proxies for topic model interpretability in applied settings?

This work has significant practical implications. It signals the need to re-develop interpretability measures and reappraise best-practice for validating and evaluating topic models and their applications. Our research contributes the following:

1. Introduces a novel human-centered qualitative framework for evaluating interpretability in model development that mimics those processes seen in applied settings.
2. Demonstrates that the ranking of topic quality using state-of-the-art coherence measures is inconsistent with those produced through validation tasks performed in an applied setting.
3. Systematically quantifies the impact of model behavior, dataset composition, and other pre-

viously reported factors ([Morstatter and Liu, 2017](#); [Lau and Baldwin, 2016](#)), on coherence measures for many topics across four variant datasets and two topic models.

4. Provide evidence to show that interpretability measures for evaluating  $T_{ws}$  and  $T_{dc}$  for applied work in specialized contexts (e.g., Twitter) are ill-suited and may hinder model development and topic selection.

The remainder of this paper is organized as follows. Section 2 provides a review of related work around the interpretability of topic models. Section 3 describes five propositions that have informed the design of interpretable topic models and their evaluation measures. This is followed by a description of the experimental framework we designed to test these propositions. Section 4 provides the results of these evaluations and Section 5 contains a discussion of findings.

## 2 Background

This section provides a brief overview of work related to interpretability evaluation, followed by a review of the challenges associated with coherence optimization for specialized contexts.

### 2.1 Topic Model Interpretability

Topic model interpretability is a nebulous concept ([Lipton, 2018](#)) related to other topic model qualities, but without an agreed-upon definition. Measures of semantic coherence influence how easily understood the top- $N$   $T_{ws}$  are ([Morstatter and Liu, 2017](#); [Lund et al., 2019](#); [Newman et al., 2010a](#); [Lau et al., 2014b](#)). This is also referred to as topic understandability ([Röder et al., 2015](#); [Aletras et al., 2015](#)).

A coherent topic is said to be one that can be easily labeled and thus interpreted ([Newman et al., 2011](#); [Morstatter and Liu, 2017](#)), but only if the label is meaningful ([Hui, 2001](#); [Newman et al., 2010b,a](#)). Some have modeled coherence measures based on topic meaningfulness ([Lau et al., 2014a](#)); others state that a meaningful topic is not necessarily a useful one ([Boyd-Graber et al., 2015](#)). Indeed, the literature remains divided over whether usefulness is a property of an interpretable topic ([Röder et al., 2015](#)), or if interpretability is a property of a useful topic ([Aletras and Stevenson, 2013](#); [Newman et al., 2010b](#)). Such terminological disagreement suggests that there are challenges to the progression of this area of research.

The ease of labeling a topic is assumed to be an expression of how coherent that topic is and thus its degree of interpretability. This assumption is challenged when annotators provide different labels for a topic. Morstatter and Liu (2017) presented interpretability from the perspective of both coherence and consensus, where consensus is a measure of annotator agreement about a topics’ representation in its  $T_{dc}$ . Alignment is how representative a topic is of its  $T_{dc}$  and is another understanding of interpretability (Ando and Lee, 2001; Chang et al., 2009; Mimno et al., 2011; Bhatia et al., 2017; Alokaili et al., 2019; Morstatter and Liu, 2017; Lund et al., 2019). However, the probabilistic nature of topic models impede this measure. The ambiguity of interpretability as a performance target raises questions about how topic models are used and evaluated.

## 2.2 Related Work

Following the seminal work of Chang et al. (2009), the development of coherence measures and the human evaluation tasks that guide their design has been actively pursued (Newman et al., 2010a; Bhatia et al., 2017, 2018; Morstatter and Liu, 2017; Lau and Baldwin, 2016; Lund et al., 2019; Alokaili et al., 2019). Newman et al. (2010a) showed that human ratings of topic coherence (observed coherence) correlated with their coherence measure when the aggregate Pointwise Mutual Information (PMI) pairwise scores were calculated over the top- $N$   $T_{ws}$ . In addition to the word intrusion task (Chang et al., 2009), Mimno et al. (2011) validated their coherence measure for modeling domain-specific corpora using expert ratings of topic quality. The measure takes the order of the top- $N$   $T_{ws}$  into account using a smoothed conditional probability derived from document co-occurrence counts. This performance was further improved by substituting PMI for Normalized PMI ( $C_{NPMI}$ ) (Aletras and Stevenson, 2013; Lau et al., 2014b). Aletras and Stevenson (2013) used crowdsourced ratings of topic usefulness to evaluate distributional semantic similarity methods for automated topic coherence. Röder et al. (2015) conducted an exhaustive study evaluating prior work and developing several improved coherence measures.

Similarly, Ramrakhiani et al. (2017) made use of the same datasets and evaluations and presented a coherence measure which is approximated with the size of the largest cluster produced from embed-

dings of the top- $N$   $T_{ws}$ . Human evaluation tasks have also been created to measure how representative a topic model is of the underlying  $T_{dc}$  (Chang et al., 2009; Bhatia et al., 2017; Morstatter and Liu, 2017; Alokaili et al., 2019; Lund et al., 2019).

## 2.3 Practical Applications

Within computer science, topic modeling has been used for tasks such as word-sense disambiguation (Boyd-Graber and Blei, 2007), hierarchical information retrieval (Blei et al., 2003), topic correlation (Blei and Lafferty, 2007), trend tracking (Al-Sumait and Domeniconi, 2008), and handling short-texts (Wang et al., 2018). Outside of computer science, topic modeling is predominantly used to guide exploration of large datasets (Agrawal et al., 2018), often with a human-in-the-loop approach. Here topics are generated before some form of qualitative method is used to gain insights into the data. These methods include exploratory content analysis (Korenčić et al., 2018), critical discourse analysis (Törnberg and Törnberg, 2016), digital autoethnography (Brown, 2019), grounded theory (Baumer et al., 2017), and thematic analysis (Doogan et al., 2020; Andreotta et al., 2019).

Qualitative techniques make use of topics in different ways. “Open labeling” of topics by Subject Matter Experts (SME) is followed by a descriptive analysis of that topic (Kim et al., 2016; Morstatter et al., 2018; Karami et al., 2018). However, this method is subjective and may fail to produce the depth of insight required. Supplementing a topic analysis with samples from the  $T_{dc}$  increases the depth of insight (Eickhoff and Wieneke, 2018; Kagashe et al., 2017; Nerghees and Lee, 2019). Alternatively, the  $T_{dc}$  alone can be used for in-depth analysis (Törnberg and Törnberg, 2016). However, human evaluation tasks that require open labeling are not generally used to validate new coherence measures (O’Callaghan et al., 2015; Korenčić et al., 2018).

## 3 Evaluating Interpretability

We have generated five propositions about the relationship between coherence scores, human evaluation of topic models, and the different views of interpretability to explore the research question. We conduct five experiments to interrogate these propositions and re-evaluate how informative coherence measures are for topic interpretability. Because we are evaluating existing coherence measures, we

do not employ automatic topic labeling techniques. Instead, we make use of human evaluation tasks that reflect those conducted in applied settings.

**Proposition 1.** *If coherence scores are robust, they should correlate.* The battery of coherence measures for evaluating novel topic models and automated labeling approaches are inconsistent across the literature. Each new measure claims superior alignment to topic model interpretability. As these measures are evolutionary (Röder et al., 2015), and there is no convention for which measure should be used, particularly as a standard measure of qualitative performance (Zuo et al., 2016; Zhao et al., 2017; Zhang and Lauw, 2020), they are considered notionally interchangeable. Thus, we would expect that there would be some correlation between these measures. However, previous studies have not considered the impact that the data type or model has on the coherence scores. Particularly for non-parametric models, these issues may be compounded by how coherence measures are presented as an aggregate, e.g., The presentation of the top-N models. Indeed, studies reporting multiple coherence measures have demonstrated inconsistencies at the model-level that are obscured during reporting (Blair et al., 2020).

**Proposition 2.** *An interpretable topic is one that can be easily labeled.* How easily a topic could be labeled has been evaluated on an ordinal scale where humans determined if they could hypothetically give a topic a label (Mimno et al., 2011; Morstatter and Liu, 2017). However, humans are notoriously poor at estimating their performance, particularly when they are untrained and do not have feedback on their performance (Dunning et al., 2003; Morstatter and Liu, 2017). Thus, a rater’s perception of whether they could complete a task is actually less informative than having them complete the task.

**Proposition 3.** *An interpretable topic has high agreement on labels.* Agreement on a topic label is considered a feature of interpretability by Morstatter and Liu (2017), who propose “consensus” as a measure of interpretability. A high level of agreement on topic labels, particularly in crowdsourcing tasks, is seen as a means to infer that a  $T_{ws}$  is interpretable. However, in applied tasks, a topic is described in a sense-making process resulting in one coherent label. Thus, the consensus task is not necessarily a reasonable means to infer inter-

pretability. A robust way to measure agreement on a topic label is needed. Inter-coder reliability (ICR) measures are an appropriate means to achieve this.

**Proposition 4.** *An interpretable topic is one where the document-collection is easily labeled.* The investigation of topic document-collections is an emerging trend in the applied topic modeling literature. In these studies, authors have either used a topics “top documents” to validate or inform the labels assigned to  $T_{ws}$  (Kirilenko et al., 2021), or have ignored the  $T_{ws}$  in favor of qualitative analysis of the richer  $T_{dc}$  (Doogan et al., 2020). The use of topic modeling for the exploration of document-collections requires a  $T_{dc}$  to be coherent enough that a reader can identify intertextual links between the documents. The label or description given to the  $T_{dc}$  results from the readers’ interpretation of individual documents relative to the other documents in the collection.  $T_{dc}$  that have a high degree of similarity between their documents will be easiest to interpret and therefore label. The ease of labeling a  $T_{dc}$  decreases as the documents become more dissimilar.

**Proposition 5.** *An interpretable topic word-set is descriptive of its topic document-collection.* The alignment of  $T_{ws}$  to  $T_{dc}$  is an expected property of a “good” topic (Chang et al., 2009), which human evaluation tasks have been developed to assess. Typically these tasks ask annotators to choose the most and/or least aligned  $T_{ws}$  to a given document (Morstatter and Liu, 2017; Lund et al., 2019; Alokaili et al., 2019; Bhatia et al., 2018), identify an intruder topic (Chang et al., 2009; Morstatter and Liu, 2017), rate their confidence in a topic-document pair (Bhatia et al., 2017), or select appropriate documents given a category label (Aletras et al., 2017). However, none of these methods address the need for the topic *document-collection* to be evaluated and labeled. Furthermore, they generally use one document and/or are not comparable to applied tasks.

### 3.1 Data

The Auspol-18 dataset was constructed from 1,830,423 tweets containing the hashtag #Auspol, an established Twitter forum for the discussion of Australian politics. The diminutives, slang, and domain-specific content provide a realistic example of a specialized context. Four versions of the dataset were constructed from a subset of 123,629 tweets; AWH (contains the 30 most frequent hash-

tags), AWM (contains the 30 most frequent mentions of verified accounts), AWMH (contains the 30 most frequent hashtags and 30 most frequent mentions of verified accounts), and AP (contains neither hashtags nor mentions). Pre-processing included stopword removal, POS-tagging, lemmatization, exclusion of non-English tweets, duplicate removal, removal of tokens with a frequency  $n < 10$ , and removal of tweets with  $n < 5$  tokens, and standardization of slang, abbreviations (Agrawal et al., 2018; Doogan et al., 2020) <sup>1</sup>.

### 3.2 Models and Parameters

To investigate interpretability in an applied setting, we compare LDA to MetaLDA (Zhao et al., 2017), a recent non-parametric topic model designed to improve short-text topic modeling by leveraging the incorporation of the document and word meta-information using word embeddings as well as non-parametrically learning topic proportions. Despite the many extensions to LDA, the vanilla model maintains popularity among applied researchers (Sun et al., 2016), and as the baseline model, it is necessary to compare LDA with a model purpose-built for short-text applications. MetaLDA is one reasonable representative of such models and has demonstrated effectiveness on Twitter data for applied work (Doogan et al., 2020). The extensive effort of human labeling in our experiments (see Section 3.4) precludes us from adding more models. LDA and MetaLDA are available in the MetaLDA package<sup>2</sup>, which is implemented on top of Mallet (McCallum, 2002).

Default parameter settings were used for both LDA and MetaLDA. We use Glove2Vec embeddings trained on the Wikipedia corpus (Pennington et al., 2014) for MetaLDA. We constructed topic sets with the number of topics  $K = \{10, 40, 20, 60, 80, 100, 150, 200\}$ .

### 3.3 Coherence Measures

Several coherence measures were evaluated. These were  $C_{Umass}$  (Mimno et al., 2011),  $C_V$ ,  $C_P$  (Röder et al., 2015),  $C_A$  and  $C_{NPMI}$  (Aletras and Stevenson, 2013). These were calculated for each topic using the Palmetto package<sup>3</sup> using the top ten most frequent words. Along with the default  $C_{NPMI}$ , which is calculated using Wikipedia, we introduced

$C_{NPMI-ABC}$ , which is calculated using a collection of 760k Australian Broadcasting Company (ABC) news articles<sup>4</sup> with 150 million words (enough to make the  $C_{NPMI}$  scores stable), and  $C_{NPMI-AP}$  calculated using the AP dataset and is used to test  $C_{NPMI}$  but with statistics drawn from the training data. We report the average scores and the standard deviations over five random runs.

### 3.4 Qualitative Experiments

A primary concern in machine learning research is the need to establish model performance. Following the recent trend to analyze  $T_{dc}$ , we devised qualitative tests for the assessment of whether the  $T_{ws}$  and  $T_{dc}$  were adequately aligned and whether current performance measures are informative of this alignment. We also tested to see if there is a relationship between topic alignment and the topic diagnostic statistics; effective number of words<sup>5</sup>, and topic proportion, denoted  $D_{ew}$  and  $D_{tp}$ , respectively.

**Topic Word-sets:** Four SMEs were recruited from a multidisciplinary pool of researchers who were representative of the political-ideological spectrum and who were Australian English speakers. They were shown the same topics consisting of the top-10 words ranked by term frequency that were generated by LDA and MetaLDA on AP, AWH, and AWM for  $K=10-60$  topics<sup>6</sup>, producing a total of 3,120 labels (780 for each SME) generated for the 390 topics (130 per model-dataset combination). Their task was to provide a descriptive label for each  $T_{ws}$  and to use ‘NA’ if they were unable to provide a label. Appendix A provides an example of this task. Two measures were constructed from these labels. The first was the number of raters able to label the topic, a count between 0–4 denoted  $Q_{nbr}$ . The second was a simple ICR measure, Percentage Agreement denoted  $Q_{agr}$ , which calculated as the number of times a set of annotators agree on a label, divided by the total number of annotations, as a percentage.

**Topic Document-collections:** Two SMEs analyzed the  $T_{dc}$ s of the 60 topics each modeled by LDA and MetaLDA on the AP dataset, referred to hereafter as the *qualitative set*. Samples of  $T_{dc}$  generated by each model ( $K=10-60$ ) were reviewed, and those generated from both models 60-topic sets

<sup>1</sup>Tweet IDs and pre-processing details are available at: <https://github.com/wbuntine/auspoldata>

<sup>2</sup><https://github.com/ethanhezhaio/MetaLDA>

<sup>3</sup><http://aksw.org/Projects/Palmetto.html>

<sup>4</sup><http://www.abc.net.au/news/archive>

<sup>5</sup>For word proportion vector  $\vec{p}$ , this is  $e^{-Entropy(\vec{p})}$ .

<sup>6</sup>The AWMH dataset was not included.

were found to be of equal or higher quality than those produced by other values of  $K$ .

The SMEs reviewed the top-30 tweets representative of a topic and provided a label for each tweet. They then inductively determined a label or phrase describing that  $T_{dc}$ . They noted any key phrases, names, or other terms that were consistent across the collection. The SMEs were experienced at annotating such datasets and were familiar with the online #Auspol community. The SMEs then discussed the results together and agreed on a final label for each  $T_{dc}$ .

The SMEs were asked to rate on a scale of 1–3 how difficult it was to label each  $T_{dc}$ , where 1 was difficult, 3 was easy, and 0 was where a label could be assigned. This qualitative statistic is denoted  $Q_{dif}$ . The researchers then scored, on a scale of 1–5, the degree of alignment between topic labels and the labels assigned to their corresponding collections. A score of 5 indicated the labels were identical, and a score of 0 indicated the  $T_{ws}$  and/or  $T_{dc}$  was incoherent. This statistic is denoted  $Q_{aln}$ . Examples of these tasks are in Appendix A.

### 3.5 Statistical Tests

We measure the strength of the association between variables using Pearson’s  $r$  correlation coefficient in evaluation 1 (see section 4.1) and Spearman’s  $\rho$  correlation coefficient in evaluations 2–5 (see sections 4.2, 4.3, 4.4, and 4.5). Pearson’s  $r$  is used in the few papers that evaluate coherence scores over the same datasets (Röder et al., 2015; Lau et al., 2014b). The practical reason for using Pearson’s  $r$  for our evaluation of proposition 1 was to make valid comparisons with these studies. The statistical justification for using Pearson’s  $r$  (rather than Spearman’s  $\rho$ ) is that the datasets are continuous (neither is ordinal, as Spearman’s  $\rho$  requires) and believed to have a bivariate normal distribution.<sup>7</sup> Spearman’s  $\rho$  is only appropriate when the relationship between variables is monotonic, which has not been consistently demonstrated for coherence (Röder et al., 2015; Bovens and Hartmann, 2004). Spearman’s  $\rho$  is appropriate to assess the association between coherence scores and human judgments in evaluations 2–5<sup>8</sup>. It is a preferred method

<sup>7</sup>We confirmed this with a Kolmogorov-Smirnov test for normality on the coherence scores.

<sup>8</sup>Although Kendall’s  $\tau$  has been used for similar evaluations (Rosner et al., 2013), it is unreliable when the range of each dataset varies significantly as in these experiments (Sanderson and Soboroff, 2007).

for such tasks (Aletras and Stevenson, 2013; Newman et al., 2010a) as it is unaffected by variability in the range for each dataset (Lau et al., 2014b).

## 4 Results

Here we detail the results of our analysis of the five propositions about interpretability evaluation.

### 4.1 Evaluation 1: Coherence Measure Correlations

As per proposition 1, coherence measures should be robust and highly correlated. To test this proposition, we conducted a Pearson’s correlation analysis of paired coherence measures calculated for  $K=10$ –60 for each model-dataset combination. Pooling the results for  $K$  and the three datasets, we calculate the  $\bar{x}_r$  for LDA and MetaLDA.

$C_{NPMI}$  and  $C_P$  scores were strongly correlated for all datasets. Ranging from  $\bar{x}_r=0.779$ –0.902 for LDA, and  $\bar{x}_r=0.770$ –0.940 for MetaLDA.  $C_{NPMI}$  and  $C_{NPMI-ABC}$  also showed a moderate-to-strong correlation for all datasets with LDA ranging from  $\bar{x}_r=0.719$ –0.769, and MetaLDA from  $\bar{x}_r=0.606$ –0.716.  $C_{NPMI-ABC}$  appears more sensitive to changes in  $K$  than  $C_P$ . No significant trends were seen between other coherence measures calculated for any dataset. These results are reported in Appendix B.

Methods to aggregate coherence scores may mask any differences in the models’ behaviors as  $K$  increases. To test this, aggregate coherence measures, typical of the empirical evaluation of topic models, were calculated per value of  $K$ . These were the mean of all topics (Average), the mean for all topics weighted by the topic proportion (WeightedAverage), and the mean of the Top- $N$  percent of ranked topics by coherence score (TopNpcnt), where  $N = \{25, 50, 80\}$ .

Both models showed trends in aggregated coherence scores calculated on the AP dataset. As shown in Figure 1, the peak for each measure varies according to different values of  $K$  and between models. For instance, aggregates of both models  $C_{NPMI}$  and  $C_{NPMI-ABC}$  peak at 60 and 10 topics, respectively. However,  $C_V$  aggregate peaks are completely divergent between models,  $K=200$  for MetaLDA and  $K=50$  for LDA. Indeed, the two models favored different coherence measures and aggregate methods. Generally, MetaLDA exhibits superior performance across all aggregates for  $C_V$  and  $C_A$ , while LDA is superior for  $C_{Umass}$ . No-

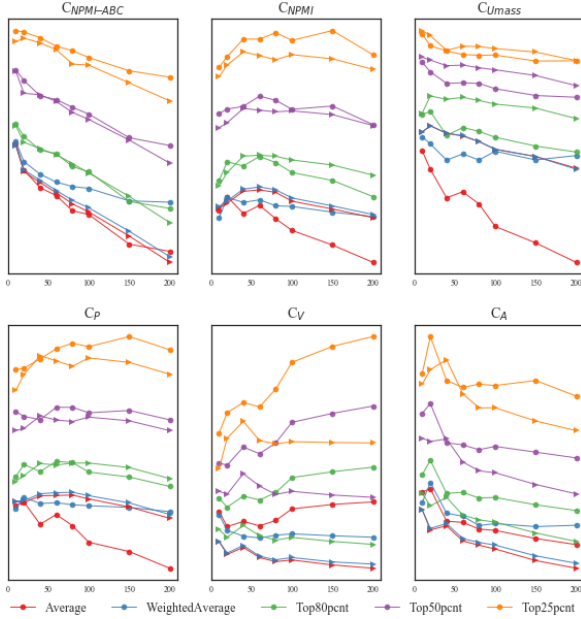


Figure 1: Comparison of LDA (triangle) and MetaLDA (circle) aggregated coherence scores for the AP dataset. Scores are shown on the y-axis, and  $K$  is shown on the x-axis. Individual points are averaged across five runs, where the typical sample standard is 0.005, but up to 0.010 for  $K=20$ .

tably, MetaLDA shows superior  $C_{NPMI}$ ,  $C_{NPMI-ABC}$ ,  $C_{NPMI-AP}$  scores for Top20pct, Top50pct, and Top80pct aggregations, but is inferior when the full average of these scores is calculated. Other datasets are broadly similar and shown in Appendix B.

We also compare MetaLDA with LDA. Pooling the results for  $K=10-200$  for each of the four datasets, we get a set of differences in the scores and compute the  $p$ -value for a one-sided student  $t$ -test to determine whether LDA has higher average coherence scores than MetaLDA. MetaLDA yields significantly higher  $C_{NPMI}$  scores calculated using the Top20pct ( $p<0.01$ ) and Top50pct of topics ( $p<0.05$ ). Conversely, LDA yields significantly higher  $C_{NPMI}$  scores for the other aggregates ( $p<0.01$ ). Except for the full average, MetaLDA achieves significantly higher ( $p<0.01$ )  $C_{NPMI-ABC}$ ,  $C_{NPMI-AP}$ , and  $C_V$  scores than LDA for the other aggregate methods.

Disturbingly, the “best” models, or optimal  $K$  varies depending on the coherence measure and the aggregate measure used to calculate it. This has implications for topic model selection in applied settings, where coherence is used to inform  $K$  (Kirilenko et al., 2021). When repeating the analysis using different  $K$ , a second trend emerges: Met-

aLDA significantly outperforms LDA in  $C_{NPMI}$  for smaller  $K$  on average but loses out for larger  $K$ . Results from our qualitative analysis confirmed this occurred because LDA had many less frequent topics (e.g., when  $K = 60$ , all topics occur about 1/60 of the time), unlike MetaLDA, which mixes more and less frequent topics.

## 4.2 Evaluation 2: Labeling Topic Words-sets

Proposition 2 states that if topics can be labeled they are interpretable. Coherence as a measure of interpretability should then be predictive of topics that can be labeled. To evaluate this proposition, a Spearman’s  $\rho$  correlation coefficient was used to assess the relationship between coherence measures and the number of raters able to label the  $T_{ws}$ ,  $Q_{nbr}$ , for each of the 130 topics produced per model-dataset combination. These results are available in Appendix C. There was no significant correlation between any coherence measure and  $Q_{nbr}$ . Interestingly, the SMEs reported several topics they could not label despite their high coherence scores. For instance, the LDA modeled topic “red, wear, flag, blue, gold, black, tape, tie, green, iron” could not be labeled despite being the 9<sup>th</sup>/60 highest ranked topic for  $C_{NPMI}$ .

## 4.3 Evaluation 3: Topic Label Agreement

Proposition 3 states an interpretable topic is one where there is high agreement between annotators on its label. As such, coherence should align to measures of consensus or agreement. To evaluate this proposition, we calculate the gold-standard ICR measures, Fleiss’ kappa ( $\kappa$ ) (Fleiss, 1971) and Krippendorff’s alpha ( $\alpha$ ) (Krippendorff, 2004). Both allow for multiple coders and produce a chance-corrected estimate of ICR but do not facilitate the isolation of low-agreement topics. For this, we also calculated the Percentage Agreement  $Q_{agr}$  for each topic, as shown in Appendix D.

Generally,  $\alpha$ ,  $\kappa$ , and  $Q_{agr}$  improved as  $K$  increased. As shown in Table 1, LDA consistently outperformed MetaLDA when  $K=60$  across all three datasets and generally attained higher  $\alpha$ ,  $\kappa$ , and  $Q_{agr}$  scores than MetaLDA. There was a moderate-to-strong agreement between SMEs, a reliable result for an open labeling task (Landis and Koch, 1977). However, the performance of each model was notably affected by the datasets. LDA outperformed MetaLDA on the AP dataset across all three measures except for  $\kappa$  when  $K=20$ , and for  $Q_{agr}$  when  $K=10$ . Except for  $\kappa$  when  $K=40$ ,

MetaLDA achieved higher or comparable scores to LDA on the AWH dataset when  $K=20-40$ , but outperformed LDA only when  $K=10-20$  on the AWM dataset.

	Kripp. $\alpha$		Fleiss' $\kappa$		Pcnt. $Q_{agr}$	
	LDA	Meta	LDA	Meta	LDA	Meta
AP	0.584	0.486	0.578	0.485	0.503	0.492
AWH	0.512	0.498	0.527	0.515	0.439	0.411
AWM	0.513	0.447	0.535	0.492	0.428	0.369

Table 1: Krippendorff’s  $\alpha$ , Fleiss’  $\kappa$ , and  $Q_{agr}$  ICR statistics for topic labeling when  $K=60$ .

Spearman’s  $\rho$  was calculated to measure the strength of the relationship between  $Q_{agr}$  and the generated coherence measures. As shown in Appendix D, results were random with no significant correlations. As shown in Table 2, there was a statistically significant correlation between  $Q_{agr}$  and  $Q_{nbr}$  when  $K=60$ .

LDA	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
AP	0.240	0.504	0.548	0.012	0.431	<0.01	0.475	<0.01
AWH	0.000	0.000	0.458	0.042	0.561	<0.01	0.644	<0.01
AWM	0.506	0.136	0.345	0.136	0.490	<0.01	0.697	<0.01
Meta LDA	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
AP	0.544	0.104	0.147	0.535	0.445	<0.01	0.690	<0.01
AWH	0.532	0.113	0.478	0.033	0.147	0.366	0.629	<0.01
AWM	0.548	0.101	0.414	0.069	0.743	<0.01	0.700	<0.01

Table 2: The Spearman’s  $\rho$  correlation coefficients for pairwise combinations of  $Q_{agr}$  and  $Q_{nbr}$  for all learned models.

Coherence measures did not correlate with  $Q_{agr}$ , and in some cases, were contradictory. For example,  $Q_{agr}$  generally increases with  $K$  (and our experts reported that labeling was often easier for smaller topics), but coherence measures such as  $C_A$  and  $C_{NPMI-ABC}$  tended to decrease (in Figure 1). These results show that the two models show different sensitivities to dataset preparation and the value of  $K$ .

#### 4.4 Evaluation 4: Labeling Topic Document-collections

Proposition 4 states that topics that are interpretable have a  $T_{dc}$  that is easily labeled. To evaluate this proposition, a Spearman’s  $\rho$  was used to assess the relationship between coherence measures and SME ratings of  $T_{dc}$  labeling difficulty,  $Q_{dif}$ . The full set, Top25pct, top50pct, and bottom 15% (Bot15pct) of ranked  $Q_{dif}$  scores were analyzed. The only notable correlation was between the Bot15pct of LDA  $T_{dc}$  for  $C_{NPMI-ABC}$  ( $\rho=-0.817$ ,

$p=<0.01$ ). Interestingly, when ranked by topic diagnostic  $D_{ew}$ , the Top25pct and Top50pct of  $T_{dc}$ s showed moderate correlation with  $Q_{dif}$  for MetaLDA ( $\rho=-0.764$ ,  $p<0.01$ ;  $\rho=-0.630$ ,  $p<0.01$ ).

A repeat analysis with topic diagnostic  $D_{tp}$  did not yield any statistically significant results. However, we observed that for  $T_{dc}$ s produced by MetaLDA, the three largest and three smallest topics could not be labeled. By contrast, the LDA  $T_{dc}$ s that were not interpretable were from the smallest 20% of topics. We hypothesize that this distinction results from MetaLDA’s broadly distributed  $D_{tp}$  ( $0.017\pm 0.155$ ), which features several very large and very small topics. By comparison, LDA  $D_{tp}$  is approximately uniformly distributed ( $0.017 \pm 0.001$ ).

#### 4.5 Evaluation 5: Topic Label Alignment

Proposition 5 states that an interpretable topic is one that is descriptive of the  $T_{dc}$ . To test this proposition, we constructed an alignment score  $Q_{aln}$ , which rate the similarity between the standardized topic label from  $T_{ws}$  and the label from  $T_{dc}$ . Similar to the evaluation of Proposition 4, we conducted a Spearman’s  $\rho$  to test for a relationship between  $Q_{aln}$ , coherence measures, and diagnostic scores.

The following illustrates a high scoring, but poorly aligned topic with a  $C_{NPMI}$  of 0.073.  $T_{ws}$ : “law, bill, power, gun, democracy, control, freedom, rule, protect, legislation” was labeled “Gun control”, but the  $T_{dc}$  was labeled “Foreign Interference Act”. Appendix F contains additional examples.

LDA showed a strong relationship between  $Q_{aln}$  and  $C_{NPMI-ABC}$  for the Top25pct of topics ( $\rho=0.825$ ,  $p<0.01$ ), but the relationship was weak for other coherence measures. No coherence measures were correlated with MetaLDA  $Q_{aln}$  scores.

As per section 4.4, we repeated the analysis by ranking topics by  $D_{ew}$ . MetaLDA showed a strong-to-moderate correlation between  $D_{ew}$  and  $Q_{aln}$  for the Top25pct ( $\rho=-0.776$ ,  $p=<0.01$ ), Top50pct ( $\rho=-0.646$ ,  $p<0.01$ ), and Bot15pct ( $\rho=0.693$ ,  $p=0.039$ ) of topics, making  $D_{ew}$  a potentially useful proxy for alignment for MetaLDA.

## 5 Discussion

We repeated the work of Zhao et al. (2017), who demonstrated that when the top-ranked topics by  $C_{NPMI}$  are considered, MetaLDA produces higher  $C_{NPMI}$  scores than LDA. However, when  $C_{NPMI}$  was measured using alternative aggregate methods, we



discovered that LDA outperformed MetaLDA. This is likely to be because the smaller topics in MetaLDA can be effectively ignored or scrapped, while in LDA, all topics are of comparable size and are used by the model. Other non-parametric topic models are believed to behave similarly. While MetaLDA generated higher  $C_{NPMI-ABC}$  scores than LDA for all aggregates, it was highly dependent on dataset heterogeneity and the value of  $K$ . This should indicate that MetaLDA is more adaptive to specialized language, an effect expected in other topic models supported by word embeddings.

The comparative performance of coherence measures can vary significantly depending on the aggregate calculation method used and the way the data has been prepared. This latter point has been well established in the literature, most notably for Twitter data (Symeonidis et al., 2018), but is often overlooked when evaluating novel topic models. This is a cause for concern, given the growing reliance on coherence measures to select the optimal model or  $K$  in applied settings (Xue et al., 2020; Lyu and Luli, 2021).

Propositions 2 and 3 addressed  $T_{ws}$  interpretability. We have demonstrated the difference between comprehending a topic and providing a topic label that is both informative and reliable. However, coherence measures may not be informative of these qualities. Propositions 4 and 5 addressed  $T_{dc}$  interpretability. We have demonstrated that the ease of labeling a  $T_{dc}$  and the alignment between  $T_{ws}$  and  $T_{dc}$  does not correlate with coherence measures. Additionally, we identified several areas for future research into the use of diagnostic statistics in applied settings. We observed unexpected behaviors in the distributions of  $D_{ew}$  and  $D_{tp}$  after a comparative analysis between LDA and the non-parametric model MetaLDA, affecting the interpretability of both  $T_{ws}$  and  $T_{dc}$ . Correlations between  $Q_{dif}/Q_{aln}$  and  $D_{ew}/D_{tp}$  for MetaLDA, for example, indicate that these topic diagnostics could assist in evaluating  $T_{dc}$  interpretability.

## 6 Conclusion

We have shown that coherence measures can be unreliable for evaluating topic models for specialized collections like Twitter data. We claim this is because the target of “interpretability” is ambiguous, compromising the validity of both automatic and

human evaluation methods<sup>9</sup>.

Due to the advancements in topic models, coherence measures designed for older models and more general datasets may be incompatible with newer models and more specific datasets. Our experiments show that non-parametric models, such as MetaLDA, which employs embeddings to improve support for short-texts, behaves differently to LDA for these performance and diagnostic measures. This is critical because recent research has focused on sophisticated deep neural topic models (Zhao et al., 2021), which make tracing and predicting behaviors more challenging. Abstractly, we may compare the use of coherence measures in topic modeling to the use of BLEU in machine translation. Both lack the finesse necessary for a complete evaluation, as is now the case with BLEU (Song et al., 2013).

Additionally, our study demonstrated that an examination of the  $T_{dc}$  could provide greater insights into topic model behaviors and explained many of the observed problems. We argue for the representation of topics as a combination of thematically related  $T_{dc}$  and  $T_{ws}$ , and the further adoption of empirical evaluation using specialized datasets and consideration of  $T_{dc}$  interpretability. To date, few papers have attempted this combination (Korenčić et al., 2018).

However, we believe coherence measures and automated labeling techniques will continue to play a critical role in applied topic modeling. Contextually relevant measures like  $C_{NPMI-ABC}$  and topic diagnostics like  $D_{ew}$  can be key indicators of interpretability. Aside from the empirical evaluation of novel topic models, new automated labeling techniques, having proven themselves useful for labeling  $T_{tw}$ , should be extended for  $T_{dc}$ .

## Acknowledgments

We thank Callum Waugh, Elliot Freeman and Elizabeth Daniels for conducting the expert annotations. We also thank Henry Linger for providing feedback on earlier drafts. The first author discloses the following financial support for the research: an Australian Government Research Training Program (RTP) Stipend and RTP Fee-Offset Scholarship, and an Australian Government Defence Science and Technology Group Research scholarship.

<sup>9</sup>Specifically, construct validity, which confirms if the operational definition of a variable (interpretability) reflects the true theoretical meaning of a concept (O’Leary-Kelly and Vokurka, 1998).

## Ethics and Impact

This project has been reviewed and approved by the Monash University Human Research Committee (Project ID: 18167), subject to abidance with legislated data use and protection protocols. In particular, the Twitter Inc. developers policy prohibits the further distribution of collected tweets and associated metadata by the authors group, with the exception of tweet IDs which may be distributed and re-hydrated. The subject matter of the tweets collected is Australian Politics. We have forgone the use of material included in the paper that would be offensive or problematic to marginalized groups in the Australian political context.

## References

- A. Agrawal, W. Fu, and T. 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.
- N. Aletras, T. Baldwin, J. H. Lau, and M. Stevenson. 2017. [Evaluating topic representations for exploring document collections](#). *Journal of the Association for Information Science and Technology*, 68(1):154–167.
- N. Aletras, J. H. Lau, T. Baldwin, and M. Stevenson. 2015. [TM 2015–topic models: Post-processing and applications workshop](#). In *Proc. of the 24th ACM Int. on Conf. on Information and Knowledge Management (CIKM)*, pages 1953–1954.
- N. Aletras and M. Stevenson. 2013. [Evaluating topic coherence using distributional semantics](#). In *Proc. of the 10th Int. Conf. on Computational Semantics (IWCS)*, pages 13–22.
- A. Alokaili, N. Aletras, and M. Stevenson. 2019. [Re-ranking words to improve interpretability of automatically generated topics](#). In *Proc. of the 13th Int. Conf. on Computational Semantics (IWCS)*, pages 43–54, Gothenburg, Sweden. ACL.
- D. AlSumait, L. and Barbará and C. Domeniconi. 2008. [On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking](#). In *Proc. of the 8th IEEE Int. Conf. on Data Mining (ICDM)*, pages 3–12. IEEE.
- R.-K. Ando and L. Lee. 2001. [Iterative residual rescaling](#). In *Proc. of the 24th annual int. ACM SIGIR conf. on Research and development (SIGIR)*, SIGIR '01, pages 154–162, New York, NY, USA. ACM.
- M. Andreotta, R. Nugroho, M. Hurlstone, F. Boschetti, S. Farrell, I. Walker, and C. Paris. 2019. [Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis](#). *Behavior research methods*, 51(4):1766–1781.
- C. Arnold, A. Oh, S. Chen, and W. Speier. 2016. [Evaluating topic model interpretability from a primary care physician perspective](#). *Computer Methods and Programs in Biomedicine*, 124:67–75.
- E. Baumer, D. Mimno, S. Guha, E. Quan, and G. Gay. 2017. [Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?](#) *Journal of the Association for Information Science and Technology*, 68(6):1397–1410.
- S. Bhatia, J. H. Lau, and T. Baldwin. 2017. [An automatic approach for document-level topic model evaluation](#). In *Proc. of the 21st Conf. on Computational Natural Language Learning (CoNLL)*, pages 206–215.
- S. Bhatia, J. H. Lau, and T. Baldwin. 2018. [Topic intrusion for automatic topic model evaluation](#). In *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 844–849.
- S J. Blair, Y. Bi, and Maurice D Mulvenna. 2020. [Aggregated topic models for increasing social media topic coherence](#). *Applied Intelligence*, 50(1):138–156.
- D. Blei and J. Lafferty. 2007. [A correlated topic model of science](#). *The Annals of Applied Statistics*, 1(1):17–35.
- D. Blei, A. Ng, and M. Jordan. 2003. [Latent Dirichlet allocation](#). *Journal of Machine Learning Research*, 3(Jan):993–1022.
- L. Bovens and S. Hartmann. 2004. *Bayesian Epistemology*. Oxford University Press.
- J. Boyd-Graber and D. Blei. 2007. [PUTOP: Turning predominant senses into a topic model for word sense disambiguation](#). In *4th SemEval-2007*, pages 277–281.
- J. Boyd-Graber, D. Mimno, and D. Newman. 2015. [Care and feeding of topic models: Problems, diagnostics, and improvements](#). In Edoardo M. Airolidi, David Blei, Elena A. Erosheva, and Stephen E. Fienberg, editors, *Handbook of Mixed Membership Models and Their Applications*. CRC Press Boca Raton, FL.
- N. Brown. 2019. [Methodological cyborg as black feminist technology: constructing the social self using computational digital autoethnography and social media](#). *Cultural Studies ↔ Critical Methodologies*, 19(1):55–67.
- J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei. 2009. [Reading tea leaves: How humans interpret topic models](#). In *Proc. of the 23rd Annu. Conf. Neural Information Processing Systems (NeurIPS)*, pages 288–296.
- W. Chen, J. Wang, Y. Zhang, H. Yan, and X. Li. 2015. [User based aggregation for biterm topic model](#). In *Proc. of the 53rd Annu. Meeting of the Association*

- for *Computational Linguistics and the 7th Int. Joint Conf. on Natural Language (ACL-IJCNL)*, pages 489–494.
- X. Cheng, X. Yan, Y. Lan, and J. Guo. 2014. **BTM: Topic modeling over short texts**. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.
- C. Doogan, W. Buntine, H. Linger, and S. Brunt. 2020. **Public perceptions and attitudes toward covid-19 nonpharmaceutical interventions across six countries: A topic modeling analysis of twitter data**. *Journal of Medical Internet Research*, 22(9).
- D. Dunning, K. Johnson, J. Ehrlinger, and J. Kruger. 2003. **Why people fail to recognize their own incompetence**. *Current directions in psychological science*, 12(3):83–87.
- M. Eickhoff and R. Wieneke. 2018. **Understanding topic models in context: A mixed-methods approach to the meaningful analysis of large document collections**. In *Proc. of the 51st Hawaii Int. Conf. on System Sciences (HICSS)*, pages 903–912.
- J. Fleiss. 1971. **Measuring nominal scale agreement among many raters**. *Psychological Bulletin*, 76(5):378–382.
- T. Hecking and L. Leydesdorff. 2019. **Can topic models be used in research evaluations? reproducibility, validity, and reliability when compared with semantic maps**. *Research Evaluations*, 28(3):263–272.
- K. Hui. 2001. *Automatic Topic Detection From News Stories*. Ph.D. thesis, CUHK.
- I. Kagashe, Z. Yan, and I. Suheryani. 2017. **Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data**. *Journal of Medical Internet Research*, 19(9):e315.
- A. Karami, A. Dahl, G. Turner-McGrievy, H. Kharrazi, and G. Shaw Jr. 2018. **Characterizing diabetes, diet, exercise, and obesity comments on Twitter**. *International Journal of Information Management*, 38(1):1–6.
- A. Karami, M. Lundy, F. Webb, and Y. Dwivedi. 2020. **Twitter and research: a systematic literature review through text mining**. *IEEE Access*, 8:67698–67717.
- E. Kim, Y. Jeong, Y. Kim, K. Kang, and M. Song. 2016. **Topic-based content and sentiment analysis of ebola virus on Twitter and in the news**. *Journal Information Science*, 42(6):763–781.
- A. P. Kirilenko, S. O. Stepchenkova, and X. Dai. 2021. **Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply?** *Tourism Management*, 83:104241.
- D. Korenčić, S. Ristov, and J. Šnajder. 2018. **Document-based topic coherence measures for news media text**. *Expert Systems With Applications*, 114:357–373.
- K. Krippendorff. 2004. **Measuring the reliability of qualitative text analysis data**. *Quality and quantity*, 38:787–800.
- J. Landis and G. Koch. 1977. **The measurement of observer agreement for categorical data**. *Biometrics*, pages 159–174.
- J. H. Lau and T. Baldwin. 2016. **The sensitivity of topic coherence evaluation to topic cardinality**. In *Proc. 17th Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 483–487.
- J. H. Lau, P. Cook, D. McCarthy, S. Gella, and T. Baldwin. 2014a. **Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models**. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 259–270.
- J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. 2011. **Automatic labelling of topic models**. In *Proc. of the 49th Conf. on Human Language Technologies (HLT)*, pages 1536–1545, Portland, Oregon, USA. ACL.
- J. H. Lau, D. Newman, and T. Baldwin. 2014b. **Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality**. In *Proc. of the 9th Conf. European Chapter of the Association for Computational Linguistics (EACL)*, pages 530–539.
- X. Li, C. Li, J. Chi, and J. Ouyang. 2018. **Short text topic modeling by exploring original documents**. *Knowledge and Information Systems*, 56(2):443–462.
- Kar Wai Lim, Changyou Chen, and Wray L. Buntine. 2013. **Twitter-Network Topic Model: A full Bayesian treatment for social network and text modeling**. In *Proc. of the 27th Annu. Conf. on Neural Information Processing Systems (NeurIPS): Topic Models Workshop*, NeurIPS Workshop 2013, pages 1–5.
- Z. C. Lipton. 2018. **The mythos of model interpretability**. *Communications of the ACM*, 61(10):36–43.
- J. Lund, P. Armstrong, W. Fearn, S. Cowley, C. Byun, J. Boyd-Graber, and K. Seppi. 2019. **Automatic evaluation of local topic quality**. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 788–796.
- J. Lyu and G. Luli. 2021. **Understanding the public discussion about the centers for disease control and prevention during the covid-19 pandemic using twitter data: Text mining analysis study**. *Journal of Medical Internet Research*, 23(2):e25108.
- D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, and T. Häussler. 2018. **Applying LDA topic modeling in communication research: Toward a**

- valid and reliable methodology. *Communication Methods and Measures*, 12(2–3):93–118.
- A. McCallum. 2002. **MALLET: A machine learning for language toolkit**.
- D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. **Optimizing semantic coherence in topic models**. In *Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP '11, pages 262–272, USA. ACL.
- F. Morstatter and H. Liu. 2017. **In search of coherence and consensus: Measuring the interpretability of statistical topics**. *Journal of Machine Learning Research*, 18(1):6177–6208.
- F. Morstatter, Y. Shao, A. Galstyan, and S. Karunasekera. 2018. **From alt-right to alt-rechts: Twitter analysis of the 2017 German federal election**. In *Proc. of the 2018 World Wide Web Conference (TheWebConf)*, pages 621–628. IW3C2.
- A. Nerghes and J. S. Lee. 2019. **Narratives of the refugee crisis: A comparative study of mainstream-media and Twitter**. *Media and Communication*, 7(2):275–288.
- D. Newman, E. Bonilla, and W. Buntine. 2011. **Improving topic coherence with regularized topic models**. In *Proc. of the 25th Annu. Conf. on Neural Information Processing Systems (NeurIPS)*, pages 496–504.
- D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. 2010a. **Automatic evaluation of topic coherence**. In *Proc. of the 11th Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 100–108. ACL.
- D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin. 2010b. **Evaluating topic models for digital libraries**. In *Proc. of the 10th ACM/IEEE-CS Joint Conf. on Digital Libraries, (JCDL)*, pages 215–224.
- D. O’Callaghan, D. Greene, J. Carthy, and P. Cunningham. 2015. **An analysis of the coherence of descriptors in topic modeling**. *Expert Systems With Applications*, 42(13):5645–5657.
- S. O’Leary-Kelly and R. Vokurka. 1998. **The empirical assessment of construct validity**. *Journal of Operations Management*, 16(4):387–405.
- J. Pennington, R. Socher, and C. Manning. 2014. **GloVe: Global vectors for word representation**. In *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. ACL.
- N. Ramrakhiani, S. Pawar, S. Hingmire, and G. Palshikar. 2017. **Measuring topic coherence through optimal word buckets**. In *Proc. of the 15th Conf. European Chapter of the Association for Computational Linguistics (EACL)*, pages 437–442.
- M. Röder, A. Both, and A. Hinneburg. 2015. **Exploring the space of topic coherence measures**. In *Proc. of the 8th ACM Int. Conf. on Web Search and Data Mining (WSDM)*, pages 399–408.
- F. Rosner, A. Hinneburg, M. Roder, M. Nettling, and A. Both. 2013. **Evaluating topic coherence measures**. In *Proc. 27th Annu. Conf. on Neural Information Processing Systems (NeurIPS)*, pages 1–4.
- M. Sanderson and I. Soboroff. 2007. **Problems with kendall’s tau**. In *Proc. of the 30th annual int. ACM SIGIR conf. on Research and development (SIGIR)*, pages 839–840.
- L. Sinnenberg, A. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. Merchant. 2017. **Twitter as a tool for health research: a systematic review**. *American Journal of Public Health*, 107(1):e1–e8.
- X. Song, T. Cohn, and L. Specia. 2013. **BLEU deconstructed: Designing a better MT evaluation metric**. *International Journal of Computational Intelligence and Applications*, 4(2):29–44.
- X. Sun, X. Liu, B. Li, Y. Duan, H. Yang, and J. Hu. 2016. **Exploring topic models in software engineering data analysis: A survey**. In *Proc. of the 17th IEEE/ACIS Int. Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 357–362.
- S. Symeonidis, D. Effrosynidis, and A. Arampatzis. 2018. **A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis**. *Expert Systems with Applications*, 110:298–310.
- Y. Teh, M. Jordan, M. Beal, and D. Blei. 2006. **Hierarchical Dirichlet processes**. *Journal of the American Statistical Association*, 101(476):1566–1581.
- A. Törnberg and P. Törnberg. 2016. **Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum**. *Discourse & Society*, 27(4):401–422.
- J. Wang, L. Chen, L. Qin, and X. Wu. 2018. **Astm: An attentional segmentation based topic model for short texts**. In *Proc. of the 18th IEEE Int. Conf. on Data Mining (ICDM)*, pages 577–586.
- J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu. 2020. **Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter**. *PLoS one*, 15(9):e0239441.
- C. Zhang and H. W. Lauw. 2020. **Topic modeling on document networks with adjacent-encoder**. volume 34, pages 6737–6745.

- H. Zhao, L. Du, W. Buntine, and G. Liu. 2017. [Met-  
alda: A topic model that efficiently incorporates  
meta information](#). In *Proc. of the 17th IEEE Int.  
Conf. on Data Mining (ICDM)*, pages 635–644.
- H. Zhao, Dinh P., v. Huynh, y. Jin, L. Du, and W. Bun-  
tine. 2021. [Topic modelling meets deep neural net-  
works: A survey](#). *arXiv preprint arXiv:2103.00498*.
- W. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and  
X. Li. 2011. [Comparing twitter and traditional me-  
dia using topic models](#). In *Proc. of the 33rd Euro-  
pean Conf. on Information Retrieval (ECIR)*, pages  
338–349, Berlin, Heidelberg. Springer Berlin Hei-  
delberg.
- M. Zhou and L. Carin. 2015. [Negative binomial pro-  
cess count and mixture modeling](#). *IEEE Transac-  
tions on Pattern Analysis and Machine Intelligence*,  
37(2):307–320.
- Y. Zuo, J. Zhao, and K. Xu. 2016. [Word network topic  
model: a simple but general solution for short and  
imbalanced texts](#). *Knowledge and Information Sys-  
tems*, 48(2):379–398.

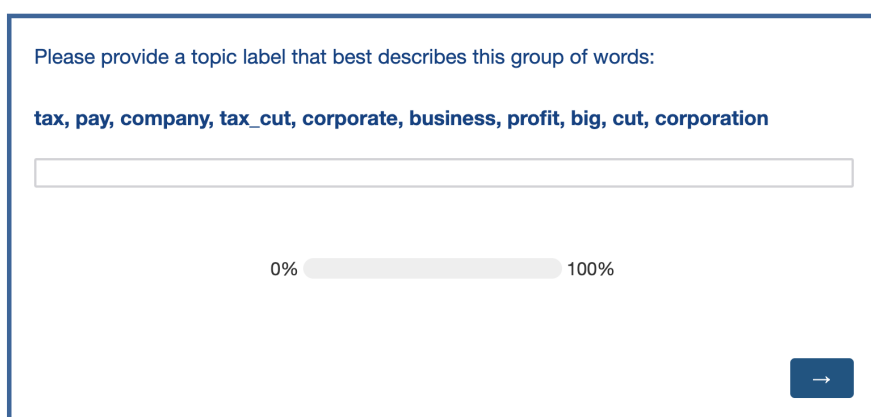
# APPENDIX: Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures

Caitlin Doogan and Wray Buntine

NAACL-HLT 2021

## A Examples of tasks for Qualitative Experiments

### Topic word-set labeling



Please provide a topic label that best describes this group of words:

**tax, pay, company, tax\_cut, corporate, business, profit, big, cut, corporation**

0%  100%

Figure 2: Example of topic word-set labeling task. Topic 40 from AP modeled on LDA.

## Topic document-collection labeling

@ScottMorrisonMP @KellyODwyer #Budget2018 #Auspol have a read. No benefits to company tax cuts
...and not increasing executive salaries or share buybacks. #TaxCuts #auspol
Can anyone explain who RTPDS Aus Pty Ltd is and how out of \$14,000,000,000 (\$14B if you can't keep up with the zeros) taxable income they paid only \$6,375 in corporate tax? Is that even possible?? #auspol
@SwannyQLD why should multinationals and corporations that pay ZERO tax get a tax cut? #Insiders #AskingForAFriend #auspol
Corp tax cuts, just another gift, nothing to do with stimulating the economy, especially when so many don't even pay tax as it is. #auspol
#Fukushima #auspol AVOID #thorium #nuclear SHILLS John Quakes Quade Malloys Gus Rawles Leo Sutton Marcelina Thomas Hypermetropia
#auspol Profits up 25% wages up 0% ... give my corporation a tax cut!!!. via @watoday
So Treasurer, tell us again how a tax cut for corporations who don't pay tax anyway will mean they will increase salaries. #auspol
Cut out the corporate tax cuts and do not even give them trickle #auspol
Well how else could they afford to pay him so much if they had to pay tax and workers as well!! #auspol

Please provide a topic label that best describes the collection of tweets above.

Your label should describe a central theme, concept, event, or other unifying feature of the collection. If one or more tweets do not align with a label that is suitable to describe all other tweets, then you may disregard them.

You may click on the tweet to see it in context.

If you cannot identify an appropriate label, please enter 'NA'.

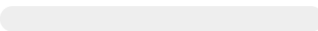
0%  100%

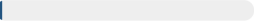
Figure 3: Example of topic document-collection labeling task. Only the top 10 tweets have been shown for brevity.

## Difficulty of topic document-collection labeling

Please rate how difficult was it to identify a central theme, concept, event, or other unifying feature of the the collection above, and then provide a descriptive label for this?

Your answer should reflect how coherent the collection was, not the complexity of the concept.

- Extremely easy
- Neither easy nor difficult
- Difficult
- I could not label the collection and wrote 'NA'

0%  100%



 

Figure 4: Example question asking SME to rate how difficult it was to label a topic document-collection.



## Topic word-set and topic document-collection label alignment

Please rate how aligned the following topic-wordset label to the topic-document collection label?

**Identical:** Labels are nearly identical or synonymous.  
*Example:* Dual citizenship crisis AND Section 44 crisis.

**Closely Aligned:** Labels are related but describe different aspects of the matter.  
*Example:* Change-the-date campaign AND Australia Day debate.

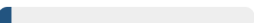
**Somewhat Aligned:** The topic label is very general acting as a category the collection label would fit under.  
*Example:* Political scandal AND Politicians extramarital affairs.

**Loosely Aligned:** Labels are related, but the topic label does not accurately describe the document collection.  
*Example:* Adani coal mine AND Anti-fracking protests.

**Not Aligned:** The labels are not aligned.  
*Example:* Racial Vilification AND Climate Change

Note that common words or synonyms which are contextually dissimilar do not align topics.

	Not Aligned	Loosely Aligned	Somewhat Aligned	Closely Aligned	Identical
Election AND Opinion Poll results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Climate Science AND Arctic warming	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Media AND Investigative Journalism	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Industrial Relations AND Train driver strikes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
National Identity AND Republican movement	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Addiction services AND Cannabis decriminalisation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

0%  100%



 

Figure 5: Example question asking SME to rate how aligned a topic word-set label was to topic document-collection label.

## B Evaluation 1: Coherence Measures

### Pearson’s correlation coefficients between paired coherence measures

LDA AP	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI}$ and $C_{NPMI-ABC}$	0.671	0.034	0.749	< 0.01	0.762	< 0.01	0.727	< 0.01
$C_{NPMI}$ and $C_{U_{mass}}$	0.661	0.037	0.752	< 0.01	0.579	< 0.01	0.447	< 0.01
$C_{NPMI}$ and $C_{NPMI-AP}$	0.383	0.275	0.394	0.086	0.258	0.108	0.256	0.048
$C_{NPMI}$ and $C_P$	0.897	< 0.01	0.912	< 0.01	0.919	< 0.01	0.878	< 0.01
$C_{NPMI}$ and $C_A$	0.692	0.027	0.523	0.018	0.547	< 0.01	0.556	< 0.01
$C_{NPMI}$ and $C_V$	0.127	0.726	0.117	0.624	-0.071	0.663	0.089	0.498
$C_{NPMI-ABC}$ and $C_{NPMI-AP}$	0.538	0.109	0.541	0.014	0.343	0.030	0.375	0.003

MetaLDA AP	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$r_p$	$p$
$C_{NPMI}$ and $C_{NPMI-ABC}$	0.778	0.008	0.783	< 0.01	0.625	< 0.01	0.651	< 0.01
$C_{NPMI}$ and $C_{U_{mass}}$	0.834	0.003	0.738	< 0.01	0.715	< 0.01	0.590	< 0.01
$C_{NPMI}$ and $C_{NPMI-AP}$	0.523	0.121	0.096	0.686	-0.084	0.606	0.005	0.972
$C_{NPMI}$ and $C_P$	0.929	< 0.01	0.959	< 0.01	0.949	< 0.01	0.923	< 0.01
$C_{NPMI}$ and $C_A$	0.788	0.007	0.435	0.055	0.623	< 0.01	0.505	< 0.01
$C_{NPMI}$ and $C_V$	-0.471	0.170	-0.078	0.742	-0.284	0.075	-0.145	0.269
$C_{NPMI-ABC}$ and $C_{NPMI-AP}$	0.100	0.784	0.190	0.423	0.125	0.443	0.172	0.188

Table 3: Pearson’s  $r$  and  $p$ -values reported for the analysis of correlations between coherence measures for the AP dataset

### Aggregate mean for coherence measures

LDA	AP	AWH	AWM
$C_{NPMI}$ and $C_{NPMI-ABC}$	0.727±0.040	0.719±0.050	0.769±0.115
$C_{NPMI}$ and $C_{U_{mass}}$	0.601±0.129	0.395±0.406	0.419±0.088
$C_{NPMI}$ and $C_{NPMI-AP}$	0.323±0.076	0.507±0.189	0.387±0.123
$C_{NPMI}$ and $C_P$	0.902±0.018	0.779±0.101	0.855±0.043
$C_{NPMI}$ and $C_A$	0.578±0.076	0.391±0.094	0.626±0.069
$C_{NPMI}$ and $C_V$	0.066±0.092	-0.108±0.053	0.253±0.118
$C_{NPMI-ABC}$ and $C_{NPMI-AP}$	0.449±0.105	0.565±0.056	0.423±0.073

MetaLDA	AP	AWH	AWM
$C_{NPMI}$ and $C_{NPMI-ABC}$	0.709±0.083	0.606±0.126	0.716±0.104
$C_{NPMI}$ and $C_{U_{mass}}$	0.719±0.100	0.539±0.086	0.272±0.249
$C_{NPMI}$ and $C_{NPMI-AP}$	0.135±0.269	0.258±0.267	0.181±0.153
$C_{NPMI}$ and $C_P$	0.940±0.017	0.770±0.149	0.884±0.037
$C_{NPMI}$ and $C_A$	0.588±0.154	0.360±0.183	0.285±0.187
$C_{NPMI}$ and $C_V$	-0.245±0.174	-0.138±0.273	0.087±0.217
$C_{NPMI-ABC}$ and $C_{NPMI-AP}$	0.147±0.042	0.362±0.223	0.390±0.159

Table 6: The aggregate mean Pearson’s correlation coefficient for LDA and MetaLDA across all topics.

LDA AWH	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI}$ and $C_{NPMI-ABC}$	0.794	0.006	0.702	0.01	0.693	< 0.01	0.688	< 0.01
$C_{NPMI}$ and $C_{Umass}$	0.714	0.020	-0.193	0.414	0.454	0.003	0.606	< 0.01
$C_{NPMI}$ and $C_{NPMI-AP}$	0.746	0.013	0.545	0.013	0.438	0.005	0.297	0.021
$C_{NPMI}$ and $C_P$	0.628	0.052	0.816	< 0.01	0.846	< 0.01	0.826	< 0.01
$C_{NPMI}$ and $C_A$	0.315	0.375	0.316	0.174	0.421	0.007	0.511	< 0.01
$C_{NPMI}$ and $C_V$	-0.093	0.798	-0.176	0.459	-0.112	0.490	-0.049	0.709
$C_{NPMI-ABC}$ and $C_{NPMI-AP}$	0.586	0.075	0.619	0.004	0.567	< 0.01	0.488	< 0.01

MetaLDA AWH	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$r_p$	$p$
$C_{NPMI}$ and $C_{NPMI-ABC}$	0.719	0.019	0.613	0.004	0.428	0.006	0.663	< 0.01
$C_{NPMI}$ and $C_{Umass}$	0.593	0.071	0.459	0.042	0.473	0.002	0.631	< 0.01
$C_{NPMI}$ and $C_{NPMI-AP}$	0.450	0.192	0.476	0.034	0.204	0.208	-0.098	0.458
$C_{NPMI}$ and $C_P$	0.547	0.102	0.844	< 0.01	0.855	< 0.01	0.835	< 0.01
$C_{NPMI}$ and $C_A$	0.089	0.808	0.407	0.075	0.469	0.002	0.476	< 0.01
$C_{NPMI}$ and $C_V$	0.034	0.925	-0.491	0.028	0.115	0.479	-0.209	0.108
$C_{NPMI-ABC}$ and $C_{NPMI-AP}$	0.489	0.151	0.580	0.007	0.303	0.058	0.076	0.565

Table 4: Pearson's  $r$  and  $p$ -values reported for the analysis of coherence measures correlations for the AWH dataset

### Graphs of aggregate coherence measures for LDA vs MetaLDA

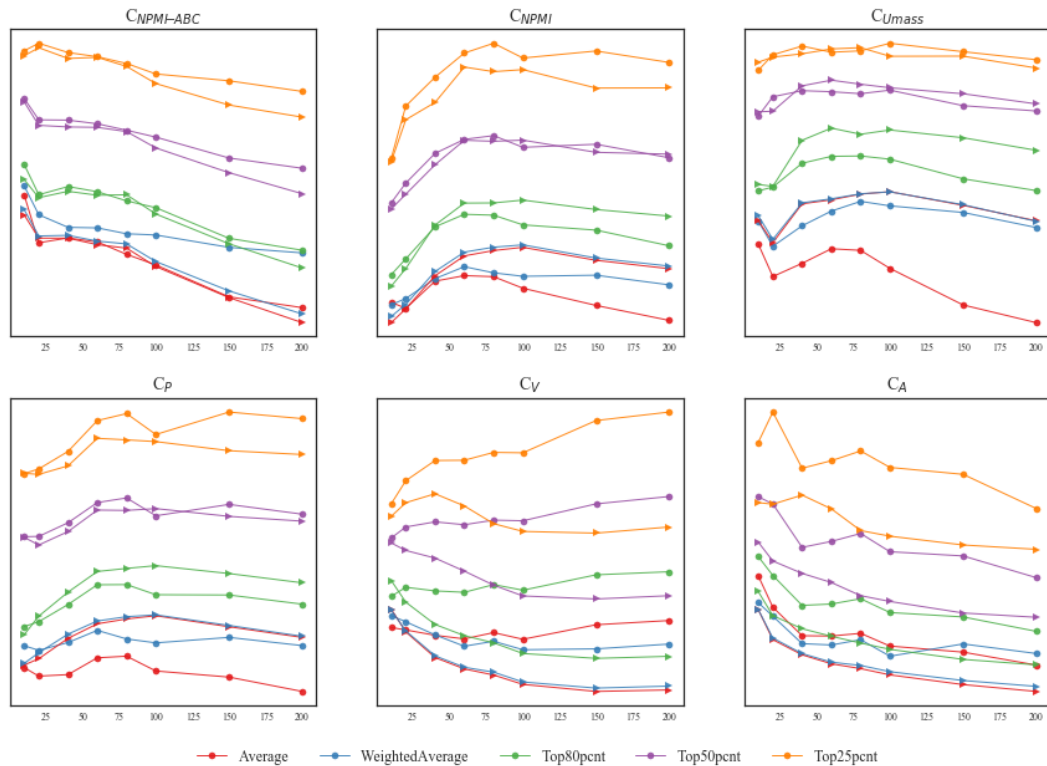


Figure 6: Comparison of LDA (triangle) and MetaLDA (circle) aggregated coherence scores for the AWH dataset. Scores are shown on the y-axis and  $K$  is shown on the x-axis.

LDA AWM	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$r_p$	$p$
$C_{NPMI}$ and $C_{NPMI-ABC}$	0.894	< 0.01	0.821	< 0.01	0.734	< 0.01	0.627	< 0.01
$C_{NPMI}$ and $C_{Umass}$	0.512	0.131	0.450	0.047	0.303	0.058	0.410	< 0.01
$C_{NPMI}$ and $C_{NPMI-AP}$	0.294	0.409	0.389	0.090	0.560	< 0.01	0.304	0.018
$C_{NPMI}$ and $C_P$	0.898	< 0.01	0.886	< 0.01	0.816	< 0.01	0.821	< 0.01
$C_{NPMI}$ and $C_A$	0.725	0.018	0.584	0.007	0.620	< 0.01	0.576	< 0.01
$C_{NPMI}$ and $C_V$	0.341	0.335	0.105	0.660	0.354	0.025	0.210	0.108
$C_{NPMI-ABC}$ and $C_{NPMI-AP}$	0.364	0.301	0.391	0.088	0.528	< 0.01	0.407	< 0.01

MetaLDA AWM	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI}$ and $C_{NPMI-ABC}$	0.866	< 0.01	0.687	0.001	0.684	< 0.01	0.625	< 0.01
$C_{NPMI}$ and $C_{Umass}$	0.302	0.396	-0.080	0.736	0.500	0.001	0.366	< 0.01
$C_{NPMI}$ and $C_{NPMI-AP}$	0.289	0.418	0.133	0.576	0.315	0.048	-0.015	0.912
$C_{NPMI}$ and $C_P$	0.939	< 0.01	0.860	< 0.01	0.861	< 0.01	0.875	< 0.01
$C_{NPMI}$ and $C_A$	0.080	0.825	0.333	0.151	0.207	0.201	0.518	< 0.01
$C_{NPMI}$ and $C_V$	-0.109	0.765	0.384	0.094	-0.033	0.839	0.105	0.425
$C_{NPMI-ABC}$ and $C_{NPMI-AP}$	0.417	0.231	0.516	0.020	0.469	0.002	0.160	0.222

Table 5: Pearson's  $r$  and  $p$ -values reported for the analysis of correlations between coherence measures for the AWM dataset

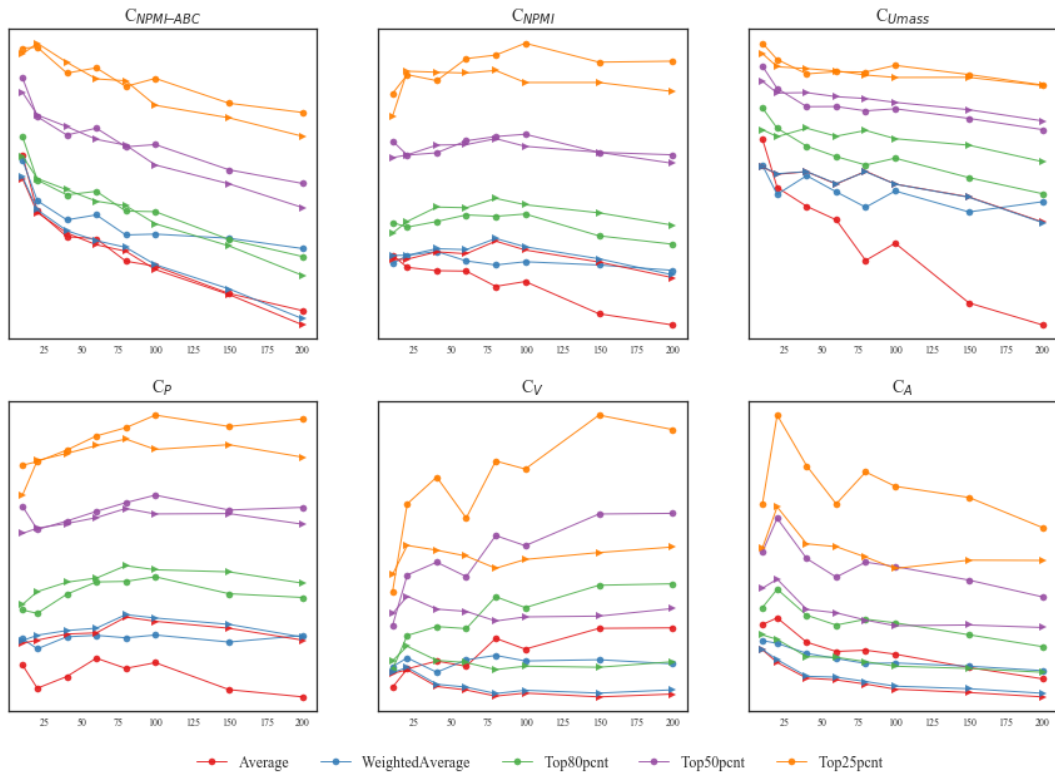


Figure 7: Comparison of LDA (triangle) and MetaLDA (circle) aggregated coherence scores for the AWM dataset. Scores are shown on the y-axis and  $K$  is shown on the x-axis.

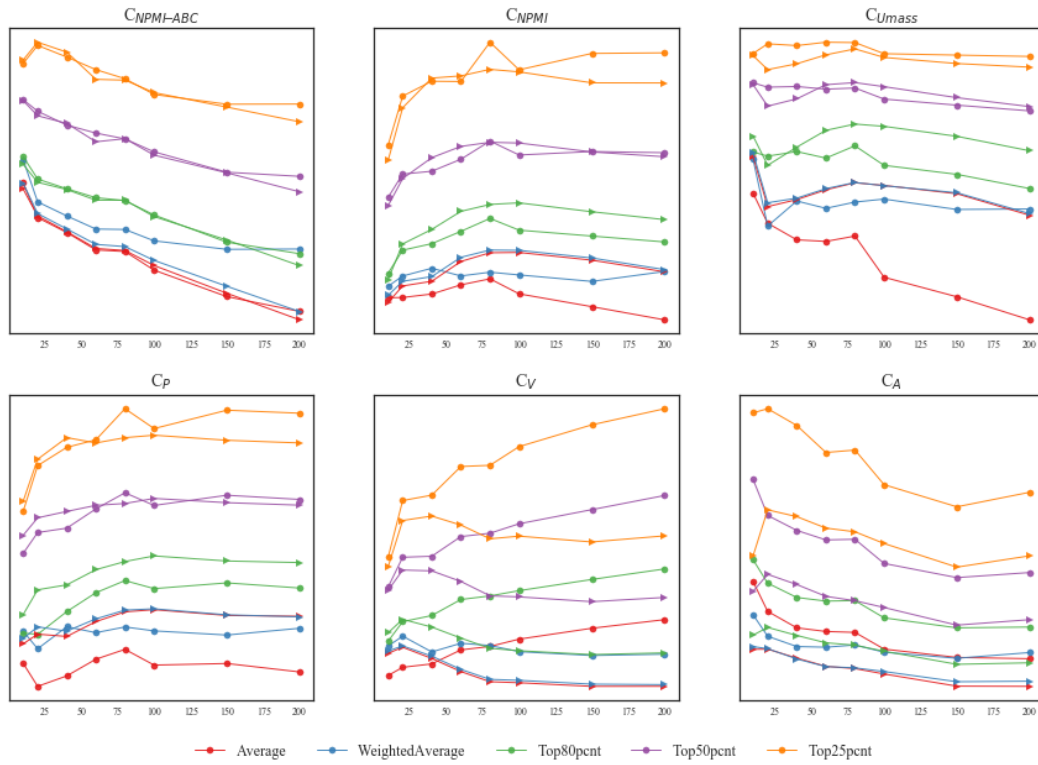


Figure 8: Comparison of LDA (triangle) and MetaLDA (circle) aggregated coherence scores for the AWHM dataset. Scores are shown on the y-axis and  $K$  is shown on the x-axis.

## C Evaluation 2: Labeling Topics

The Spearman's  $\rho$  correlation coefficients for pairwise combinations of  $Q(nbr)$  and coherence measures for all learned models.

LDA AP	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI}$	0.058	0.873	0.264	0.261	0.144	0.374	-0.165	0.209
$C_{NPMI-ABC}$	0.522	0.122	0.379	0.099	0.245	0.128	-0.038	0.774
$C_{Umass}$	0.522	0.122	0.104	0.663	-0.022	0.892	-0.384	< 0.01
$C_{NPMI-AP}$	0.406	0.244	0.626	0.003	0.638	< 0.01	0.217	0.096
$C_P$	0.174	0.631	0.355	0.125	0.245	0.128	-0.062	0.636
$C_A$	0.406	0.244	0.409	0.073	0.383	0.015	0.246	0.058
$C_V$	0.058	0.873	0.138	0.562	0.254	0.114	0.472	< 0.01
$D_{tp}$	-0.174	0.631	0.046	0.846	0.017	0.919	0.056	0.671
$D_{ew}$	-0.522	0.122	-0.814	< 0.01	-0.535	< 0.01	-0.223	0.087

MetaLDA AP	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI}$	0.623	0.054	-0.069	0.774	0.253	0.115	0.336	< 0.01
$C_{NPMI-ABC}$	0.337	0.340	0.096	0.686	0.098	0.547	0.403	< 0.01
$C_{Umass}$	0.623	0.054	-0.302	0.195	0.099	0.544	0.348	< 0.01
$C_{NPMI-AP}$	0.450	0.192	0.054	0.822	0.220	0.172	0.329	0.010
$C_P$	0.623	0.054	-0.088	0.714	0.339	0.032	0.307	0.017
$C_A$	0.701	0.024	-0.022	0.927	0.600	< 0.01	0.304	0.018
$C_V$	-0.545	0.103	0.320	0.169	0.172	0.289	0.118	0.371
$D_{tp}$	0.078	0.831	-0.250	0.288	0.069	0.670	0.230	0.077
$D_{ew}$	-0.017	0.962	-0.211	0.372	-0.082	0.616	-0.099	0.45

Table 7: Spearman’s  $\rho$  and  $p$ -values reported for the analysis of correlations between coherence measures and  $Q_{(nbr)}$  for the AP dataset

LDA AWH	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI}$	0.000	nan	0.371	0.107	-0.045	0.784	0.101	0.441
hline $C_{NPMI-ABC}$	0.000	nan	0.328	0.158	0.160	0.325	0.181	0.165
$C_{Umass}$	0.000	nan	-0.280	0.232	-0.078	0.634	-0.003	0.980
$C_{NPMI-AP}$	0.000	nan	0.182	0.443	0.355	0.025	0.216	0.097
$C_P$	0.000	nan	0.375	0.104	-0.016	0.921	0.066	0.615
$C_A$	0.000	nan	0.103	0.665	0.179	0.270	0.319	0.013
$C_V$	0.000	nan	0.043	0.857	0.192	0.235	0.235	0.071
$D_{tp}$	0.000	nan	-0.131	0.581	0.024	0.884	0.193	0.140
$D_{ew}$	0.000	nan	-0.589	< 0.01	-0.345	0.029	-0.157	0.230

MetaLDA AWH	10		20		40		60	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI}$	-0.522	0.122	0.074	0.758	-0.051	0.753	0.394	< 0.01
$C_{NPMI-ABC}$	-0.406	0.244	0.217	0.357	0.401	0.010	0.305	0.018
$C_{Umass}$	-0.290	0.416	0.127	0.594	-0.039	0.813	0.297	0.021
$C_{NPMI-AP}$	-0.174	0.631	0.676	< 0.01	0.490	< 0.01	0.011	0.934
$C_P$	-0.522	0.122	-0.031	0.897	-0.087	0.595	0.392	< 0.01
$C_A$	-0.406	0.244	0.322	0.166	0.240	0.135	0.404	< 0.01
$C_V$	0.290	0.416	-0.088	0.713	0.338	0.033	-0.013	0.920
$D_{tp}$	0.174	0.631	0.200	0.399	0.132	0.417	0.321	0.0120
$D_{ew}$	0.058	0.873	-0.255	0.279	-0.374	0.017	-0.049	0.708

Table 8: Spearman’s  $\rho$  and  $p$ -values reported for the analysis of correlations between coherence measures and  $Q_{(nbr)}$  for the AWH dataset

<b>LDA</b>	10		20		40		60	
<b>AWM</b>	$\rho, p$	$\rho, p$	$\rho, p$	$\rho, p$	$\rho, p$	$\rho, p$	$\rho, p$	$\rho, p$
$C_{NPMI}$	0.142	0.696	-0.039	0.871	0.092	0.573	0.014	0.914
$C_{NPMI-ABC}$	0.321	0.366	0.063	0.792	0.141	0.386	0.207	0.112
$C_{Umass}$	-0.142	0.696	0.015	0.952	0.083	0.609	-0.314	0.015
$C_{NPMI-AP}$	0.321	0.366	0.285	0.223	0.323	0.042	0.295	0.022
$C_P$	0.306	0.390	-0.119	0.619	-0.007	0.968	-0.045	0.730
$C_A$	0.350	0.321	-0.056	0.816	0.237	0.140	0.286	0.027
$C_V$	0.634	0.049	-0.099	0.677	0.149	0.357	0.228	0.080
$D_{tp}$	0.007	0.984	0.312	0.180	0.059	0.719	0.432	< 0.01
$D_{ew}$	-0.500	0.141	-0.508	0.022	-0.408	0.009	-0.203	0.120

<b>MetaLDA</b>	10		20		40		60	
<b>AWM</b>	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI}$	0.151	0.678	0.187	0.429	0.492	< 0.01	0.407	< 0.01
$C_{NPMI-ABC}$	0.243	0.499	-0.002	0.993	0.228	0.158	0.159	0.224
$C_{Umass}$	0.125	0.732	-0.289	0.216	-0.067	0.680	-0.015	0.908
$C_{NPMI-AP}$	0.321	0.365	-0.099	0.678	-0.055	0.734	0.047	0.721
$C_P$	0.282	0.430	0.153	0.520	0.527	< 0.01	0.481	< 0.01
$C_A$	-0.164	0.651	-0.085	0.722	0.053	0.747	0.267	0.039
$C_V$	-0.164	0.651	0.332	0.153	0.053	0.745	0.220	0.091
$D_{tp}$	0.085	0.815	0.004	0.986	0.594	< 0.01	0.372	< 0.01
$D_{ew}$	-0.125	0.732	-0.040	0.868	0.274	0.087	0.224	0.085

Table 9: Spearman’s  $\rho$  and  $p$ -values reported for the analysis of correlations between coherence measures and  $Q_{(nbr)}$ , for the AWM dataset

## D Evaluation 3: Topic Label Agreement

### Inter-coder Reliability results

Kripp. alpha	10		20		40		60	
	LDA	Meta	LDA	Meta	LDA	Meta	LDA	Meta
AP	0.398	0.363	0.250	0.361	0.402	0.361	0.584	0.486
AWH	0.283	0.391	0.294	0.327	0.368	0.405	0.512	0.498
AWM	0.267	0.344	0.323	0.322	0.366	0.361	0.513	0.447
Fleiss kappa	10		20		40		60	
	LDA	Meta	LDA	Meta	LDA	Meta	LDA	Meta
AP	0.387	0.363	0.156	0.332	0.411	0.344	0.578	0.485
AWH	0.265	0.406	0.290	0.305	0.381	0.371	0.527	0.515
AWM	0.258	0.394	0.321	0.353	0.362	0.356	0.535	0.492
$Q_{agr}$	10		20		40		60	
	LDA	Meta	LDA	Meta	LDA	Meta	LDA	Meta
AP	0.417	0.433	0.167	0.292	0.342	0.283	0.503	0.492
AWH	0.283	0.400	0.258	0.258	0.286	0.296	0.439	0.411
AWM	0.217	0.250	0.275	0.292	0.296	0.279	0.428	0.369

Table 10: The ICR for labels of each topic set using Krippendorff’s  $\alpha$ , Fleiss’  $\kappa$ , and Percentage Agreement  $Q_{agr}$

## E Evaluation 4: Ease of Labeling Collections

### Difficulty labeling document-collections

This section presents all the correlations with  $Q_{aln}$  and  $Q_{dif}$ .

	LDA		MetaLDA		LDA		MetaLDA	
	$Q_{aln}$		$Q_{aln}$		$Q_{dif}$		$Q_{dif}$	
<b>All</b>	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI-ABC}$	-0.0423	0.7486	0.141	0.283	-0.1454	0.2676	0.1688	0.197
$C_{NPMI-AP}$	0.3394	0.008	0.39	0.002	0.2224	0.0877	0.3694	0.004
$C_{NPMI}$	-0.2156	0.0981	0.148	0.259	-0.1291	0.3257	0.2342	0.072
$C_A$	0.123	0.3491	0.297	0.021	-0.0672	0.61	0.2276	0.080
$C_P$	-0.1016	0.4397	0.226	0.082	-0.0712	0.5887	0.1334	0.310
$C_V$	0.2376	0.0676	0.147	0.264	-0.079	0.5485	0.2321	0.074
$C_{Umass}$	-0.397	0.0017	0.029	0.827	-0.1671	0.2018	0.0131	0.921
Proportion	0.0717	0.5861	0.09	0.493	0.0037	0.9774	0.0536	0.684
Effwords	-0.2657	0.0402	-0.239	0.066	-0.1528	0.2438	-0.285	0.027
<b>Top25pct</b>	LDA		MetaLDA		LDA		MetaLDA	
	$Q_{aln}$		$Q_{aln}$		$Q_{dif}$		$Q_{dif}$	
	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI-ABC}$	0.8245	0.0002	0.08	0.778	0.1643	0.5586	-0.1956	0.485
$C_{NPMI-AP}$	0.4838	0.0677	0.11	0.697	0.2918	0.2914	-0.2999	0.278
$C_{NPMI}$	0.5904	0.0205	0.347	0.206	0.0545	0.8469	0.2962	0.284
$C_A$	-0.3135	0.2552	-0.014	0.961	-0.3557	0.1932	-0.3181	0.248
$C_P$	0.568	0.0272	0.236	0.397	0.1609	0.5667	0.24	0.389
$C_V$	-0.3442	0.209	0.119	0.674	-0.2046	0.4645	0.4085	0.131
$C_{Umass}$	-0.2858	0.3017	-0.326	0.235	-0.2283	0.4132	-0.3965	0.143
$D_{tp}$	-0.0897	0.7505	-0.592	0.02	0.2163	0.4388	-0.5307	0.042
$D_{ew}$	-0.1391	0.621	-0.776	0.001	-0.1171	0.6778	-0.7638	0.001



	LDA		MetaLDA		LDA		MetaLDA	
	$Q_{aln}$		$Q_{aln}$		$Q_{dif}$		$Q_{dif}$	
<b>Top50pct</b>	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI-ABC}$	0.079	0.6781	0.213	0.259	-0.0166	0.9308	0.0862	0.651
$C_{NPMI-AP}$	0.2038	0.2801	0.045	0.815	0.2194	0.2441	-0.1221	0.521
$C_{NPMI}$	0.3603	0.0505	0.227	0.228	0.0184	0.9233	0.1811	0.338
$C_A$	-0.0986	0.6041	0.12	0.528	-0.1881	0.3194	0.0143	0.940
$C_P$	-0.0628	0.7415	0.342	0.065	-0.1011	0.5949	0.221	0.241
$C_V$	0.0694	0.7157	-0.217	0.248	-0.1216	0.5223	0.019	0.921
$C_{Umass}$	-0.4059	0.026	-0.023	0.904	-0.3119	0.0933	0.1051	0.581
$D_{tp}$	0.2278	0.2261	-0.338	0.068	0.0306	0.8726	-0.256	0.172
$D_{ew}$	-0.2545	0.1748	-0.65	0	-0.2132	0.2581	-0.6298	0.000

	LDA		MetaLDA		LDA		MetaLDA	
	$Q_{aln}$		$Q_{aln}$		$Q_{dif}$		$Q_{dif}$	
<b>Bot15pct</b>	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI-ABC}$	0.5317	0.1407	0.037	0.924	0.8165	0.0072	-0.2282	0.555
$C_{NPMI-AP}$	0.1581	0.6845	-0.091	0.815	0.1862	0.6315	-0.0797	0.839
$C_{NPMI}$	-0.0851	0.8276	0	1	0.2294	0.5527	-0.0913	0.815
$C_A$	0.0769	0.844	-0.159	0.682	-0.1491	0.7019	-0.01	0.980
$C_P$	-0.4473	0.2274	-0.169	0.663	-0.1101	0.778	0	1.000
$C_V$	0.2946	0.4416	0.356	0.347	-0.2092	0.5891	0.2926	0.445
$C_{Umass}$	0.4873	0.1833	-0.186	0.631	-0.3119	0.0933	0	1.000
$D_{tp}$	-0.523	0.1486	0.523	0.149	-0.3486	0.3579	0.2739	0.476
$D_{ew}$	0.2305	0.5507	0.693	0.039	0.3578	0.3444	0.2635	0.493

	LDA		MetaLDA		LDA		MetaLDA	
	$Q_{aln}$		$Q_{aln}$		$Q_{dif}$		$Q_{dif}$	
<b>Bot10pct</b>	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$	$\rho$	$p$
$C_{NPMI-ABC}$	0.6377	0.1731	0.463	0.355	0.6768	0.1398	-0.0926	0.862
$C_{NPMI-AP}$	-0.0304	0.9545	0	1	-0.206	0.6954	0.1014	0.848
$C_{NPMI}$	-0.3189	0.5379	0.44	0.383	-0.0304	0.9545	0.8783	0.021
$C_A$	0.239	0.6483	0.101	0.848	-0.2	0.6059	0	1.000
$C_P$	0.7537	0.0835	-0.44	0.383	0.8024	0.0547	0.0976	0.854
$C_V$	0.1543	0.7704	0.741	0.092	-0.3719	0.4679	0.7407	0.092
$C_{Umass}$	0.7537	0.0835	-0.216	0.681	0.8024	0.0547	0.414	0.414
$D_{tp}$	-0.8452	0.0341	0.828	0.042	-0.8452	0.0341	0.6831	0.135
$D_{ew}$	-0.0883	0.8679	0.82	0.046	-0.4414	0.3809	0.4938	0.320

## F Examples

Examples of poorly aligned topics are shown in Table 11.

Topic Label	Collection Label	Topic	NPMI
Gun Control	Foreign interference act	law, bill, power, gun, democracy, control, freedom, rule, protect, legislation	0.0734
Cost of Living	Politician's rental property	house, free, property, home, rent, pay, live, buy, move, money	0.0814
Addiction help	Legalization of drugs	health, drug, care, test, medical, doctor, access, alcohol, live, death	0.0702

Table 11: Topics which did not align well with the document-collection despite having a high coherence.