Labeled Interactive Neural Topic Models: No Longer Take It or Leave It

Anonymous ACL submission

Abstract

Topic models help understand document collections, but they don't always identify the most relevant topics. Classical probabilistic and anchor-based topic models offer interactive versions that allow users to guide the models towards better topics. However, such interactive features have been lacking in neural topic models. To correct this lacuna, we introduce a user-friendly interaction for neural topic models. This interaction permits users to assign a word label to a topic, leading to an update in the topic model where the words in the topic become closely aligned with the given label. Our approach encompasses two distinct kinds of neural topic models. The first includes models where topic embeddings are trainable and evolve during the training process. The second involves models where topic embeddings are integrated post-training. To facilitate user interaction with these neural topic models, we have developed an interactive interface that enables users to engage with and re-label topics. We evaluate our method through a human study, where users can relabel topics to find relevant documents. Using our method, user labeling improves document rank scores, helping to find more relevant documents to a given query when compared to no user labeling.

Topic: Dengue outbreak in Asia **Query**: What countries are seeing an outbreak?
No topic labeling
Topic 0: 'dengue', 'vaccine', 'sanofi', 'dengvaxia', 'phillipines', 'vaccination' Topic 1: 'virus', 'countries', 'new', 'according', 'dr', 'pandemic' Topic 2: 'time', 'get', 'however', 'gonaives', 'haiti', 'town', 'stud'
After topic labeling
Topic 0: 'dengue', 'vaccine', 'sanofi', 'dengvaxia', 'phillipines', 'vaccination' Topic 1: 'virus', 'countries', 'new', 'according', 'dr', 'pandemic' **Topic 2: 'india', 'genotype', 'denv', 'asian', 'study', 'singapore'**

Table 1: This figure demonstrates the capability of interactive topic modeling in refining topics. Initially, 'Topic 2' does not align with the query. Before the labeling, the topic words, as generated by the ETM, show that while the first two topics correlate with the task, 'Topic 2' is unrelated. After the labeling, the updated 'Topic 2' now closely aligns with the user-specified label, 'india', showcasing how I-NTM adapts in real-time to user input, giving greater relevance and accuracy in topic representation.

1 Topic Models Need Help

Topic modeling is an unsupervised machine learning method for analyzing a set of documents to learn meaningful clusters of related words (Boyd-Graber et al., 2017). Despite decades of new models that purport to improve upon it, the most popular method remains Latent Dirichlet Allocation (Blei et al., 2003a, LDA), which is two decades old.

This venerable model is still the workhorse for those who use unsupervised analysis to discover the structure of document collections in digital humanities (Meeks and Weingart, 2012), bioinformatics (Liu et al., 2016), political science (Grimmer and Stewart, 2013), and social science (Ramage et al., 2009b). However, if you look at the computer science literature, topic modeling has been taken over by neural approaches (Zhao et al., 2021), such as the embedded topic model (ETM) (Dieng et al., 2020) and contextualized topic models (CTM) (Bianchi et al., 2020). We review LDA and neural topic models in Section 2.

So what explains this discrepancy? A sceptic would posit that there is not sufficient evidence to support the claims that neural topic models are substantially better either in terms of runtime, ease-of-use, or on human-centric methods (Hoyle et al., 2021). In addition to these legitimate concerns, there are also functional lacunae: abilities "classic"

topic models have that neural models lack. Neural models are often a "take it or leave it" proposition: if the results do not match what you want, a user (particularly a non-expert in machine learning) has little recourse. In contrast, the probabilistic topic modeling literature has a rich menu of options to improve topic models: works involving labeling topics through images using neural networks, using a sequence-to-sequence model to automatically generate topics, or using unsupervised graphical methods to label topics (Aletras and Mittal, 2016; Aletras and Stevenson, 2014; Alokaili et al., 2020). Pleple (2013) designed an interactive framework that allows the user to give live feedback on the topics, allowing the algorithm to use that feedback to guide the LDA parameter search. Choo et al. (2013) developed an interactive interface for LDA for user-driven topic modeling. Unfortunately, these improvements are not currently available for neural topic models.

Making neural models interactive requires two things: models to support interactivity and an interface to allow users to make changes to the model. This paper provides both and applies them both to models by directly updating topic embeddings (ETM, NVDM) or by adding topic embeddings in after training CTM. To use I-NTM interactively—based on the topic label from the user—we embed the label in the embedding space and *move* the corresponding topic embedding closer to the label. We detail the two different types of "moving" in Section 3.1. This adjusts the center of the topic embedding: throwing out unrelated words, prioritizing words that are "close" to the users' label.

While there have been many previous works for interactive labeling, our work introduces a way of improving topics through a natural way of labeling that is typically done a posteriori. We call this method *Interactive Neural Topic Modeling* or I-NTM. Additionally, we provide a user-friendly interface that allows for such interactions.

To demonstrate the efficacy of our interactive labeling method and interface, we conduct a human study using the CTM backend of I-NTM. CTM was chosen since it showed to find the most diverse and coherent topics out of the three models we provide support for. We find that if a user has a specific task for a corpus, I-NTM quantitatively helps users quickly identify more documents relevant to their information needs, as we will see in Section 4.2

2 Best of Both Worlds: Neural Word Knowledge and Bayesian Informative Priors

This section reviews topic models: how they are useful to practitioners, their shortcomings, and motivate our attempt to ameliorate this with embedding-based interactions.

2.1 Latent Dirichlet Allocation

Topic models are exemplified by latent Dirichlet allocation LDA (Blei et al., 2003b). LDA posits a generative story for how the data came to be and uses probabilistic inference to find the best explanation for the dataset (Griffiths and Steyvers, 2004a). Often, one of the first steps of using the output of a topic model is to *name* the topics. Either by selecting top words through a Markov chain Monte Carlo algorithm (Griffiths and Steyvers, 2004b; Hofmann, 2017) or through manual generation of descriptive topics (Mei et al., 2006; Wang and McCallum, 2006).

For probabilistic models, however, this is not the end of the story. The Bayesian framework—through the use of informed priors—encourages the incorporation of expert knowledge into interactive topic models. This can either represent a dictionary (Hu et al., 2014b), word lists from psychology (Zhai et al., 2012), or the needs of a business organization (Hu et al., 2014a). This feedback to a model helps match a user's information needs or reflect world knowledge and common sense. Of course, one could move to a fully supervised model (Blei and McAuliffe, 2007), where every training document has a topic label. But this requires substantially more interaction with the user than giving feedback on a handful of topics—full supervision requires hundreds or thousands of labeled examples. But these interactive models are not without their faults. First, they're slow; probabilistic inference—whether with MCMC methods or variational inference—struggles to update in the seconds required to satisfy the best practices of an interactive application. Second, while one of their goals is to incorporate the knowdge of users, they completely ignore the vast world knowledge available "for free" from representations trained on large text corpora.

2.2 Neural Topic Models

Neural topic models have emerged as a powerful alternative to probablistic models. These models

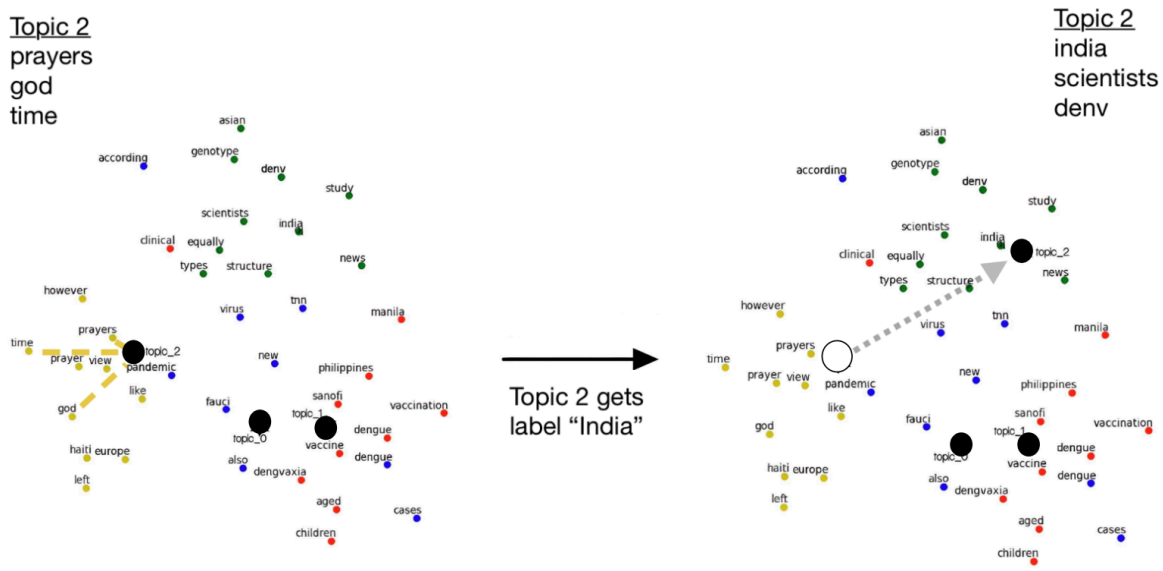

Figure 1: Visual representation labeling a new topic with out method, like in Table 1. Our method moves the embedding center for the topic closer to the new label word, in this case, *India*.

leverage deep learning techniques to capture complex relationships and representations within textual data, offering several advantages over traditional methods. One of the key strengths of neural topic models is their proficiency in generating coherent and interpretable topics. This is primarily due to their use of nonlinear functions, which are more adept at closely matching the observed distribution of words and topics in the data.

One popular architecture for neural topic modeling is the Variational Autoencoder (VAE) based topic model. VAE-based models, such as NVDM (Miao et al., 2016) encode documents into continuous latent spaces, enabling a smoother and more expressive representation of topics. For this model, we add topic embeddings directly to the latent space of the model, creating learnable topic embeddings similar to the topic embeddings inherently found inETM.

Neural models capture data nuances by learning distributed representations of words and topics. This leads to topics that are not only more semantically meaningful but better aligned with human interpretations. ETM takes advantage of these representations by associating each topic with an embedding. These embeddings can be learned by the model or pre-trained word embeddings may be used. Like traditional topic models, each document has a vector connecting it to the K latent topics. While a traditional topic model would have a full distribution over the vocabulary, the k^{th} topic in ETM is a vector $\alpha_k \in R^L$—just like words in the embedding space. ETM induces a per-topic distribution over the vocabulary from this representation.

Moreover, neural models can handle large-scale text corpora efficiently and can adapt to different domains or utilize the knowledge of large language models (LLM). In CTM researchers sought to leverage the knowledge of LLM for better word representations. One such method is to combine the traditional BOW method with word embeddings from a LLM to develop contextualized embeddings which lead to better topic models (Bianchi et al., 2020). The symbolic meaning of traditional BOW is lost after a single neural layer, so they hypothesize that contextualized embeddings would improve this. As CTM are one of the best neural topic models, we decide to extend CTM to be interactive as well and use it as the focus for our human study.

3 Interactive Neural Topic Modeling

In this section, we explore the rationale and methodology behind modifying labels in neural topic models, focusing on two primary mechanisms: learnable topic embeddings and topic embeddings added in post-training. These methods offer similar, yet distinct approaches to refining topic models. Traditional topic models often suffer from the absence of explicit labels, leading to potential mismatches between documents and topics or the generation of incoherent topics. This can lead to situations

	Vocab Size	Coherence	Diversity
ETM	2565	0.19	0.81
	3572	0.17	0.85
	10830	0.11	0.92
I-NTM (ETM)	2565	0.14	0.84
	3572	0.10	0.88
	10830	0.10	0.94
I-NTM (CTM)	2565	**0.21**	0.91
	3572	0.18	0.92
	10830	0.15	**0.95**

Table 2: Interactivity improves downstream classification tasks and the overall diversity of topics. In some cases, topic coherence decreases since coherence improves with general topics and we are labeling topics. Topic coherence and topic diversity, varying vocabulary sizes for ETM and various I-NTM models on the BETTER dataset. Our both models under I-NTM outperform standard ETM in terms of topic diversity and topic coherence

where documents are associated with topics that they should not be (Ramage et al., 2009a) or topics that just do not make sense (Newman et al., 2010). Also, they require users to manually analyze the topics found and then use labels such as the Business topic. Non-technical users also use a similar process when using topic models: they inspect the topics, find the topics relevant to their use case, and label them accordingly. Thus, since labeling is a natural way people have already been interacting with topic models, we use labeling to both improve topics and help guide the model to relevant topics for the users. We will dissect two key methods in I-NTM for updating topics, depending on the underlying model used. The first method involves learnable topic embeddings, that is models that have or can have learnable topic embeddings. The second method is post-training adjustments, where topic embeddings are added in and modified after the model has been trained. These methods offer a suite of neural topic models to use interactively. By combining learnable topic embeddings and post-training adjustments, I-NTM provides a robust and flexible framework for users to interact with and guide the development of neural topic models.

3.1 Adjusting Learnable Topic Embeddings

In this section, we explore the first of two primary methods for updating topics in neural topic models: models that have or can have learnable topic embeddings. As discussed above, for ETM and NVDM we induce a topic distribution from word representations and a topic embedding. These models fall under the type of neural models where topic embeddings are or can be directly represented in the model and therefore changed. To make the topic modeling interactive, we allow for the users to adjust the underlying embedding for each topic, thus "moving" the topic closer to the word embeddings they desire. We will discuss what this looks like in terms of users' actions in a moment, but for the moment we assume that this can be expressed as a vector

$$\vec{\alpha_k}^{new} = \lambda(\vec{w_k} - \vec{\alpha_k}^{old}) + (1 - \lambda)\vec{\alpha_k}^{old} \quad (1)$$

where α_k^{old} is the topic embedding generated by the model and w_k is the word embedding associated with the topic the user inputs. That is, if the user wants a topic of food, the topic embedding is moved toward the word embedding corresponding to food. The weight of adjusting the topic embedding towards the new label, can be tuned through the parameter λ, which determines how close the topic embedding is moved.

Following the example in Table 1, Figure 1 shows the topic and word embeddings before and after the adjustment of Topic 2. The words surrounding Topic 2 before adjusting the label, do not at first seem to be relevant to the question. After labeling of Topic 2, as India, we see the topic embedding, is close to the words "india", "denv", and "scientist", which are more likely to be relevant to the question and to reveal more relevant documents.

3.2 Adding Adjustable Topic Embeddings After Training

ETM and NVDM have trainable embeddings, but what about models that cannot or adding them negatively affects training? In such cases, the idea is to introduce a form of topic embedding post-training, to enhance the model's performance and interpretability. We can simulate the effect of an embedding by creating a weighted average over the words that constitute a topic. This weighted average essentially serves as a stand-in for a physical topic embedding. Our interactive framework supports these types of models. In this case, the topic embedding is a proxy for a physical embedding to change the topics. CTM falls into this category. Here, given a new label w_l for a topic t_i, the distribution over words for t is updated to have higher

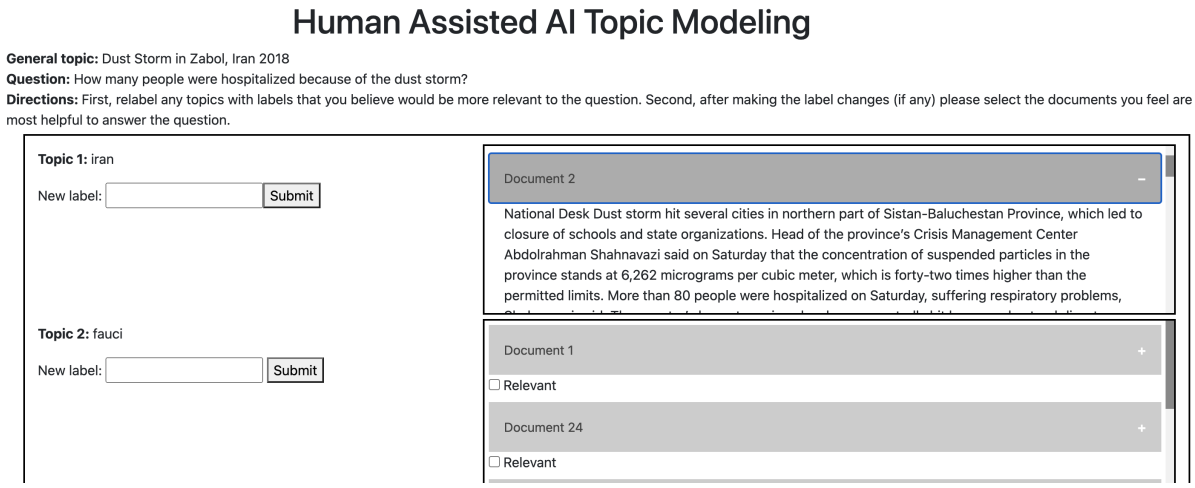

Figure 2: Human study interface for I-NTM, using CTM as the neural model. Users can see the given topics that are found for a set of tasks/requests and can change the label to better fit their needs. Additionally, the assigned documents for each topic are shown and users can select which documents are most relevant.

probability for w_l and for similar words, w_s:

$$P_{\text{update}}(w_l \mid t_i) = P_{\text{orig}}(w_l \mid t_i) + \Delta P(w_l \mid t_i) \quad (2)$$

and for similar words,

$$\Delta P(w_s \mid t_i) = \lambda \cdot \text{sim}(w_l, w_s) \cdot \Delta P(w_l \mid t_i) \quad (3)$$

where $\Delta P(w \mid t)$ is the amount by which you increase the probability of word, w, in topic, t.

In neural topic models, topics are typically represented as distributions over words. Each topic is a blend of various words, with certain words having more weight or influence in defining the topic. Thus, when a user assigns a label to a topic, they are providing a semantic point of reference for that topic. The model is prompted to adjust the weights of words in the topic distribution to align more closely with the semantics of the label.

3.3 User Interface

While Equation 1 outlines a theoretical framework for labeling topics, its practical application hinges on a user-friendly interface that allows for real-time interaction. To address this, we have developed an interface, as depicted in Figure 2, which not only makes interactive topic modification feasible for neural models but also enhances user engagement beyond existing NTM visualizations. Our interface is designed with low-latency interactions in mind, a crucial feature for ensuring efficient topic refinement. There is immediate feedback when users label or re-label topics, fostering a dynamic interaction where users can intuitively understand the impact of their inputs on the model.

Furthermore, the interface is tailored to accommodate users without technical expertise. It allows them not only to assign labels to topics but also to observe, in real-time, how such labeling alters the document-topic assignments. This level of interaction is an advancement over traditional NTM visualizations, which typically offer static or less responsive user experiences. Users can delve into the topics, peruse associated documents, and input new labels via the interface.

The underlying system seamlessly handles the complex tasks: adjusting topic embeddings, recalculating document-topic distributions, and updating the display to reflect these changes. This backend processing ensures that the interface remains user-friendly and effective. A key feature of our interface is its ability to support continuous topic updates. Users can modify a topic multiple times, and there is flexibility to update several topics concurrently. To maintain the coherence and distinctiveness of topics, the interface incorporates safeguards against creating duplicate topics or topics with terms not found in the existing vocabulary.

3.4 Automatic Metrics

Since Dieng et al. (2020) improve topic coherence and diversity compared to LDA, to check if our method negatively affects them, we compare coherence and topic diversity for varying vocabulary sizes between I-NTM models. Topic coherence is an automated method for evaluating the semantic similarity of top words in a given topic. We measure the normalized Pointwise Mutual Information

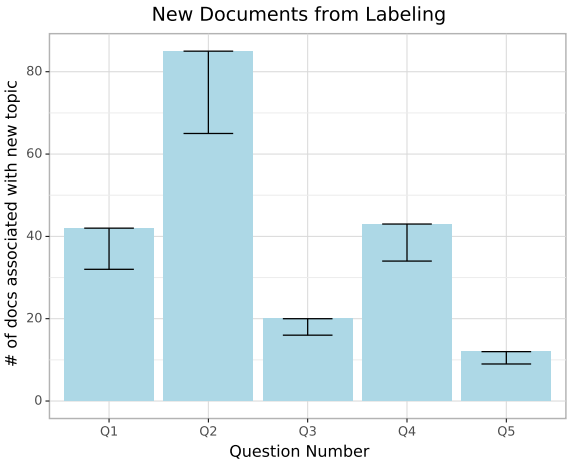

Figure 3: Labeling topics leads to, otherwise missed, documents to be revealed. The maximum number of new documents, that is, a document that was not previously associated with the topic, found for each question across all users. The range of the number of documents found across all users is shown by the black bars.

(NPMI). NPMI is just an extension of PMI, where the vectors are weighted (Aletras and Stevenson, 2013).

For our user study, we use information retrieval (IR) as a metric for evaluating human labeling. IR focuses on retrieving documents that are relevant to a given query. By using IR as a metric, one can objectively assess how well user labeling improves the model's ability to retrieve relevant documents. If the user-labeled topics lead to more relevant documents being retrieved in response to a query, this indicates that the labeling process is effective. Additionally, IR focus on retrieving relevant documents mirrors real-world use cases of topic models. By using IR as an evaluation metric, you ensure that your assessment reflects practical scenarios where users rely on the model to find information quickly and accurately.

3.5 Human Study

To validate the efficacy of I-NTM, we recruit participants to test our model in finding more relevant documents for different information needs, ex. *"Find documents that relate to foreign intervention in Cuba."* Information retrieval tasks are an intuitive way to measure the success of our method, since they involve finding relevant information specific to a need. To verify that user labeling uncover more relevant documents, we compare document ranking scores before and after labeling and between a control group, where no labeling is done and the test group, where users can label topics.

Setup We recruited 20 participants through the online platform Prolific.

1. Our model I-NTM generates topics on the Text REtrieval Conference (TREC) Question Classification dataset. We randomly selected approximately 1500 documents from the Foreign Broadcast Information Service (FBIS).
2. Participants see an information need with topics generated by our model. They can label topics as they deem best
3. After labeling topics, they select a maximum of five documents that they believe best answer the information need

We limit users to five minutes per question. We limit the users to select five documents to normalize the results across all users and limit outliers, i.e. a user taking an hour to comb through hundreds of documents to achieve maximum ranking score.

We want to mimic real-world scenarios where thousands of documents and possibly hundreds of questions need to be answered, where users would not have time to spend hours on each question. For each user we collect the topic information and document distribution before and after the human interaction. Then, using the B25 algorithm (Robertson et al., 1994), an information retrieval ranking function, we compare the estimated relevancy of topics before and after the human interaction. We use BM25 since no gold relevance annotations are available for the TREC dataset and since BM25 works by using a bag-of-words retrieval function that ranks a set of documents based on query terms present in the document, this is an effective way to compare retrieval performance between our two groups.

4 I-NTM Experimental Results

We evaluate I-NTM on standard evaluation metrics and through a human study. Our experiments confirm that an interactive topic modeling interface greatly improves users' ability to find relevant documents in a timely manner.

4.1 Labeling Improves Coherence

Initially, we tested I-NTM with automatic metrics without human intervention, to understand how interaction changes the coherence and diversity of topics. Looking at the ETM backend, topic coherence drops with our method, but diversity is

Topic	Type	Avg Time	Docs
Cuba	Control	5* min	3
	Interactive	2 min	5
South Korea	Control	5 min	3
	Interactive	4 min	5
Taiwan	Control	5* min	3
	Interactive	2 min	5
Balkans	Control	5 min	3
	Interactive	3 min	5
China	Control	5 min	5
	Interactive	4 min	5

Table 3: Our interactive method led to document selection, with more relevant documents being selected, on average. For the 5 different questions, the general topic of that question, the average amount of time a user spent on each question, and the average number of document selected are reported. A time of 5* indicates they hit the set time limit of 5 minutes per question.

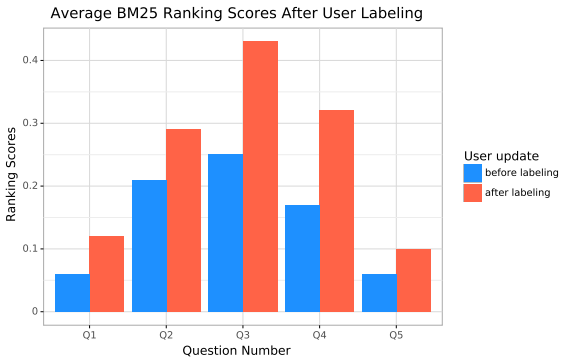

Figure 4: Average BM25 document ranking scores for each of the 5 questions averaged, over the 20 users. User inputted topic labels find more relevant documents and significantly improve document ranking scores

higher (Table 2). This effect is dataset dependent. For Wikipedia, adjusting six of the topics to have distinct labels for classification results in a more diverse topic words. However, coherence typically improves with more general clustering topics, since it measure co-occurence of words in the documents with the topic words. So, with distinct topics, this can result in lower topic coherence. In contrast, the documents in the BETTER dataset (Table 1 and Figure 1) are curated to be related to disaster situations. In this case, when topics are labeled to better fit the request at hand, the topic words tend to have more overlap, since the request is so specific. With the BETTER dataset, I-NTM decreases topic diversity but increases in topic coherence.

Regardless, topic coherence is an imperfect metric for neural topic modeling evaluation (Hoyle et al., 2021). Nevertheless, we report these scores for coherence and diversity since this is the current standard for topic model evaluations.

Human validation is viewed as the gold standard when it comes to topic model evaluation, thus we report those results in the next section.

4.2 Human Study

To evaluate the effectiveness of I-NTM, we conducted a human study using both a control (no labeling) and interactive (allows for labeling) scenario. For both treatments, CTM backend of I-NTM is used. The control group and interactive group are given the same question, set of documents, and topic model. However, the control group is asked to find relevant documents without the ability to label any topics. In contrast, the interactive group can label topics and then select relevant documents. When comparing the BM25 document ranking scores of the control and interactive group, we find labeling topics leads to more representative documents being revealed and chosen, when averaged across all 20 users (Figure 4).

In all cases except for Q5, there is a stark increase in ranking scores after the update. Question 5 was *"Find document related to Chinese economic intervention in other countries"* and due to a large portion of the documents came from Chinese news sources, it was easier to find documents discussing China's economic relations in comparison to the other questions. We see further evidence of this in Table 3 where Q5 had the highest average number of documents selected across all control scenarios and matched the interactive scenario in average time taken. In contrast, the other four questions took more time and selected less documents than the interactive case. In the case of Q1 and Q3 where the time limit was reached, the users were not able to find 5 related documents in time. While a high number of new associated documents does not necessarily correlate with an increase in document ranking score, as some of the new documents might be related to the general topic but not the specific information need, we find that labeling does reveals a significant amount of documents that were not previously in that topic (Figure 3). Again, we see a significantly smaller number for Q5, which we believe is due to the prevalence of documents related to China in the dataset.

5 Related Work

Topic modeling covers a wide range of methods for discovering topics within a corpus and there has been extensive research across these different methods. We discuss these similar methods and contrast them with our own in the following seciton.

Neural topic models With the recent developments in deep neural networks (DNNS, there has been work to use these advancements to increase performance of topic models. One of the most common frameworks for neural topic models (NTMS), described in (Zhao et al., 2021), as VAE-NTMS. Much research was focused on adapting VAE's for topic modeling; Zhang et al. (2018); Srivastava and Sutton (2017) focus on developing different prior distributions for the reparameterization step of VAE, such as using hybrid stochastic-gradient MCMC and approximating Dirchelt samples with Laplace approximations. VAE-NTM also were extended to work with different architectures, Nallapati et al. (2017) developed a sequential NTM where the model generates documents by sampling a topic for one whole sentence at a time and uses a RNN decoder. ETM and therefore, I-NTM use these advancements in VAE to update the neural model parameters.

Interactive topic modeling. Interactive labeling of topics has been thoroughly explored for probabilistic topic models. Smith et al. (2017) compared labels generated by users after seeing topic visualizations with automatically generated labels. Hu et al. (2014a) provides a method for iteratively updating topics by enforcing constraints. Mei et al. (2007) make the task of labeling into an optimization problem, to provide an objective probabilistic method for labeling. But there has yet to be work that extends this iterative process to neural-based topic models in an intuitive and natural sense such as I-NTM. There has been extensive work in the area of anchor-based topic modeling—where a single word is used to identify a topic. Lund et al. (2017) present "Tandem Anchors" where multi-word anchors are used to interactively guide topics. Yuan et al. (2018) developed a framework for interactively establishing anchors and alignment across languages. Dasgupta et al. (2019) introduces a protocol that allows users to interact with anchor words to build interpretable topic. The most similar and recent work to outs is (Fang et al., 2023) which simultaneously developed a user-interface for interactive and guided topic modeling, based on Gibbs sampling. While it has obvious similarities, we developed the first interactive interface for neural topic models and have an interface that users can see in real-time their changes to the model.

Automatic topic modeling For a similar purpose, but through a different process, many works have sought to automatically generate labels (Alokaili et al., 2020). Where they re-rank labels from a large pool of words to label topics in a two-stage method. Lau et al. (2011) uses top terms from titles and subwords from Wikipedia articles to rank and label topics based on lexical features. Mao et al. (2012) exploit the parent-sibling relationship of hierarchical topic models to label the topics. Unsupervised methods that differ from topic models but with the same goal of clustering data also exist. LLM can be prompted to cluster data with or without labels in an intelligent way (Wang et al., 2023)

6 Conclusion and Future Work

We introduce I-NTM: a method and interface for users to interactively update topics given by neural topic models. While there have been previous efforts to improve probabilistic topic modeling through labeling, this is the first work to our knowledge that allows interactive updating of neural topic models. Especially in real-world situations, such as disaster relief, the ability to improve topics through labeling allows non-technical users to tailor the topics to their specific needs.

Additionally, our user study verifies that giving users the ability to label topics improves performance on downstream information retrieval tasks in less time, validating that more relevant documents are being found.

To take this work further and give as much flexibility to the user as possible adding the ability to guide the training of topic models by interactive labeling throughout the training, multi-word labeling instead of single, vocabulary based labels, stronger encoders, and direct access to making the adjustments in the embedding space through embedding visualizations would improve upon this presented method. Finally, while we present a suite of three different neural models, interactive topic modeling and by extension our interface, could be extended to other models such as LLMs..

Limitations

This work we seeks to solve a key limitation in traditional topic models— guiding the topics of a model in a way that is relevant to the user. Along the lines of what it means to "help" identify more relevant topics, (Hoyle et al., 2021) discusses the limitations of coherence, an automatic metric for topic model evaluation. Topic coherence is an automatic metric that is not validated by human experiments and thus its validity of evaluating topic models is limited. While our method is an attempt to improve interpretability of topic models, it still suffers from many of the problems that topic models in general do. Topic models do not conform to well-defined linguistic rules and due to the non-compositionality of labels, from a linguistic viewpoint, can be viewed as not actually modeling topics (Shadrova, 2021).

We recognize that with any study there are limitations, while topics are meant to be representative labels of the corpus, users tended to use words directly in the query or general task, treating it more as a keyword match. While this is not how topic models are meant to be used and most likely due to a lack of knowledge about topic models, this process did work in most cases at improving the relevancy scores for the questions.

Finally, the BM25 requires a query to calculate the scores. We used the scenario and corresponding question as the query (removing stopwords), however a variation in query could lead to different BM25 scores. While this does not change the fact that labeling topics on average improved BM25 scores, it means a good query is required to effectively rank documents.

Ethical Considerations

The data that we used for the experiments in this paper was all human gathered by others and ourselves. If I-ETM was to be used in a real-word situation, where identifying key documents or tweets about a time-sensitive issue was paramount, any failures in the system could result in a negative outcome if the wrong information is disseminated. We went through the appropriate IRB pipeline to receive approval for our human conducted study. The users were paid based on the recommendation of the Prolific platform, which bases its' recommendation based on the time of the study and other studies. This was a rate of $12 an hour. No personal identification information was collected from the users, so there poses no threat to the participants of exposure of personal information.

References

Nikolaos Aletras and Arpit Mittal. 2016. Labeling topics with images using neural networks.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Nikolaos Aletras and Mark Stevenson. 2014. Labelling topics using unsupervised graph-based methods. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 631–636, Baltimore, Maryland. Association for Computational Linguistics.

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2020. *Automatic Generation of Topic Labels*, page 1965–1968. Association for Computing Machinery, New York, NY, USA.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020. Cross-lingual contextualized topic models with zero-shot learning. *CoRR*, abs/2004.07737.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

David M. Blei, A. Ng, and Michael I. Jordan. 2003a. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

David M. Blei, Andrew Ng, and Michael Jordan. 2003b. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. *Applications of Topic Models*, volume 11 of *Foundations and Trends in Information Retrieval*. NOW Publishers.

Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park. 2013. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001.

Sanjoy Dasgupta, Stefanos Poulis, and Christopher Tosh. 2019. Interactive topic modeling with anchor words.

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

Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and Rob Procter. 2023. A user-centered, interactive, human-in-the-loop topic modelling system. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 505–522, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Thomas L. Griffiths and Mark Steyvers. 2004b. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235.

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

Thomas Hofmann. 2017. Probabilistic latent semantic indexing. *SIGIR Forum*, 51(2):211–218.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken?: The incoherence of coherence.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014a. Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Yuening Hu, Ke Zhai, Vladimir Edelman, and Jordan Boyd-Graber. 2014b. Polylingual tree-based topic models for translation domain adaptation. In *Association for Computational Linguistics*.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *Proceedings of the Association for Computational Linguistics*, pages 1536–1545.

Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22.

Jeff Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Association for Computational Linguistics*.

Xian-Ling Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, and Xiaoming Li. 2012. Automatic labeling hierarchical topics. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2383–2386, New York, NY, USA. ACM.

Elijah Meeks and Scott B Weingart. 2012. The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1):1–6.

Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web*, page 533–542, New York, NY, USA. Association for Computing Machinery.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 490–499, New York, NY, USA. Association for Computing Machinery.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Ramesh Nallapati, Igor Melnyk, Abhishek Kumar, and Bowen Zhou. 2017. Sengen: Sentence generating neural variational topic model.

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.

Quentin Pleple. 2013. *Interactive Topic Modeling*. University of California, San Diego.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009a. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore. Association for Computational Linguistics.

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher Manning, and Daniel Mcfarland. 2009b. Topic modeling for the social sciences.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Anna Shadrova. 2021. Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, 2021.

Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, and Leah Findlater. 2017. Evaluating visual representations for topic understanding and their effects on manually generated topic labels. *Transactions of the Association for Computational Linguistics*, 5:1–16.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 424–433, New York, NY, USA. Association for Computing Machinery.

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10626–10649, Singapore. Association for Computational Linguistics.

Weiwei Yang, Jordan L. Boyd-Graber, and Philip Resnik. 2019. A multilingual topic model for learning weighted topic links across corpora with low comparability. In *EMNLP*.

Michelle Yuan, Benjamin Van Durme, and Jordan L. Ying. 2018. Multilingual anchoring: Interactive topic modeling and alignment across languages. In *NeurIPS*.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan L. Boyd-Graber. 2020. Interactive refinement of cross-lingual word embeddings. In *EMNLP*.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the World Wide Web Conference*.

Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. 2018. Whai: Weibull hybrid autoencoding inference for deep topic modeling.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.

A Example Appendix

B Datasets

We used the BETTER dataset and a curated Wikipedia dataset.[1] To preprocess the data, we removed English stopwords and used the 0.01 and 0.85 as the minimum and maximum document frequency, respectively.

[1]https://github.com/forest-snow/mtanchor_demo

B.1 Training details

For all the results presented in this paper, our model was trained using 4 NVIDIA RTX2080ti The I-NTM model was trained for 200 epochs using 20 topics. The ADAM optimizer is used with a learning rate of 0.005.[2] The rest of the details can be found in the appendix. For our human study, we trained a model using only 5 topics. This was due to not wanting to overwhelm users with a lot of topics and the limited number of documents in the dataset.

C Models

We used the PyTorch implementation of ETM to build our code off of.[3] We used an embedding space size and rho size of 300 and a hidden layer size of 800. The rest of the hyperparameters are the default and can be found in the original code or our own. To greatly improve training time, we used the pre-trained fasttext embeddings (Mikolov et al., 2018).

D Code

The code will be publicly made available on our Github page.

[2]we followed the other default parameters in the original paper and can be found in our code as well.

[3]https://github.com/lffloyd/embedded-topic-model