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

Anonymous EMNLP submission

Abstract

Topic models are essential for understanding document collections but often fail to identify all relevant topics. While classical probabilistic and anchor-based models offer interactive features for user guidance, such capabilities are missing in neural topic models. To address this, we introduce a user-friendly interaction for neural topic models, allowing users to assign word labels to topics. This interaction updates the model, aligning topic words with the given label. Our approach covers two types of neural topic models: those with trainable topic embeddings that evolve during training, and those with embeddings integrated post-training. We develop an interactive interface for user engagement and re-labeling of topics. A human study shows that user labeling improves document rank scores on average by at least 30% and helps users find more relevant documents compared to no user labeling.

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 numerous new models, Latent Dirichlet Allocation (Blei et al., 2003, LDA) remains the most popular method, even two decades after its introduction.

LDA continues to be the workhorse for unsupervised analysis in fields like 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, neural approaches, such as the embedded topic model (Dieng et al., 2020, ETM) and contextualized topic models (Bianchi et al., 2020, CTM), dominate the computer science literature (Zhao et al., 2021). We review LDA and neural topic models in Section 2.

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: [asia, singapore, denv, india, study, genotype]**

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, ‘asia’, showcasing how I-NTM adapts in real-time to user input, giving greater relevance and accuracy in topic representation.

There is insufficient evidence that neural topic models are better in runtime, ease-of-use, or human-centric methods (Hoyle et al., 2021). Moreover, while neural models excel at capturing complex relationships in text, they often lack the functional flexibility of classic probabilistic topic models, which users can readily improve through interaction.

Probabilistic models, such as Latent Dirichlet Allocation (LDA), offer rich interactive features that enable user-guided refinement. Examples include live feedback on topics (Pleple, 2013) and user-driven interfaces for topic adjustment (Choo et al., 2013). These capabilities allow probabilistic models to adapt to user needs dynamically, yet similar functionality has been largely absent in neural topic models. Bridging this gap requires developing neural models that support interactivity and designing interfaces that enable users to refine topics. We address these challenges, introducing methods and tools that bring the benefits of interactivity to neural topic modeling for the first time. We introduce a novel approach to interactive neural topic modeling, applying it to models by directly updating topic embeddings (ETM, NVDM) or by adding topic embeddings post-training (CTM).

Our method embeds user-provided topic labels in the embedding space and moves the corresponding topic embedding closer to the label, adjusting the center of the topic embedding to prioritize relevant words. These adjustments are detailed in Section 3.1.

While interactive techniques exist for probabilistic topic models, they do not have the rich, distributed representations inherent in neural models. Probabilistic models, such as Latent Dirichlet Allocation (LDA), rely on word-count representations. Adjustments in these models typically involve altering priors or incorporating user feedback as constraints on the distributions, which is straightforward due to their interpretable structure. However, this rigidity limits their ability to adapt to more nuanced, semantic representations of text. In contrast, neural topic models represent topics and documents in latent spaces, capturing complex relationships that probabilistic models cannot. This fundamental difference necessitates interactive methods tailored specifically to neural architectures.

Our method, *Interactive Neural Topic Modeling* (I-NTM), bridges this gap by introducing an interactive framework that takes full advantage of the flexibility and semantic richness of neural models. Unlike probabilistic methods, I-NTM allows real-time adjustments of topic embeddings, enabling users to directly influence the deep semantic space in which topics reside. This allows for more personalized and contextually relevant topic refinement. We evaluated this interactivity on traditional metrics such as topic coherence and diversity (3.1), as well as user-centric information retrieval tasks (4.3).

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 motivates embedding-based interactions.

2.1 Latent Dirichlet Allocation

Topic models, such as latent Dirichlet allocation (Blei et al., 2003, LDA), generate explanations for datasets using probabilistic inference (Griffiths and Steyvers, 2004a). One of the initial steps in using topic model output is to name the topics, either by selecting top words via algorithms (Griffiths and Steyvers, 2004b; Hofmann, 2017) or manually generating descriptive names (Mei et al., 2006; Wang and McCallum, 2006). For probabilistic models, this is just the beginning. The Bayesian framework allows for the incorporation of expert knowledge through informed priors, such as dictionaries (Hu et al., 2014b), word lists from psychology (Zhai et al., 2012), or specific organizational needs (Hu et al., 2014a) .

This feedback helps align the model with users' information needs and common sense. While fully supervised models, where each document has a topic label, offer another option (Blei and McAuliffe, 2007), they require extensive user interaction and many labeled examples. Interactive models, however, have limitations: they are slow, as probabilistic inference methods struggle to update quickly, and they often overlook vast world knowledge available from large text corpora.

2.2 Neural Topic Models

Neural topic models have emerged as a powerful alternative to probabilistic models.

One of the key strengths of neural topic models is their coherent and interpretable topics. This is primarily due to their use of non-linear functions, which are more adept at closely matching the observed distribution of words and topics in the data.

I-NTM is designed to work with any neural topic model that incorporates an embedding space. By either integrating topic embeddings directly or creating them as proxies by averaging word embeddings. We explore I-NTM with three specific nerual models, CTM, ETM, NVDM. Using the embedding space of these models, I-NTM allows users to interactively adjust and relabel topics, thus improving

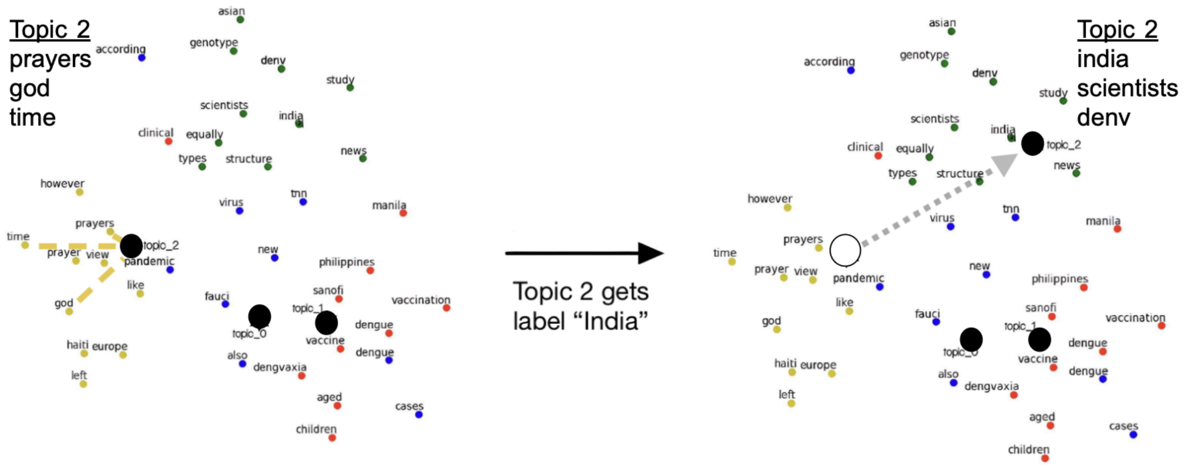


Figure 1: Visual representation labeling a new topic with out method. Our method moves the embedding center for the topic closer to the new label word, in this case, *India*.

the relevance and precision of retrieved documents based on user labels.

One popular architecture for neural topic modeling is the Variational Autoencoder (VAE). VAE-based models, such as NVDM 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 intrinsically found in ETM.

Neural topic models excel at capturing semantic nuances through distributed representations of words and topics. This capability results in topics that are more semantically meaningful and better aligned with human interpretations. ETM uses such representations by associating each topic with an embedding, which can be learned by the model or derived from pre-trained word embeddings. Unlike traditional topic models that rely on a full distribution over the vocabulary, ETM induces a per-topic distribution, offering a better representation of topics.

The adaptability of neural models makes them well-suited for handling large-scale text corpora and diverse domains. Notably, CTM uses the knowledge embedded in large language models (LLMS) to improve word representations. By combining the traditional BOW approach with embeddings from LLMS, CTM creates contextualized embeddings that enhance the quality of topic models (Bianchi et al., 2020). This integration of advanced language models highlights the versatility of neural topic models in capturing rich contextual information.

	Vocab Size	Coherence	Diversity
ETM	2565	0.19	0.81
	3572	0.17	0.85
	10830	0.11	0.92
CTM	2565	0.21	0.90
	3572	0.20	0.91
	10830	0.17	0.94
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. For some vocabulary sizes, 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 CTM and various I-NTM models on the BETTER dataset. Both models under I-NTM outperform standard ETM in topic diversity and topic coherence.

3 Interactive Neural Topic Modeling

Our interactive neural topic model enables users to actively participate in refining and adjusting the topics generated by the model. The core mechanism of this interactivity uses labels provided by users that are used to refine the topic representations. 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 dis-

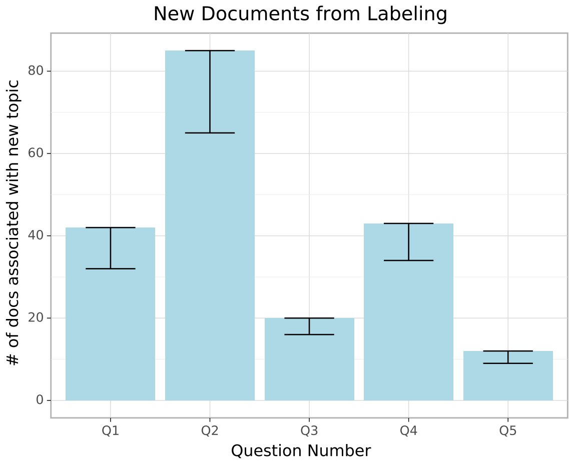


Figure 2: Labeling topics reveals, otherwise unseen, documents to be revealed. The maximum number of new documents, that is, a document that was not previously associated with the topic, found for each question across all users. The range of the number of documents found across all users is noted by the black bars.

tinct approaches to refining topic models.

Traditional topic models 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 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).

When using topic models, non-technical users use a process where 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 been interacting with topic models, we use labeling to both improve topics and help guide the model to relevant topics for the users.

We propose two key methods in I-NTM for updating topics. 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 that can be used to propagate the user inputted labels to update the model accordingly.

Both these methods follow a similar structure: Given a label, that label has a word embedding, and I-NTM moves the topic embedding closer to the new label (Equation 1).

3.1 Adjusting Learnable Topic Embeddings

This section explores 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. 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 for users' actions in a moment, but for now 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 food topic, the topic embedding is moved toward the word embedding corresponding to food. The weight of this adjustment can be tuned through the parameter λ, which determines how close the topic embedding moves to the new label.

This process is shown in Figure 1, which complements Table 1 by showing the topic and word embeddings before and after the labeling of Topic2. Initially, the words associated with Topic2 are irrelevant to the query. However, after labeling Topic 2 with India, the topic embedding shifts closer to words such as ”india”,”denv”, and “scientist,” aligning the topic more closely with the query and uncovering more relevant documents

3.2 Adding Adjustable Topic Embeddings After Training

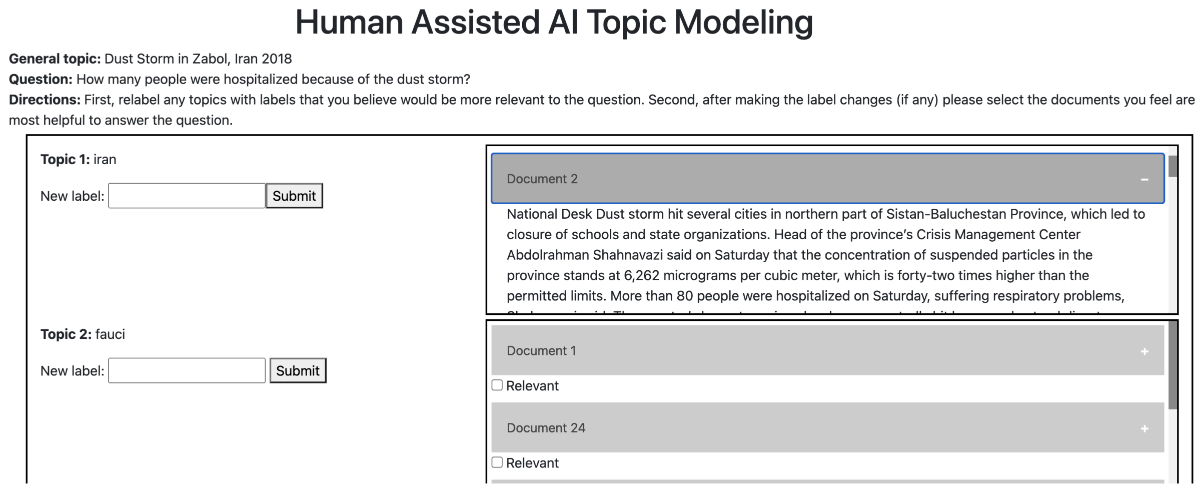
ETM and NVDM have trainable embeddings, but what about models that cannot or adding them negatively affects training? In such cases, we introduce a topic embedding post-training. We can simulate the effect of an embedding by creating a weighted average over the words that constitute a topic. This weighted average is a stand-in for a physical topic embedding, 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 probability for w_l and for similar words, w_s:

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

and for similar words,

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

where C is the amount by which you increase the probability of word, w, in topic, t.



Human Assisted AI Topic Modeling

General topic: Dust Storm in Zabol, Iran 2018

Question: How many people were hospitalized because of the dust storm?

Directions: First, relabel any topics with labels that you believe would be more relevant to the question. Second, after making the label changes (if any) please select the documents you feel are most helpful to answer the question.

Topic 1: iran

New label: Submit

Document 2 –

National Desk Dust storm hit several cities in northern part of Sistan-Baluchestan Province, which led to closure of schools and state organizations. Head of the province's Crisis Management Center Abdolrahman Shahnavazi said on Saturday that the concentration of suspended particles in the province stands at 6,262 micrograms per cubic meter, which is forty-two times higher than the permitted limits. More than 80 people were hospitalized on Saturday, suffering respiratory problems,

Topic 2: fauci

New label: Submit

Document 1 +

□ Relevant

Document 24 +

□ Relevant

Figure 3: 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.

Neural topic models often represent topics as distributions over words. Thus, when a user assigns a label to a topic they encourage the topic to pick up new words to align the topic distribution with the label.

3.3 User Interface

Although Equation 1 provides a theoretical framework for labeling topics, it cannot be used without an interface. Our low-latency interface, which is important for the user experience (Weiser, 1999), ensures efficient topic refinement with immediate feedback when users label or re-label topics (Figure 3). This fosters a dynamic interaction where users can intuitively see the result of their inputs on the model.

The interface is user-friendly, accommodating users without technical expertise. It allows them to assign labels to topics and observe in real-time how these labels alter document-topic assignments. Users can explore topics, review associated documents, and input new labels through the interface. The system handles the complex tasks of adjusting topic embeddings, recalculating document-topic distributions, and updating the display seamlessly, ensuring the interface remains effective. A key feature is its support for continuous topic updates. Users can modify a topic multiple times and update several topics concurrently. The interface incorporates safeguards to maintain topic coherence and prevent duplicates or irrelevant topics.

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 five different questions, the general topic of that question, the average 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 five minutes per question.

4 I-NTM Experimental Results

We evaluate I-NTM on standard evaluation metrics, comparison with an interactive probabilistic topic model, and through a human study. Our experiments confirm that an interactive topic modeling interface greatly improves users' ability to find relevant documents quickly.

4.1 Labeling Improves Coherence

To evaluate I-NTM, we first tested its performance using automatic metrics without human interaction. Following Dieng et al. (2020), who demonstrated that ETM improves topic coherence and diversity compared to LDA, we sought to ensure I-NTM does not degrade these metrics across different vocabulary sizes. Additionally, these experiments help identify the optimal configuration for our user study (Section B).

Automatic evaluations were conducted on two datasets: Wikipedia, a general corpus, and the BETTER dataset, which contains disaster-related news articles paired with topics and queries (details in the appendix). Using the ETM backend, we observed dataset-dependent trends in topic coherence and diversity (Table 2). For the Wikipedia dataset, labeling topics to distinguish clusters reduced topic coherence while increasing diversity. This trade-off arises because coherence measures word co-occurrence within documents, favoring broader clusters, whereas distinct labels promote diversity at the expense of coherence. In contrast, the BETTER dataset—curated for specific disaster scenarios—exhibited increased topic coherence and reduced diversity when topics were labeled more closely to match the focused nature of the queries (Table 1, Figure 1).

These findings show how dataset characteristics influence automatic metrics. While coherence generally drops when topics are distinct, labeling for specific queries can lead to improved coherence and overlapping topic words, as seen with the BETTER dataset. However, it is important to note that topic coherence, while a standard metric, is not always a reliable measure for neural topic modeling evaluation (Hoyle et al., 2021).

Human validation remains the gold standard for evaluating topic models. While we report coherence and diversity metrics for comparison, the practical utility of I-NTM is better captured through user studies, as detailed later in this section.

4.2 Probablistic Comparison

To complement our evaluation of I-NTM's interactive capabilities, we conducted a comparative analysis inspired by the iterative constraint experiment in the (Hu et al., 2014a). These experiments, performed on the 20 Newsgroups (20NG) dataset, provide a more controlled and systematic validation to assess the richness of neural model representations in contrast to probabilistic methods. While the earlier evaluations highlights practical usability and interactivity, this experiment serves as a sanity check to confirm that neural models maintain their representational power and adaptability under class-specific constraints.

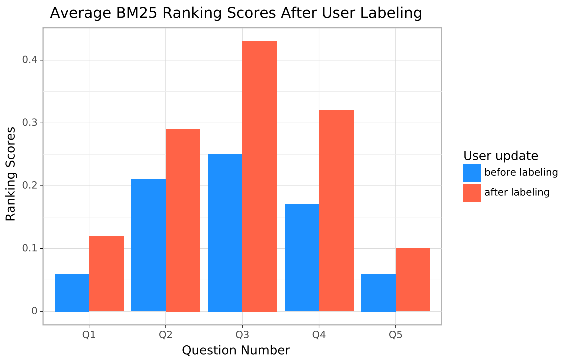


Figure 4: Average BM25 document ranking scores for each of the five questions averaged, over the 20 users. User inputted topic labels find more relevant documents and significantly improve document ranking scores

Our approach focused on incorporating class-specific constraints and using the resulting topic distributions as features for a classifier.

In Hu et al.'s ITM, the model iteratively refines topics by adding coherent sets of correlated words as constraints, to improve semantic consistency of topics. In contrast, I-NTM uses a single label per topic, iteratively updating this label by swapping it with the next word from a pre-compiled list of the 18 most relevant terms for each 20NG class. While Hu et al.'s method builds topics by aggregating sets of related terms, our approach uses distributed representations in neural models.

It is worth noting, however, that the setup used for this comparison is somewhat unfair to I-NTM. Hu et al.'s ITM benefits from iteratively adding multiple constraints, which naturally better captures the semantic space. In contrast, I-NTM's single-label approach does not aggregate constraints but rather relies on the flexibility of distributed representations to maintain topic coherence and relevance. This difference inherently puts I-NTM at a disadvantage in of topic breadth, a limitation of the chosen setup that should be considered when interpreting these results.

For both ITM and I-NTM, the topic distributions generated at each iteration become feature vectors in a classifier trained on 20NG. As a benchmark, we compared these results against a baseline classifier with no additional topic features. Figure 5 reveals a consistent trend: I-NTM demonstrated a lower error rate than ITM across all iterations, regardless of the specific labels used as constraints.

This shows that a well-chosen label can yield greater interpretability and relevance than a set of multiple constraints.[1]

4.3 Human Study

To validate the efficacy of I-NTM, we conducted a human study to assess its ability to help users find relevant documents for various information needs (e.g., *Find documents that relate to foreign intervention in Cuba*). Participants were divided into two groups: a control group (no labeling) and an interactive group (with labeling). Both groups used the CTM backend of I-NTM and were given the same queries, documents, and topic model. The goal was to identify documents that best answered the query, with the interactive group having the added ability to label topics to refine their search.

The study involved 20 participants recruited through Prolific. Topics were generated on the TREC Question Classification dataset, with 1500 documents randomly selected from the Foreign Broadcast Information Service (FBIS). Participants were allowed up to five minutes per query to label topics (if applicable) and select a maximum of five documents they believed answered the query. This setup mimicked real-world scenarios where users must quickly locate information without excessive time per query. Document relevance was assessed using BM25 (Robertson et al., 1994). We use BM25 to evaluate I-NTM in helping users find relevant documents for specific queries. BM25 serves as a benchmark to measure how well the topics refined through user labeling guide document retrieval. BM25 evaluates relevance by scoring documents against the same query presented to the user. This allows us to simulate the user's ability to recreate the steps of a classical information retrieval system.

While a BM25 system running directly on the dataset would theoretically achieve perfect retrieval, the purpose of this experiment is not to outperform BM25 but to assess how the interactive topic model supports users in finding relevant documents, especially when a fully automated retrieval system may not be available or when human judgment and iterative refinement are necessary. By comparing BM25 scores before and after user interaction, we quantify how much labeling topics improves the user's ability to retrieve documents that align with the query. This approach highlights I-NTM's utility in scenarios where user involvement is essential, such as in cases of ambiguous queries or domains requiring expert judgment.

Across most queries, the interactive group outperformed the control group, with higher BM25 ranking scores indicating the retrieval of more representative documents (Figure 4). For instance, labeling helped users streamline their search, leading to faster query resolution and broader retrieval of relevant documents. In contrast, in cases like Q5 (*"Find documents related to Chinese economic intervention in other countries"*), the control group performed well due to a dataset bias toward Chinese economic topics, reducing the marginal benefit of labeling. Q5 also had the highest average documents selected and time taken, highlighting how dataset-specific factors can impact interactivity's effectiveness.

Interactive labeling revealed many new documents previously unrelated to a topic, demonstrating the model's ability to uncover a broader range of relevant information (Figure 2). While some new documents may have aligned more with general topics rather than the specific query, the capability to dynamically adjust topics allowed users to refine results based on evolving information needs. This breadth of retrieval is particularly valuable in scenarios demanding comprehensive exploration, making the model responsive and adaptive.

These results underscore the practical utility of I-NTM and its user-friendly interface. Traditional metrics, such as topic coherence and diversity, assess internal topic quality but fail to capture real-world applicability. By emphasizing user interaction, I-NTM enables dynamic adjustments that align the model's output with user goals, improving the relevance, efficiency, and utility of the search process.

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.

[1] Since I-NTM swaps each label iteratively, the model's topic embedding gradually centers around the midpoint among the relevant terms.

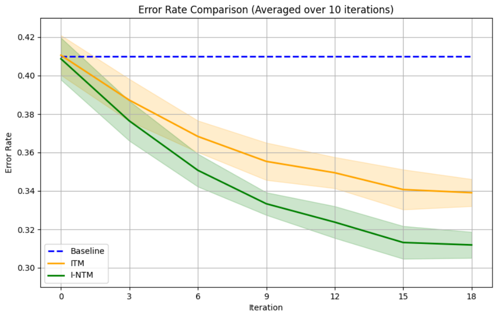


Figure 5: Error rates across 18 iterations for the baseline, ITM, and I-NTM models on the 20 Newsgroups dataset. I-NTM consistently demonstrates lower error rates than ITM and the baseline classifier, even when using only single-label constraints.

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) compare 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 ours 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, their work only works for one type of probabilistic models. We developed the first interactive interface for a suite of neural topic models and have an interface that users can see in real-time their changes to the model. Contemporaneously, Pham et al. (2024); Lam et al. (2024) developed models to study unstructed data using prompting of large language models, however, this interface is not interactive.

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 generated by neural topic models. While previous efforts have improved probabilistic topic modeling through labeling, this is the first work to enable interactive updating of neural topic models. In real-world situations this allows non-technical users to tailor topics to their specific needs.

Our user study verifies that allowing users to label topics improves performance on downstream information retrieval tasks in less time, demonstrating that more relevant documents are being found without performing worse on traditional topic model metrics.

To improve I-NTM, we aim to guide topic model training through interactive labeling, support multi-word labels, integrate stronger encoders, and enable direct embedding space adjustments via visualizations.

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-NTM 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

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. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

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*.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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.

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.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework.

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.

Mark Weiser. 1999. The computer for the 21st century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3(3):3–11.

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 Datasets

We used the BETTER dataset[2] and a version of the Wikipedia dataset.[3] To preprocess the data, we removed English stopwords and used the 0.01 and 0.85 as the minimum and maximum document frequency, respectively. A very similar version to our TREC dataset can be found at https://huggingface.co/datasets/CogComp/trec.

B Training details

For all the results presented in this paper, our model was trained using a 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. 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.

B.1 Models

For our baseline models we used the PyTorch implemenetiation of ETM.[4] 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).

For both NVDM[5] and CTM,[6] we followed the author implementations, using their recommended default parameters.

We then built upon these implementations to develop I-NTM.

C Code

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

D Results

In our study, participants were tasked with finding relevant documents based on specific queries. One such query was, *“How many people were killed by the floods in Peru?”* Initially, the topic model generated a cluster labeled "water", while these documents were broadly related to water and some did relate to the floods in Peru, they were not all relevant and did not specifically address the query about fatalities caused by floods in Peru.

To refine the topics, a user labeled the topic from “water" to “flood." This labeling improved the relevance of the documents retrieved. By specifying the label “flood," the model was directed to focus more narrowly on flood-related documents. Consequently, this adjustment brought forth a highly relevant document that directly addressed the query: *“Flooding, landslides kill at least eight in Peru and Chile Downpours typical of El Nino continue to cause destruction in South America. 09 Feb 2019 10:27 GMT Eight people have been killed across the Pacific coast of South America after heavy rain caused flooding and landslides..."* This document provided information about the number of fatalities caused by floods in Peru, thereby answering the query effectively (Table 4). The success of this labeling process shows the effectiveness of I-NTM.

[2] https://ir.nist.gov/better/portfolio

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

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

[5] https://github.com/YongfeiYan/Neural-Document-Modeling

[6] https://github.com/MilaNLProc/contextualized-topic-models/tree/master

Query	How many people were killed by the floods in Peru?
Before Labeling	**Topic 1: water** **Document 1:** Water in development Working in development As water scarcity deepens across Latin America, political instability grows Ecuadorians clash with police 30km from Quito in 2010 in a protest over proposed water privatisation. Photograph: Pablo Cozzaglio/AFP/Getty Images ... **Document 2:** Flood torrents devastate Peru and Chile Heavy rains have brought torrential floods and mudslides to parts of Peru and Chile - including the Atacama desert, one of the driest regions of the world. Several people have been killed, and homes destroyed. Thousands are without electricity and clean water and the clean-up operation is being hampered by the scale of the destruction. Bill Hayton reports ... **Document 3:** Image for representation only. Credit: Thinkstock Nikhil Chand, who heads the chocolate and confectionery portfolio for Nestle in India has been with the Swiss major for over 18 ...
After Labeling	**Topic 1: flood** New label: flood (user input) **Document 1:** March 17, 2017 / 11:12 PM / 2 years ago Abnormal El Nino in Peru unleashes deadly downpours; more flooding seen Mitra Taj 4 Min Read LIMA (Reuters) - A sudden and abnormal warming of Pacific waters off Peru has unleashed the deadliest downpours in decades, with landslides and raging rivers sweeping away people, clogging highways and destroying crops. ... **Document 2:** Peru floods put thousands in danger Unseasonal rain in Peru has forced one region to declare a state of emergency. Police have closed bridges for fear of collapse in the city of Piura, the regional capital. 12 Mar 2017 09:41 GMT ... **Document 3:** *Flooding, landslides kill at least eight in Peru and Chile Downpours typical of El Nino continue to cause destruction in South America. 09 Feb 2019 10:27 GMT Eight people have been killed across the Pacific coast of South America after heavy rain caused flooding and landslides. Three people died in Peru when mudslides hit two towns in the southern Arequipa region. ...*

Table 4: A real example of labeling bringing forth more relevant documents. After a user labeled Topic 1 as "flood", a document answering the query was found. A topic of "flood" is more relevant to the query than a general topic of "water".