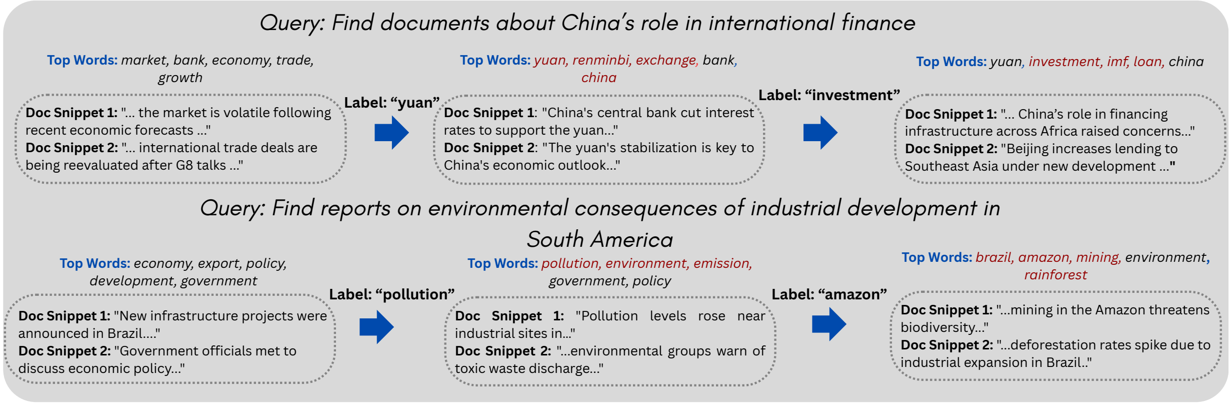
Chapter 3: Interactive Neural Topic Modeling

3.1 Introduction

As we saw in Section 2.3, while there are a suite of interactive methods for probabilistic models, but these have yet to extend to neural topic models. We fill that gap in the following section.

3.2 Introduction

Making neural models interactive requires two things: models to support interactivity and an interface to allow users to make changes to the model. This line of work provides both and applies them to models through an interaction that permits users to assign a word label to a topic, leading to an update in the topic model where the words in the topic become closely aligned with the given label. There are two modes of interaction; directly updating topic embeddings or by adding topic embeddings to the model after training. We call this framework *interactive neural topic models* (I-NTM).



Query: Find documents about China's role in international finance

Top Words: market, bank, economy, trade, growth
Label: "yuan"
Top Words: yuan, renminbi, exchange, bank, china
Label: "investment"
Top Words: yuan, investment, imf, loan, china

Doc Snippet 1: "... the market is volatile following recent economic forecasts ..."
Doc Snippet 2: "... international trade deals are being reevaluated after G8 talks ..."

Doc Snippet 1: "China's central bank cut interest rates to support the yuan..."
Doc Snippet 2: "The yuan's stabilization is key to China's economic outlook..."

Doc Snippet 1: "... China's role in financing infrastructure across Africa raised concerns..."
Doc Snippet 2: "Beijing increases lending to Southeast Asia under new development ..."

Query: Find reports on environmental consequences of industrial development in South America

Top Words: economy, export, policy, development, government
Label: "pollution"
Top Words: pollution, environment, emission, government, policy
Label: "amazon"
Top Words: brazil, amazon, mining, environment, rainforest

Doc Snippet 1: "New infrastructure projects were announced in Brazil...."
Doc Snippet 2: "Government officials met to discuss economic policy..."

Doc Snippet 1: "Pollution levels rose near industrial sites in..."
Doc Snippet 2: "...environmental groups warn of toxic waste discharge..."

Doc Snippet 1: "...mining in the Amazon threatens biodiversity..."
Doc Snippet 2: "...deforestation rates spike due to industrial expansion in Brazil.."

Figure 3.1: A user explores the TREC using I-NTM to answer the query, After assigning two successive labels, "yuan" and "investment", the corresponding topic embedding is updated, shifting the topic toward relevant terms. This adjustment reshapes the topic-word distribution, brings forth new relevant documents, demonstrating how user-guided interaction enables better information retrieval

3.3 Interactive Modeling

Topic modeling aims to uncover latent thematic structure in document collections. However, unsupervised models, whether probabilistic or neural like variational autoencoder based methods[**?** **?**], are typically trained without regard to user goals, domain-specific knowledge, or task-specific relevance. As a result, these models often produce topics that are incoherent, redundant, or misaligned with the intended use case. In practical settings, such as social science analysis, digital humanities, or interactive search, users frequently need the ability to correct, refine, or guide the model's behavior.

Interactive topic modeling (ITM) emerged to address this mismatch between unsupervised topic inference and user needs. The central idea is to bring humans into the loop: allowing users to iteratively inspect the topics, provide feedback (e.g., by indicating that certain words should or should not belong to the same topic), and have the model update accordingly. This not only improves topic interpretability and utility but also aligns model outputs with human expectations.

The original ITM framework by **(author?)** [76] introduced a method for incorporating user-specified word constraints into LDA, ensuring that certain sets of words appeared together in a topic. This approach was later refined to allow iterative feedback and model updating[20], enabling users to gradually shape topics over time. Other techniques, such as Dirichlet forest priors [90], seeding topics through lexical priors [91], or word correlation [92] offered more flexible mechanisms for encoding soft or structured constraints on word co-occurrence.

These approaches share a common goal: to reduce the burden on users to interpret poorly formed topics and to allow users to influence model outputs in a controllable, interpretable way. Moreover, interactive systems such as UTOPIAN [93] and TopicPanorama [94] combined interactive modeling with visual analytics, empowering users not only to modify model internals but to do so through intuitive visual interfaces.

Despite the success of these probabilistic frameworks, interactive modeling has yet to be fully extended to neural topic models. Neural topic models offer advantages such as improved topic coherence [62] and flexibility in incorporating deep text representations, but they also come with challenges: their parameters are typically less interpretable, and their inference processes are less amenable to direct intervention. Thus, building interactive mechanisms into neural architectures remains an open and necessary direction.

3.3.1 Topic Representations in Neural Models

In traditional neural topic models, such as the Neural Variational Document Model (NVDM) [18] and ProdLDA [19], topics are formed implicitly via a parameter matrix $\beta \in R^{K\times V}$, where K is the number of topics and V is the vocabulary size. Each row β_k corresponds to the unnormalized logits for a topic-word distribution, typically passed through a softmax to yield:

$$p(w \mid z = k) = \text{softmax}(\beta_k).$$

Here, β is not learned as a standalone set of topic representations; instead, it is often simply the decoder weight matrix in a variational autoencoder framework. The document-topic distribution $\theta_d \in R^K$ is obtained by transforming a sampled document latent variable z_d through a neural network and a softmax layer. Word generation is then modeled as:

$$p(w \mid d) = \theta_d^\top \text{softmax}(\beta),$$

which approximates a mixture over topics.

This formulation lacks a direct, interpretable representation of topics as standalone entities. The topic vectors β_k are not constrained to form an embedding space, nor are they regularized to be semantically coherent. Topics are instead emergent properties of the decoder weights, and while the model may achieve good perplexity or coherence, there is no mechanism to inspect or manipulate topics directly.

To address this, some recent work introduces *explicit topic embeddings* $\alpha \in R^{K \times d}$, where d is the embedding dimensionality [62]. These topic vectors are designed to live in the same space as word embeddings $W \in R^{V \times d}$, enabling topic-word distributions to be defined by similarity:

$$p(w \mid z = k) = \text{softmax}(\alpha_k^\top W^\top).$$

This modification has several important implications:

- **Interpretability**: Each topic α_k becomes a vector in semantic space, whose nearest neighbors in W define the most representative words. This makes topics directly inspectable and interpretable.

- **Manipulability**: Because topics are embeddings, they can be steered or adjusted using gradient updates, regularization, or user-provided vectors (e.g., label-based guidance). This opens the door to interactive or guided modeling.

- **Expressivity**: The dot-product similarity between α_k and word embeddings enables richer, context-sensitive word distributions. Unlike a raw linear decoder, topic embeddings can capture fine-grained semantic relationships.

In our framework, we build on this embedding-based design to enable interactivity. By allowing users to assign labels or guide topics toward particular words, we can update the topic embeddings α to reflect these constraints, thereby shifting the overall topic-word distribution. This formulation also enables post hoc modifications, where new topic embeddings can be added or blended into the space without retraining the full model.

3.4 Interactive Neural Topic Modeling

Topic models are widely used as an exploratory tool, helping users uncover latent themes in document collections. However, while neural topic models offer improved coherence and flexibility over traditional probabilistic approaches, they remain static once trained and can generate topics that are misaligned with user needs. Errors in topic assignments, incoherent topics, or topics that do not match the user's expectations are common [56, 95, 96]. Without interactivity, users are forced to either accept these imperfections or manually curate results, limiting the effectiveness of neural topic modeling in real-world applications.

We introduce two methods for interactive topic updates: (1) learnable topic embeddings, where models incorporate trainable topic representations that evolve during training, and (2) post-training adjustments, where topic embeddings are modified after training to reflect user-provided labels. In both cases, the adjustment follows a common structure: given a label, I-NTM retrieves its word embedding and moves the corresponding topic embedding closer to that label. These updates allow for more intuitive and precise topic refinement, ensuring that users can adapt topics to better reflect their evolving information needs.

Our interactive neural topic model (I-NTM) addresses this limitation by allowing users to actively refine and adjust topics after training, instead of relying solely on the model's output. While models like BERTopic [30] introduce guided topic modeling through techniques such as seed words, they still suffer from many of the same issues as traditional static models, once trained, adjusting topics requires retraining or significant post-processing. I-NTM, in contrast, enables real-time modifications, allowing users to iteratively refine topic distributions without the need for retraining (we discuss this more in Appendix **??**).

This interactivity is particularly important because topic models are most useful in exploratory settings, where users might refine topics as they uncover patterns in the data. When topic models generate incoherent topics or associate documents with the wrong themes, interactive labeling provides a corrective mechanism that adapts topics on the fly. Since labeling has always been a natural way users engage with topic models, I-NTM formalizes this process by embedding user-provided labels into the model's latent space, allowing dynamic updates that reshape topic-word and document-topic distributions.

3.5 Trainable Topic Embeddings

Updating learnable topic embeddings directly: Neural topic models such as ETM and ECRTM associate each topic t_k with a dense vector $\vec{\alpha}_k \in R^d$, where d is the dimensionality of the embedding space. These vectors act as the semantic centers of topics and are used to define the topic-word distribution, typically through similarity-based functions (inner product or Euclidean distance) with word embeddings.

In traditional training, the topic embeddings are optimized using reconstruction-based objectives like ELBO, often via backpropagation. However, once trained, these models are static, limiting user control. To support interactivity, we propose allowing users to adjust a topic embedding based on a semantic label. That is, the user selects a word they believe better represents the topic, and the model updates the topic embedding to reflect this guidance.

Let's assume a user is trying to find documents about China's role in international finance. Initially, the most relevant topic, $\vec{\alpha}_k^{\text{old}}$, may include vague or unrelated words such as *agreement*, *region*, and *policy*. The user labels the topic with "yuan", a term more directly tied to monetary policy and Chinese finance. Using Eq. 3.1, the topic embedding $\vec{\alpha}_k$ shifts closer to the embedding of "yuan", updating the semantic center of the topic.

We do this by computing an updated embedding via:

$$\vec{\alpha}_k^{\text{new}} = \lambda(\vec{w}_k - \vec{\alpha}_k^{\text{old}}) + (1 - \lambda)\vec{\alpha}_k^{\text{old}} \tag{3.1}$$

Here, $\vec{w}_k$ is the word embedding of the label provided by the user (e.g. "yuan"), and $\vec{\alpha}_k^{\text{old}}$ is the original topic embedding. The parameter $\lambda \in [0, 1]$ controls the strength of the user feedback, where $\lambda = 0$ preserves the original topic. Unlike a convex combination (i.e., $\vec{\alpha}_k^{\text{new}} = \lambda\vec{w}_k + (1 - \lambda)\vec{\alpha}_k^{\text{old}}$), our formulation centers the update on the *directional shift* from the current embedding to the target embedding. Mathematically, this can be interpreted as a directional gradient step toward the user-desired concept. The use of $(\vec{w}_k - \vec{\alpha}_k^{\text{old}})$ ensures that the update reflects the semantic change, preserving the structure of the embedding space while adapting the topic in a smooth and controlled manner.

After the adjustment, the topic-word distribution $\boldsymbol{\beta}$ can be recomputed as:

$$\boldsymbol{\beta}_k = \text{softmax}(\vec{\alpha}_k^{\text{new}} \cdot \mathbf{W}^{\top}) \tag{3.2}$$

where $\mathbf{W}$ is the matrix of word embeddings. This recomputation reflects the updated topic semantics in downstream document-topic and word-topic distributions. If needed, the model can resume training using these updated embeddings.

Subsequent labeling refines the topic to capture documents about China's financial policy and foreign investments. Each labeling operation acts as a controlled semantic shift, moving the cluster center (i.e., the topic embedding) and redistributing relevance across the document set.

Unlike a simple overwrite, this update preserves some of the semantic context introduced by the first label. Because each update is a directional step rather than a hard replacement, the resulting topic becomes a semantic blend between the two labels, capturing themes of both monetary (yuan) and economic engagement (investment). This produces a more nuanced and focused topic that surfaces documents discussing China's role in global finance, as seen in Figure 3.1.

Adding Adjustable Topic Embeddings After Training: Some neural topic models, such as CTM, do not define topic embeddings explicitly. Instead, topics are indirectly represented through latent distributions or contextual encoders. In such models, direct manipulation of embeddings is not feasible. However, we can simulate this effect using a proxy embedding derived from the word distribution associated with a topic.

To incorporate user feedback, we introduce a label-driven reweighting of the topic-word distribution. Suppose a user assigns the label w_l to topic t_i. We update the topic-word distribution $P(w \mid t_i)$ by increasing the probability of the labeled word and those semantically similar to it.

Formally, for the labeled word w_l, we define:

$$P_{\text{update}}(w_l \mid t_i) = P_{\text{orig}}(w_l \mid t_i) + \Delta P(w_l \mid t_i) \tag{3.3}$$

For other words w_s similar to w_l, we compute:

$$\Delta P(w_s \mid t_i) = \lambda \cdot \text{sim}(\vec{w}_l, \vec{w}_s) \cdot C \tag{3.4}$$

Here, $\lambda \in [0, 1]$ controls the strength of the update, $\text{sim}(\vec{w}_l, \vec{w}_s)$ is a similarity function, such as cosine similarity, between the user label and another word in the vocabulary, and C is a normalization or scaling constant that ensures the distribution remains valid.

The update $\Delta P(w_s \mid t_i)$ adjusts the topic distribution by increasing the probability of words semantically related to the user's label. This enables real-time, refinement of topics through labeling, even in models where embeddings are not explicitly learned.

By modifying the topic-word matrix $\boldsymbol{\beta}$ in this way, we enable post-training interactivity without retraining or re-encoding. This approach generalizes to any topic model that represents topics as distributions over words and supports the computation of similarity in a shared semantic

Method	F1 ↑	TD@15 ↑	Pur. ↑	NMI ↑
Full Text Search	0.750	–	–	–
Topic Model (Control)	0.804	0.653	0.465	0.288
Topic Model (I-NTM)	**0.874**	**0.733**	**0.501**	**0.327**

Table 3.1: Average F1 score, topic diversity (TD@15), purity (Pur.), and NMI for full-text search, topic modeling without labeling (control), and with labeling (I-NTM) on the TREC dataset. Topic metrics are not applicable to full-text search.

space.

3.6 Experiments

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

3.6.1 Labeling Improves Topics

To evaluate I-NTM in terms of topic quality, we compare standard and interactive versions of ETM, CTM, and ECRTM on two benchmark datasets (20 Newsgroups and AGNews) using three external metrics: topic diversity (TD@15), purity, and normalized mutual information (NMI), following **(author?)** [97] . Across both datasets, ECRTM

achieves the highest scores in all metrics, indicating superior topic coherence and alignment with ground-truth categories. When interactive labeling is applied via I-NTM, we observe consistent improvements in purity and NMI across all models, demonstrating that user-guided updates lead to better-aligned topic structures (Table 3.2). Notably, ECRTM with I-NTM slightly reduces topic diversity compared to its base model, likely due to topics becoming more focused through labeling, but it achieves the highest purity and NMI overall.

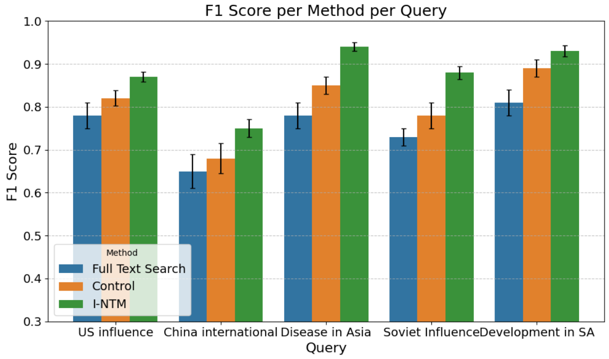


Figure 3.2: F1 scores across five queries for three methods: Full text search, standard topic modeling (Control), and I-NTM. Error bars of variability are shown. I-NTM consistently outperforms both baselines across all queries, demonstrating the benefit of interactive labeling for relevant document retrieval.

These results support the conclusion that interactivity enhances topic interpretability and alignment without significant loss in any automatic metrics, particularly when built on strong neural topic modeling backbones like ECRTM. 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.

3.6.2 Probablistic Comparison

To complement our evaluation, we compared I-NTM with an iteratively constrained probabilistic model [20], performed on the 20 Newsgroups (20NG) dataset.

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

In **(author?)** [20] ITM, topics are refined iteratively by adding coherent sets of correlated words as constraints, improving semantic consistency. In contrast, I-NTM assigns a single label per topic and updates it iteratively by replacing it with the next word from a pre-compiled list of the 18 most relevant terms for each 20NG class.

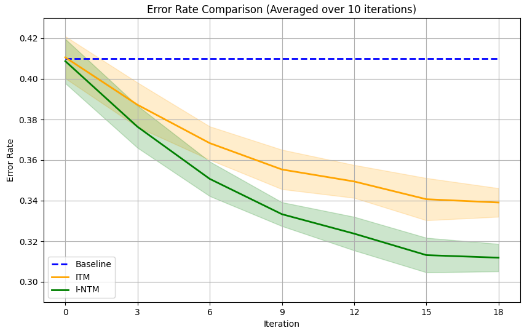


Figure 3.3: Error rates across 18 iterations for the baseline, ITM, and I-NTM models on the 20 Newsgroups dataset. I-NTM consistently demonstrates lower error rates than ITM and the baseline classifier, even when using only single-label constraints.

It is important to note that this setup inherently disadvantages I-NTM. ITM benefits from aggregating multiple constraints over time, which naturally enriches the semantic space of topics. In contrast, I-NTM's single-label approach does not accumulate constraints but instead relies on distributed representations to maintain coherence and relevance. This difference limits I-NTM in terms of topic breadth, which should be considered when interpreting results.

For both ITM and I-NTM, the topic distributions generated at each iteration are used as feature vectors in a classifier trained on 20NG. As a benchmark, we compared these results against a baseline classifier without topic-based features. Figure 3.3 shows a consistent trend: I-NTM achieved a lower error rate than ITM across all iterations, regardless of the specific labels used as constraints.

These results suggest that a well-chosen label can provide greater interpretability and relevance than multiple aggregated constraints. [1]

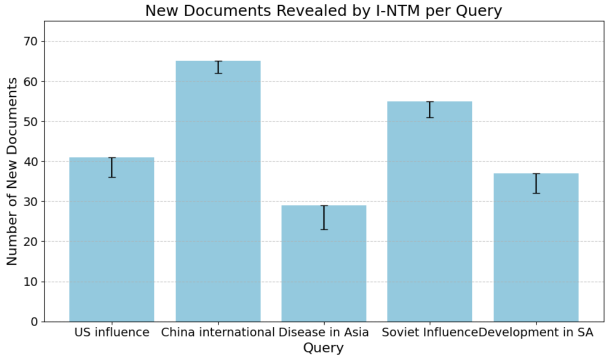


Figure 3.4: Labeling topics with I-NTM reveals, otherwise unseen, documents to be revealed. The maximum number of new documents, that is, a document that was not previously associated with the topic, found for each question across all users. The range of the number of documents found across all users is noted by the black bars.

Model	**20NG**			**AGNews**		
	TD@15	Purity	NMI	TD@15	Purity	NMI
ETM	0.731	0.356	0.301	0.833	0.663	0.225
CTM	0.773	0.411	0.314	0.871	0.698	0.247
ECRTM	**0.937**	0.605	0.560	0.960	0.804	0.361
ETM (I-NTM)	0.744	0.359	0.303	0.842	0.667	0.227
CTM (I-NTM)	0.781	0.444	0.320	0.873	0.699	0.249
ECRTM (I-NTM)	0.905	**0.615**	**0.577**	**0.968**	**0.822**	**0.387**

Table 3.2: Comparison of models on 20NG and AGNews using topic diversity (TD@15), purity (Pur.), and normalized mutual information (NMI).

3.6.3 Human Study

To evaluate I-NTM, we conducted a human study measuring its effectiveness in retrieving relevant documents. We divided 24 Prolific participants into a control group (no labeling), an interactive group (with labeling), and a full text search group, all using the ECRTM backend of I-NTM. Participants were given the same queries and documents and tasked with selecting up to five documents that best answered the query. The interactive group could relabel topics to refine search results.

[1] Since I-NTM updates the topic embedding iteratively by swapping labels, the model's representation gradually centers around the semantic midpoint of the relevant terms.

Topics were generated on the TREC Question Classification dataset, with 1500 documents randomly selected from the Foreign Broadcast Information Service (FBIS). Participants had five minutes per query, mimicking real-world constraints. To evaluate retrieval quality, we used a relevance assessment protocol based on **(author?)** [98], where LLaMA-3 [99] judged document relevance with verification from a human annotator. This provided consistent, scalable judgment across users and queries, allowing us to compute precision, recall, and F1 scores per method.

Figure 3.2 shows F1 scores across five queries for three methods: full text search, standard topic modeling (control), and I-NTM. The queries were chosen to be challenging for keyword retrieval, focusing on themes or latent topics, where text matching often fails. I-NTM consistently outperformed both baselines, with improved F1 scores for every query. Full text search and the control group sometimes performed similarly, especially on queries with high document overlap, but labeling always led to superior performance.

Lower F1 scores were observed across all methods for queries tied to dominant themes. For instance, in the query "Find documents about China's role in international finance," a substantial portion of the dataset referenced China in varying contexts. This led to low precision for both full-text search and static topic models, which pulled in documents merely mentioning China. In contrast, I-NTM allowed users to iteratively refine the topic by labeling it with "yuan" and then "investment," dynamically reshaping the topic embedding. This adjustment centered the topic around China's monetary and investment-related terms, surfacing more targeted documents and improving both relevance and retrieval efficiency.

Labeling also surfaced new relevant documents that would not have been retrieved under fixed topic assignments (Figure 3.4). This allowed users to adapt the results to better reflect their information needs. While some labeled topics drew in documents with broader relevance, the overall effect was to increase recall and reduce task completion time.

These results demonstrate the practical utility of I-NTM in exploratory retrieval tasks. Traditional topic modeling metrics like coherence and diversity measure internal structure but fail to capture downstream usefulness. In contrast, our user study shows that interactivity improves both retrieval quality and efficiency, validating I-NTM as a meaningful advance in human-centered topic modeling.

3.7 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.

Interactive labeling of topics has been thoroughly explored for probabilistic topic models. **(author?)** [66] compare labels generated by users after seeing topic visualizations with automatically generated labels. **(author?)** [20] provides a method for iteratively updating topics by enforcing constraints. **(author?)** [67] 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 work in the area of anchor-based topic modeling, where a single word is used to identify a topic. **(author?)** [71] present “Tandem Anchors” where multi-word anchors are used to interactively guide topics. **(author?)** [72] developed a framework for interactively establishing anchors and alignment across languages. **(author?)** [73] introduces a protocol that allows users to interact with anchor words to build interpretable topic. The most similar and recent work to ours is [69] 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 model. We developed the first interactive interface for a suite of neural topic models and have an interface that users can see their changes in real time. Contemporaneously, **(author?)** [100, 101] developed models to study unstructed data using prompting of large language models, however, this interface is not interactive.

From Scatter/Gather [102] to CluWords [103], interactive clustering methods have helped users organize large datasets. While these approaches offer interpretability without relying on generative assumptions, they impose rigid document assignments, forcing each document into a single cluster. In contrast, topic models allow documents to be mixtures of topics, providing greater flexibility and capturing overlapping themes more effectively.

I-NTM builds on this tradition by offering clustering-like interactivity with topic model flexibility. For example, in our user study, labeling a topic with ”investment” gradually pulled it toward finance-related terms, reshaping both the topic distribution and its associated documents. These changes would be impossible in a static clustering pipeline, highlighting the value of combining structured representations with interactive refinement.

Several works have explored automatic topic labeling through different methods. **(author?)** [104] propose a two-stage approach that re-ranks candidate labels from a large pool of words. **(author?)** [105] rank and assign labels by leveraging top terms from document titles and subwords from Wikipedia articles, using lexical features for selection. **(author?)** [106] utilize hierarchical topic models, assigning labels based on parent-sibling relationships within the topic hierarchy. More recently, **(author?)** [107] use LLMs to cluster data, either with or without predefined labels. However, they lack the refinement that I-NTM provides. We treat labeling not just as a descriptive act but as a mechanism for live topic adjustment. This is key to our approach: labels in I-NTM actively reshape the topic embedding (Eq.3.1), improving retrieval and representation quality in downstream tasks (Section **??**). Rather than replacing human-in-the-loop labeling, we use it as a guide for precise, interpretable adjustments.

3.8 Conclusion

In this chapter we present I-NTM, a framework for interactively updating topics in neural topic models through user-provided labels. While prior work has enabled interactivity in probabilistic models, I-NTM is the first to support real-time topic refinement across a suite of neural models. This enables users without retraining to shape topics toward their specific goals.

Our experiments demonstrate that interactivity improves topic diversity, cluster purity, and NMI, and outperforms both static topic modeling and full-text search in downstream document retrieval tasks. By allowing users to iteratively adjust topic embeddings, I-NTM surfaces more relevant documents more efficiently.

Useful directions to explore include guiding the topic model training through interactive labeling rather than after training, supporting more complex labeling, and enable direct embedding space adjustments via visualizations.