Evaluating Unsupervised Hierarchical Topic Models Using a Labeled Dataset

Judicael Poumay
ULiege/HEC Liege
Rue Louvrex 14, 4000 Liege, Belgium
judicael.poumay@uliege.be

Ashwin Ittoo
ULiege/HEC Liege
Rue louvrex 14, 4000 Liege, Belgium
ashwin.ittoo@uliege.be

Abstract

Topic models are often evaluated with measures such as perplexity and topic coherence. However, these methods fall short in determining the comprehensiveness of identified topics. This research introduces a complementary approach to evaluating unsupervised topic models using a labeled dataset. By training hierarchical topic models and utilizing known labels for evaluation, we found a high accuracy of 70% for expected topics. Despite having 90 labels in the dataset, even those representing only 1% of the data achieved an average accuracy of 37.9%, illustrating hierarchical topic models' effectiveness on smaller subsets. Additionally, we confirmed that this new evaluation method helps assess the topic tree quality, demonstrating that hierarchical topic models generate coherent taxonomies. Lastly, we established that coherence measures alone are insufficient for a holistic topic model evaluation.

1 Introduction

Hierarchical Topic Models such as the LSHTM(Pujara and Skomoroch, 2012), nCRP(Blei et al., 2004), nHDP(Paisley et al., 2015), and HTMOT(Poumay and Ittoo, 2021) enable the extraction of topics and sub-topics organized in a tree-like hierarchy. Topic hierarchies provide a more fine-grained view of the underlying data, which is particularly useful in applications such as ontology learning (Zhu et al., 2017) and research idea recommendation(Wang et al., 2019). Additionally, models like nCRP, NHDP, and HTMOT dynamically determine the appropriate number of topics and sub-topics during training, contrary to the traditional model of LDA(Blei et al., 2003).

Evaluating the quality of the extracted topics is crucial to ascertain their real-world utility. However, as these methods extract knowledge in an unsupervised manner, previous studies on topic model evaluation have been limited to evaluating the quality of the resulting topics. Hence, many methods have been proposed to study the performance of these models, such as perplexity and coherence measures (Newman et al., 2010; Doogan and Buntine, 2021a; Bhatia et al., 2017).

Nevertheless, these measures have proven to be unrelated to human judgment (Chang et al., 2009; Doogan and Buntine, 2021b; Bhatia et al., 2017), indicating that humans do not agree with these measures when it comes to the quality of the topics extracted. Recently, the word intrusion task has been proposed to evaluate the extracted topic quality (Chang et al., 2009). While its initial implementation relies on human annotators, it can be automated without losing the link to human judgment (Lau et al., 2014).

However, all the methods previously presented have failed to ask other essential questions about the extracted topics and the completeness of the results. For example: Do we extract every topic? How well do we extract them? Do we extract unexpected topics? And in the context of hierarchical topic models, is the hierarchy produced coherent?

Hence, in this article, we propose a method for evaluating topic models using a well-known labeled dataset (Reuters-21578 (Tekn, 2020)), but the method can be extended to another dataset. Our approach differs from previous methods by focusing on known topics that we expect to extract and their quality, providing a better understanding of the completeness of the model. Using known labels, we can automatically name extracted topics. Afterward, we can study whether the document topic distribution can predict the actual labels of the documents. We call this *label accuracy*, and it provides a quantitative assessment of how well we fit the training set. Moreover, if more topics are extracted than expected, we can study their relevance

Proceedings of Recent Advances in Natural Language Processing, pages 846–853
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_091

and unexpectedness. Finally, as the extracted topics exist in a hierarchy, we can analyze the coherence of the taxonomy produced from the known labels.

To perform our experiments, we trained 60 different models (30 hierarchical and 30 flat models) with various hyperparameters to understand how and if this new evaluation approach can help us determine quantitatively which model provides the best topics.

Results show that label accuracy provides a more conservative measure of topic quality compared to coherence. We show that while low coherence (Newman et al., 2010) is a good indicator of poor quality in topics, a high coherence score is not sufficient to determine the quality of a set of topics. We also compute the label accuracy for labels that account for less than 1% of the data and demonstrate that it is a good metric if we care about extracting small sub-topics. Precisely, we see that although we have 90 labels, the accuracy of small topics can get as high as 37.9%, while the largest topics achieve more than 70% accuracy. In that sense, we have noticed a logarithmic relationship between the number of documents per label and its accuracy, as accuracy quickly goes up with the number of documents, indicating that hierarchical topic models can extract small topics effectively.

2 Background and Related Work

2.1 Topic Models

LDA (Blei et al., 2003) is the first traditional topic model. At the core of LDA is a Bayesian generative model with two Dirichlet distributions, respectively for the document-topic distributions and for the topic-word distributions. These distributions are learned and optimized via an inference procedure which enables topics to be extracted. The main weakness of LDA is that it requires the user to specify a predefined number of topics to be extracted. The subsequent HDP (Teh et al., 2006) model uses Dirichlet Processes to determine the number of topics during training.

Since then, many hierarchical topic models have been proposed (Pujara and Skomoroch, 2012; Mimno et al., 2007; Blei et al., 2004; Paisley et al., 2015; Poumay and Ittoo, 2021). These are models that extract topics and sub-topics resulting in a topic hierarchy that provides a deeper understanding of the underlying themes inside a corpus. Simple approaches like LSHTM (Pujara and Skomoroch, 2012) recursively apply LDA to a corpus. Therefore, it suffers from the same weakness as LDA, as the topic tree dimension must be decided in advance. Models like nCRP, nHDP, and HTMOT (Blei et al., 2004; Paisley et al., 2015; Poumay and Ittoo, 2021) use Dirichlet Processes to automatically decide the number of topics to extract during training. Each model is an improvement over the previous one. The nCRP model only allowed documents to sample topics in one branch of the topic tree, while the nHDP lets documents sample from any number of branches. HTMOT followed suit by integrating temporality into the model to extract specific events at the deeper level of the topic tree. Finally, hPAM (Mimno et al., 2007) proposes another approach using a directed acyclic graph structure instead of a tree to model topic hierarchy.

2.2 Evaluating Topic Models

Perplexity has been the standard for comparing topic models for a long time. It defines how likely it is that the training data would have been generated by the trained topic model. However, it has been discovered that this method does not correlate with human judgment (Chang et al., 2009). Hence, new methods for evaluating topics have been proposed, but none have provided a new standard.

Topic coherence (Newman et al., 2010) was also proposed as a method of topic evaluation. This method consists of computing some similarity scores between the top N topic words. Specifically, it is computed as (where w_i is more frequent than w_j): $\sum_{i<j} score(w_i, w_j)$. Topic coherence is a modular evaluation method as it allows for many different scoring functions. The most popular are UCI and UMass, which use word co-occurrence to score word sets. UCI is an extrinsic measure based on Wikipedia articles, while UMass is intrinsic and uses the training corpus. However, other score functions such as the cosine similarity of word embeddings can also be used. The topic coherence score of a model is the average coherence score of the topics. Nevertheless, a recent study puts into question whether coherence measures themselves correlate with human ratings (Newman et al., 2010; Doogan and Buntine, 2021a; Bhatia et al., 2017).

The Word Intrusion task is the latest evaluation method devised. For each topic, it involves inserting an intruder word in the topic top word list and then asking people to find it (Chang et al., 2009). This intruder is selected at random from a pool of words with a low probability in the current topic

but a high probability in some other topic to avoid rare words. The idea is that in good topics, the annotators would easily find this intruder. With this evaluation method, the final score corresponds to the average classification accuracy made by humans.

Finally, all topic modeling methods presented provide a qualitative analysis of the extracted topics. Compared to opaque measures such as coherence and perplexity, the qualitative analysis provides a direct understanding of the model's performance. However, such an evaluation method is prone to cherry-picking, especially when many topics are extracted.

Hence, all of the methods presented have been demonstrated to be unreliable on their own. Moreover, none of these methods here answers our research questions: Do we extract every topic we expect to extract? How well do we extract them? Do we extract unexpected topics? Is the hierarchy produced coherent? Hence, it is clear that we need new tools to evaluate topic models, especially hierarchical ones.

3 Methodology

4 Overview

Our evaluation methodology consists of multiple steps. We aim to assess the sensitivity of the topic models and compare the performance of hierarchical and flat models. To achieve this, we extract topics from our corpus using 60 variations of topic models (30 hierarchical and 30 flat models with different parameters as shown in table 1) by training them on the Reuters dataset. The varying parameters include basic LDA parameters that control the topic-word and document-topic prior distributions, as well as the dynamic parameters controlling the creation of new topics during training.

Following this, we automatically assign labels to the topics by using the known labels from the corresponding dataset, based on the document-topic distribution. Next, for each document with n labels, we compare the top n+k labeled topics for that document to calculate label accuracy. Finally, we evaluate the results.

5 Corpus

For our experiments, we will employ the Reuters-21578 corpus (Tekn, 2020), a widely used dataset in the literature on topic models. Composed of English news articles primarily focused on business and politics, this corpus was used as it has detailed and multiple labels for each document.

We preprocessed the corpus by filtering relevant tokens using Spacy's Named Entity Recognition and Part-of-Speech tags and applied lemmatization. Consequently, our training set consists of 10,788 documents, each labeled with one or more of the 90 tags in the corpus (e.g. wheat, gold, money-fx, etc.).

The label distribution is highly uneven, resembling a power-law distribution, with labels such as 'earn' or 'acq' constituting approximately 36% and 22% of the documents, respectively. In contrast, labels like 'rye' and 'castor-oil' appear only in a single document each.

6 Constructing and Training the Models

In our experiments, we utilized the nHDP and HDP topic models albeit with a distinct training procedure. While the original implementation of these models used Stochastic Variational Inference (SVI), we employ a fast implementation of Gibbs sampling for training (Poumay and Ittoo, 2021). According to (Blei et al., 2017), Gibbs sampling outperforms SVI for small topics. Small topics are crucial since they may represent weak signals in the data, and hierarchical topic models tend to generate more small topics compared to their flat counterparts.

We explored 48 distinct models, training 24 hierarchical models (nHDP) and 24 flat models (HDP). Each hierarchical/flat model pair shares the same set of parameters (refer to table 1).

The parameters that we vary in each model are defined as follows: α: the rate at which we create new topics in the document trees. β: the rate at which we create new topics in the corpus tree. ϕ: the prior for the topic-word distribution. ϵ: the prior for the corpus and document-topic distributions.

These 30 pairs of models are grouped as follows:

- 6 pairs of models with different values for alpha
- 6 pairs of models with different values for beta
- 6 pairs of models with different values for epsilon
- 6 pairs of models with different values for phi

Models	alpha	beta	phi	epsilon
A1	0.000005	0.02	0.1	0.5
A2	0.00001	0.02	0.1	0.5
A3	0.00005	0.02	0.1	0.5
A4	0.0005	0.02	0.1	0.5
A5	0.001	0.02	0.1	0.5
A6	0.005	0.02	0.1	0.5
B1	0.0001	0.001	0.1	0.5
B2	0.0001	0.002	0.1	0.5
B3	0.0001	0.004	0.1	0.5
B4	0.0001	0.1	0.1	0.5
B5	0.0001	0.2	0.1	0.5
B6	0.0001	0.4	0.1	0.5
E1	0.0001	0.02	0.1	0.001
E2	0.0001	0.02	0.1	0.01
E3	0.0001	0.02	0.1	0.02
E4	0.0001	0.02	0.1	0.1
E5	0.0001	0.02	0.1	2.
E6	0.0001	0.02	0.1	5.
P1	0.0001	0.02	0.001	0.5
P2	0.0001	0.02	0.01	0.5
P3	0.0001	0.02	0.02	0.5
P4	0.0001	0.02	0.5	0.5
P5	0.0001	0.02	1.	0.5
P6	0.0001	0.02	5.	0.5

Table 1: Sets of parameters for the models trained

7 Automatic Titling

To automatically assign a label *l* to a topic we used a simple heuristic. For each trained model, we compute the label-topic distribution of label *l* by averaging the document-topic distribution of documents that have this label. If the model is hierarchical, this means we end up with a topic tree with topic frequencies corresponding to this label.

Starting from the root, we select the topic with the highest frequency for that label. We do the same for the sub-topic of the selected topic until we reach a leaf. In the end, we have selected a branch of the tree where the label is most frequent.

Next, we compare the known frequency of the label *l* with each topic of this branch and select the topic with the closest frequency. This topic will be given the label *l*.

This method is applied iteratively for each label. It is worth noting that a topic may have multiple labels in its title if it is selected by several labels.

This heuristic is simple by design and is an important hypothesis that has a large impact on the performance of our evaluation methodology. Nonetheless, we will show that it is sufficient to provide interesting results.

8 Computing Top n+k Label Accuracy

To calculate the top n+k label accuracy, we order labeled topics by their document-topic distribution for each document. Considering that document *d* has n labels, we choose the top n+k topics from the sorted list. We subsequently extract the labels given to these topics. Finally, using the set of extracted labels from the topics T and the use of known labels of the document L, we determine the label accuracy for document *d* using the formula $\frac{|L \cap T|}{|L|}$. The overall top n+k label accuracy of the model is calculated as the average across all documents. The overall top n+k label accuracy of each label *l* is calculated as the average across all documents with that label *l*.

In addition to the overall top n+k label accuracy, we compute the small topic label accuracy, which excludes labels that correspond to more than 1% of the dataset. This exclusion accounts for 80% of the tags, or 72 tags in total.

9 Results

In this section, we will review the results of our experiments. We will start by comparing the coherence measure to the label accuracy measure. Next,

we will compare the performance of the flat and hierarchical models. Finally, we will study the hyperparameters' importance.

9.1 Coherence vs Label Accuracy

In table 2, we display the metrics computed for six of the 7 models. Three were the worst in at least one metric and four were the best in at least one metric. The metrics are the average topic coherence and the top 3 label accuracy. The topic 3 label accuracy is computed for all the labels in each hierarchical model, in their flat counterpart (F), for small topics (S), and for both small topics in the flat model (F/S).

We observe that the model with the worst coherence (P1) did produce topics that are difficult to interpret. However, the model with the highest coherence (E1) is decisively not the best model. The label tree it produces is incoherent and most of the labels are pushed to the leaves of the tree. Consequently, this model has many topics sharing multiple labels indicating that the model could not separate the labels properly. Specifically, 81% of labels share a topic, and one topic shares as many as 34 labels. Moreover, this model created many duplicate topics, with the majority of the topics being similar if not the same. Finally, we can observe that this model also has poor accuracy being the second worst.

The best-performing model is (B5) with the highest small topic accuracy. Although its coherence is lower than (E1), its label tree is much more coherent and detailed. Most labels do not share co-labels meaning that the model is better at separating the labels into specific topics. Specifically, 34% of labels share a topic, and one topic shares as many as 5 labels. B5 being the highest small topic accuracy, we also observed that small labeled topics are easily interpretable.

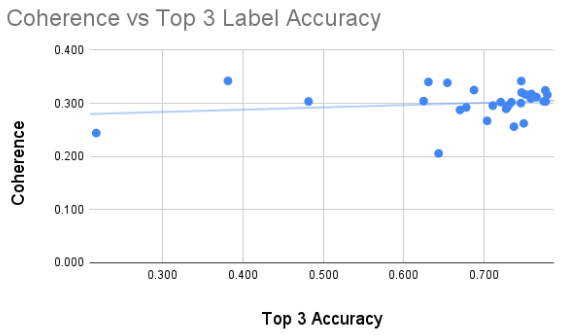

Figure 1: Coherence vs label accuracy across all models

Id	A	A (F)	A (S)	A (S/F)	C
P2	*.218*	*.247*	*.057*	.006	.244
P1	.643	.543	.178	.004	*.206*
A4	.711	.338	.323	*.003*	.296
A2	**.778**	**.590**	.271	.012	.316
E1	.382	.542	.128	.005	**.342**
B5	.727	.350	**.379**	.006	.290
E2	.631	.373	.267	**.018**	.340

Table 2: Comparing best and worst models for each measure. A corresponds to the top 3 label accuracy and C corresponds to the UMass coherence. (F) corresponds to the equivalent flat model performance. (S) corresponds to the small topics' performance.

Tags	Real	B5	P2
nat-gas	proportion	0.89	16
gnp	1.19%	2.15	1.02
coffee	1.49%	1.41	12.09
trade	1.6%	2.25	1.06
crude	5.31%	3.13	6.62
money-fx	6.01%	1.53	6.49
acq	6.91%	20.57	8.6
MSE	24.56%	10.499	63.932

Table 3: Comparing the worst hierarchical topic model (P2) with the best small accuracy topic model on a set of random topics. We compare the real proportion of the tags in the data with the proportion of the topics with that label. We then compute the Mean Square Error (MSE) of this difference for both models.

Hence, the coherence measure is good at determining if a set of topics is of bad quality. However, it is not sufficient in itself to determine if the topics are of good quality. A set of coherent but duplicate topics will yield a high coherence score even if this results in bad topic extraction overall. Moreover, high coherence does not guarantee that topics are well separated or that the inferred hierarchical structure of topics makes sense. Figure 1 shows that both label accuracy and coherence are not highly correlated which indicates they measure a different aspect of a model's performance.

Another way to ensure that the label accuracy represents the model's performance is to look at the discrepancy between the actual label size and the size of the topic with that label. In table 3, we compare the worst and best models for small label accuracy. We see that for the best model, labels correspond to topics with a size that is closer to the actual label size.

We can also compare how the coherence and

label accuracy metrics compare depending on the size of labels or topics. Since coherence is computed for each topic and label accuracy is computed for each label we cannot make a direct comparison. In the figures 3 and 2, we plot these results and observe that there is a logarithmic relationship between label accuracy and size. Indicating that the quality of topics greatly increases with a small increase in the number of documents. This implies that topic models could detect weak signals and emerging trends early as a few documents can produce relatively decent topics. However, for coherence, there is not such a clear relationship between topic size and coherence; the bigger topics do not seem to gain in coherence either. Nonetheless, a qualitative analysis of topics reveals that bigger topics are much easier to interpret.

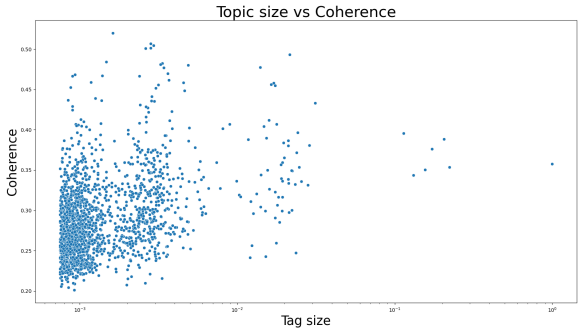

Figure 2: Topic coherence vs size. The x-axis uses a logarithmic scale.

Hence, we have demonstrated that while coherence is good at avoiding bad topics it is not sufficient to select good topic trees. The accuracy of small labels on the other hand provides us with a better understanding of the quality of a topic tree as a whole.

9.2 Flat vs Hierarchical Models

In table 2, we can observe the label accuracy for the flat topic model for all the labels and the small ones. While the label accuracy can get close to 60%, it is mostly a reaction to the highly unbalanced labels in the corpus. Once, we focus on the smaller labels, this accuracy nearly drops to zero. This demonstrates the power of the hierarchical topic model to uncover smaller topics.

As we automatically label topics in a topic tree, we can also observe the coherence of the hierarchy produced. While the original labels are not structured in a hierarchy, we observe that the taxonomy created from the topic makes sense (see figure 4 for a sample). Thus, indicating that hierarchical topic models can produce coherent taxonomy from labeled documents.

9.3 Hyper-Parameter Importance

Finally, we can study the hyper-parameter importance. We observe that ϵ and ϕ are positively correlated with label accuracy which controls document-topic and word-topic distributions, indicating that a more uniform distribution provides a better prior for this dataset. Nonetheless, for coherence higher values for ϕ and lower values for ϵ are preferable. For ϵ this discrepancy is interesting, although we have discussed that the model (E1) with the lowest value for ϵ is one of the worst models qualitatively and in terms of label accuracy.

If we believe in label accuracy, we may conclude that it is better to start with a uniform prior which does not set up the model in any specific local minimum. Indeed, lower values of ϵ or ϕ will lead the model to select some random configuration for these distributions early on before it has been able to see the whole data; this is called the burn-in phase of the Gibbs procedure. On the other hand, starting with a uniform prior distribution forces the model to remain uniform until it has seen enough data that the empirical distribution in the data takes precedence over the prior. However, even higher values for these priors eventually lead to degrading performance since it will eventually have a higher weight than the data itself.

Considering the parameters that control the creation of topics during training. We see that higher β, which controls the rate at which we create new topics in the corpus tree, does not significantly impact label accuracy but does negatively impact coherence. We observe similar results for α: the rate at which we create new topics in the document trees. Except that higher values for α are correlated with higher small label accuracy. Once again, these priors mostly impact the model during the burn-in phase of the Gibbs procedure.

9.4 Do we Extract Unexpected Topics?

While quantitative analysis of topic models is important, it is necessary to remember that such models are not predictive. Hence, part of the reason we use topic models is to discover unexpected topics. It is important to note that while we have 90 labels in the dataset, we extract about 1500 topics on average. Meaning that on average less than 5% of topics receive a label.

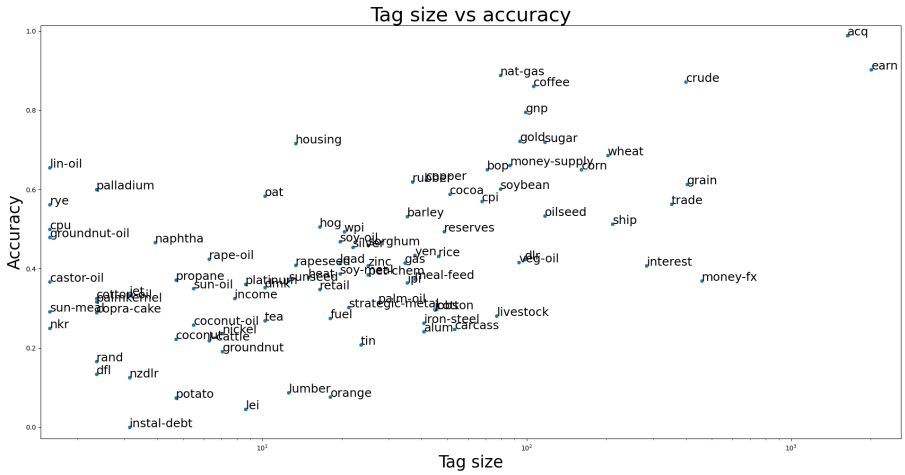

Figure 3: Label accuracy vs size. The x-axis uses a logarithmic scale.

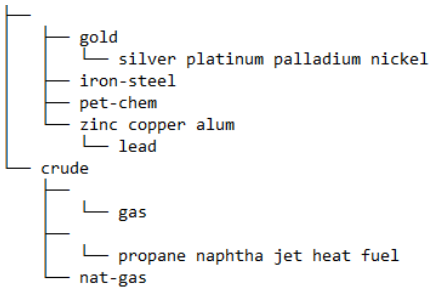

Figure 4: Selected sample of the label hierarchy produced. The entire label tree is too large to be shown entirely.

Ship attack	Ore reserves	Trade dispute
iranian	estimate	semiconductor
attack	reserve	tariff
tanker	property	pact
missile	exploration	sanction
platform	total	impose
war	mining	market
oil	development	japanese
protect	prove	failure
ship	result	chip
shipping	program	computer

Table 4: A selection of small unexpected topics. These topics have a frequency of 0.49%, 1.11%, 0.49% respectively.

Hence, other unexpected topics have been extracted as well. We can look at the small unexpected topics extracted by the B5 model; these topics are displayed in table 4. These topics are not specifically described by any of the labels present in the original dataset.

10 Conclusion

Our study introduces a novel method for evaluating hierarchical topic models based on labeled data. We trained hierarchical topic models on the Reuters-21578 dataset and used the known labels to evaluate the quality of the resulting topics. Our approach differs from previous methods by focusing on known topics that we expect to extract, providing a better understanding of the completeness of the model.

We found that labels with a large number of documents yielded high accuracy above 70%, while smaller labels (1% of the data) had lower accuracy, but remained relatively high for multi-class accuracy with 90 labels at 37.9%. Additionally, we observed a logarithmic relationship between label accuracy and size, indicating that even a small increase in the number of documents could greatly improve the quality of the extracted topics. This suggests that topic models can detect weak signals and emerging trends early, with just a few documents producing relatively decent topics.

Furthermore, we demonstrated that coherence alone is not sufficient to select a good topic tree, and the accuracy of small labels provides a better understanding of the quality of the topic tree. Our approach also allowed us to discover unexpected topics, such as trade disputes or ore reserves, that would have been missed by traditional evaluation methods. Lastly, we have shown that hierarchical topic models produce relatively coherent label taxonomy.

Future research could build on our approach by developing better evaluation methods that consider

not only the quality of topics extracted but also the ability to extract expected topics. Another direction for future research is to measure the unexpectedness of extracted topics since topic models are often used to discover unknown patterns in the data.

References

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2017. An automatic approach for document-level topic model evaluation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 206–215, Vancouver, Canada. Association for Computational Linguistics.

David M Blei, Thomas L Griffiths, Michael I Jordan, and Joshua B Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16(16):17–24.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22, pages 288–296. Curran Associates, Inc.

Caitlin Doogan and Wray Buntine. 2021a. 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.

Caitlin Doogan and Wray Buntine. 2021b. 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.

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.

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, pages 633–640.

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*, HLT '10, page 100–108, USA. Association for Computational Linguistics.

J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. 2015. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270.

Judicael Poumay and Ashwin Ittoo. 2021. HTMOT : Hierarchical Topic Modelling Over Time.

Jay Pujara and Peter Skomoroch. 2012. Large-scale hierarchical topic models. In *NIPS Workshop on Big Learning*, volume 128.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Yaşar Tekn. 2020. Optimization of lda parameters. In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Hei-Chia Wang, Tzu-Ting Hsu, and Yunita Sari. 2019. Personal research idea recommendation using research trends and a hierarchical topic model. *Scientometrics*, 121(3):1385–1406.

Xiaofeng Zhu, Diego Klabjan, and Patrick N. Bless. 2017. Unsupervised terminological ontology learning based on hierarchical topic modeling. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 32–41.