Improving the TENOR of Labeling: Re-evaluating Topic Models for Content Analysis

Zongxia Li[1] **Andrew Mao**[1] **Daniel Stephens**[3] **Pranav Goel**[1]
Emily Walpole[2] **Alden Dima**[2] **Juan Fung**[2] **Jordan Boyd-Graber**[1]

[1]University of Maryland, {zli12321, amao, pgoel}@cs.umd.edu, jbg@umiacs.umd.edu
[2]NIST, {emily.walpole, alden.dima, juan.fung}@nist.gov
[3]Morgan State University, dstephens@morgan.edu

Abstract

Topic models are a popular tool for understanding text collections, but their evaluation has been a point of contention. Automated evaluation metrics such as coherence are often used, however, their validity has been questioned for neural topic models (NTMs) and can overlook a model's benefits in real-world applications. To this end, we conduct the first evaluation of neural, supervised and classical topic models in an interactive task-based setting. We combine topic models with a classifier and test their ability to help humans conduct content analysis and document annotation. From simulated, real user and expert pilot studies, the Contextual Neural Topic Model does the best on cluster evaluation metrics and human evaluations; however, LDA is competitive with two other NTMs under our simulated experiment and user study results, contrary to what coherence scores suggest. We show that current automated metrics do not provide a complete picture of topic modeling capabilities, but the right choice of NTMs can be better than classical models on practical tasks.

1 Introduction

Establishing a label set to organize a collection of documents is a fundamental task in many fields such as social science, and, linguistics, education. For example, in the social sciences, grounded theory emphasizes *structural coding* as a framework for discovering similarities and differences in large-scale experimental data and assigning meaning to it (Glaser and Strauss, 2017; Lindstedt, 2019; Krommyda et al., 2021). Such a process is difficult and time-consuming, partly because it requires a global understanding of the entire dataset, and local knowledge to accurately label individual documents. We emphasize that this strictly more general than document classification: classification presumes *a priori* a label set; while we will use classifiers in our method, we first need a user's help to determine the label set and the training data.

Topic modeling (Boyd-Graber et al., 2017) has emerged as a popular tool to help with the *coding* process to discover the label set (Section 2.4). These models treat documents as admixtures of latent topics, each represented by a distribution over words. The most popular topic model, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has over 40,000 citations with numerous extensions and variants (Churchill and Singh, 2022).

Previously, Active Learning with Topic Overviews (Poursabzi-Sangdeh et al., 2016, ALTO) demonstrated that combining LDA with an active learning classifier could help people create label sets more efficiently. After topic models provide a global overview of the data, exposing the broad themes of the corpus, active learning selects documents that direct the annotator's attention to topically distinct examples to label. Together, these two ingredients train a classifier to automatically label the documents more efficiently.

However, a gap remains in the literature, given recent advancements in topic modeling. Neural topic models (NTM), which use continuous text embeddings to capture contextual and semantic relationships in high-dimensional data, have gained prominence, besting classical probabilistic topic models on automatic evaluation metrics such as coherence (Aletras and Stevenson, 2013). However, automated evaluation metrics have been called into question; Hoyle et al. (2021b) show they do not necessarily correlate with human ratings on topic model outputs and call for task-centered evaluations, such as helping users analyze content.

We aim to fill this gap, and evaluate the effectiveness of neural, supervised, and classical topic models to help social scientists with content analysis and label set creation. We do this by taking the starting point of ALTO—classicial topic models applied to this problem—and probe "deeper" to create Topic-Enabled Neural Organization and Recommendations (TENOR), an interactive tool that

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 840–859

supports various topic models with active learning to speed up the process of content analysis.[1] We conduct synthetic experiment on LDA, supervise LDA and three NTMs with followup user study and expert user study and show that the choice of Contextualized Topic Model (Bianchi et al., 2021) (CTM) helps users create higher quality label sets than using classical LDA, as measured by both cluster metrics (Section 4.3) and user ratings. However, LDA is still competitive or better when compared with two other popular NTMs. Thoughtfully using topic models as part of a larger system with human interactions gives a more comprehensive evaluation and understanding of their real-world usage (Section 4.5).

2 Background

Manually sorting thousands of documents to establish a label set to create is mentally challenging and time-consuming. Baumer et al. (2017) compare grounded theory with topic modeling: although the two methods are from distinct fields, they produce similar insights on large-scale data. Topic models cluster documents and extract meaningful themes and can help users induce labels.

For content analysis, machine learning and NLP focus on developing NTMs (Hoyle et al., 2021b), because they win nearly every automatic coherence metric. However, most of the computational social science community remains focused on older probabilistic models (Abdelrazek et al., 2023). Thus, we explore this open question: should we use classical or neural topic models for label induction and content analysis?

One of the reasons that NTMs might be better is that ALTO showed the benefits of active learning (Settles, 2012): start with a dataset with an undefined label set; users add labels to the set by going through individual documents (guided by topic overviews); once the users establish at least two distinct labels for the label set, a classifier trained on the labeled documents can point users to documents that are either challenging for the current label set or that might require new labels. One of the criticisms of NTMs is that they are too granular and specific (Hoyle et al., 2021b), but this may be a boon for label induction: it can find candidates for a new label.

In addition to ignoring neural models (which had not reached maturity when ALTO was proposed), another lacuna of (Poursabzi-Sangdeh et al., 2016) is that it ignores supervised topic models that can combine classification with topics. Supervised topic models (Mcauliffe and Blei, 2007) *change* as labels are added and can adapt—for instance—when a user associates two labels with a topic. Thus we evaluate neural and classical topic models that tasks humans with creating a label set and annotating a document collection, with the assistance of topic models and a text classifier on the a dataset of US congressional bills (Adler and Wilkerson, 2008).

We delve into specific topic models, active learning, and evaluation metrics for the rest of this section.

2.1 Topic Models

Topic models identify latent themes within a corpus, providing a snapshot of its overall narrative. Given a set of documents and a specified topic count K, these models divide documents into K clusters. Each cluster represents a topic defined by key terms, denoting its core theme (examples in Appendix 3). Users can explore the corpus's main themes and label individual documents with the topics and keywords.

Supervised Latent Dirichlet Allocation. sLDA retains the generative process of LDA but also adds a step to generate labels for each document *given* its empirical distribution over topic assignments in a document. For example, for movie comment reviews, LDA generates general topics people discuss movies that are unlikely to correlated with users' star ratings. In contrast, sLDA can: an LDA topic about romance films would split into "good" and "bad" versions with sLDA. We use the classifier's predictions as surrogate response variables, and update sLDA constantly as users label more documents. We expect sLDA's topics to better reflect user inputs by interacting with the classifier trained with user input labels.[2]

Neural Topic Models Current popular neural topic models include Contexualized topic models (Bianchi et al., 2021, CTM), BERTopic (Grootendorst, 2022), and Embedded topic model (Dieng et al., 2020, ETM). Theses neural models take ad-

[1] https://github.com/zli12321/TENOR.git

[2] Suppose a user creates 15 unique labels for 80 documents, we train the classifier on the 80 documents with the user input labels. Then we use the classifier to make predictions for all the documents and use the predictions as response variables for sLDA

vantage of pre-trained word embeddings with rich contextual information to enhance the quality of discovered topics. CTM builds on pre-trained language models like SBERT (Reimers and Gurevych, 2019) to generate sentence embeddings concatenated with Bag-of-Word (BoW) representations and runs a variational autoencoder (VAE) on the representation, while BERTopic uses UMAP (McInnes et al., 2020) and HDBSCAN (McInnes et al., 2017) create and refine topics from encoded word embeddings. ETM retains the same generative process as LDA but the topics are learned from word embeddings that contain rich semantic meanings instead of pure word distributions.

2.2 Active Learning

Active learning (Settles, 2012) guides users' attention to examples that would be the most beneficial to label for a classifier, using techniques such as uncertainty sampling. By directing users to annotate uncertain documents first, active learning is valuable in situations constrained by time or budget.

2.3 Preference Functions

During the initial stages of training, a classifier must generalize to unseen data quickly. A rapid improvement facilitates high-quality data analysis and optimizes time and costs, especially for large datasets (Muthukrishna et al., 2019). Mathematically, "preference functions" are the tool that allows this early generalization in active learning generally and in TENOR specifically to get a good set of labels with representative documents as quickly as possible.

A preference function uses uncertainty and diversity sampling to pick the most beneficial document and guide users' local attention to that document to label. According to the preference function, the classifier favors documents with the highest confusion scores that are most likely to be in the boundaries between multiple labels, which are documents that users are most likely to make new labels–uncertainty and diversity. For our baseline classifier, when it does not incorporate topic models, let L be the label set probability distribution for document d, the preference function for d is :

$$\mathbb{H}_d(L) = -\sum_{i=1}^{n} P(l_i) \log P(l_i). \tag{1}$$

Here, $\mathbb{H}_d$ represents the cross-entropy (Shannon, 1948) of the classifier. The "most beneficial" document is the one whose label distribution (as defined by a classifier) is most confused: more mathematically, has the highest entropy. If the user can resolve that confusion by providing a new or existing label (or remove the document from the set), it will most benefit the next iteration of the classifier.

We follow the insight of ALTO and interleave topic models and active learning to make the preference function topic-dependent. This is important for real-world scenarios where context-switching can impede human labeling throughput (Raeburn, 2022). First, the most confusing topic by the classifier is selected, and then, within this topic, the document with the highest preference function score (the most confusing document) is chosen.

Given K topics from topic models, each document is characterized by a topic distribution vector $\theta^d \equiv \{\theta_1^d, \theta_2^d, \ldots, \theta_K^d\}$. For a particular document, its predominant topic is:

$$\theta_{\max}^d = \max_{i=1}^{K} \theta_i^d. \tag{2}$$

We also adopt hierarchical sampling for active learning (Dasgupta and Hsu, 2008) and incorporates vector representation of topic models and users' label inputs to match their individual preferences (Zhang et al., 2019)

$$\mathbb{H}_d^t(L) = \mathbb{H}_d(L) \cdot \theta_{\max}^d. \tag{3}$$

With a clearly defined preference function, we choose a topic k^* first based on the following criterion: Given K topics, let $\mathcal{D}_k$ denote the set of all documents that are most prominently associated with topic k. The classifier selects a topic k^* such that its documents' median preference score, $\mathbb{H}_d^t$, is maximized. Formally, this is

$$k^* = \underset{k \in \{1,2,\ldots,K\}}{\arg\max} \text{ median } \left\{ \mathbb{H}_d^t(L) : d \in \mathcal{D}_k \right\}. \tag{4}$$

2.4 Evaluation Metrics

Our objective is for users to establish new label sets for a common dataset. This is a hard problem: indeed, Kleinberg (2002) proves that it is impossible to satisfy multiple reasonable clustering properties simultaneously. We thus, like ALTO we use tree of reasonable metrics—described below—to compare how far user-induced labels deviate from a gold label set (in this case, the consensus labels of political scientists on the congressional bills dataset).

In addition to these standard cluster evaluation metrics, we also measure the coherence for each topic of LDA, sLDA, NTMs (more detail in Appendix A).

Purity Purity evaluates how *pure* an induced cluster is: in other words, what proportion of documents in a cluster are not commingled with documents with a different gold label (Zhao, 2005). As we will see with many of these metrics, there is a clear failure mode: the purity metric can be easily manipulated by assigning a unique label to each document. We mitigate this risk by not disclosing these metrics to labelers and limiting the time users have to create labels.

Adjusted Normalized Mutual Information (ANMI) Normalized Mutual Information (Strehl and Ghosh, 2003, NMI) assesses clustering quality by measuring the interdependence between true and predicted labels. One can gain insights of the true labels by understanding the predicted labels. The ANMI (Amelio and Pizzuti, 2016), an enhancement of NMI, corrects for the chance alignment of clusters.

Adjusted Rand Index (ARI) The Rand Index (RI)(Rand, 1971, RI) measures for any pair of documents the probability that their gold labels and their assigned labels match. The Adjusted Rand Index (ARI) (Sundqvist et al., 2022, ARI) refines this measure by adjusting for chance, which can yield negative values if the new labeling actively contradicts the gold labeling.

Coherence Normalized pointwise mutual information (NPMI) measures how semantically similar the top words of a topic are, which was proposed for classical topic models, but can also be used for NTMs (Aletras and Stevenson, 2013).[3] (Chang et al., 2009) uses large-scale of user study to show coherence creates a computational proxy that simulates human judgments for classical topic models. We use NPMI to evaluate the quality of topics generated by classical and neural topic models.

The clustering metrics evaluate the alignment, quality, and information overlap between two clusters. A higher value in these metrics indicates greater similarity and alignment between the induced labels and the gold labels. However, using just one of them to measure user label quality has limitations. If users assign every document a different label, they will reach a perfect purity score, but that violates the task. ARI does not measure the quality of individual clusters. For example, two clusters might have high ARI, but both are very poor quality. ANMI is sensitive to the number of clusters, where a significant difference in the number of clusters between the standard cluster and classifier predictions can lead to a reasonable ANMI score, but the clusters have a high mismatch. By using all of them to complement each other, we are more confident in comparing the quality of classifier predictions.

[3] NPMI and ANMI are over different evaluation metrics over different probability spaces.

3 Study Setup

3.1 Groups

For the simulated user study, we use the following models in combination with active learning:

1. (NONE);
2. Latent Dirichlet Allocation–(LDA);
3. Supervised LDA–(sLDA);
4. Bertopic–(BERTopic);
5. Embedded Topic Model–(ETM);
6. Contextualized Topic Model–(CTM).

Our baseline (1) NONE gives users access to a classifier with active learning, but no topic model organization to help them first establish a "big picture". The rest of the groups provides the users topic overview and a classifier that has access to topic model probability vectors and active learning. More implementation details of our study groups are in Appendix B.

3.2 Dataset

Our simulated experiment uses the 20newsgroups (Mitchell, 1999) and the Congressional bills dataset. Both datasets have hierarchical labels; the first level is a general category, such as *Health* or *Education* for the Bills; and recreation (*rec*) or science (*sci*) for 20newsgroups. Under each of the first layer labels, there are more specific labels; for example, under *Health*, there are *Health Insurance*, *Mental Health and Cognitive Capacities*, *Children and Prenatal Care*, etc.

Since we want to test our system theoretically and in a user study setting, having datasets with hierarchical labels enables us to use more specific labels as user input labels and more general labels as standard labels in simulated experiments. In real-world settings, users are more likely to make

more specific labels that are more closely related to the contents of individual documents.

3.3 Simulated Experiment

Before conducting a real-world user study, we run simulated experiments on both datasets. We choose $K = 35$ topics for all five topic models.[4] Since users are more likely to create more detailed labels for each document, we use sub-labels as pseudo-user labels, while using the more general labels as our gold standard. We use logistic regression as our classifier and unigram tf-idf as input features for the classifier.[5] We also concatenate topic probability distributions for all the documents with tf-idf features, which encodes topic information to the classifier for settings with topic models. We use incremental leaning (Rosenblatt, 1958) to fit and update the classifier after applying a synthetic label to each document.[6] The clustering quality is assessed by the classifier's predictions with the more general labels using the three evaluation metrics. We run the experiment for 400 documents since we expect it to be the maximum for a participant to label within an hour.

Coherence and simulated experiment results do not have a direct relationship CTM does the best on all cluster metrics on both datasets (Figure 1), while LDA and sLDA remain competitive with other NTMs. Topic models with higher NPMI in Table 1 do not necessarily have better simulated experiment results shown in Figure 1. ETM does the worst among all the groups—despite having high coherence—and CTM does the best, where LDA and sLDA are even better than BERTopic and ETM on the 20newsgroup dataset.

While our synthetic data can serve as partial proxy, relying solely on automated evaluation metrics does not capture how much the users find the topic model helpful in helping them conduct content analysis. Thus, our next section investigates this question and surveys users' ratings on how they find topic models useful.

[4]We choose $K = 35$ because it optimizes average coherence for all topic models (details and hyperparameter selections are in Appendix C).

[5]Using sentence transformer features produces similar results but takes much longer to update.

[6]With two exceptions…we reinitialize the classifier: if a new label class is introduced to the classifier; if SLDA is updated with surrogate response variables, we rebuild the features by concatenating tf-idf features with new topic information.

Dataset	LDA	sLDA	CTM	ETM	BERTopic
Bills	0.07	0.09	0.09	0.13	0.15
Newsgroup	0.06	0.05	-0.09	0.09	0.10

Table 1: On average, NTMs have higher NPMI coherence than LDA, where BERTopic has the highest coherence, followed by ETM. However, the NTMs with higher coherence are not better than CTM and LDA under a task-based experiment (Figure 1).

4 User Study

We conduct a general user study and expert study to compare topic models in the rest of the paper. For the general user study, we compare settings (1) NONE, (2) LDA, (3) sLDA, and (6) CTM since CTM is the best-performing neural model in the simulated experiment. We use the Bills dataset to conduct a 60-user study with our interface, with 15 people in each group. Our Bills dataset's topics are accessible to lay annotators and allows us to quantitatively understand users' acclimation to the dataset as they explore the corpus. We then run a smaller, more qualitative followup expert study on an expert dataset, where the experts are familiar with the topics in the dataset with the best two models from our user study results. The goal of the expert study is to ensure that our user study results can generalize to experts with deeper knowledge of US federal policy.

4.1 User Study Interface

For the groups using topic models, users are shown documents grouped by their top topic, with topic keywords. The document selected by the active learning preference function is highlighted and displayed both at the top of its topic and at the top of the interface. When users click a document, they are presented with its full text, label options, top five topics, and top ten keywords per topic. Words above a 0.05 threshold in the primary topic are highlighted. In NONE group, users see unsorted documents with the recommended one at the top. Clicking a document shows its contents, without topic keywords or highlights (detailed interface in Appendix F).

4.2 Participant Recruitment

We sourced participants via Prolific, restricting our selection to individuals from the US with an approval rate exceeding 95% with at least ten previous participations on Prolific. Participants were randomly assigned to one of four groups, each ac-

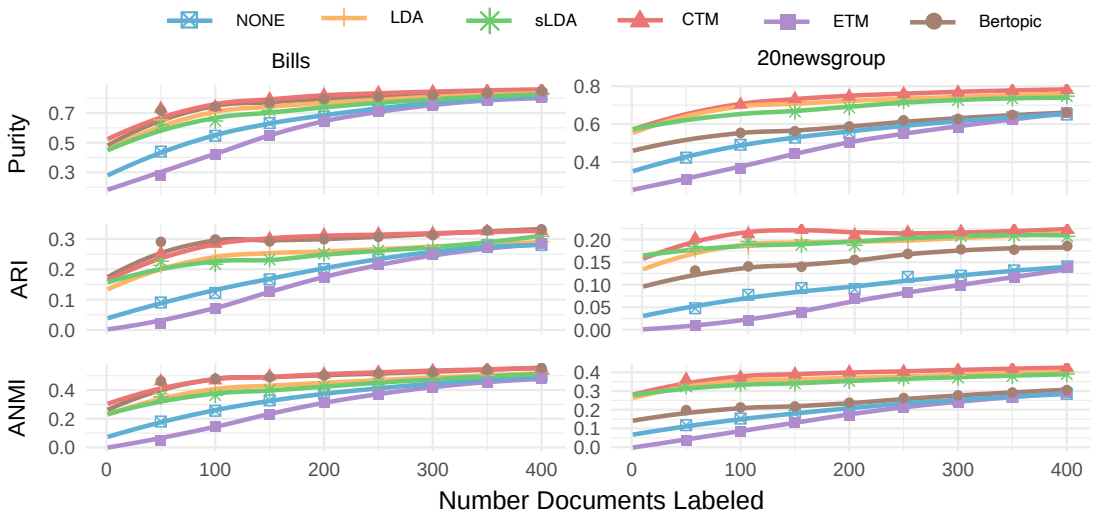

Figure 1: Cluster scores of simulated labeling experiments, median of 15 runs. CTM with active learning has the highest score across all metrics and datasets. LDA and sLDA are better than or competitive with the other NTMs (ETM, BERTopic). Given these results on synthetic data, we use CTM for the human experiments.

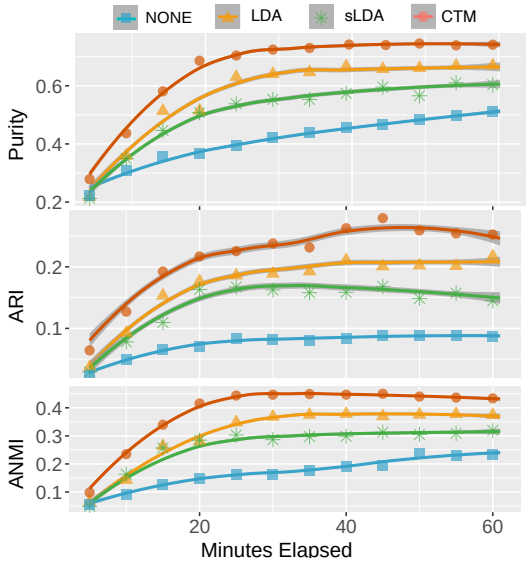

Figure 2: User study label cluster metrics plotted against time. For each group, we take the median of each metric for every minute passed. The user study results are similar to the simulated experiment; CTM does the best on all three clustering metrics.

commodating a maximum of fifteen participants.[7] Participants first reviewed the task instructions and completed a brief tutorial to familiarize themselves with the process. Participants complete a follow-up survey to receive a 20-dollar compensation after the one-hour session.

[7]We use the same trained models from the simulated experiment. We update sLDA in the backend once the previous training is complete.

4.3 Cluster Quality Evaluation Metrics

We record the purity, ARI, and ANMI for every minute passed during each session. For each group, we plot the median of each metric for every minute passed (Figure 2).

Topic model groups do better than NONE Throughout the 60-minute study session, the classifier has a wide gap between groups with topic models and NONE. Topic model groups have faster early gains on all three metrics than NONE, confirming the results from Poursabzi-Sangdeh et al. (2016).

CTM does the best on cluster metrics, followed by LDA, sLDA, and NONE. In real-world user applications, CTM is the best for classification. The classifier with neural topic features, trained on user input labels, can generalize unseen data better than classical generative topic probability features. Although CTM is the best, having the classifier have access to topic model features is better for the classifier to generalize and predict unseen data than not. We later manually evaluate the validity of the user labels by random sampling (Appendix E), where 98.38% of the selected examples are qualified under evaluations of two authors.

sLDA falters on compared to LDA and CTM This is partly attributed to inaccuracies in the classifier's predictions. For instance, when a user labels 30 documents midway through the session, the classifier, in turn, predicts labels for the entire dataset.

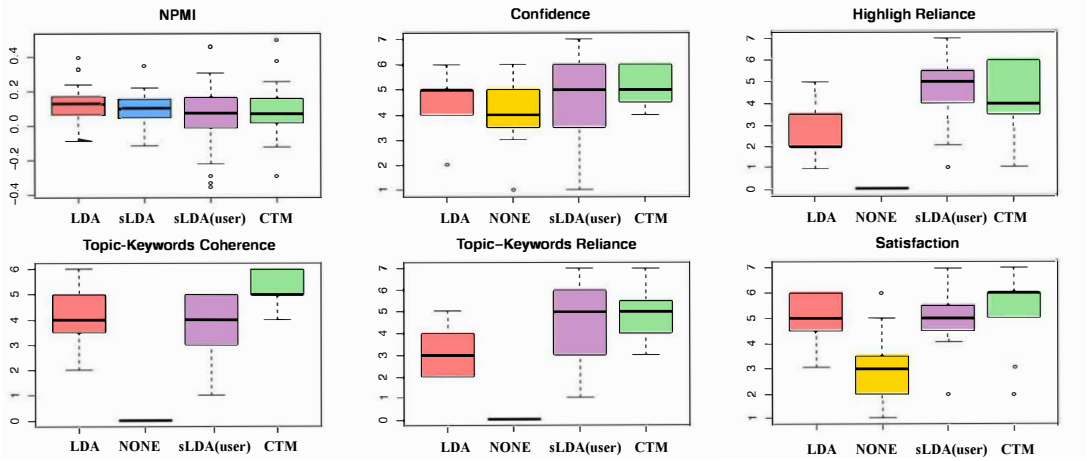

Figure 3: The first Plot shows NPMI Coherence for all topics on the Bills dataset, where sLDA(user) is trained on user input labels, and sLDA is the initial model used for all sLDA users. The rest of the plots shows users' rating on different questions on a scale 1 to 7, which the higher is better. Although sLDA is worse than LDA and CTM on clustering evaluations, most of the median of user ratings do not differ from CTM, and surpass LDA in some ratings. For ratings 2 to 4, NONE groups users all rate 0 because they do not have access to those features

However, if the user only creates two label categories for the 30 documents, the lack of diversity of response variables can generate document topic probability as features that confuses the classifier. Nonetheless, sLDA can align certain topics with user intent labels, which means that sLDA might be capable of generating topic keywords that are semantically similar to user labels, thus improving users' overall experience. Subsequent survey analyses will investigate whether sLDA supports this hypothesis in user survey ratings.

Examining coherence, quality of document clusters, and quality of topic keywords We go through the topics with top two, middle two, and bottom two coherence scores for the models we use for user study (including sLDA trained on user labels), and show the NPMI, topic keywords, and a randomly selected passage from the topic in Tables 3 and 5.[8] Although the coherence scores vary for different topics, the top keywords are representative of the documents, but a low median coherence score does not necessarily show lower median user ratings (Figure 4). CTM has the highest top coherence scores but the median coherence score is lower than sLDA and LDA. However, CTM is still better on clustering evaluations and user ratings.

[8]We load the saved sLDA model trained on user labels predicted by the classifier at the end of the session, we call it sLDA(user).

4.4 User Ratings

Our survey comprises five questions aimed at gauging user judgment and evaluating topic models, using a scale ranging from 1 to 7.[9]

CTM and sLDA users rely more on topic models than LDA Figure 4, the second to sixth plot show a summary of users' ratings for question 1 to 5. The median of user ratings on CTM and sLDA are similar for most of the questions except for **Topic-Keyword Coherence**, which sLDA falls short. Based on the median user ratings, users generally rely more on topic keywords and highlights to create labels for documents if they are assigned to the CTM or sLDA group. Users also rate the topic keywords they use to label documents as more coherent for CTM and sLDA. Although the classifier in sLDA falls short on the three cluster metrics among the three topic models, users generally have better overall experience with sLDA than LDA users.

Automatic coherence likes LDA topics, users do not Although the top topics for CTM, sLDA and sLDA(user) have higher coherence scores than LDA (Figure 4), LDA's coherence scores are quite tight in

[9]**Confidence** asks how confident the users feel about their created labels. **Highlight Reliance** asks how much the users rely on the *highlight* functionality to make labels. **Topic-Keywords Coherence** asks whether users find the topic keywords coherent while they explore topics and peruse keywords to assist them in label creation. **Topic-keyword Dependence** investigates the frequency at which users consult the most related topic keywords while creating labels for documents. **Satisfaction** assesses the users' overall satisfaction with the tool, exploring whether users find the tool likable and helpful.

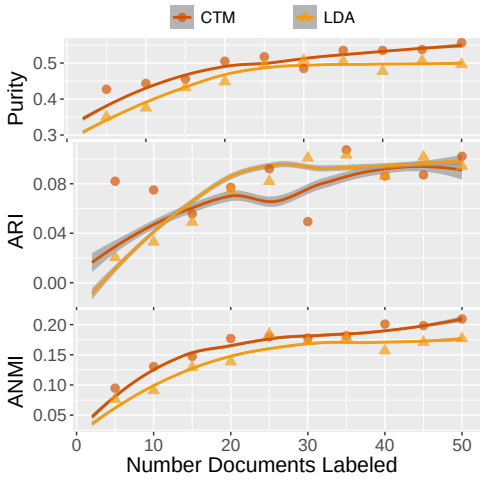

Figure 4: We run a followup pilot study with six social science experts (three in each group) on their internal social science dataset (800 documents). They are familiar with the topics in the dataset. Up to the 50th document labeled, CTM still generalizes well for expert datasets and expert users.

Metric	p-Value	Significant Pair
Confidence	0.327	None
HighlightReliance	0.035	sLDA vs. LDA
topicCoherence	0.017	CTM vs. sLDA
topicReliance	0.034	CTM vs. LDA
satisfaction	0.002	NONE vs. Other 3

Table 2: Significance test results across subjective ratings for three groups at a 0.05 significance level. There is no significant difference in user ratings between NTM and LDA except for *Topic-Keyword Reliance*. For rows 2-4, we exclude NONE to do testing. The third column shows the group pairs that are statistically significant. For example, the significant pair for *satisfaction* is between NONE and other three groups with topic models, and it indicates a difference of *user satisfaction* rating between NONE and other three groups under a 95% confidence level, where the NONE users are less satisfied with their experience from the sixth plot in Figure 4.

the boxplot and LDA has higher median coherence than the other two models. sLDA(user) has diverse coherence scores for its topics. However, when looking at the median user rating of all the five questions, LDA does not surpass CTM and sLDA: there is not a strong and direct relationship between coherence and human usability. This is a task-specific confirmation of Hoyle et al. (2021a).

Different topic models, different purposes We run ANOVA (Fisher, 1935) and posthoc Turkey-Kramer for pairwise comparison between ratings of any two of the user groups. Users are less likely to rely on the topic keywords generated by LDA to label documents, compared to CTM and sLDA based on significance results (Table 2) because LDA generates overly general topic keywords that are less useful to label individual documents. For specific tasks, such as label set establishment and tasks involving understanding individual documents, CTM is a better choice.

4.5 Expert Verification

The expert conditions were LDA and CTM, the two winning conditions in our general user study. Six experts all hold at least a graduate degree in community resilience related field that focuses on assisting communities and stakeholders on issues related to anticipated hazards conditions and disaster preparedness field.[10] We use the same user interface described in Section 4.1 with the given expert dataset on 800 documents. The documents are collected from local governments across the United States providing structured ways to set community-scale goals and developing plans for recovery of community functions after natural or human-caused hazards (U.S. Department of Commerce, 2020). Experts conduct analysis and assign labels to this dataset so they can understand different categories of hazards and develop plans for a community to prepare for anticipated hazards, adapt to changing conditions, and withstand and recover rapidly from disruptions. The dataset has been previously labeled by multiple experts using Cohen's Kappa agreement (McHugh, 2012) over a six-month period. CTM surpasses LDA on two out of three clustering metrics and has similar ARI at the 50th document (Figure 4).

Experts rely less on keywords but still like them Since all the experts are quite familiar with the topics in the dataset, one expert using LDA mentions that the topic keywords are not helpful but the highlighted texts are more helpful for individual document annotation. LDA produces topics that are too general, so experts already prefer the more specific keywords from CTM.

5 Related Work

Applications of topic models are important, as exemplified by previous work by Fang et al (Fang et al., 2023), which addresses the human-centric applications of topic models. Bakharia et al. (2016)

[10] https://www.nist.gov/community-resilience

shows that interactive topic models have gained traction among social science researchers and data analysts. Nevertheless, classical topic models dominate most applications in social science research (Boyd-Graber et al., 2017; Lin, 2009). Despite their theoretical advantages, this persistent preference for classical models underscores the need for comprehensive studies on the practical utility of NTMs.

As one of the most popular topic models, LDA has been widely applied and tested in diverse fields from health (Paul and Dredze, 2011) to political opinion analysis (Chen et al., 2010), social media data analysis (Zhao et al., 2011), etc. Thus, LDA has already proved itself as a useful tool for real applications.

For supervised models, most work focuses on sLDA's power to predict response variables from text (Xu and Eguchi, 2022). Few works have study whether the induced topics align with user intents such as labeling. Using sLDA interactively for document recommendation and annotation is more intuitive and straightforward than using unsupervised classical LDA.

Beyond connecting a single response variable to topic assignments, neural models offer even more flexibility and have over a hundred variants, but the evaluation of NTMs is mainly based on topic coherence, topic diversity, and classification applications (Zhao et al., 2021). The major framework of NTMs are mostly sequential NTMs, which leverages the architecture power of Recurrent Neural Network (RNN); NTMs with pre-trained language models, such as BERT, that already learns the semantic relationship and association of words from a large corpus of texts. NTMs have the advantage of producing higher automatic evaluation scores, and classification abilities, along with other more extensive applications that classical topic models cannot do, which includes texts generation (Tang et al., 2019; Wang et al., 2019), summarization (Zhao et al., 2020; Wang et al., 2020).

However, with the new popularity of NTMs, to the best of our knowledge, there are still few works using NTMs for social science due to their complex architecture and more computing resource demands. Our work examines this gap to study the trade-off between using neural, supervised, or classical topic models. While some recent studies have compares NTM and LDA with human analysis of the topic outputs, they still predominantly rely on automatic evaluation metrics, with limited emphasis on analyzing the quality of models from a human perspective or task-based utility of topic models (Doan and Hoang, 2021). Papadia et al. (2023) concludes that LDA is better than NTM in metrics on coherence(Röder et al., 2015) and classification (Phan et al., 2008). However, this conclusion is for non-English datasets. Our research intends to bridge this gap by conducting an English-language topic model quality evaluation, incorporating human interaction to help content analysis.

Our approach differs from previous studies, which compares NTMs and classical models' stability and alignment with stationary, pre-determined ground truth labels (Hoyle et al., 2021a). In the former, LDA was better; in the latter, LDA was better than many NTMs (Hoyle et al., 2021b). However, Hoyle et al. (2021a)'s approaches only evaluate topic models by analyzing human ratings on topic keywords with labels without any task applications. In contrast, for the tasks of content analysis and building a label set, the overly specific NTM keywords are actually helpful for people to come up with labels more easily than more general and dispersed keywords. While the overall topics may not look as "pretty" to a user, they are useful.

6 Conclusion

We provide an interactive task-based evaluation of neural, supervised, and classical topic models, using the task of content analysis and label set creation. Using CTM with an active learning classifier helps both expert and non-expert annotators produce higher quality label sets more quickly, according to cluster metrics and human ratings, validating that the right choice of NTMs can be better than LDA for content analysis. However, LDA is still competitive with two other NTMs, contrary to what coherence scores would suggest. We show that current automated metrics do not provide a complete picture of topic modeling capabilities, but the right choice of NTMs can still be better than classical models on practical tasks. With the popularity of large language models (LLMs), future work can include exploring more effective ways to use TENOR combined with LLMs for content analysis, where experts have a set of pre-defined research question and hypothesis, and use TENOR to actively select documents to prompt an LLM to build up a label set for the dataset quickly to answer their research questions and verify their research hypothesis.

7 Limitations

We provide a human-in-the-loop framework to evaluate topic models, extending beyond automated evaluation metrics. Yet, our experiment only focuses on a very narrow and specific task to evaluate topic models. In addition, although our work shows that the right choice of NTM can bee more powerful than LDA for specific tasks, the debate about evaluation of topic models is still present. From a language perspective, our experiments are based on English dataset only. Our conclusions theoretically can be generalized to some other languages but need to be practically tested. It might come to a different conclusion for languages with completely different structures than English. Furthermore, with the rise of LLMs that can complete various tasks close to human level, the use of LLM to help with the process of label set generation, classification (Zhou et al., 2024), and content analysis is a more efficient and cost-effective approach that can simulates our human study compared with our human study in Section 4.5. A comparative analysis of the quality of labels created by actual human users and LLM would be valuable for the social science and NLP community to confirm the validity of using LLMs to simulate actual user studies to speed up their research process. We will conduct further comparative analysis between human created labels and LLM created labels in our future work.

8 Ethics

We received approval from the Institutional Review Board before initiating the user study. All participants are based in the United States. Users are required to review an instruction and consent statement before participation commitment. They have the option to withdraw if they disagree with the terms. Throughout the study, no personal information that could reveal identities is collected. To the best of our knowledge, our study presents no known risks.

9 Acknowledgement

We thank anonymous reviewers and Alexander Hoyle and Kyle Seelman for their insightful comments for helping us make our paper experiments and arguments more solid. We thank Emily Walpole and Juan Fung's community resilience groups for taking their time participating our expert verification experiment and providing valuable qualitative comments and feedback. Zongxia Li, Andrew Mao, Daniel Stephens, and Pranav Goel's contributions are supported by the NIST Professional Research Experience Program.

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

E. Scott Adler and John Wilkerson. 2008. Congressional Bills Project. http://www.congressionalbills.org. Accessed: insert access date here.

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.

Alessia Amelio and Clara Pizzuti. 2016. Correction for closeness: Adjusting normalized mutual information measure for clustering comparison: Correction for closeness: Adjusting nmi. *Computational Intelligence*, 33.

Aneesha Bakharia, Peter Bruza, Jim Watters, Bhuva Narayan, and Laurianne Sitbon. 2016. Interactive topic modeling for aiding qualitative content analysis. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, page 213–222, New York, NY, USA. Association for Computing Machinery.

Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *J. Assoc. Inf. Sci. Technol.*, 68(6):1397–1410.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jordan Boyd-Graber, Yuening Hu, and David Minmo. 2017. *Applications of Topic Models*. Now Foundations and Trends.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea

leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Bi Chen, Leilei Zhu, Daniel Kifer, and Dongwon Lee. 2010. What is an opinion about? exploring political standpoints using opinion scoring model. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, page 1007–1012. AAAI Press.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s).

Sanjoy Dasgupta and Daniel Hsu. 2008. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 208–215, New York, NY, USA. Association for Computing Machinery.

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.

Thanh-Nam Doan and Tuan-Anh Hoang. 2021. Benchmarking neural topic models: An empirical study. In *Findings*.

Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and Rob Procter. 2023. A user-centered, interactive, human-in-the-loop topic modelling system.

Ronald A. Fisher. 1935. *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021a. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033.

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

Jon Kleinberg. 2002. An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Maria Krommyda, Anastasios Rigos, Kostas Bouklas, and Angelos Amditis. 2021. An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media. *Informatics*, 8(1).

Minchul Lee. 2022. bab2min/tomotopy: 0.12.3.

Jimmy Lin. 2009. Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, 10:46.

Nathan C. Lindstedt. 2019. Structural topic modeling for social scientists: A brief case study with social movement studies literature, 2005–2017. *Social Currents*, 6(4):307–318.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Leland McInnes, John Healy, and S. Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2:205.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Tom Mitchell. 1999. Twenty Newsgroups. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5C323.

Daniel Muthukrishna, Gautham Narayan, Kaisey S. Mandel, Rahul Biswas, and Renée Hložek. 2019. Rapid: Early classification of explosive transients using deep learning. *Publications of the Astronomical Society of the Pacific*, 131(1005):118002.

Gabriele Papadia, Massimo Pacella, Massimiliano Perrone, and Vincenzo Giliberti. 2023. A comparison of different topic modeling methods through a real case study of italian customer care. *Algorithms*, 16(2).

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. *Proceedings of the International AAAI Conference on Web and Social Media*.

Xuan Hieu Phan, Minh Le Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *The Web Conference*.

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. ALTO: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1158–1169, Berlin, Germany. Association for Computational Linguistics.

Alicia Raeburn. 2022. Context switching is killing your productivity. `https://asana.com/resources/context-switching`. Accessed: insert access date here.

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Burr Settles. 2012. Active learning (synthesis lectures on artificial intelligence and machine learning). In *Findings*.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3(null):583–617.

Martina Sundqvist, Julien Chiquet, and Guillem Rigaill. 2022. Adjusting the adjusted rand index: A multinomial story. *Comput. Stat.*, 38(1):327–347.

Hongyin Tang, Miao Li, and Beihong Jin. 2019. A topic augmented text generation model: Joint learning of semantics and structural features. 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 5090–5099, Hong Kong, China. Association for Computational Linguistics.

U.S. Department of Commerce. 2020. Community Resilience Planning Guide for Buildings and Infrastructure Systems. Technical Report NIST SP 1190GB-16, National Institute of Standards and Technology.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online. Association for Computational Linguistics.

A. T. Wilson and P. A. Chew. 2010. Term weighting schemes for latent dirichlet allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 465–473. Association for Computational Linguistics.

W Xu and K Eguchi. 2022. A supervised topic embedding model and its application. *PLoS One*, 17(11):e0277104. PMID: 36331905; PMCID: PMC9635756.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019. Sp-10k: A large-scale evaluation set for selectional preference acquisition.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

He Zhao, Piyush Rai, Lan Du, Wray Buntine, Dinh Phung, and Mingyuan Zhou. 2020. Variational autoencoders for sparse and overdispersed discrete data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1684–1694. PMLR.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*.

Ying Zhao. 2005. *Criterion Functions for Document Clustering*. Ph.D. thesis, USA. AAI3180039.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024. Explore spurious correlations at the concept level in language models for text classification.

10 Appendix

A Clustering Evaluation Metric Details

We list and show the calculation details of automated evaluation metrics discussed in Section 2.4 for easy of reproducing our work in this section. Suppose the classifier is trained on existing documents with user input labels (5% of the documents), and the classifier predicts labels for all the documents, and they are partitioned into clusters denoted as $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$. The official gold clusters are denoted as $C = \{c_1, c_2, \ldots, c_J\}$.

Purity It is calculated by assigning each cluster to the class which is most frequent in the cluster, and counting the correctly assigned points in that cluster. The formula to calculate the purity between the predicted and the gold clusters is:

$$\text{Purity}(\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j|. \quad (5)$$

N is the total number of points, ω_k is the kth cluster, c_j is the jth class. $\omega_k \cap c_j$ is the number of points in cluster ω_k that belongs to class c_j, and $\max_j$ is maximum number of class c_j intersection with cluster ω_k (Zhao, 2005).

Adjusted Normalized Mutual Information The Adjusted Normalized Mutual Information (ANMI) is an improved version of the Normalized Mutual Information (NMI) metric used for comparing the similarity between two clusterings that adjusts for chance to make the score more robust and comparable across different situations:

$$\text{ANMI} = \frac{2 \times (\text{MI} - \text{E}[\text{MI}])}{(H(\text{C}) + H(\text{K})) - 2 \times \text{E}[\text{MI}]}. \quad (6)$$

The mutual information (MI) measures how much information we know about the gold clustering by knowing about the predicted clustering. The expected mutual information E[MI] is calculation of what the MI would be if the predicted clusters were completely at random, but still considering the size of the clusters. H(K) measures the randomness or disorder within the gold clustering and H(C) measures the randomness or disorder within the predicted clustering–entropy. A higher entropy means higher randomness for the clusters (Amelio and Pizzuti, 2016).

Adjusted Rand Index Rand Index (RI) computes the similarity between two clustering by considering pairs that are assigned in the same or different clusters in the predicted and true clustering (Rand, 1971). The formula for RI is:

$$RI = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (7)$$

TP is the number of pairs that are in the same set in both the predicted and gold clusters, and TN is the number of pairs that are in different sets in the predicted and gold clusters. Otherwise, the pairs are either FP or FN.

The Adjusted Rand Index (ARI) is the corrected-for-chance version of the RI. It accounts for the fact that the RI score will increase as the number of clusters increases, even if the clustering is random:

$$RI = \frac{\text{RI} - \text{Expected RI}}{\text{Max RI} - \text{Expected RI}}. \quad (8)$$

Expected RI is the expected value of the RI under random labeling, respecting the marginal distributions of cluster sizes. Max RI is the highest possible value that the RI could take, given the constraints of the clustering problem. A Max RI of 1.0 indicates two clusterings are identical, but when adjusting it for chance, Max RI can be less than 1 depending on the distribution of cluster sizes.

Normalized Pointwise Mutual Information NPMI evaluates how semantically related the top words in each topic are to the documents in that topic, which in turn reflects the quality of the generated topics by a topic model:

$$\text{NPMI}(x, y) = \frac{\log \frac{P(x,y)}{P(x) \cdot P(y)}}{-\log P(x, y)}. \quad (9)$$

P(x, y) represents the probability of words x and y co-occuring together in a set of documents, where P(x) and P(u) are probabilities of observing words x and y independently in the set of documents.

B Study Group Details

We provide more details of implementation about our 6 study groups introduced in Section 3.1 with two components– user experience and classifier training.

B.1 User Interface Experience

The baseline group (1) NONE users only has access to a list of documents in the initial page shown in Figure 5. Active learning picks the most informative document and place it on top of the page so users can quickly selects it. Groups (2)-(6) with topic models have access to both active learning and topic overview shown in Figure 6. Users can explore the overall themes of the document sets then start labeling documents. After a user selects a document, topic model group users have access to the most related topics for the document, keywords, and highlighted keywords that are above 0.05 threshold for a selected topic shown in Figure 7. Group (1) NONE users do not have access to the topic keywords and highlighted texts, but still retain the active learning basic features- the

top three most relevant labels of the document predicted by the classifier. In all groups, users can click *submit & next* button to automatically go to the next document selected by active learning or they can go back to the list of documents to select other documents.

B.2 Classifier Training

(1) NONE group users has a logistic regression classifier trained with their labeled documents. The classifier picks the next document based on the preference function with only *tf-idf* as its input features. For group (2)-(6), we first compute the topic model probability features, where each document has an associated vector that contains probabilities it belongs to each topic. We encode the raw text features using *tf-idf* first, and concatenate the topic vector with each encoded document features and train a classifier with user labeled documents. Classifiers in Group (2)-(6) have additional features generated by different topic models that can help classification to generalize better to unseen documents. Different topic models generate different features that can have diverse performance in downstream classification tasks.

C Simulated Experiment Details

Training Topic Models We preprocess the dataset by tokenizing and filtering stopwords; we use a tf-idf threshold of three to remove rare and too-common words.

For LDA and sLDA, we use the Tomotopy library (Lee, 2022), which uses Gibbs sampling to train classical topic models. To compare two datasets fairly, we chose $K = 35$ topics for all five topic models in our group, which optimized average coherence. For LDA and sLDA, we use the term weight scheme ONE (Wilson and Chew, 2010). sLDA takes more extra hyperparameters than LDA does. For sLDA, we also use binary-type response variables to indicate user input labels. Otherwise, LDA and sLDA use the default hyperparameter values. sLDA initially does not take in any response variables. We train LDA and sLDA with 2000 iterations until a smaller change of log-likelihood and NPMI coherence.

For CTM, we use SBERT `paraphrase-distilroberta-base-v2` to fetch sentence embeddings for the dataset, then concatenate them with BoW representation. We used *CombinedTM* (Bianchi et al., 2021) with a 768 contextual size, with $K = 35$ topics, and trained it with 250 epochs. We also use `paraphrase-distilroberta-base-v2` to fetch sentence ebmeddings to train Bertopic. For ETM, we use Word2Vec (Mikolov et al., 2013) to encode documents and train it with 250 epochs.

Classifier Initialization and Features Since users are more likely to create more granular label specifications for each document. We used sublabels as pseudo user-entered labels while using the more general labels as our gold standard.

We use `sklearn` SGD as our classifier for active learning document selection.[11] We transform our raw dataset using unigram tf-idf as input features for the classifier. For LDA, sLDA, and CTM groups, we also concatenate topic probability distributions for all the documents with unigram tf-idf features that also encode topic information to the classifier. Since the classifier requires at least two classes to be fitted, we pick random documents, and use sublabels as surrogate user input labels, and activate the preference function until the classifier has at least two class labels. We use incremental leaning (Rosenblatt, 1958) to fit and update the classifier, retaining originally learned parameters.[12] The classifier's predictions with the more general labels assess the clustering quality.

Simulated Experiment Upon analyzing the document lengths in our dataset, we deduced that considering individual reading speed variances, a user can feasibly label between 90 to 400 documents within an hour. For our simulated user study, we automatically run our algorithm for each group to input labels for 400 documents, constantly updating the classifier for every document labeled, and sLDA for every 50 documents labeled. Each group underwent 15 iterations of the experiment. For consistency, we aggregated the results by taking the median value for each document in each group.

Validity of Simulated Experiments Of all the methods, CTM consistently does better on purity, ARI, and ANMI, which underscores the right choice

[11] We use hyperparameters: loss='log_loss', penalty='l2', tolerance=10e-3, random_state=42, learning_rate='optimal', eta0=0.1, validation_fraction=0.2, and alpha=0.000005.

[12] There are two exceptions we reinitialize the classifier: if a new label class is introduced to the classifier, we reinitialize the classifier and train it with labeled documents; if sLDA is updated with surrogate response variables, we rebuild the features by concatenating tf-idf features with new topic probability distributions, and restart the classifier with new features.

of NTM can generate topic probability features that do better on classification. Such features, rooted in pre-trained embeddings, are perceived by compact machine learning models as more intuitive than the generative topic probabilities yielded by classical models like LDA and sLDA. sLDA and ETM, on the other side, is worse than LDA, where LDA remains competitive against two other NTMs. The classifier without topic information falls short behind the classifier with topic information except for ETM.

Our simple simulated experiments serve as a reliable proxy, allowing us to expect similar trends when actual human labeling is in play and to track the evolution of classifier predictions as more documents are labeled over time. However, we acknowledge that relying solely on simulated evaluation metrics has limitations. The classifier does not consider using topic keywords and topic overviews to create labels. Other factors, including fatigue and loss of attention, might also affect the quality of labels created by real users. Such metrics also do not capture the complete essence of user preferences, especially concerning the keywords produced by topic models, the highlighted keywords, or the specific documents recommended by the preference function.

D Dataset Details

The Bills have over 400,000 bills spanning from 1947 to 2009, where each bill is meticulously labeled with primary and secondary topics, as detailed in a comprehensive codebook.[13] The latest iteration of this dataset has seen its topics labeled by adept human coders, who were trained using the preceding dataset version. The inter-annotator agreement was observed to be an impressive 95% for primary topics and 75% for secondary ones. Such extensive and refined labeling, carried out by trained annotators over numerous years, assures the dataset's label quality. The 20newsgroup is a popular benchmark dataset that has 6 major labels and 20 sub-labels. We remove duplicate documents, documents that are shorter than 30 tokens, documents that contain sensitive topics, and documents that the general public is not familiar with the Bills and 20newsgroup dataset.

[13] https://comparativeagendas.s3.amazonaws.com/codebookfiles/Codebook_PAP_2019.pdf.

E User Label Evaluations

We do a sanity check on the 800 randomly selected labeled documents, to ensure users are creating meaningful labels. Within each group, we sort the users based on the summation of purity, ARI, ANMI at the end of the 61st minute in ascending order. We take the middle 8 users and randomly pick 200 labeled documents from each group. We have two annotators manually judge the user labels based on the following two criteria: 1. Can the user label be considered equivalent or a subfield of the gold label (major label and sub label)? 2. Does the user label reflect the contents of the passage? If the annotator rates 'yes' for criteria 1, criteria 2 will be skipped. Otherwise, the annotator will need to read the actual passage to judge the quality of the user labels. Among 800 labeled documents, we have 787 documents that satisfy at least one of the two criteria, which ensures most users are making meaningful labels and carefully conducting the study.

F User Interface

Figure 6 and Figure 7 show a basic layout of CTM used in our user study. The keywords and document clusters will not be displayed to NONE group users. Instead, a random list of documents are displayed to them in Figure 6 page. In Figure 7 page, NONE users are not displayed with the *Top Topic Keywords* and the highlighted texts.

G Topic Model Keywords

Table 3, 4, and 5 show the 2 topics with highest, median, and lowest NPMI coherence scores for LDA, sLDA, CTM, and sLDA trained with user input labels as response variables. The topic keywords generated by LDA are more general and inclusive while the topic keywords generated by CTM are more specific and related to the top passages.

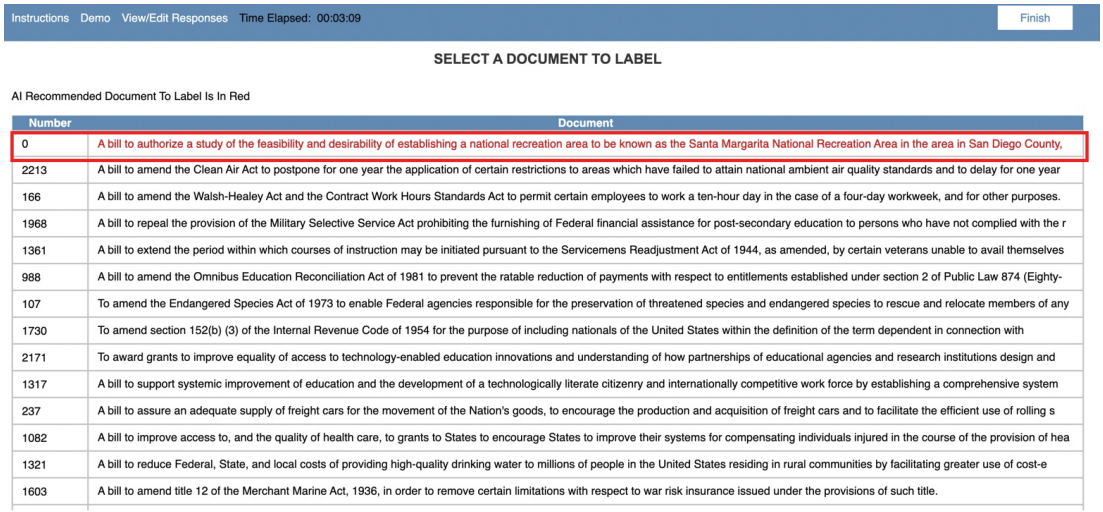

Figure 5: This is the overview (1) NONE group. Users are not presented with topic overview, but active learning classifier picks the document based on the preference function and place it on top of the page.

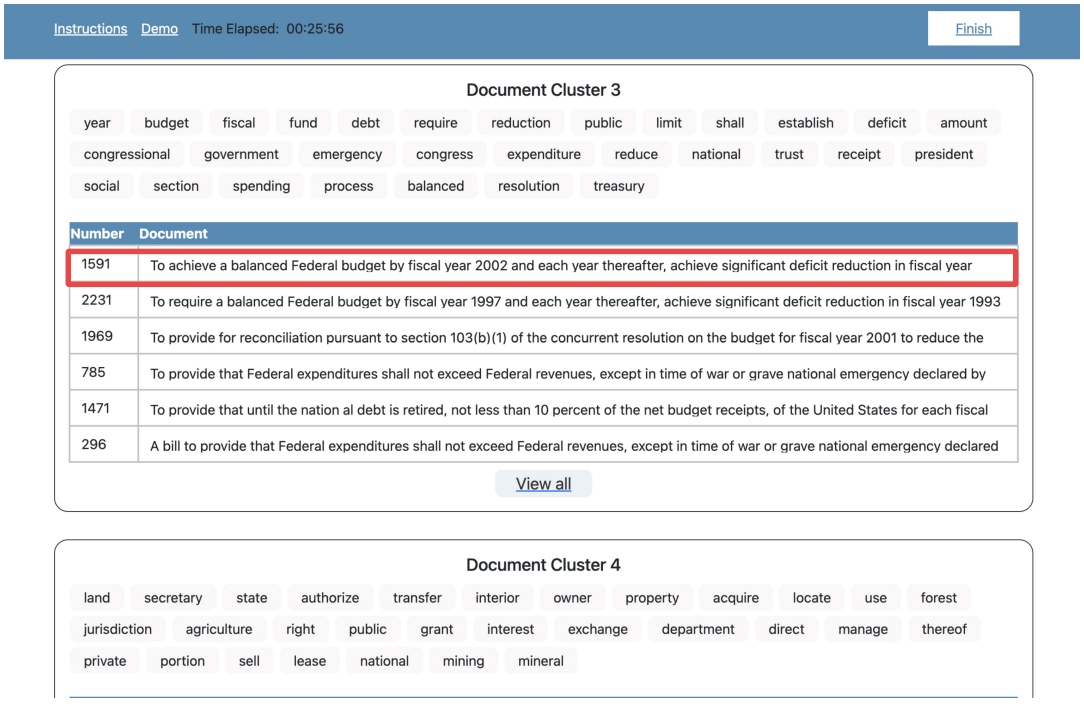

Figure 6: Under topic model settings, users are displayed all topics, keywords, and documents in each topic. If active learning picks a document, the topic and the document cluster containing that document will be displayed at the very top of this page. The document is also displayed on the top of the document cluster. For example, the document marked red is an example of a document picked by active learning. For the baseline, NONE group, topic keywords, and document clusters are not displayed. All documents are displayed in one block, and the recommended document is always on top of the page above other documents.

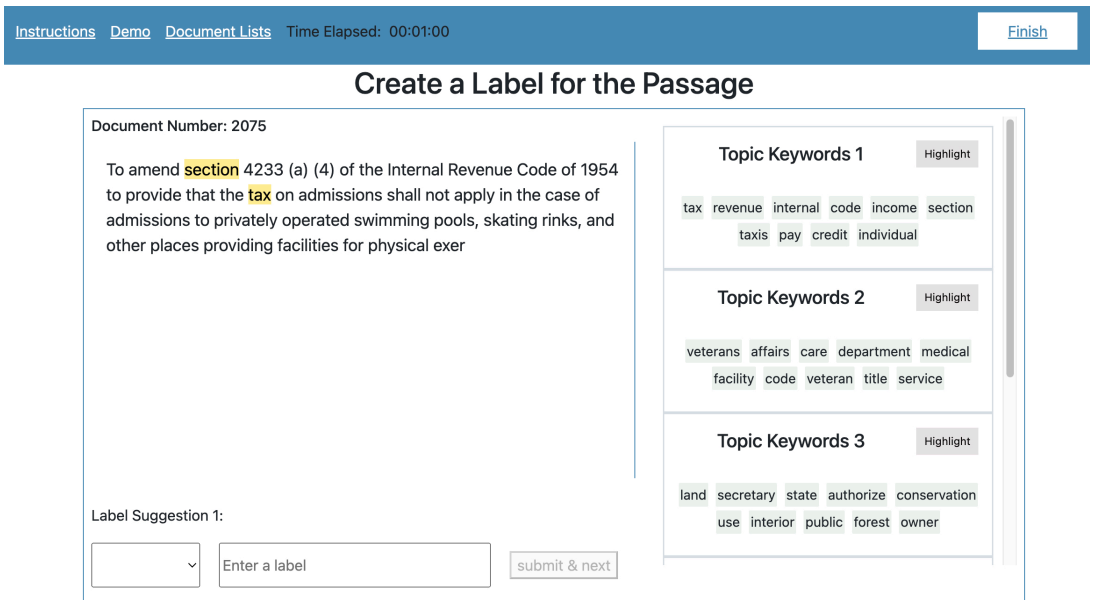

Figure 7: For a user-selected document, a user can either make a label for the document or skip the document. The top 5 most relevant topics and top keywords for the selected document are displayed on the right side. The highlight function helps users quickly find words that are above the 0.05 threshold for a chosen topic. Users could also select a label from the dropdown box, which the labels are ranked by softmax probabilities of the classifier, and the dropdown labels are what the users have created so far. For NONE, the highlights and topics will not be available to the users.

Model	NPMI	Keywords	Passage
LDA	0.39	exemption, income, dependent, increase, taxpayer, tax, spouse, personal, additional, include	To provide that certain survivor benefits received by a child under public retirement systems shall not be taken into account in determining whether the child is a dependent for income tax purposes.
LDA	0.24	tax, revenue, internal, code, income, section, taxis, pay, credit, individual	To amend the Internal Revenue Code of 1954 to include the sintering and burning of clay, shale, and slate used as lightweight aggregates as a treatment process considered as mining.
sLDA	0.35	rescind, control, authority, budget, president, special, impoundment, propose, transmit, section	To rescind certain budget authority proposed to be rescinded (R92-66) in a special message transmitted to the Congress by the President on March 20, 1992.
sLDA	0.22	tax, revenue, income, internal, code, exemption, section, individual, taxis, shall	To amend the Internal Revenue Code to provide that gain or loss from the sale or exchange of certain real estate shall be treated as a capital gain or loss.
CTM	0.50	president, authority, propose, rescind, congress, special, impoundment, march, accordance, trasmit, message	A bill to rescind certain budget authority contained in the message of the President of January 27, 1978 (H. Doc. 95-285), transmitted pursuant to the Impoundment Control Act of 1974.
CTM	0.38	exemption, include, taxpayer, personal, additional, increase, dependent, spouse, income, old	To increase from $600 to $750 the personal income tax exemptions of a taxpayer (including the exemption for a spouse, the exemption for a dependent, and the additional exemption for old age, or blindness).
sLDA(user)	0.42	budget, rescind, control, president, authority, impoundment, congress, transmit, message, section	To amend part C of the Balanced Budget and Emergency Deficit Control Act of 1985 to extend the discretionary spending limits and pay-as-you-go through fiscal year 2009.
sLDA(user)	0.26	education, school, student, loan, program, secondary, institution, elementary, educational, teacher	To amend the Higher Education Act of 1965 to expand the loan forgiveness and loan cancellation programs for teachers, to provide loan forgiveness and loan cancellation programs for nurses, and for other purposes.

Table 3: Topic models automatically discover topics and themes in the Bills dataset. These topics give users a global sense of probable stories and themes in a dataset. We show the top 2 topics for each topic model and their relevant keywords and relevant passages. sLDA is the initial model without fitting with response variables, which is used for all users in sLDA group. sLDA(user) uses a pre-saved model, which is derived from the median calculations (median of summation of purity, ARI, ANMI among 15 users) across 15 users in sLDA. sLDA(user) generates top topics with higher top coherence scores than other models. The keywords also appear more often and are more related to passages.

Model	NPMI	Keywords	Passage
LDA	0.13	water, wildlife, conservation, fish, establish, management, resource, national, development, coastal	To create a joint commission of the United States and the State of Alaska to make administrative determinations of navigability of inland nontidal waters in the State of Alaska for State selections.
LDA	0.12	food, drug, use, cosmetic, respect, human, child, information, intend, manufacturer	A bill to amend Sections 403 and 405 of the Federal Food, Drug, and Cosmetic Act to require that foods intended for human consumption be labeled to show the amount of sodium and potassium they contain.
sLDA	0.10	labor, section, employee, national, organization, fair, provision, relations, right, railway	To amend the Railroad Retirement Act of 1937 and the Social Security Act to eliminate those provisions which restrict the right of a spouse or survivor to receive benefits simultaneously under both acts.
sLDA	0.07	highway, title, section, amend, national, code, fund, system, construction, stat	A bill to supplement the Federal Aid Road Act, approved July 11, 1916, as amended and supplemented, to authorize appropriations for the construction of greatly needed rural local roads, and for other purposes.
CTM	0.07	contract, standards, work, wage, contractor, cause, hour, fair, employer, employee	A bill to provide for the creditability of certain service in determining the order of retention for competing employees in a reduction in force affecting the Federal Grain Inspection Service.
abrctm	0.06	revenue, internal, code, section, estate, sale, admission, value, treatment, relate	To amend section 112 (b) of the Internal Revenue Code (relating to recognition of gain in certain corporate liquidations) so that it will apply to cases where the transfer of all the property under the liquidation occurs within 1 calendar month in 1953.
sLDA(user)	0.03	program, establish, improve, development, system, promote, assist, provide, national, encourage	A bill to improve existing tertiary eye centers, to examine the delivery of eye care to the general public, and to study the feasibility of implementing a system of tertiary eye care centers throughout the United States.
sLDA(user)	0.02	state, fund, program, year, title, establish, assistance, construction, facility, authorize	To amend the National Housing Act to authorize the Secretary of Housing and Urban Development to insure mortgages for the acquisition, construction...

Table 4: The table shows the 18th and 19th coherent topics discovered by different topic models. The bottom 2 topics for sLDA(user) only have a few passages associated with each of them.

Model	NPMI	Keywords	Passage
LDA	-0.10	person, foreign, prohibit, business, engage, country, trade, domestic, enable, stock	To provide an exception from certain group health plan requirements to allow small businesses to use pre-tax dollars to assist employees in the purchase of policies in the individual health insurance market, and for other purposes.
LDA	-0.08	vessel, coast, guard, marine, specie, merchant, port, law, academy, endangered	To amend the Merchant Marine Act of 1936 and the Maritime Academy Act of 1958 to enlarge the mission of the U.S. Merchant Marine Academy and to assist in enlarging the mission of the State maritime academies.
sLDA	-0.12	meat, product, inspection, state, continental, shelf, outer, poultry, import, land	A bill to modify the method of determining quantitative limitations on the importation of certain articles of meat and meat products, to apply quantitative limitations on the importation of certain additional articles of meat, meat products, and livestock, and for other purposes.
sLDA	-0.11	fla, know, value, historic, shall, national, site, use, fort, dam	A bill to provide that the reservior formed by the lock and dam referred to as the Millers Ferry lock and dam on the Alabama River, Alabama, shall hereafter be known as the William Bill Dannelly Reservior.
CTM	-0.29	locate, convey, transfer, territory, memorial, historical, washington, smithsonian, city, conveyance	To provide for the conveyance of certain excess real property of the United States to the city of Mission, the city of McAllen, and the city of Edinburg, all situated in the State of Texas.
CTM	-0.12	highway, aid, interstate, road, alaska, system, fund, fla, commission, transportation	To amend section 5 of the Department of Transportation Act to authorize the National Transportation Safety Board to employ 5,000 investigators to carry out its powers and duties under that act.
sLDA(user)	-0.36	gas, purpose, greenhouse, wheat, red, cheese, cheddar, operate, exist, standards	To provide that the rules of the Environmental Protection Agency entitled National Emission Standards for Hazardous Air Pollutants for Reciprocating Internal Combustion Engines...
sLDA(user)	-0.31	gram, trans, drugs, deadline, intervention, temple, manatees, plains, ombudsman, leaseholder	To direct the Commissioner of Food and Drugs to revise the Federal regulations applicable to the declaration of the trans fat content of a food on the label and in the labeling of the food when such content is less than 0.5 gram.

Table 5: The table shows the least two coherent topics discovered by different topic models. The bottom 2 topics for sLDA(user) only have a few passages associated with each of them.