Labeled Interactive Topic Models

Anonymous ACL submission

Abstract

Topic models help users understand large document collections; however, topic models do not always find the "right" topics. While classical probabilistic and anchor-based topic models have interactive variants to guide models toward better topics, such interactions are not available for neural topic models such as the embedded topic model (ETM). We correct this lacuna by adding an intuitive interaction to ETM: users can label a topic with a word, and topics are updated so that the topic words are close to the label. This allows user to refine topics based on their information need. We evaluate our method through a human study, where users can relabel topics to find relevant documents. We find that 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.

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., 2007). 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), policitical 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), exemplified by the embedded topic model (ETM), which amalgamates topic models with word embeddings (Dieng et al., 2020). We review LDA and ETM 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). We are sympathetic to these arguments, and we discuss them in detail at the end of this paper (Section 7).

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: incorporating rich priors, incorporating syntactic information, or structural priors from covariates (Mcauliffe and Blei, 2007; Griffiths et al., 2004; Boyd-Graber and Blei, 2008). Richer interactions are also possible through tree-based priors and through spectral methods (Hu and Boyd-Graber, 2012; Arabshahi and Anandkumar, 2017). Unfortunately, these improvements are not currently available for neural topic models.

In an effort to expand the options that neural topic models have and fill in the lacunae between probabilistic and neural models. Our goal is to extend the neural topic modeling method, ETM—making it interactive—to give users the capability to change the topics to fit their request or needs better.

ETM improves upon LDA by introducing topic embeddings, where each topic embedding is a distributed representation in the semantic space of words, inducing a per-topic distribution over the vocabulary. This is in contrast to traditional topic models, where each topic is a full distribution over the vocabulary. Here, to use ETM interactively—based on the topic label from the user—we embed

Task: Dengue outbreak in Asia **Request**: What countries are seeing an outbreak?	
No topic labeling	After 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'	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: (Right) The closest words to the different topic embeddings after running ETM. While the first two topics are related to the task/request, Topic 2 is not. (Left) shows the updated Topic 2 after moving the topic embeddings towards the word "india".

the label in the embedding space and *move* the corresponding topic embedding closer to the label. This adjusts the center of the topic embedding: throwing out unrelated words, prioritizing words that are "close" to the users' label.

There have been many previous works for interactively labeling probabilistic topic models, which humans already do post-process. For example, UTOPIAN (Choo et al., 2013), enables users to interact with the topic model and steer the result in a user-driven manner, such as topic merging or document-induced topic creation. Our work, in contrast, introduces a way of improving topics through labeling in a natural way that is typically done a posteriori. We call this method *Interactive Embedded Topic Modeling* or I-ETM.

We conducted a human study to demonstrate the efficacy of our interactive labeling method over base ETM. Additionally, if a user has a specific task when running a topic model on a corpus, our interactive labeling method qualitatively helps users quickly identify documents relevant to their information needs. While I-ETM can be used in any setting where topic modeling is useful, our method can be especially helpful in an urgent setting where relevant documents need to be quickly identified, giving the users the ability to direct the model to the most relevant documents for their request.

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

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

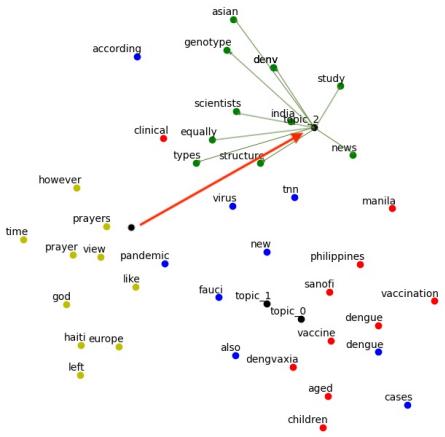

Figure 1: The topic and word embeddings that corresponds with Table 1 before and after the labeling of Topic 2. The topic embedding moves towards a new cluster of word embeddings after a label is used.

2.1 Latent Dirichlet Allocation

Topic models are exemplified by latent Dirichlet allocation LDA (Blei et al., 2003b). Given a large collection of documents and an integer parameter K, topic models like LDA find the K topics that best describe the collection.

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). While we do not fully recapitulate the LDA generative story here—our focus is on neural models after all—the key is that one part of the story is a distribution over words for each of the K topics. 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). This is common especially in the social sciences, where topics are given sentences to describe the documents that make it up such as "topic is associated with articles on the life and works of Goethe" (Riddell, 2012).

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), the needs of a business organization (Hu et al., 2014a). This feedback to a model helps correct word sense issues, 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. Our next model incorporates this world knowledge through word embedding vectors that learn this representation through the corpora, however it replaces the Bayesian priors and lacks the interactivity of probabilistic models.

2.2 Embedded Topic Modeling

The embedded topic model (Dieng et al., 2020, ETM) takes advantage of these representations by associating each topic with an embedding. In addition, each token in the vocabulary also has a L-dimensional 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, in ETM the k^{th} topic 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 that representation. More concretely, ETM induces this distribution from a log-linear model that takes the inner product of the word embedding matrix ρ and the k^{th} topic embedding vector α_k:

$$\beta \equiv \text{softmax}\left(\rho^\top \alpha_k\right). \quad (1)$$

ETM assigns high probability to a word v in topic k if its representation is close to the topic embedding: In the next section, we take advantage of this by allowing users to adjust the topic embedding by assigning a label to a topic.

3 Interactive Embedded Topic Modeling

Why is changing the labeling of a topic model a good way for interaction? Previous topic models do not have explicit 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). Also, they require users to manually analyze the topics found to 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.

This could look like say for a humanist, instead of giving topics sentence descriptions a posteriori, labels such as "works of Goethe" can actually be used to improve the model itself.

Topic models can be used in time-sensitive situations such as identifying key areas that need relief supplies through social media postings (Resch et al., 2018; Zhang et al., 2021). Table 1 details an example of why our labeling method can help improve topics in these types of situations. The scenario and scenario specific questions, as well as the data used is from a dataset that focuses on disaster relief situations (Mckinnon and Rubino, 2022).

In this case, the corpus is from an information retrieval (IR) query for the given scenario of "Dengue outbreak in Asia". Also there is the question of

	Vocab Size	Coherence	Diversity	F1
ETM	2565	0.19	0.81	0.86
	3572	0.17	0.85	0.85
	10830	0.11	0.92	0.76
Interactive ETM	2565	0.14	0.84	**0.93**
	3572	0.10	0.88	0.89
	10830	0.10	**0.95**	0.77

Table 2: Topic coherence, topic diversity, and classification accuracy for varying vocabulary sizes for regular ETM and our interactive ETM. While the F1 scores drop as the vocabulary size increases, our method outperforms ETM in terms of topic diversity and F1 score.

"What countries are seeing an outbreak?" which gives a user more specific information to find regarding the given scenario.

After running I-ETM, the left side of the table shows the top three topics and their corresponding words. The first two topics have words relevant to the scenario and question, such as "dengue", "virus", and "vaccine" ("dengvaxia" is the name of the vaccine created by "sanofi"). However, the third topic is less relevant, with words such as "gonaives" which is a commune in Haiti. While dengue, a mosquito-borne virus, might be present in Haiti, the scenario is focused on Asia. So perhaps a user knows there has been an outbreak in India, they are able to label Topic 2 (chosen to be least relevant to the scenario/question) with a word that might bring forth more relevant documents to the request, such as "india". This shifts the Topic 2 embedding towards the word embedding for "india" and now the nearest words that make up the new topic help to focus on documents more related to Asia.

3.1 Adjusting topic embeddings

As discussed above, ETM induces a topic distribution from word representations and a topic embedding (Equation 1). 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

$$\alpha_k^{new} = \lambda(\vec{w_k} - \vec{\alpha_k}^{old}) + (1 - \lambda)\vec{\alpha_k}^{old} \quad (2)$$

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, in (Figure 1) the topic and word embeddings are shown before and after the adjustment of Topic 2. We can see the words surrounding Topic 2 before adjusting the label, at first read do not seem to be relevant to the task or request. After the labeling of Topic 2, we see the topic embedding, is close to the words "india", "singapore", and "asian", which are more relevant to the request and could bring forth more relevant documents.

3.2 Cross-lingual topic modeling

While I-ETM adds interactivity to a neural-based topic model and improves the relevancy of documents, our model initally, like ETM lags behind the state-of-the-art in cross-lingual capabilities. (Bianchi et al., 2020) found that replacing the traditional bag-of-words (BOW) input for contextualized embeddings improved the cross-lingual capabilities of their neural topic model. We follow a similar structure, replacing the BOW input with pre-trained multilingual representations from SBERT (Reimers and Gurevych, 2019).

4 Findings

4.1 Training details

For all the results presented in this paper, our model was trained using 4 NVIDIA RTX2080ti The I-ETM model was trained for 200 epochs using 20 topics. The ADAM optimizer is used with a learning rate of 0.005.[1] 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

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

Language	Sentence	Predicted Topics
EN	Philippines halts sale of dengue vaccine...	dengue, vaccine, sanofi, dengvaxia
ES	Filipinas suspende la venta de la vacuna...	dengue, vaccine, sanofi, dengvaxia
EN	Flood torrents devastate Peru...	rains, peru, heavy, floods
FR	Des torrents d'inondations dévastent le Pérou...	rains, peru, heavy, floods
EN	The earthquake caused buildings to collapse...	earthquake, rubble, quake, aug
IT	Il terremoto ha causato il crollo di edifici...	earthquake, rubble, quake

Table 3: Predicted topics for English documents translated into various languages. With the addition of multilingual embeddings, I-ETM is able to predict similar topics across languages.

to not wanting to overwhelm users with a lot of topics and the limited number of documents in the dataset.

4.2 Cross-lingual capabilities

While the focus of this paper is the interactivity of our model, Table 3 demonstrates the cross-lingual capabilities of I-ETM. The predicted topics for different English sentences and their corresponding translations are shown. I-ETM is able to predict similar topics for the same sentence in different languages. Continuing with our running example of using our model in disaster relief situations, having the capability to group documents in different languages with understandable representative topics is vital for quick and effective responses. It is important to note, that since our formulation was adapted from (Bianchi et al., 2020), this is done in a zero-shot setting. The model is not trained on documents in the other languages, this capability is just due to the pre-trained multilingual embeddings.

4.3 Topic coherence and diversity

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 (NPMI). Given a set vectors of the top words in a topic, $\{w_1, w_2, ..., w_N\}$, the PMI is

$$\text{PMI}(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)}.$$

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

Since (Dieng et al., 2020) showed improvements in topic coherence and diversity over LDA, to check if our method negatively affects them, we looked at coherence, diversity, and F1 scores for varying vocabulary sizes between ETM and I-ETM. Topic coherence drops with our method, but diversity and F1 scores are 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-ETM actually results in a decrease in topic diversity and an increase in topic coherence. In either situation it is important to note, that topic coherence has been found to be a poor metric for topic modeling evaluation (Hoyle et al., 2021). We report scores for coherence and diversity since this is the current standard for topic model evaluations.

5 Human Study

To validate the efficacy of I-ETM, we recruited participants to test our model in finding more relevant documents for different scenarios and information needs. Information retrieval tasks are an intuitive way to measure the success of our method, since they involve finding relevant information specific to a query. Comparing the scores from these scenarios, before and after labeling, is used to verify that user labeling brings forth more relevant documents. We find users do relabel at least one topic on average and the labeling does improve ranking scores on a common information retrieval algorithm.

5.1 Setup

To conduct this study, we recruited 20 participants through the online platform Prolific, whom were given one hour to complete the task. Our study is

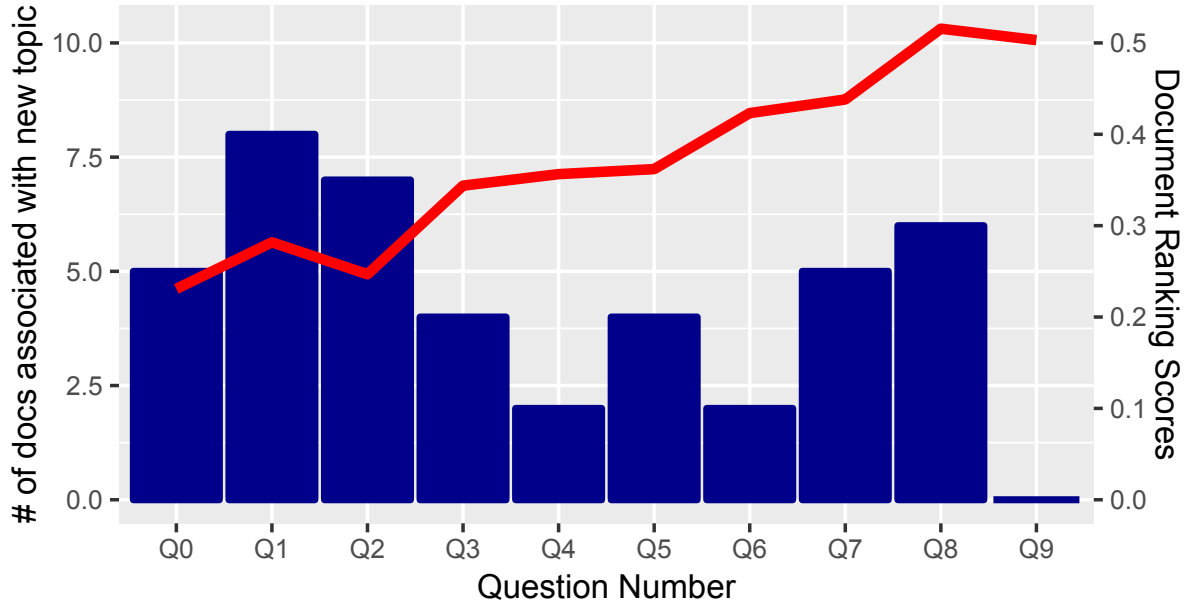

Figure 2: The increase of total document ranking score across all scenarios with the number of new documents associated with the new labeled topics

set up as follows:

1. Our model I-ETM is used to generate topics on a dataset related to disaster relief situations[2]
2. Participants are shown an information need with topics generated by our model. After consideration, they have the opportunity to label topics in a way they deem best
3. After labeling topics, they are asked to select a few documents that they believe best answer the information need

For each user we collected the topic information and document distribution before and after the human interaction. Then following the algorithm commonly known as BM25 (Robertson et al., 1994), an information retrieval ranking function. We compare the estimated relevancy of topics before and after the human interaction. BM25 works by using a bag-of-words retrieval function that ranks a set of documents based on query terms present in the document. Formally speaking, given a set of keywords $q_1, q_2, ..., q_k$ from a query, Q, the score function is defined as:

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\tilde{n}}} \quad (3)$$

where $f(q_i, D)$ is the number of times q_i appears in D, $|D|$ is the number of words in D, $\tilde{n}$ is the average document length, and k_1, b are the term frequency and document length scaling factors, respectively.

5.2 Results

In Figure 3 we see the average BM25 document ranking scores for each question averaged over the 20 users. We show the scores for the topics that were changed, before and after the users made the change. This shows that on average, changing the topic led to more representative documents being shown (when looking at up to the first 5 documents for each topic). In the case of *Q2* and *Q9*, no *Before update* bar is present because some topics initially had no documents clustered under that topic. We believe that due to the limited size of the dataset, a few topics covered a majority of the documents, causing some topics to have no documents clustered to it.

To understand how non-technical users interact with our topic model, we kept track of the average number of topic changes per question, shown in Table 4. Across all questions there were 1.24 topic changes. While there was no expected number of

[2]we chose the BETTER disaster relief dataset mentioned throughout the paper since it has documents sorted into different scenarios with questions already given.

topic changes, we believe a lower average could be due to a few reasons. First, a topic label must be a word that is present in the model vocabulary, so if a user tried to label a topic with a word not in the vocabulary, they were alerted of that. Secondly, if the user was not familiar with topic models (which we had no requirement for technical experience, so this was most likely the case), it's possible they had a hard time coming up with good topic labels. We see the best case of this on *Q10*, where the average number of topic changes was only *0.40*, with many users not changing any topics at all. However, for this question, users on average made *3.15* attempts at relabeling a topic, indicating that users had difficulty finding a representative topic within the vocabulary.

Additionally, we calculated the average rank of the relevant documents chosen by the users, scored using the BM25 algorithm. Since only up to the first five most relevant documents for each topic were shown, this gives insight to how well users are able to choose relevant documents after labeling any topics. In Table 4 the ranks are shown for each of the 10 questions, with an average rank of 8.7 across all questions. With each question having 50 documents, we see that users are able to select on average documents in the top 20% of most relevant documents. Due to there being a time restraint and some documents being quite long, we believe these ranks would improve even more if users had more time. However, we wanted to limit the time to mirror a real-life situation where a user might need to be both effective and efficient.

Figure 2 shows how over the course of the study, as the user labels topics for more scenarios, the overall document ranking score increases. This is demonstrating how for a given corpus, by labeling topics for different questions, one can incrementally increase overall document relevancy. This figure also shows the average number of new documents associated with the new topics that the user labels. A high number of new associated documents, does not necessarily correlate with a larger increase in document ranking score, as some low ranked documents could pull down the rest.

6 Related Work

Neural topic models With the recent developments in deep neural networks (DNNS, there has been work to leverage 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 usung hybrid of 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-ETM use these advancements in VAE to update the neural model parameters.

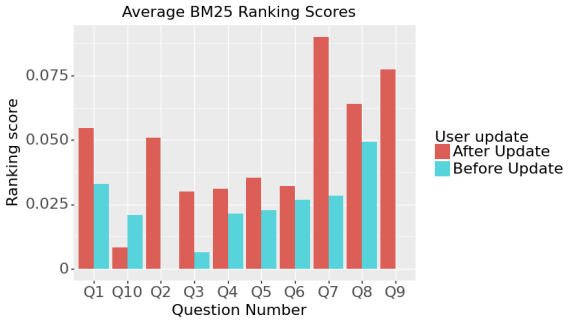

Figure 3: Average BM25 document ranking scores for each of the 10 questions averaged, over the 20 users.

Interactive topic modeling. Interactive labeling of topics has been thoroughly explored for probabilistic 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. (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-ETM. There has been

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Average # topic changes	1.85	1.45	2.0	0.82	1.0	1.05	1.25	1.95	0.65	0.40
Average rank of relevant docs	4.25	11	6	13.7	9.4	13	3.25	7.25	3.6	11.6

Table 4: Tabular view of averages across different metrics from human study. "Average # of topic changes" is the average number of topics that users gave a new label for each of the 10 questions. "Average rank of relevant docs" is the average rank of the documents selected as most relevant by the users (lower is better).

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.

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.

Cross-lingual topic modeling. (Mimno et al., 2009) were the first to introduce a multilingual topic model using LDA with Polylingual Topic Model. With many works in multilingual topic modeling to follow (Liu et al., 2015; Hao and Paul, 2018). But few works focus on cross-lingual topic modeling, with enables cross-lingual representation transfer to model topics across languages. (Heyman et al., 2016) develop C-BiLDA, a cross-lingual LDA model, which outperforms BiLDA (De Smet and Moens, 2009) and does not assume different language corpora share a common topic distribution. However, these works typically require extensive additions to the topic models to get good performance, where the methods we follow take advantage of the advancements in pretrained multilingual embeddings from large language models.

7 Conclusion and Future Work

In this work, we introduced a method 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 to improve the found topics. Especially in real-world situations, such as disaster relief, the ability to improve topics through labeling allows users to tailor the topics to their specific needs.

The interactivity can help classification accuracy without having a significantly negative (if at all) effect on topic coherence and diversity. In recent years, many works such as contextualized topic models (Bianchi et al., 2020), have taken advantage of large language models by using the pretrained multilingual embeddings from SBERT (Reimers and Gurevych, 2019) to predict the topic distributions. Adapting this method for our own model is shown to work well even in the zero-shot cross-lingual setting, bringing a neural-based interactive topic model into the cross-lingual space. Additionally, through a user study, we verified that giving users the ability to label topics improves performance on downstream information retrieval tasks, validating that more relevant documents are being found.

To take this work even further, we believe adding the ability to guide the training of topic models by interactive labeling throughout the training process would greatly improve upon this presented method. Similar work has been done in the probabilistic space, however, we leave this extension in the neural-based architectures to future work.

Limitations

In this work we sought to solve a key limitation in traditional topic models— guiding the topics of a model in a way that is relevant to the user. While we believe we provided an effective framework for interactively updating topics in a neural topic model, it does not come without limitations. 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 the study we conducted had limitations, which need to be considered in conjunction with our results. First, the average number of topic changes per question is low, an average of 1.24 across all questions. Users seem to be most likely to make one topic change and then choose relevant documents. This could be due to a few reasons; lack of knowing another good topic label or believing sufficient documents were brought forward after only one label.

Second, 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. No personal idenficiation 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.

Forough Arabshahi and Anima Anandkumar. 2017. Spectral Methods for Correlated Topic Models. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1439–1447. PMLR.

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.

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. 3.

Jordan Boyd-Graber and David Blei. 2008. Syntactic topic models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation.

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.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. pages 57–64.

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.

Thomas Griffiths, Mark Steyvers, David Blei, and Joshua Tenenbaum. 2004. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Thomas L. Griffiths and Mark Steyvers. 2004a. Finding scientific topics. 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.

Shudong Hao and Michael J. Paul. 2018. Learning multilingual topics from incomparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2595–2609, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2016. C-bilda extracting cross-lingual topics from non-parallel texts by distinguishing shared from unshared content. *Data Mining and Knowledge Discovery*, 30.

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 and Jordan Boyd-Graber. 2012. Efficient tree-based topic modeling.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014a. Interactive topic modeling. 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. 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.

Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. 2015. Multilingual topic models for bilingual dictionary extraction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 14:1–22.

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. CIKM '12, pages 2383–2386, New York, NY, USA. ACM.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Timothy Mckinnon and Carl Rubino. 2022. The IARPA BETTER program abstract task four new semantically annotated corpora from IARPA's BETTER program. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3595–3600, Marseille, France. European Language Resources Association.

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.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, Singapore. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Bernd Resch, Florian Usländer, and Clemens Havas. 2018. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, 45(4):362–376.

Allen Beye Riddell. 2012. *How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models*, pages 91–113. Camden House.

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.

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.

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

Ting Zhang, Shi Shen, Changxiu Cheng, Kai Su, and Xiangxue Zhang. 2021. A topic model based framework for identifying the distribution of demand for relief supplies using social media data. *International Journal of Geographical Information Science*, 35:1–22.

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 and a curated 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.

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

B Models

We used the PyTorch implementation of ETM to build our code off of..[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.

C Human Study Interface

We provide a sample page of the human study interface that participants saw. After a few pages of instructions and example scenarios, the user is give a set of questions to choose from and each question brings this screen (Figure 4) with different information. The general topic and corresponding question are shown, as well as a small reminder of the instructions. Users see the different topics with topic words, the the space to enter a new label. Additionally, all the associated documents are shown with dropdown bars, where the user can read the whole document. Finally, there are boxes to check for the relevant documents.

D Code

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

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

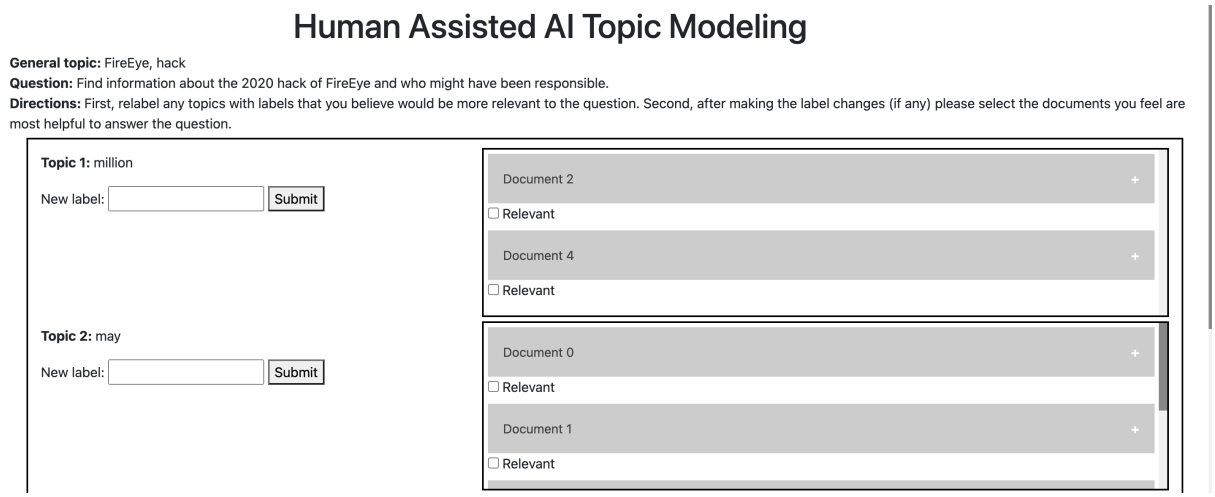

Figure 4: Human study interface for I-ETM. Users can see the given topics that are found for a set of tasks/requests and can change the label to better fit their needs. Additionally, the assigned documents for each topic are shown and users can select which documents are most relevant.