ALTO: Active Learning with Topic Overviews for Speeding Label Induction and Document Labeling

Forough Poursabzi-Sangdeh
Computer Science
University of Colorado
forough.poursabzisangdeh@colorado.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Jordan.Boyd.Graber@colorado.edu

Leah Findlater
iSchool and UMIACS
University of Maryland
leahkf@umd.edu

Kevin Seppi
Computer Science
Brigham Young University
kseppi@cs.byu.edu

Abstract

Effective text classification requires experts to annotate data with labels; these training data are time-consuming and expensive to obtain. If you know what labels you want, active learning can reduce the number of labeled documents needed. However, establishing the label set remains difficult. Annotators often lack the global knowledge needed to induce a label set. We introduce ALTO: Active Learning with Topic Overviews, an interactive system to help humans annotate documents: topic models provide a global overview of what labels to create and active learning directs them to the right documents to label. Our user study with forty annotators shows that while active learning by itself is best in extremely resource limited conditions, topic models (even by themselves) lead to better label sets, and ALTO's combination is best overall.

1 Introduction

Many fields depend on texts labeled by human experts; computational linguistics uses such annotation to determine word senses and sentiment (**?**; **?**); while social science uses "coding" to scale up and systemetize content analysis (**?**; **?**).

Classification takes these labeled data as a training set and labels new data automatically. Creating a broadly applicable and consistent label set that generalizes well is time-consuming and difficult, requiring expensive annotators to examine large swaths of the data. Effective NLP systems must measure (**?**; **?**; **?**) and reduce annotation cost (**?**). Annotation is hard because it requires both *global* and *local* knowledge of the entire dataset. Global knowledge is required to create the set of labels, and local knowledge is required to annotate the most useful examples to serve as a training set for an automatic classifier.

We create a single interface—ALTO (Active Learning with Topic Overviews)—to address both the global and local challenges using two machine learning tools: *topic models* and *active learning* (we review both in Section 2). Topic models address the need for annotators to have a *global overview* of the data, exposing the broad themes of the corpus so annotators know what labels to create. Active learning *selects* documents that help the classifier understand the differences between labels and directs the user's attention *locally* to them. We thus create four experimental conditions to compare the effects of providing users with either a topic model or a simple list of documents, with or without active learning suggestions (Section 3). Following this section we then describe our data and evaluation metrics (Section 4).

Through both synthetic experiments (Section 5) and a user study (Section 6) with 40 participants, we evaluate ALTO and its constituent components by comparing results from the four conditions introduced above. We first examine user strategies for organizing documents, user satisfaction, and user efficiency. Finally, we evaluate the overall effectiveness of the label set in a post study crowdsourced task.

Topic words	Document Title
metropolitan, carrier, rail, freight, passenger, driver, airport, traffic, transit, vehicles	A bill to improve the safety of motorcoaches, and for other purposes.
violence, sexual, criminal, assault, offense, victims, domestic, crime, abuse, trafficking	A bill to provide criminal penalties for stalking.
agricultural, farm, agriculture, rural, producer, dairy, crop, producers, commodity, nutrition	To amend the Federal Crop Insurance Act to extend certain supplemental agricultural disaster assistance programs through fiscal year 2017, and for other purposes.

Table 1: Given a dataset—in this case, the US congressional bills dataset—topics are automatically discovered sorted lists of terms that summarize segments of a document collection. Topics also are associated with documents. These topics give users a sense of documents' main themes and help users create high-quality labels.

2 ALTO: Active Learning with Topic Overviews

This section details ALTO:[1] a framework for assigning labels to documents that uses both global and local knowledge to help users create and assign document labels. We explain how ALTO uses topic models to aid label induction and document labeling. We then use active learning to direct user attention and speed document labeling.

Topic Models Topic models (**?**) automatically induce structure from a text corpus. Given a corpus and a constant K for the number of topics, topic models output (i) a distribution over words for each topic k ($\phi_{k,w}$) and (ii) a distribution over topics for each document ($\theta_{d,k}$). Each topic's most probable words and associated documents can help a user understand what the collection is about. Table 1 shows examples of topics and their highest associated documents from our corpus of US congressional bills.

Our hypothesis is that showing documents grouped by topics will be more effective than having the user wade through an undifferentiated list of random documents and *mentally sort the major themes themselves*.

[1]Code available at `https://github.com/Foroughp/ALTO-ACL-2016`

Active Learning Active learning (**?**) directs users' attention to the examples that would be most useful to label when training a classifier. When user time is scarce, active learning builds a more effective training set than random labeling: uncertainty sampling (**?**) or query by committee (**?**) direct users to the most useful documents to label.

In contrast to topic models, active learning provides local information: this is the individual document you should pay attention to. Our hypothesis is that active learning, when used as a *preference function* to direct the users to documents most beneficial to label, will not only be more effective than randomly selecting documents but will also *complement* the global information provided by topic models. Section 3.3 describes the preference functions for the experimental conditions.

3 Study Conditions

Our goal is to characterize how local and global knowledge can aid users in annotating a dataset. This section describes our four experimental conditions and outlines the user's process for labeling documents.

3.1 Study Design

The study uses a 2×2 between-subjects design, with factors of document collection *overview* (two levels: topic model or list) and document *selection* (two levels: active or random). The four conditions, with the TA condition representing ALTO, are:

1. Topic model and active selection (TA)
2. Topic model and random selection (TR)
3. List and active selection (LA)
4. List and random selection (LR)

3.2 Document Collection Overview

The topic and list overviews offer different overall structure but the same basic elements for users to create, modify, and apply labels (Section 3.4). The topic overview (Figure 1a) builds on **?**): for each topic, the top twenty words are shown alongside twenty document titles. Topic words (w) are sized based on their probability $\phi_{k,w}$ in the topic k and the documents with the highest probability of that topic ($\theta_{d,k}$) are shown. The list overview, in contrast, presents documents as a simple, randomly ordered list of titles (Figure 1b). We display the same number of documents ($20K$, where K is the total number of topics) in both the topic model and list overviews, but the list overview obviously provides no topic information.

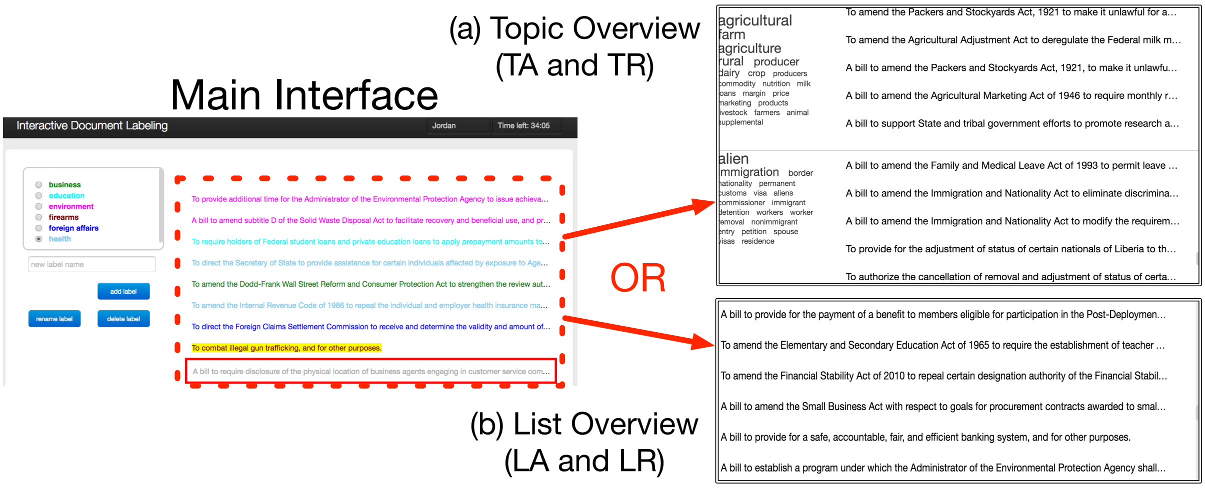

Figure 1: Our annotation system. Initially, the user sees lists of documents organized in either a list format or grouped into topics (only two topics are shown here; users can scroll to additional document). The user can click on a document to label it.

3.3 Document Selection

To provide consistency across the four conditions, all of the conditions use a document preference function U to direct the user's attention to a document to label. For the random selection conditions, TR and LR, document selection is random, within a topic or globally. We expect this to be less useful than active learning. The document preference functions are:

LA: LA uses traditional uncertainty sampling:

$$U_d^{\text{LA}} = \mathbb{H}_C\left[Y_d\right], \tag{1}$$

where $\mathbb{H}_C\left[y_d\right] = -\sum_i P(y_i|d) log P(y_i|d)$ is the classifier entropy. Entropy measures how confused (uncertain) classifier C is about its prediction of a document d's label y. Intuitively, it prefers documents that most of the labels are likely to be predicted to documents that only one of the labels is highly likely to be chosen. **LR**: LR's approach is the same as LA's except we replace $\mathbb{H}_C\left[y_d\right]$ with a uniform random number:

$$U_d^{\text{LR}} \sim \text{unif}(0, 1). \tag{2}$$

In contrast to LA, which suggests the most uncertain document, LR suggests a random document.

TA: **?**) argue that clustering should inform active learning criteria, balancing coverage against classifier accuracy. We adapt their method to flat topic models—in contrast to their hierarchical cluster trees—by creating a composite measure of document uncertainty within a topic:

$$U_d^{\text{TA}} = \mathbb{H}_C\left[y_d\right]\theta_{d,k}, \tag{3}$$

where k is the prominent topic for document d. U_d^{TA} prefers documents that are *representative* of a topic (i.e., have a high value of $\theta_{d,k}$ for that topic) and are informative for the classifier.

TR: TR's approach is the same as TA's except we replace $\mathbb{H}_C\left[Y_d\right]$ with a uniformly random number:

$$U_d^{\text{TR}} = \text{unif}(0, 1)\theta_{d,k}. \tag{4}$$

Similar to TA, this prefers documents that are representative of a topic, but not any particular such document. Incorporating the random component encourages for covering different documents in diverse topics.

In LA and LR, the preference function directly chooses a document and directs the user to it. On the other hand, U_d^{TA} and U_d^{TR} are topic dependent: To avoid the negative and misleading effect of topics, users' attention should be drawn to documents that are representative of a specific topic. Therefore, the factor $\theta_{d,k}$ appears in both. Thus, they require that a topic be chosen first and then the document with maximum preference, U, within that topic can be chosen. In TR, the topic is chosen randomly. In TA, the topic is chosen by

$$k^* = \arg\max_k(\text{median}_d(\mathbb{H}_C\left[y_d\right]\theta_{d,k}). \tag{5}$$

That is the topic with the maximum median U. Median encodes how "confusing" a topic is.[2] Intuitively, k^* is the topics that the classifier is confused about its documents' labels.

[2]Outliers skew other measures (e.g., max or mean).

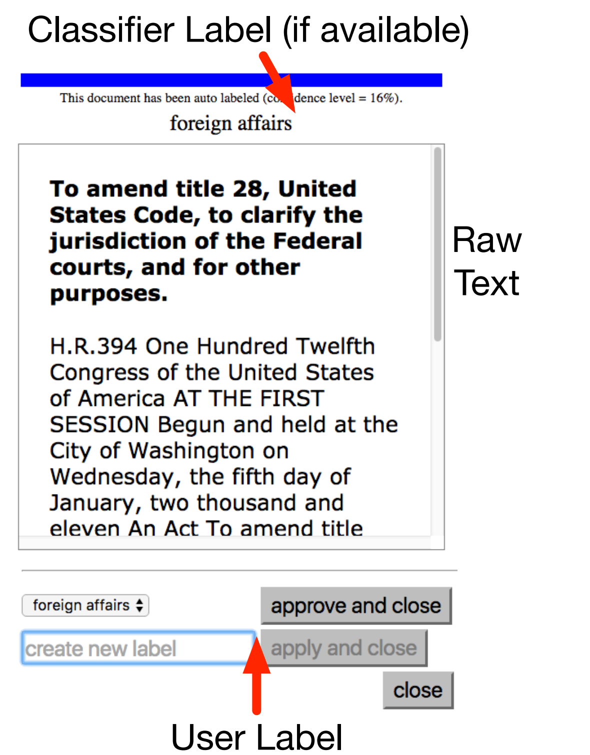

Figure 2: Document view: after clicking on a document from the list or topic overview, the user inspects the text and provides a label. If the classifier has a guess at the label, the user can confirm the guess.

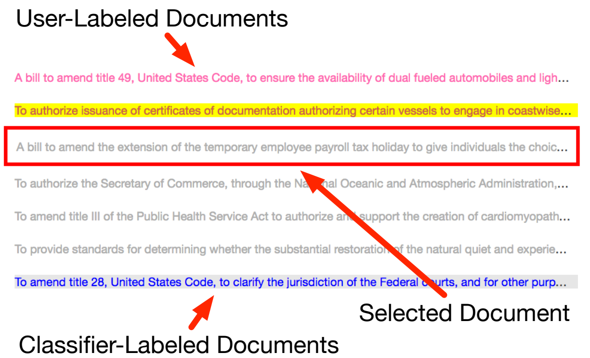

Figure 3: After the user has labeled some documents, the system can automatically label other documents and select which documents would be most helpful to annotate next. In the random selection setting, random documents are selected.

3.4 User Labeling Process

The user's labeling process is the same in all four conditions. The *overview* (topic or list) allows users to examine individual documents (Figure 1). Clicking on a document opens a dialog box (Figure 2) with the text of the document and three options:

1. Create and assign a new label to the document.
2. Choose an existing label for the document.
3. Skip the document.

Once the user has labeled two documents with different labels, the displayed documents are replaced based on the preference function (Section 3.3), every time the user labels (or updates labels for) a document. In TA and TR, each topic's documents are replaced with the top twenty most uncertain documents. In LA and LR, all documents are updated with the top $20K$ uncertain documents.[3]

The system also suggests one document to consider by auto-scrolling to it and drawing a red box around its title (Figure 3). The user may ignore that document and click on any other document. After the user labels ten documents, the classifier runs and assigns labels to other documents.[4] For classifier-labeled documents, the user can either approve the label or assign a different label. The process continues until the user is satisfied or a time runs out (forty minutes in our user study, Section 6). We use time to control for the varying difficulty of assigning documents: active learning will select more difficult documents to annotate, but they may be more useful; time is a more fair basis of comparison in real-world tasks.

4 Data and Evaluation Metrics

In this section, we describe our data, the machine learning techniques to learn classifiers from examples, and the evaluation metrics to know whether the final labeling of the complete documents collection was successful.

4.1 Datasets

Data Our experiments require corpora to compare user labels with gold standard labels. We experiment with two corpora: 20Newsgroups (?) and US congressional bills from GovTrack.[5]

For US congressional bills, GovTrack provides bill information such as the title and text, while the Congressional Bills Project (?) provides labels and sub-labels for the bills. Examples of labels are <u>agriculture</u> and <u>health</u>,

[3] In all conditions, the number of displayed unlabeled documents is adjusted based on the number of manually labeled documents. i.e. if the user has labeled n documents in topic k, n manually labeled documents followed by top $20 - n$ uncertain documents will be shown in topic k.

[4] To reduce user confusion, for each existing label, only the top 100 documents get a label assigned in the UI.

[5] https://www.govtrack.us/

while sub-labels include agricultural trade and comprehensive health care reform. The twenty top-level labels have been developed by consensus over many years by a team of top political scientists to create a reliable, robust dataset. We use the 112th Congress; after filtering,[6] this dataset has 5558 documents. We use this dataset in both the synthetic experiments (Section 5) and the user study (Section 6).

The 20 Newsgroups corpus has $19,997$ documents grouped in twenty news groups that are further grouped into six more general topics. Examples are talk.politics.guns and sci.electronics, which belong to the general topics of politics and science. We use this dataset in synthetic experiments (Section 5).

4.2 Machine Learning Techniques

Topic Modeling To choose the number of topics (K), we calculate average topic coherence (**?**) on US Congressional Bills, between ten and forty topics and choose $K = 19$, as it has the maximum coherence score. For consistency, we use the same number of topics ($K = 19$) for 20 Newsgroups corpus. After filtering words based on TF-IDF, we use Mallet (**?**) with default options to learn topics.

Features and Classification A logistic regression predicts labels for documents and provides the classification uncertainty for active learning. To make classification and active learning updates efficient, we use incremental learning (**?**, LingPipe). We update classification parameters using stochastic gradient descent, restarting with the previously learned parameters as new labeled documents become available.[7] We use cross validation on prominent topic as labels to set the parameters for learning the classifier.[8]

The features for classification include topic probabilities, unigrams, and the fraction of labeled documents in each document's prominent topic. The intuition behind adding this last feature is to allow active learning to suggest documents in a diverse range of topics if it finds this feature a useful indicator of uncertainty.[9]

4.3 Evaluation Metrics

Our goal is to create a system that allows users to quickly induce a high-quality label set. We compare the user-created label sets against the data's gold label sets. Comparing different clusterings is a difficult task, so we use three clustering evaluation metrics: purity (**?**), rand index (**?**, RI), and normalized mutual information (**?**, NMI).[10]

Purity Purity measures how "pure" user clusters are compared to gold clusters. Given each user cluster, it measures what fraction of the documents in a user cluster belong to the most frequent gold label in that cluster:

$$\text{purity}(\mathbf{U}, \mathbf{G}) = \frac{1}{N} \sum_{l} \max_{j} |U_l \cap G_j|, \tag{6}$$

where L is the number of labels user creates, $\mathbf{U} = \{U_1, U_2, \dots, U_L\}$ is the user clustering of documents, $\mathbf{G} = \{G_1, G_2, \dots, G_J\}$ is gold clustering of documents, and N is the total number of documents. The user U_l and gold G_j labels are interpreted as sets containing all documents assigned to that label.

Rand index (RI) RI is a *pair counting based* measure, where cluster evaluation is considered as a series of decisions. If two documents have the same gold label and the same/different user label (TP/FN) and if they do not have the same gold label and are not/are assigned the same user label (TN/FP) the decision is right/wrong. RI measures the percentage of decisions that are right:

$$\text{RI} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{7}$$

Normalized mutual information (NMI) NMI is an *information theoretic based* measure that measures the amount of information one gets about the gold clusters by knowing what the user clusters are:

$$\text{NMI}(\mathbf{U}, \mathbf{G}) = \frac{2\mathbb{I}(\mathbf{U}, \mathbf{G})}{\mathbb{H}_{\mathbf{U}} + \mathbb{H}_{\mathbf{G}}}, \tag{8}$$

[6] We remove bills that have less than fifty words, no assigned gold label, duplicate titles, or have the gold label GOVERNMENT OPERATIONS or SOCIAL WELFARE, which are broad and difficult for users to label.

[7] Exceptions are when a new label is added, a document's label is deleted, or a label is deleted. In those cases, we train the classifier from scratch. Also, for final results in Section 6, we train a classifier from scratch.

[8] We use `blockSize` = 1/#examples, `minEpochs` = 100, `learningRate` = 0.1, `minImprovement` = 0.01, `maxEpochs` = 1000, and `rollingAverageSize` = 5. The regression is unregularized.

[9] However, final classifier's coefficients suggested that this feature did not have a large effect.

[10] We avoided using adjusted rand index (**?**), because it can yield negative values, which is not consistent with purity and NMI. We also computed variation of information (**?**) and normalized information distance (**?**) and observed consistent trends. We omit these results for the sake of space.

Figure 4: Synthetic results on US Congressional Bills and 20 Newsgroups data sets. Topic models help guide annotation attention to diverse segments of the data.

where $\mathbf{U}$ and $\mathbf{G}$ are user and gold clusters, $\mathbb{H}$ is the entropy and $\mathbb{I}$ is mutual information (**?**).

While purity, RI, and NMI are all normalized within $[0, 1]$ (higher is better), they measure different things. Purity measures the intersection between two clusterings, it is sensitive to the number of clusters, and it is not symmetric.

On the other hand, RI and NMI are less sensitive to the number of clusters and are symmetric. RI measures pairwise agreement in contrast to purity that directly measures intersection. Moreover, NMI measures shared information between two clusterings.

None of these metrics are perfect: purity can be exploited by putting each document in its own label, RI does not distinguish separating similar documents with distinct labels from giving dissimilar documents the same label, and NMI's ability to compare different numbers of clusters means that it sometimes gives high scores for clusterings by chance. Given the diverse nature of these metrics, if a labeling does well in all three of them, we can be relatively confident that it is not a degenerate solution that games the system.

5 Synthetic Experiments

Before running a user study, we test our hypothesis that topic model overviews and active learning selection improve final cluster quality compared to standard baselines: list overview and random selection. We simulate the four conditions on Congressional Bills and 20 Newsgroups.

Since we believe annotators create more specific labels compared to the gold labels, we use sub-labels as simulated user labels and labels as gold labels (we give examples of labels and sub-labels in Section 4.1). We start with two randomly selected documents that have different sub-labels, assign the corresponding sub-labels, then add more labels based on each condition's preference function (Section 3.3). We follow the condition's preference function and incrementally add labels until 100 documents have been labeled (100 documents are representative of what a human can label in about an hour). Given these labels, we compute purity, RI, and NMI over time. This procedure is repeated fifteen times (to account for the randomness of initial document selections and the preference functions with randomness).[11]

Synthetic results validate our hypothesis that topic overview and active learning selection can help label a corpus more efficiently (Figure 4). LA shows early gains, but tends to falter eventually compared to both topic overview and topic overview combined with active learning selection (TR and TA).

However, these experiments do not validate ALTO. Not all documents require the same time or effort to label, and active learning focuses on the hardest examples, which may confuse users. Thus, we need to evaluate how effectively actual users annotate a collection's documents.

6 User Study

Following the synthetic experiments, we conduct a user study with forty participants to evaluate ALTO (TA condition) against three alternatives that lack topic overview (LA), active learning selection (TR), or both (LR) (Sections 6.1 and 6.2). Then, we conduct a crowdsourced study to compare the overall effectiveness of the label set generated by the participants in the four conditions (Section 6.3).

6.1 Method

We use the freelance marketplace Upwork to recruit online participants.[12] We require participants to have more than 90% job success on Upwork, English fluency, and US residency. Participants are randomly assigned to one of the four conditions and we recruited ten participants per condition.

Participants completed a demographic questionnaire, viewed a video of task instructions, and then interacted with the system and labeled documents until satisfied with the labels or forty minutes had elapsed.[13] The session ended with a survey, where participants rated mental, physical, and temporal demand, and performance, effort, and frustration on 20-point scales, using questions adapted from the NASA Task Load Index (**?**, TLX). The survey

[11] Synthetic experiment data available at `http://github.com/Pinafore/publications/tree/master/2016_acl_doclabel/data/synthetic_exp`

[12] `http://Upwork.com`

[13] Forty minutes of activity, excluding system time to classify and update documents. Participants nearly exhausted the time: 39.3 average minutes in TA, 38.8 in TR, 40.0 in LA, and 35.9 in LR.

Figure 5: User study results on US Congressional Bills dataset. Active learning selection helps initially, but the combination of active learning selection and topic model overview has highest quality labels by the end of the task.

also included 7-point scales for ease of coming up with labels, usefulness and satisfaction with the system, and—for TR and TA—topic information helpfulness. Each participant was paid fifteen dollars.[14]

For statistical analysis, we primarily use 2×2 (*overview* $\times$ *selection*) ANOVAs with Aligned Rank Transform (**?**, ART), which is a non-parametric alternative to a standard ANOVA that is appropriate when data are not expected to meet the normality assumption of ANOVA.

6.2 Document Cluster Evaluation

We analyze the data by dividing the forty-minute labeling task into five minute intervals. If a participant stopped before the time limit, we consider their final dataset to stay the same for any remaining intervals. Figure 5 shows the measures across study conditions, with similar trends for all three measures.

Topic model overview and active learning both significantly improve final dataset measures. The topic overview and active selection conditions significantly outperform the list overview and random selection, respectively, on the final label quality metrics. Table 2 shows the results of separate 2×2 ANOVAs with ART with each of final purity, RI, and NMI scores. There are significant main effects of *overview* and *selection* on all three metrics; no interaction effects were significant.

TR outperforms LA. Topic models by themselves outperform traditional active learning strategies (Figure 5). LA performed better than LR; while active learning was useful, it was not as useful as the topic model overview (TR and TA).

LA provides an initial benefit. Average purity, NMI and RI were highest with LA for the earliest labeling time intervals. Thus, when time is very limited, using traditional active learning (LA) is preferable to topic overviews; users need time to explore the topics and a subset of documents within them. Table 3 shows the metrics after ten minutes. Separate 2×2 ANOVAs with ART on the means of purity, NMI and RI revealed a significant interaction effect between *overview* and *selection* on mean NMI ($F(1, 36) = 5.58$, p = .024), confirming the early performance trends seen in Figure 5 at least for NMI. No other main or interaction effects were significant, likely due to low statistical power.

[14]User study data available at `http://github.com/Pinafore/publications/tree/master/2016_acl_doclabel/data/user_exp`

	F		p	
	Overview	Selection	Overview	Selection
final purity	81.03	7.18	< .001	.011
final RI	39.89	6.28	< .001	.017
final NMI	70.92	9.87	< .001	.003

df(1,36) for all reported results

Table 2: Results from 2×2 ANOVA with ART analyses on the final purity, RI, and NMI metrics. Only main effects for the factors of *overview* and *selection* are shown; no interaction effects were statistically significant. Topics and active learning both had significant effects on quality scores.

	$M \pm SD\,[median]$		
	purity	RI	NMI
TA	0.31 ± 0.08 [0.32]	0.80 ± 0.05 [0.80]	0.19 ± 0.08 [0.21]
TR	0.32 ± 0.09 [0.31]	0.82 ± 0.04 [0.82]	0.21 ± 0.09 [0.20]
LA	0.35 ± 0.05 [0.35]	0.82 ± 0.04 [0.81]	0.27 ± 0.05 [0.28]
LR	0.31 ± 0.04 [0.31]	0.79 ± 0.04 [0.79]	0.19 ± 0.03 [0.19]

Table 3: Mean, standard deviation, and median purity, RI, and NMI after ten minutes. NMI in particular shows the benefit of LA over other conditions at early time intervals.

Subjective ratings. Table 4 shows the average scores given for the six NASA-TLX questions in different conditions. Separate 2×2 ANOVA with ART for each of the measures revealed only one significant result: participants who used the topic model overview find the task to be significantly less frustrating ($M = 4.2$ and $median = 2$) than those who used the list overview ($M = 7.3$ and $median = 6.5$) on a scale from 1 (low frustration) to 20 (high frustration) ($F(1, 36) = 4.43$, $p = .042$), confirming that the topic overview helps users organize their thoughts and experience less stress during labeling.

Participants in the TA and TR conditions rate topic information to be useful in completing the task ($M = 5.0$ and $median = 5$) on a scale from 1 (not useful at all) to 7 (very useful). Overall, users were positive about their experience with the system. Participants in all conditions rated overall satisfaction with the interface positively

	$M \pm SD\,[median]$					
Condition	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
TA	$9.8 \pm 5.6\,[10]$	$2.9 \pm 3.4\,[2]$	$9 \pm 7.8\,[7]$	$5.5 \pm 5.8\,[1.5]$	$9.4 \pm 6.3\,[10]$	$4.5 \pm 5.5\,[1.5]$
TR	$10.6 \pm 4.5\,[11]$	$2.4 \pm 2.8\,[1]$	$7.4 \pm 4.1\,[9]$	$8.8 \pm 6.1\,[7.5]$	$9.8 \pm 3.7\,[10]$	$3.9 \pm 3.0\,[3.5]$
LA	$9.1 \pm 5.5\,[10]$	$1.7 \pm 1.3\,[1]$	$10.2 \pm 4.8\,[11]$	$8.6 \pm 5.3\,[10]$	$10.7 \pm 6.2\,[12.5]$	$6.7 \pm 5.1\,[5.5]$
LR	$9.8 \pm 6.1\,[10]$	$3.3 \pm 2.9\,[2]$	$9.3 \pm 5.7\,[10]$	$9.4 \pm 5.6\,[10]$	$9.4 \pm 6.2\,[10]$	$7.9 \pm 5.4\,[8]$

Table 4: Mean, standard deviation, and median results from NASA-TLX post-survey. All questions are scaled 1 (low)–20 (high), except performance, which is scaled 1 (good)–20 (poor). Users found topic model overview conditions, TR and TA, to be significantly less frustrating than the list overview conditions.

($M = 5.8$ and $median = 6$) on a scale from 1 (not satisfied at all) to 7 (very satisfied).

Discussion One can argue that using topic overviews for labeling could have a negative effect: users may ignore the document content and focus on topics for labeling. We tried to avoid this issue by making it clear in the instructions that they need to focus on document content and use topics as a guidance. On average, the participants in TR created 1.96 labels per topic and the participants in TA created 2.26 labels per topic. This suggests that participants are going beyond what they see in topics for labeling, at least in the TA condition.

6.3 Label Evaluation Results

Section 6.2 compares clusters of documents in different conditions against the gold clustering but ignores what the labels actually are. To assess how the final induced label sets compare in different conditions, we use crowdsourcing to assess the quality of documents' labels.

For completeness, we also compare labels against a fully automatic labeling method (**?**) that does not require human intervention. We assign *automatic* labels to documents based on their most prominent topic.

We ask users on a crowdsourcing platform to *vote* for the "best" and "worst" label that describes the content of a US congressional bill (we use Crowdflower and require contributors to be in the US).

Five users label each document and we use the aggregated results generated by Crowdflower. The user gets $0.20 for each task.

We randomly choose 200 documents from our dataset (Section 4.1). For each chosen document, we randomly choose a participant from all four conditions (TA, TR, LA, LR). The labels assigned in different conditions and the automatic label of the document's prominent topic construct the candidate labels for the document.[15] Identical labels are merged into one label to avoid showing duplicate labels to users. If a merged label gets a "best" or "worst" vote, we split that vote across all the identical instances.[16] Figure 6 shows the average number of "best" and "worst" votes for each condition and the automatic method. ALTO (TA) receives the most "best" votes and the fewest "worst" votes. LR receive the most worst votes. The automatic labels, interestingly, appear to do at least as well as the list view labels, with a similar number of best votes and fewer worst votes. This indicates that automatic labels have reasonable quality compared to at least some manually generated labels. However, when users are provided with topic model overview, with or without active learning selection, they can generate label sets that improve upon automatic labels and labels assigned without the topic model overview.

Figure 6: Best and worst votes for document labels. Error bars are standard error from bootstrap sample. ALTO (TA) gets the most best votes and the fewest worst votes.

7 Related Work

Text classification—a ubiquitous machine learning tool for classifying text (**?**)—requires feature extraction (**?**). These are both well-trodden areas of NLP research. The difficulty is often creating the training data (**?**; **?**); coding theory is an entire subfield of social science devoted to creating, formulating, and applying labels to text data (**?**; **?**). Crowdsourcing (**?**) and active learning (**?**), can de-

[15] Some participants had typos in the labels. We corrected all the typos using pyEnchant (`http://pythonhosted.org/pyenchant/`) spellchecker. If the corrected label was still wrong, we corrected it manually.

[16] Evaluation data available at `http://github.com/Pinafore/publications/tree/master/2016_acl_doclabel/data/label_eval`

crease the cost of annotation but only *after* a label set exists.

ALTO quantitatively shows that corpus overviews aid text understanding, building on traditional interfaces for gaining both local and global information (**?**). More elaborate interfaces (**?**; **?**) provide richer information given a fixed topic model. Alternatively, because topic models are imperfect (**?**), refining underlying topic models may also improve users' understanding of a corpus (**?**; **?**).

Summarizing document collections through discovered topics can happen through raw topics labeled manually by users (**?**), automatically (**?**), or by learning a mapping from labels to topics (**?**). When there is not a direct correspondence between topics and labels, classifiers learn a mapping (**?**; **?**; **?**). Because we want topics to be consistent between users, we use a classifier with static topics in ALTO. ALTO combines active learning with topic models to provide both global and local knowledge document labeling requires.

8 Conclusion and Future Work

We introduce ALTO, an interactive framework that combines active learning *selection*s with topic model *overview*s to help users induce a label set and label documents. We show that users can more effectively and efficiently induce label set and create training data using ALTO in comparison with other conditions, which lack either topic *overview* or active *selection*.

We can further improve ALTO (the TA condition) to help users gain better and faster understanding of text corpora. Our current system limits users to view only $20K$ documents at a time and allows for one label assignment per document. Moreover, the topics are static and do not adapt to better reflect users' labels. Users should have better support for browsing documents and assigning multiple labels. Topics can also improve via SLDA (**?**) or LLDA (**?**) as users add labels.

Finally, with slight changes to what the system considers a document, we believe ALTO can be extended to NLP applications other than classification, such as named entity recognition or semantic role labeling, to reduce the annotation effort.

Acknowledgments

We thank the anonymous reviewers, Philip Resnik, and Burr Settles for their insightful comments. We also thank Nikolaos Aletras for providing the automatic topic labeling code. Boyd-Graber and Poursabzi-Sangdeh's contribution is supported by NSF Grant NCSE-1422492; Findlater, Seppi, and Boyd-Graber's contribution is supported by collaborative NSF Grant IIS-1409287 (UMD) and IIS-1409739 (BYU). Any opinions, findings, results, or recommendations expressed here are of the authors and do not necessarily reflect the view of the sponsor.