Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6271–6279
Marseille, 20-25 June 2022

A Bayesian Topic Model for Human-Evaluated Interpretability

Justin Wood, Corey Arnold, Wei Wang
University of California, Los Angeles
{juwood03, cwarnold, weiwang}@ucla.edu

Abstract

One desideratum of topic modeling is to produce interpretable topics. Given a cluster of document-tokens comprising a topic, we can order the topic by counting each word. It is natural to think that each topic could easily be labeled by looking at the words with the highest word count. However, this is not always the case. A human evaluator can often have difficulty identifying a single label that accurately describes the topic as many top words seem unrelated. This paper aims to improve interpretability in topic modeling by providing a novel, outperforming interpretable topic model. Our approach combines two previously established subdomains in topic modeling: nonparametric and weakly-supervised topic models. Given a nonparametric topic model, we can include weakly-supervised input using novel modifications to the nonparametric generative model. These modifications lay the groundwork for a compelling setting—one in which most corpora, without any previous supervised or weakly-supervised input, can discover interpretable topics. This setting also presents various challenging sub-problems of which we provide resolutions. Combining nonparametric topic models with weakly-supervised topic models leads to an exciting discovery—a complete, self-contained and outperforming topic model for interpretability.

Keywords: Topic modeling, Weak supervision, Topic interpretability

1. Introduction

Topic modeling is an effective way to analyze unstructured textual data. Although recent topic modeling research is often performed using deep neural networks (Duan and others, 2021; Chen and others, 2021; Rezaee and others, 2020), an alternative technique for topic discovery is based on a Bayesian graphical model (Blei and others, 2003b). The basic assumption of these Bayesian models consists of a generative model for the input text. Words are generated by first sampling a topic assignment from a document-level topic distribution (θ). Then for the topic assignment, a word is generated from the corresponding topic-level word distribution (ϕ). This process is completed over the entire length of the corpus. Inference of the two hidden distributions is made using Bayesian inference techniques such as Gibbs sampling (Griffiths and Steyvers, 2004; Griffiths, 2002).

The topics themselves consist of word assignments from the corpus to each topic. These word assignments are clustered together to form distributions. Since a topic is only a word assignment distribution, it is not always the case that a single n-gram can describe the topic. A topic without a single n-gram description represents a divergence from how a layperson might think of what a *topic* is—which could be: *the subject of a discourse or of a section of a discourse* (Topic., 2021). This divergence is at the center of interpretability. Interpretable topics bridge this gap by providing the cluster of words that can be described with a single label that best explains the topic[1]. For example, if the top 3 words for a topic are: *pitcher, batter, and outfielder*, an interpretable topic may label this topic as *baseball*—which could easily match what a layperson would say the topic is given the top words. This labeling concept is applied to documents as well. It is conceivable for a layperson to list a set of the highest occurring topic-labels in a document when describing what the document is about.

To evaluate interpretability in an empirical manner we argue the best approach is conduct experiments with human evaluators. Other methods that seek to score interpretability utilize a scoring metric based on pointwise mutual information (PMI) and are shown to correlate with human-evaluated tasks on coherence and interpretability (Newman and others, 2010). However, recent work has challenged the goodness of approximate scoring metrics in predicting interpretability (Doogan and Buntine, 2021). Regardless, since the core of PMI-based scoring methods is to approximate how well a layperson can interpret a topic, the approximation will always be weaker than a direct result (human evaluated tasks).

Do Bayesian and neural topic models produce interpretable results? Even though most topic modeling methods do not provide a label comprising their most popular word assignments, one would assume that there would be a semantic coherence among the most assigned words. However, this is not always the case (Chang and others, 2009). A significant reason for this is that the models tend to assign words together that are not semantically connected (Wood and others, 2017). For our baseball example, this may lead to a topic discovered whose top words are: *carrots, batter, and galaxy*. It is hard to place a single n-gram over the topic from this example. From an intuitive perspective, this is not unexpected given the nature of the generative model. No condition is placed upon the words to assure semantic relatedness. These same pitfalls are applicable to neural topic models as well. Additionally neural topic models

[1] Equivalently, an interpretable topic can be thought of as a vector of words that are semantically related.

should theoretically perform poorly on document-level interpretability; since in the implementation of many neural topic models θ is associated with a batch parameter which is often much less than the number of documents (Duan and others, 2021; Chen and others, 2021; Rezaee and others, 2020). Therefore, the same θ distribution is used for different documents, breaking the assumption of the generative model.

Nonparametric topic models do not serve to resolve the deficiencies of interpretability. However, they allow for topic models to be defined over an infinite parameter size. The infinite parameter size represents another advantage of Bayesian topic models over the neural topic model[2]. The unbounded parameter space allows for previous input parameters to be omitted. In nonparametric topic modeling, the left-out parameter is often the number of topics (K). Excluding K can be advantageous since it is somewhat unreasonable to assume the known number of topics a generative model used to create a corpus. Traditionally used numbers are often used (Blei and others, 2003b; Yang and Wang, 2021) by default without much analysis of different topic numbers. Moreover, evaluating models learned with differing number of topics, such as with a log-likelihood comparison, is too time-consuming and thus different topic number consideration is discarded.

The connection between nonparametric topic modeling and interpretability lies with weakly-supervised topic modeling. Weakly-supervised topic models concern themselves with assigning labels to topics. By consequence of its method, it also shapes the discovered topics to its weakly-supervised topic. Weakly-supervised topic models differ from previous approaches that seek to assign a topic label after inference (Jiang and others, 2020; He and others, 2021a; He and others, 2021b; Alokaili and others, 2020). Assignment after inference can lead to somewhat uninterpretable topics as the word assignment cluster representing the topic tend to combine semantically different words (Wood and others, 2017). Another approach is to utilize supervised topic labels (Blei and McAuliffe, 2007; Wang and others, 2021). However, a precise labeled input requirement can be expensive or time-consuming to obtain. A fusion of these two approaches is advanced by weakly-supervised techniques (Wood and others, 2017; Song and others, 2020)—which allows for an easier to obtain labeled input set and can help shape the topics to the labeled input set. A standard weakly-supervised input set involves a *knowledge source* that is a collection of previously labeled articles. These articles are then turned into distributions. The distributions are referred to as *knowledge source topics*.

Weakly-supervised topic models have already been shown in some cases to lead to better topic discovery (Wood and others, 2017). Additionally, there is a foundation for interpretability. If we assume that a topic drawn directly from a confirmed knowledge source is highly interpretable, then it follows that topics discovered by a topic model that are biased by the interpretable knowledge source topics would be interpretable as well. One drawback of weakly-supervised topic models is knowing how many knowledge source-topics to discover. Indeed, the models are not well defined in this matter, resorting to some heuristic for topic elimination during inference (Wood and others, 2017; Hansen and others, 2013). Additionally, the model is such that as the number of knowledge source topics increase, so does the computation time. At knowledge source input levels of just 1,000, the running time is infeasible (Wood and others, 2017).

Hence the context for combining weakly-supervised topic models with nonparametric topic models. If nonparametric models can be constructed to execute for an infinite number of topics, 1,000 topics should be easily attainable. Additionally, we can remove the need to specify the number of known topics in advance, resulting in a more flexible topic model. We further extend this combination by removing the requirement to specify the knowledge source beforehand.

Upon fusing these two domains we notice another discovery: we create a topic model for interpretability. Combining the two models, introduces a parameter that specifies the likelihood that a knowledge source topic is chosen over a regular topic model topic. This parameter acts to increase or decrease the knowledge source topics inferred as a percentage. If we assume knowledge source topics to be highly interpretable, the parameter becomes a way to increase or decrease interpretability by a pre-specified amount. By combining weakly-supervised topic models and nonparametric topic models we contend to have a way to specify the desired level of interpretability.

To reemphasize the novelties of this work, this paper claims three important contributions: (1) a novel combination of weakly-supervised and nonparametric topic models (Section 3.1)—leading to an outperforming (human-evaluated) interpretable topic model (Section 4), (2) the ability to incorporate an exponential size knowledge source (Section 3.2)—existing methods are limited to about 1,000 knowledge source topics (Wood and others, 2017), and (3) a preliminary yet groundbreaking technique to build a knowledge source without any user input (Section 3.3)—thus allowing the application of our method a much more comprehensive set of corpora.

2. Background

2.1. Weakly-supervised Topic Models

Weakly-supervised topic models are a subclass of topic models that are mostly extensions of the Dirichlet-based Bayesian topic model, latent Dirichlet allocation (LDA) (Blei and others, 2003b). One desideratum of Weakly-supervised models is to use a large collection of

[2] Some neural topic models claim to be non-parametric but are actually bounded with a condition to hide a subset of topics (Chen and others, 2021; Ning and others, 2020)

documents that have already been labeled. These documents are formed into topics and serve to bias some subset of the existing LDA topics. The methods for biasing the topic to the LDA topic can be done by setting the topic-level word distributions (ϕ) to the labeled topic document histogram (Hansen and others, 2013), using a distance metric to compute document similarity (Song and others, 2020) or by using the labeled topic histogram as the hyperparameters for a Dirichlet distribution prior over the vocabulary of the corpus (Wood and others, 2017).

Weakly-supervised methods generally assume a modification to only that of the ϕ-distributions. Therefore, when building a Gibbs sampler each model considers some existing posterior density calculation alongside a posterior density that utilizes a predetermined knowledge source. This Gibbs sample takes a general form of:

$$\mathcal{P}_1 = x, i, j, n_z, n_d, \beta, \alpha, K \tag{1}$$

$$P(\vec{z}_i{=}j|\vec{z}_{\text{-}i}, \vec{w}, x, y) \propto f_L(\mathcal{P}_1) \tag{2}$$

which is the traditionally used posterior density (f_L) that can take on different forms (Griffiths and Steyvers, 2004; Wallach, 2008) and for all $j > K$:

$$\hat{T}_j = (\hat{L}_j, \hat{w}_{1,j}, \hat{w}_{2,j}, \ldots, \hat{w}_{\hat{V}_j,j}) \tag{3}$$

$$\hat{X}_j = t_{\hat{X}}(\hat{T}_j) \tag{4}$$

$$\mathcal{P}_2 = x, i, j, n_z, n_d, \beta, \alpha, K, \hat{X}, \hat{T} \tag{5}$$

$$P(\vec{z}_i{=}j|\vec{z}_{\text{-}i}, \vec{w}, x, y) \propto f_S(\mathcal{P}_2) \tag{6}$$

where: $\vec{z}$ is a vector of topic assignments for document x, i is the index of the current token in document x, $\vec{w}$ the vector of words for document x, n_z is the count matrix for each word and each topic, n_z is the count matrix for each topic and each document, β is the symmetric hyperparameter for the word to topic mixtures, α is the symmetric hyperparameter for the topic to document mixture, K is the number of all non-labeled topics, y becomes the index of w_i, $\hat{L}_j$ is the label associated with knowledge source topic j, $\hat{w}_{1,j}$ is the count of word w_1 in knowledge source topic j, V the size of the vocabulary, with $t_{\hat{X}}$ and f_S as a transformation and density function specific to the model.

2.2. Nonparametric topic modeling

Nonparametric topic modeling is based off the hierarchical Dirichlet process (Blei and others, 2003a). These initial techniques interpret the Dirichlet process as a Chinese restaurant franchise, which is an alternate view of the hierarchical Dirichlet process. Inference can be made in a similar manner to Dirichlet distribution topic modeling, using Markov chain Monte Carlo techniques. Later techniques have shown inference between nonparametric and parametric topic modeling to be close to the same (Wood et al., 2021).

A subfield of nonparametric topic modeling is that of hierarchical topic modeling. These techniques seek to find semantically hierarchal topics in a corpus (Blei and others, 2003a; Manouchehri and others, 2021; Chen and others, 2021). These can be based on the Dirichlet process or a similar method such as using a directed acyclic graph (Li and McCallum, 2006; Mimno et al., 2007). These generalizations have been shown to discover meaningful relations among topics.

2.3. Interpretable topic modeling

The interpretability problem in topic models was established by asking humans to find relationships among words comprising topics (Chang and others, 2009). Alternative methods have used more mathematically based methods (PMI) to arrive at a similar conclusion (Newman and others, 2010). With this deficiency established methods have been developed to increase the interpretability of topics. Existing methods can use visualization, careful selection of displayed words or interacting periodically with annotators to increase semantics (Deng and others, 2020; Prasad et al., 2021). Recent methods seem to have shifted to neural based topic models and represent the state-of-the art approach for obtaining high PMI-based scores (Duan and others, 2021; Chen and others, 2021; Rezaee and others, 2020; Ning and others, 2020; Bianchi and others, 2021; Tomasi and others, 2020).

3. Methods

3.1. Nonparametric weakly-supervised model

To introduce our technique of combining nonparametric and weakly-supervised topic modeling we begin with the generative model of a hierarchical Dirichlet process-based topic model of:

$$\theta_d = \sum_{i=1}^{\infty} q_{d,i} \cdot \prod_{\ell=1}^{i-1} (1 - q_{d,\ell}) \delta_{\phi_{d,i}} \tag{7}$$

$$q_{d,i} \sim Beta(1, \gamma) \tag{8}$$

$$\phi_{d,i} \sim P \tag{9}$$

$$P = \sum_{i=1}^{\infty} r_i \cdot \prod_{\ell=1}^{i-1} (1 - r_\ell) \delta_{\phi_i} \tag{10}$$

$$r_i \sim Beta(1, \zeta) \tag{11}$$

$$\phi_i \sim Dir(\alpha) \tag{12}$$

We see that we can easily inject weakly-supervised topic model information into the base distribution ϕ_i. We can simply place a mixture over the alpha-Dirichlet distribution and each labeled topic distribution. If we define B to be the number of all labeled topics, and ω_i as the vector of knowledge source weights for topic i, this transforms ϕ_i to:

$$\phi_i \sim M \tag{13}$$

$$M = (1 - \xi) \cdot \delta_{\mathrm{A}} + \frac{\xi}{B} \cdot \sum_{i=1}^{B} \delta_{\Omega_i} \tag{14}$$

$$\mathrm{A} \sim \mathrm{Dir}(\alpha) \tag{15}$$

$$\Omega_i \sim \mathrm{Dir}(\omega_i) \tag{16}$$

With a newly formulated base distribution established we are now able to build a Gibbs sampler for inference. Following previous work, we seek to find the appropriate topic assignment for each token. In our model this takes the form:

$$P(z = i|\beta, \omega, \vec{w}, \xi) \tag{17}$$

Each topic assignment is dependent on the assignment of a local stick break ($\hat{q}$) and mapping of that stick break to the parent stick break ($\hat{r}$). We formalize this as:

$$\mathcal{P}_3 = P(M_{\hat{r}}|\beta, \omega, \vec{w}, \xi, \hat{r}) \tag{18}$$

$$\mathcal{P}_4 = P(\hat{q} = \hat{r}|\beta, \omega, \vec{w}, \xi, M_{\hat{r}}) \tag{19}$$

$$P(z = i|\beta, \omega, \vec{w}, \xi) = \mathcal{P}_3 \cdot \sum \mathcal{P}_4 \tag{20}$$

However, with the change in the underlying distribution this will need to be factored into the posterior distribution and then marginalized out. Letting $\tilde{o}$ be a shorthand for the observables: $\beta, \omega, \vec{w}, \xi, \hat{r}$, our posterior calculation becomes:

$$\mathcal{P}_5 = P(M_{\hat{r}} = \mathrm{Dir}(\alpha)|\tilde{o}) \tag{21}$$

$$\mathcal{P}_6 = \sum_{j=1}^{B} P(M_{\hat{r}} = \mathrm{Dir}(\omega_j)|\tilde{o}) \tag{22}$$

$$P(M_{\hat{r}}|\tilde{o}) = (1\text{-}\xi) \cdot \mathcal{P}_5 + \frac{\xi}{B} \cdot \mathcal{P}_6 \tag{23}$$

The addition of the new underlying distribution does complicate things but we can reuse existing inference calculations for $\sum P(\hat{q} = \hat{r}|\beta, \omega, \vec{w}, \xi, M_{\hat{r}})$ since this is the basis that every hierarchical Dirichlet process based topic model must calculate. Here we will borrow the calculation from "Infinite LDA" (Heinrich, 2011) which reduces our calculation to:

$$\mathcal{P}_7 = \beta, \omega, \vec{w}, \xi, M_{\hat{r}} \tag{24}$$

$$\mathcal{P}_8 = \vec{z}_{\text{-}i}, \vec{w}, x, y, M_{\hat{r}} \tag{25}$$

$$\sum P(\hat{q} = \hat{r}|\mathcal{P}_7) \propto p(\vec{z}_i{=}j|\vec{z}_{\text{-}i}) \cdot p(\vec{z}_i{=}j|\mathcal{P}_8) \tag{26}$$

$$\mathcal{P}_9 = \vec{z}_{\text{-}i}, \vec{w}, x, y, M_{\hat{r}} \tag{27}$$

$$p(\vec{z}_i{=}j|\mathcal{P}_9) \propto \begin{cases} \text{Equation 2} & \text{if } M_{\hat{r}} = \mathrm{Dir}(\alpha) \\ \text{Equation 6} & \text{otherwise} \end{cases} \tag{28}$$

$$p(\vec{z}_i{=}j|\vec{z}_{\text{-}i}) \propto n_{\text{-}i,j}^{d_i} + \gamma \cdot \tau_z \tag{29}$$

τ represents a sample from the Antoniak distribution; for further details we refer to the "Infinite LDA" publication (Heinrich, 2011). The last step is to marginalize out all the possibilities for $P(M_{\hat{r}}|\tilde{o})$. With this probability being:

$$P(M_{\hat{r}} = m|\tilde{o}) \propto \prod p(\vec{z}_i{=}j|\vec{z}_{\text{-}i}, \vec{w}, x, y, m) \tag{30}$$

We now have the basis for our nonparametric weakly-supervised topic model (IntTM). We see that we can take an existing nonparametric model and marginalize the underlying distribution representing the weakly-supervised topics. The interesting observation to note is that the parameter ξ becomes the likelihood that a weakly-supervised topic is chosen over a "regular" topic being chosen. If we take a weakly-supervised model to be an interpretable topic, then ξ becomes a parameter specifying the level of interpretability.

3.2. Knowledge source topic approximation

Discovering topics using a large knowledge source can lead to severe degradation of execution time. The addition of the weakly-supervised topic model constraints onto the nonparametric Bayesian model imposes a $\mathcal{O}(B \times N_d \times D)$ increase in execution time. One technique to minimize the impact of this time increase is to sample $P(M_{\hat{r}}|\tilde{o})$ at different timesteps than $P(\vec{z}_i{=}j|\vec{z}_{\text{-}i}, \vec{w}, x, y, M_{\hat{r}})$; such as assigning the appropriate $P(M_{\hat{r}}|\tilde{o})$ at the document timestep as opposed to the token timestep. Another approach we take is to order the most likely knowledge source topics and take only the top s ordered topics. We can then approximate the sum of the remaining $B - s$ topics using an approximation function. If we assume a good ordering, and that each lower ordered function decreases the probability value by a constant, ρ, in the range $(0, 1)$ then we can calculate the remaining probability as:

$$\mathcal{P}_j^* = P(M_{\hat{r}} = \mathrm{Dir}(\omega_j)|\tilde{o}) \tag{31}$$

$$\sum_{i=s}^{B} \mathcal{P}_i^* = \mathcal{P}_{s-1}^* \cdot \int \rho^b db \approx \text{-}\frac{\mathcal{P}_{s-1}^*}{\ln \rho} \tag{32}$$

By sampling from this remaining probability chunk we can find the appropriate ordered item. To initially order the topics we partition each knowledge source topic by each of its top words. Then for the topic ordering we sort each topic by the top words and search for knowledge source topics that match the topic's top words. After we acquire a sufficiently sized superset ($\approx 10 \times s$) we can order the knowledge source topics using Equation 30.

3.3. Knowledge source discovery

Our solution to knowledge source discovery involves obtaining the entirety of Wikipedia as a superset of knowledge source topics. We filter out unpopular Wikipedia articles (measured by page views). Because a good match for a knowledge source topic is dominated by token assignments to that topic, it would make sense that words in the corpus that show up in a knowledge source topic many times would be a good fit. To further reduce the knowledge source, we take the top 100,000 topics scored using term frequency-inverse document frequency (tf-idf) and cosine similarity together with a specialized knowledge source ranking (Wood et al., 2022). The 100,000-topic set is then used as input into our interpretable topic model (IntTM).

3.4. Parameter updating

Due to the Bayesian nature of our model, it may be the case that the ξ guarantee is not met. Ultimately it will be the data that decides the number of interpretable topics to choose and ξ will act more as a guide. To enforce a ξ ratio of interpretable topics, we provide techniques for parameter updating.

A simple approach is to use the previous observations of the knowledge source/unlabeled topic ratio to update ξ. If we suppose a linear relationship to the number of topics and ξ, then we can model the expected number of topics given ξ as:

$$E = \mathcal{B}_1 \cdot \hat{\xi} + \mathcal{B}_0$$

The parameters $\mathcal{B}_1$ and $\mathcal{B}_0$ can be updated using linear regression and $\hat{\xi}$ can be determined by setting E to the total number of topics multiplied by the original value of ξ.

4. Results

To evaluate the effectiveness of our methodology we set up two human evaluated tasks to measure interpretability. For all experiments we use the datasets given in Table 1. The baseline methods used are: Infinite-LDA (InfTM) (Heinrich, 2011), Hierarchical LDA (hLDA) (Blei and others, 2003a), the Nonparametric Topic Model (NTM) (Wallach, 2008), the Sawtooth Factorial Topic Embeddings Guided Gamma Belief Network (SawETM) (Duan and others, 2021), the Nonparametric Tree-Structured Neural Topic Model (nTSNTM) (Chen and others, 2021), and the Variationally-Learned Recurrent Neural Topic Model (VRTM) (Rezaee and others, 2020). VRTM was also implemented to utilize outside information in the form of word embeddings (Mikolov and others, 2020) and is evaluated as a separate model (VRTM+W2V). All baseline methods were parametrized according to their experiment descriptions in their respective papers. For the Interpretable Topic Model (IntTM) we use (Wallach, 2008) for Equation 29 and (Wood and others, 2017) for Equation 6 with their respective default parameters. To maximize interpretability, we set $\xi = 1$ for IntTM. For weakly-supervised input we take the discovered knowledge source described in Section 3.3.

4.1. Word Intrusion

In the word intrusion task (Chang and others, 2009) we run each topic model against a dataset and sample an output ϕ_i. We take the 5 highest scoring words from ϕ_i as our "key" words. From the least scoring 5% of words of ϕ_i we take the word which is the highest scoring in ϕ_j where $j \neq i$ as the "intruder" word. We take this last step intentionally to allow for a more competitive "intruder." We repeat this process for a total of 20 samples across all datasets and models. Next, we shuffle the "intruder" and "key" words and create a form which asks a human evaluator to choose the "intruder" word. The exact directions submitted were: *Find the word that does not belong to the set of words.* The form was placed on Amazon Mechanical Turk[3] and each question was assigned 5 different "workers." The only requirement for the selection of the workers was proficiency in the English language. There were no additional filtration criteria outside of the standard requirements imposed by Amazon for the "workers" as we constructed the task for the identification of general topics and general words. It is our presumption that any English-speaking "worker" should be able to comprehend and identify general topics and words without difficulty.

We aggregated the 100 answers for each dataset and computed a t-statistic against the null hypothesis of random selection. Additionally, we compute the associated 95% confidence intervals of both the hypothesis mean (μ_1) and mean difference (MD) between the hypothesis and the null (μ_0) means. Table 2 shows the computed values along with the associated p-value.

4.2. Topic Intrusion

The topic intrusion task (Chang and others, 2009) is similar to the word intrusion task in that we give a set of "key" items mixed in with an intrusive item and ask the human evaluator to find the intrusive item. After topic modeling was complete for all models chose a random document d_i and the corresponding θ_i distribution. From θ_i we take the highest 3 scoring topics as the "key" topics and from the lowest scoring 5% topics we choose the topic which is the highest scoring in document d_j where $j \neq i$. The intuition behind this selection is the same as in Section 4.1. Each topic is represented by 8 of its highest scoring words and shuffled (only the topic order is shuffled, not the top words in the topic). We then create a form which presents the first 100 words of document d_i along with a selection to choose the "intruder" topic among the 3 total topics. The form also allows the user to click a button to see the full text of the document. We repeat the process for a total of 20 samples for each dataset. The form and samples are placed on Amazon Mechanical Turk and assigned to 5 workers each for a total of 100 questions per dataset. The worker selection was the same as in Section 4.1. For the Wiki-20 dataset both hLDA and nTSNTM did not output enough topics to conduct the experiment and were left off the evaluation of the Wiki-20 dataset.

After all questions were answered we compute the t-statistic and other statistical measures as we did in Section 4.1. The results are placed alongside the word

[3] https://www.mturk.com

	Description	**D**	**Topics**
CUL-180	Manually tagged scholarly papers	182	1,660
SE-2010	Scientific articles with manually assigned key phrases	244	3,107
NLM500	A collection of PubMed documents and MeSH terms	203	1,740
R21K	Labeled documents from the 1987 Reuters newswire	21,578	2,700
Wiki-20	20 academic papers annotated from Wikipedia articles	20	564
FAO-30	Annotated documents from the FAO of the UN.	30	650

Table 1: Datasets (Medelyan, 2009) used for evaluation.

intrusion topic in Table 2. Additionally, we seek to evaluate how well the models compare among themselves. Post-hoc analysis is conducted using the Tukey-Kramer method which represents the mean difference and 95% confidence intervals in Figure 1.

4.3. Effect of ξ

We seek to determine how ξ effects the interpretability of our model using human-aided evaluation.

4.3.1. Experimental Setup

For each baseline model against each dataset, we run the model with the default scaling parameter α as 1 and β as $200/V$ (V being the size of the vocabulary) for 1,000 iterations. After inference was complete, we are able to calculate the document to topic mixture (θ) and topic to word mixture (ϕ) using the end result of the topic assignments. With the θ and ϕ mixtures we can easily determine the most and least popular word for a given topic and the most and least popular topic for a given document. We then repeat the same process for all models with weakly-supervised topic modeling appended as described by Section 3.1. We run the baseline models ($\xi = 0$) against the ξ values of 0.5 and 1. After all runs were completed, we can set a word intrusion and topic intrusion task to be given for evaluation (Chang and others, 2009). To reiterate, word intrusion involves giving a person 6 words, 5 being from the most popular words in a topic and 1 being among the least popular—the least popular word is referred to as the intrusive word—and asking them to identify the intrusive word. For our evaluation we filter out topics that have 3 or more words that are not in common usage (as determined by showing up in a dictionary word list) or are numeric. Additionally, we restrict intrusive words to the same criteria. We do this because topics which contain all numbers or obscure words may be hard for non-domain experts to understand and so an evaluation with a high percentage of these words might not be meaningful. Topic intrusion is like word intrusion only applied to topics. Each user is given a block of text (100 words) that begin a document and are then given 4 topics—3 topics being the most popular in the document and 1 being among the least popular—and asking them to identify the intrusive topic. We utilize Amazon Mechanical Turk as the platform to obtain human evaluation. Each question was given to 3 different Amazon Mechanical Turk users. For subsequent questions, the users were redrawn from the Amazon Mechanical Turk pool of users which re-

	Word Intrusion				**Topic Intrusion**			
	N	μ_1	MD	p-value	N	μ_1	MD	p-value
hLDA	600	0.15 ± 0.03	-0.02 ± 0.04	0.830	500	0.27 ± 0.04	0.02 ± 0.05	0.236
InfTM	600	0.15 ± 0.03	-0.01 ± 0.04	0.736	600	0.27 ± 0.04	0.02 ± 0.05	0.215
IntTM	600	$\mathbf{0.31 \pm 0.04}$	$\mathbf{0.14 \pm 0.05}$	**2.2e-09**	600	$\mathbf{0.36 \pm 0.04}$	$\mathbf{0.11 \pm 0.05}$	**1.3e-05**
NonTM	600	0.12 ± 0.03	-0.04 ± 0.04	0.987	600	0.26 ± 0.03	0.01 ± 0.05	0.421
nTSNTM	600	0.15 ± 0.03	-0.01 ± 0.04	0.709	500	0.28 ± 0.04	0.03 ± 0.05	0.175
SawETM	600	0.15 ± 0.03	-0.02 ± 0.04	0.808	600	0.28 ± 0.04	0.03 ± 0.05	0.107
VRTM	600	0.11 ± 0.03	-0.05 ± 0.04	0.996	600	0.28 ± 0.04	0.03 ± 0.05	0.163
VRTM+W2V	600	0.12 ± 0.03	-0.04 ± 0.04	0.984	600	0.24 ± 0.03	-0.01 ± 0.05	0.656

Table 2: The p-value, mean (μ_1), mean difference (MD) and associated 95% confidence intervals for each model aggregated the datasets for both the word intrusion and topic intrusion tasks.

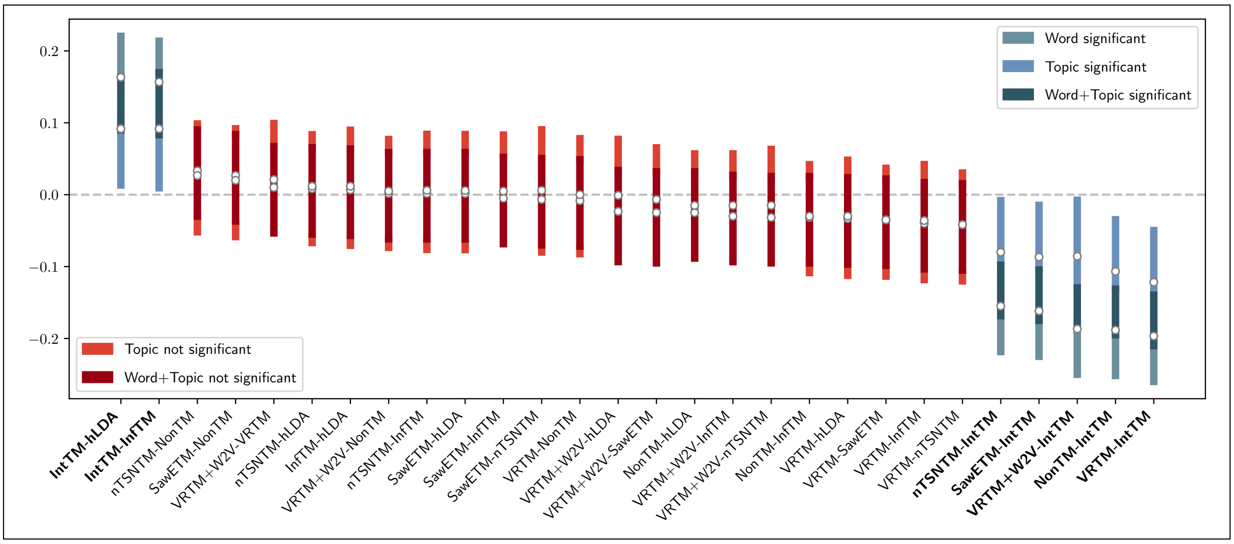
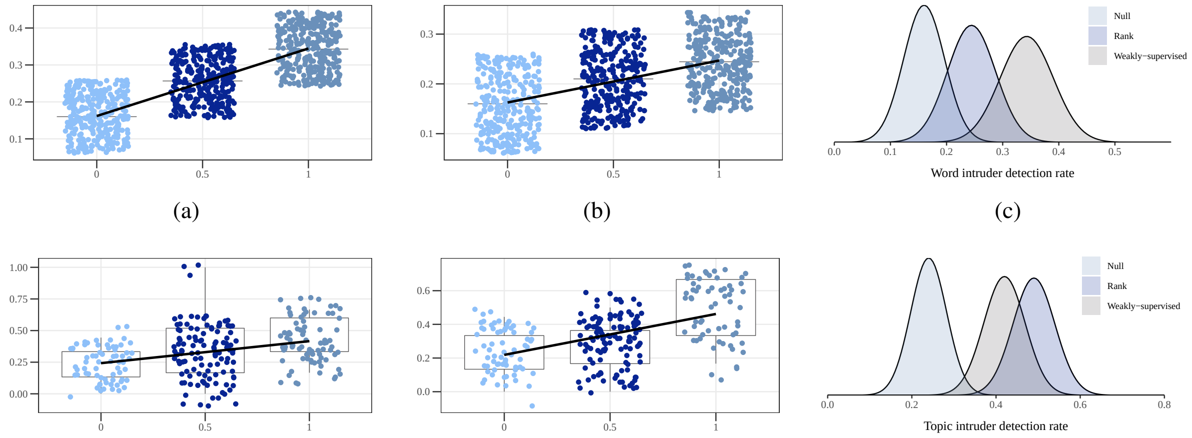


Figure 1: The Tukey-Kramer pairwise difference of means and associated 95% confidence intervals for the word and topic intrusion tasks.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2: Word detection results for all models and datasets with the provided constructed knowledge source (a) and the discovered knowledge source (b). The topic detection task is shown in (d) and (e) for the provided and discovered knowledge sources respectively. The expected distributions for our method at $\xi = 1$ show significance against the null hypothesis distribution for both word detection (c) and topic detection tasks (f).

duces the probability that any one single user answered multiple questions. As in previous experiments, the worker selection is general and described in more detail in Section 4.1.

4.3.2. Experimental Results

After the users submitted their answers to all questions for the word and topic intrusion task, we evaluated their effectiveness. Each submitted answer was assigned the value of accuracy for its group and plotted in Figure 2. The groupings were based on ξ, dataset, and model. We can clearly see the trend between ξ and interpretability for both the topic and word intrusion tasks. In both, ξ is positively associated with interpretability. We show the regression line in each task box plot. Each regression line shows a significance above 0.1. As expected, we see an increase in detection of intrusive words when using the predefined knowledge source (Weakly-supervised) versus the discovered knowledge source (Rank). However, this is not the case for the intrusive topic. We suppose the topic discrepancies may be due to randomness and does not represent a significant difference. Still, this may represent an interesting point to examine. While the pre-defined knowledge sources are human curated topic labels suggested by reading each document, the discovered ones are more numerical. Numerical in the sense that the only criteria for selecting them are using established methods for information retrieval. It then makes sense that for certain tasks the discovered knowledge source performs better.

Additionally, we calculate whether the models with $\xi = 1$ represent a significant increase in interpretability. The expected distributions, plotted in Figure 2(c) and Figure 2(f), show significance above 0.1.

5. Discussion

In both the word intrusion and topic intrusion tasks IntTM is the only model to achieve significance at the 0.01 level. In the word task we see that all other models perform worse than the null hypothesis. We suspect this has to do with the experiment design. Among the "key" words to select from there may be a mixture of coherence along with more esoteric words. With a non-consistent coherence the human evaluator is not able to discern the overall topic and the intruder word becomes more favorable (of not being chosen as the intruder) than one or more of the esoteric words. One could argue that injecting outside information into the neural topic models could produce similar results to the IntTM. We do not deny this possibility however we see that the addition of word embeddings does not significantly improve performance for VRTM. This suggests that more recent word embeddings, such as BERT (Devlin and others, 2019) may not necessarily lead to outperforming results to the IntTM.

Also of interest was the non-significant difference between the Bayesian and neural topic models outside of IntTM. For the topic intrusion task, one could expect neural topic models to perform poorly since they tend to reuse individual θ distributions (see Section 1 for more details). However, that Bayesian models outside of IntTM perform similarly to neural topic models is a surprise. We hypothesis this similarity is due more to poor performance from the Bayesian models as opposed to good performance from the neural topic models. The non-significance between Bayesian and neural models for the word task introduces an interesting area for investigation since the neural topic models produce topics with better perplexity and PMI-based scores. The inconsistency with perplexity and PMI-based metrics to our measure of interpretability is consistent with other interpretability studies (Chang and others, 2009; Doogan and Buntine, 2021). Our results may indeed add to the challenge of PMI-based methods being the most appropriate method for measuring interpretability (Doogan and Buntine, 2021). However, we contend that PMI-based methods are still valid and useful. Especially since human evaluation is both costly and time-consuming.

In our experimental design many components were stochastic and thus exact results may be difficult to reproduce. The topic modeling process itself relies on Gibbs sampling which uses randomness to infer the hidden variables. Without seeding and randomness implementation details, subsequent runs using the same parameters and datasets may yield slightly different results. Other random components were in the selection of topics and intrusion word for the word intrusion task and the selection of documents and intrusion topic for the topic intrusion task. Even after the release of the source code for our method alongside the task selection implementations—including seeding and randomness implementations, some results may still hard to reproduce exactly. The baseline Bayesian topic models were implemented by the authors and may differ slightly in random seeding to the original implementations. Additionally, the human-evaluated results cannot be reproduced exactly. However, the experiment design was intentionally constructed to include a large number of tasks to a number of different human evaluators. In power analysis of our experiments, we find the sample size to be sufficient for confidence to be maintained in the conclusions.

The technique presented here is not without limitations. One such limitation is the inability to discover general topics. For example, suppose an author constructs a corpus with one topic being "introductory college subjects." The words drawn from this general topic may consist of seemingly unrelated words, such as "proof", "Napoleon", and "chromosome". In this case it is unlikely the knowledge source topic would be detailed enough to contain those words and thus they would not be biased towards the "introductory college subjects" topic. However, they may end up in the more refined child topics of "mathematics", "history of France", and "genetics." An interesting area for future research is the cumulation of child topics into parent topics—a hierarchical adaption of our interpretable topic model. This hierarchical adaption model may be able to discover the example words above as belonging to the "introductory college subjects" topic.

6. Conclusion

This paper investigates a novel combination of nonparametric Bayesian and weakly-supervised topic models. In this combination we discover a fascinating result—a self-contained, outperforming nonparametric interpretable topic model. As we show with empirical results, this topic model discovers topics that are significantly more interpretable than both Bayesian and recent neural topic models. This novel method highlights a new approach to topic modeling—one in which human-evaluated topic interpretability is at the forefront of topic discovery.

7. Bibliographical References

Alokaili, A. et al. (2020). Automatic generation of topic labels. In *SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1965–1968.

Bianchi, F. et al. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 759–766, Online. Association for Computational Linguistics.

Blei, D. M. and McAuliffe, J. D. (2007). Supervised topic models. In *Advances in Neural Information Processing Systems 20*, pages 121–128.

Blei, D. M. et al. (2003a). Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16*.

Blei, D. M. et al. (2003b). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chang, J. et al. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pages 288–296.

Chen, Z. et al. (2021). Tree-structured topic modeling with nonparametric neural variational inference. In *ACL/IJCNLP 2021*, pages 2343–2353. Association for Computational Linguistics.

Deng, Q. et al. (2020). Detecting information requirements for crisis communication from social media data: An interactive topic modeling approach. *International Journal of Disaster Risk Reduction*, 50:101692.

Devlin, J. et al. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics.

Doogan, C. and Buntine, W. L. (2021). Topic model or topic twaddle? re-evaluating semantic interpretability measures. In Kristina Toutanova et al., editors, *NAACL-HLT 2021, Online, June 6-11, 2021*. Association for Computational Linguistics.

Duan, Z. et al. (2021). Sawtooth factorial topic embeddings guided gamma belief network. In Marina Meila et al., editors, *ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research. PMLR.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April.

Griffiths, T. (2002). Gibbs sampling in the generative model of latent dirichlet allocation.

Hansen, J. A. et al. (2013). Probabilistic explicit topic modeling using wikipedia. In *GSCL 2013*.

He, D. et al. (2021a). Automatic topic labeling model with paired-attention based on pre-trained deep neural network. In *IJCNN 2021*, pages 1–9. IEEE.

He, D. et al. (2021b). Automatic topic labeling using graph-based pre-trained neural embedding. *Neurocomputing*, 463:596–608.

Heinrich, G. (2011). Infinite lda implementing the hdp with minimum code complexity.

Jiang, H. et al. (2020). Explaining a bag of words with hierarchical conceptual labels. *World Wide Web*, 23(3):1693–1713.

Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006)*, pages 577–584.

Manouchehri, N. et al. (2021). Batch and online variational learning of hierarchical dirichlet process mixtures of multivariate beta distributions in medical applications. *Pattern Anal. Appl.*, 24(4):1731–1744.

Medelyan, O. (2009). *Human-competitive automatic topic indexing*. Ph.D. thesis, The University of Waikato.

Mikolov, T. et al. (2020). Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*.

Mimno, D. M., Li, W., and McCallum, A. (2007). Mixtures of hierarchical topics with pachinko allocation. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007)*, pages 633–640.

Newman, D. et al. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. The Association for Computational Linguistics.

Ning, X. et al. (2020). Nonparametric topic modeling with neural inference. *Neurocomputing*, 399:296–306.

Prasad, K. R., Mohammed, M., and Mohammed, N. R. (2021). Visual topic models for healthcare data clustering. *Evol. Intell.*, 14(2):545–562.

Rezaee, M. et al. (2020). A discrete variational recurrent topic model without the reparametrization trick. In *Advances in Neural Information Processing Systems 33*.

Song, D. et al. (2020). Knowledge base enhanced topic modeling. In Enhong Chen et al., editors, *2020 IEEE International Conference on Knowledge Graph, ICKG 2020, Online, August 9-11, 2020*, pages 380–387. IEEE.

Tomasi, F. et al. (2020). Stochastic variational inference for dynamic correlated topic models. In Ryan P. Adams et al., editors, *UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*. AUAI Press.

Topic. (2021). Merriam-webster.

Wallach, H. M. (2008). *Structured topic models for language*. Ph.D. thesis, University of Cambridge Cambridge, UK.

Wang, W. et al. (2021). Robust supervised topic models under label noise. *Mach. Learn.*, 110(5):907–931.

Wood, J. et al. (2017). Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In *33rd IEEE ICDE*.

Wood, J., Wang, W., and Arnold, C. (2021). The biased coin flip process for nonparametric topic modeling. In *International Conference on Document Analysis and Recognition*, pages 68–83. Springer.

Wood, J., Arnold, C., and Wang, W. (2022). Knowledge source rankings for semi-supervised topic modeling. *Information*, 13(2):57.

Yang, Y. and Wang, F. (2021). Author topic model for co-occurring normal documents and short texts to explore individual user preferences. *Inf. Sci.*, 570:185–199.