NIST Technical Note
NIST TN 2287

Human-in-the-loop Technical Document Annotation

Developing and Validating a System to Provide Machine-Assistance for Domain-Specific Text Analysis

Juan F. Fung
Zongxia Li
Daniel Kofi Stephens
Andrew Mao
Pranav Goel
Emily Walpole
Alden Dima
Jordan Lee Boyd-Graber

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2287

NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Technical Note
NIST TN 2287

Human-in-the-loop Technical Document Annotation

Developing and Validating a System to Provide Machine-Assistance for Domain-Specific Text Analysis

Juan F. Fung
Applied Economics Office
Engineering Laboratory

Zongxia Li, Andrew Mao,
Pranav Goel, Jordan Lee Boyd-Graber
University of Maryland
College Park, MD

Daniel Kofi Stephens
Morgan State University
Baltimore, MD

Emily Walpole*
**Former NIST employee; all work for this publication was done while at NIST*

Alden Dima
Information Systems Group
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2287
May 2024

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Publication History
Approved by the NIST Editorial Review Board on 2024-05-06

How to cite this NIST Technical Series Publication:
Fung JF, Li Z, Stephens DK, Mao A, Goel P, Walpole E, Dima A, Boyd-Graber JL (2024) Human-in-the-loop Technical Document Annotation. (National Institute of Standards and Technology, Gaithersburg, MD), NIST TN 2287. https://doi.org/10.6028/NIST.TN.2287

Author ORCID iDs
Juan F. Fung: 0000-0002-0820-787X
Zongxia Li: 0009-0001-1437-5132
Daniel Kofi Stephens: 0009-0007-3122-7004
Andrew Mao: 0009-0002-0344-9973
Pranav Goel: 0000-0003-1037-2687
Emily Walpole: 0000-0003-0401-1535
Alden Dima: 0000-0003-0547-3117
Jordan Lee Boyd-Graber: 0000-0002-7770-4431

Contact Information
juan.fung@nist.gov

Abstract

In this report, we address the following question: to what extent can machine learning assist a human with traditional text analysis, such as content analysis or grounded theory in the social sciences? In practice, such tasks require humans to review and categorize (e.g., by manually annotating the text with labels) a large sample of documents. We do not expect nor necessarily desire the machine to automate the tasks the human would otherwise perform, but rather want to find ways to help the human to perform the tasks more efficiently. We present a modular implementation of a system that incorporates supervised (active learning) and unsupervised (topic modeling) methods to assist humans with technical document annotation. The implemented system allows us to conduct user studies to evaluate the usefulness of machine assistance. We present results from two such user studies and highlight directions for future research.

Keywords

Text analysis; text mining; natural language processing; machine learning; human-in-the-loop; social science; content analysis.

Acknowledgments

We are grateful to Christina Gore (NIST), Timothy Blattner (NIST), David Butry (NIST), and Nicos Martys (NIST) for their review and feedback. All errors are our own.

Table of Contents

List of Tables

List of Figures

1. Introduction

Not all data is numerical. Unstructured text in a variety of media potentially contain hidden treasure troves of information. This observation is obvious across a broad array of domains and applications, including the analysis of maintenance work orders to improve manufacturing operations;[2] analysis of legal documents and case-law decisions;[3] and the analysis of open-ended survey responses.[4] The common problem across these and many similar applications is how best to transform the text found "in the wild" into structured data for analysis.

While the information age has resulted in an explosion of unstructured text, the analysis of text as data is not new and there exist many well-established methods in the social sciences for this task. Content analysis is the study of documents to examine patterns in media in a replicable and systematic manner.[5] The process gives meaning to unstructured documents, such as text or pictures, and allows researchers to quantify and analyze the presence, meanings, and relationships of certain words, themes, or concepts. As such, content analysis is often described as deductive, beginning with a set of research questions that guide the extraction of information from text. Grounded theory shares many similarities to content analysis, but is inductive. Grounded theory represents a set of methods for iteratively developing a theory from the source material to provide a meaningful representation of the observed patterns.[6]

What such methods have in common is that they are labor intensive and, therefore, very time consuming. This presents a limiting factor in the modern era where the availability of text as data is potentially unbounded, as existing social science methods are not scalable. On the other hand, a key advantage of such methods is that they embed domain expertise into the analysis in order to derive meaningful insights from the text.

It is reasonable to wonder whether computational methods offer an advantage over traditional social science methods. A key takeaway from the computational social science literature is that computational methods initially offered the promise of scalability, but did not provide an "objective" or even transparent alternative to traditional social science methods. This led many social science researchers to either adopt the use of "off-the-shelf" methods (for instance, topic modeling with Latent Dirichlet Allocation, or LDA, using default parameterizations) or else to remain skeptical of computational methods.[7–9]

There has been an increasing recognition among both social scientists and computer scientists that the key to computational methods is to blend them with social science methods. In other words, we should not expect that machine learning can replace a human but rather we should strive to use machine learning to assist a human in domain-specific tasks.[8] This paradigm shift is characterized as human-in-the-loop, or mixed-initiative, Natural Language Processing (NLP),[10] or sometimes as Technical Language Processing (TLP).[2]

In this report, we address the following question: to what extent can machine learning assist a human with traditional text analysis, such as content analysis or grounded theory? In practice, such tasks require humans to review and categorize (e.g., by manually annotating the text with labels) a large sample of documents. We do not expect nor necessarily desire the machine to automate the tasks the human would otherwise perform, but rather want to find ways to help the human to perform the tasks more efficiently (with respect to time spent on annotation).[1]

It is worth noting that in terms of human-in-the-loop frameworks, machine assistance can come in many forms, including: curating raw data; processing raw data for computation; labeling raw data; and analyzing data.[10] While document labeling is but one task, we focus on it both to illustrate how machine learning can assist humans as well as because it is an integral task to many social science methods, including content analysis and grounded theory. Thus, it is a realistic case study for human-in-the-loop social science research.

The report is structured as follow. In Sec. 2, we discuss the motivating example of analyzing community planning documents within the domains of Resilience, Climate Adaptation, and Sustainability (RAS).[11] In particular, researchers at NIST studied a corpus of guidance documents aimed at providing communities with best practices for developing particular planning documents. The goal was to explore similarities and differences across different guidance documents and the approach used content analysis. As we discuss, this approach is very labor intensive. Framing the problem as one of knowledge extraction, we explore to what extent a human-in-the-loop approach can be used for this problem.

In Sec. 3, we review the literature on applications of human-in-the-loop computational social science as well as applications of text mining and natural language processing (NLP) to the domain for our motivating example (community planning). We then summarize takeaways for the generic problem of knowledge extraction from technical text as well as for our motivating example of community planning documents.

In Sec. 4, we present the methodology we developed for the generic problem of extracting knowledge from technical documents. Specifically,

- We use supervised and unsupervised methods, combining active learning with topic models to suggest labels for clusters and recommend documents for labeling
- We develop a modular human-in-the-loop system that allows us to “plug in“ machine learning methods for testing and validation (through user studies)

One of the key contributions of this research is demonstrating the value of human-centered model evaluation relative to relying on standard “automated” evaluation metrics. Our results show how human-in-the-loop task-based evaluation can provide a more complete picture of the value of different topic models for assisting with text analysis tasks such as document annotation for content analysis.

[1] This integration is variously characterized as human-computer interaction (HCI), human-in-the-loop, or human-machine teaming and requires broader interdisciplinary collaboration,[10] as discussed in Sec. 3.

In Sec. 5, we return to our motivating example of analyzing community planning documents. We present results from a mini-study to validate the results from Sec. 4. In particular, a set of experts in community resilience use the tool on “real data” (the RAS data from the motivating example content analysis). The validation study focuses on two competing models and the results are consistent with the user study results, which suggests that the results from Sec. 4 generalize to “real data.” Moreover, we review expert feedback on the tool deployed for the user study, as well as an efficiency gains (in terms of time to label documents) relative to manually hand-annotating text for content analysis.

Finally, Sec. 6 outlines gaps and next steps for future research. We note that the advent of large language models (LLMs) has changed the landscape of machine assistance, especially for text analysis. As the field is rapidly evolving, we are working on evaluating LLMs to support humans in the analysis of technical text. Finally, additional details on the implementation of a human-in-the-loop interface for user studies is provided in Appendix A.

1.1. Terminology and definitions

Before we proceed, it is important to discuss the range of terminology across disciplines and to provide consistent terminology and definitions for what follows. One of the key differences in terminology is the use of the word “document.” For instance, an urban planner studying a community’s climate adaptation plan would consider the entire (potentially hundred-pages long) plan as a “document.” In contrast, computational approaches to text analysis often refer to the unit of analysis as a “document.” For instance, if the unit of analysis is a paragraph, then the hundred-pages long climate adaptation plan is broken down to a collection of paragraphs and each paragraph would be a “document.”

Our goal is not to debate the merits of any particular use of terminology, but rather to select a clear set of terms and to be consistent in our use throughout the report, in order to avoid confusion. In the process, we highlight potential terms that may create confusion or introduce ambiguity across domains. Future research might look to establishing consistent terminology to enable interdisciplinary collaboration.

The following terms are presented in order of complexity, to the extent possible, rather than alphabetically. In addition to the key terminology listed below, acronyms are defined in Appendix B.

Text Written human (natural) language in structured or unstructured form.

Document Text unit of analysis (e.g., sentence, paragraph, social media post).

Note 1: Alternatives include **segment**, **chunk**, or **example**.

Note 2: Not to be confused with the unit of observation in text analysis, which is often ambiguous. Tidy text principles suggest that the unit of observation (e.g., the row in a data frame) is a term (e.g., word, n-gram, or sentence).[12]

Source Original source of document example (e.g., essay, community plan, webpage).

Corpus Collection of documents.

Sample A subset of a corpus.

Label Metadata assigned to document to provide context, meaning, or other information (e.g., "cat" or "dog").

Note: Alternative terms include **code** or **annotation**.

Annotation The process of adding a label to a document.

Note: Alternatively referred to as "coding" in social science.

Annotator The human conducting document annotation.

Note: Alternatively referred to as "coder" in social science.

Term Document-level unit of analysis (e.g., word, n-gram, or sentence) used to build a lexicon (or vocabulary) for a corpus.

Cluster A group of documents that are similar with respect to some metric.

Topic A cluster of documents that share latent semantic structure (e.g., similar terms).

Theme Often used interchangeably with topic, but refers to latent semantic structure; typically, a human will attach meaning to a topic to identify the underlying theme.

Keyword Relevant terms in documents that tend to co-occur within a topic.

2. Motivating Example: RAS

As a motivating example, consider the process of planning for resilience to natural hazards within a community. Communities facing natural hazard risk are often concerned with other risks as well, including sea-level rise, resource depletion, and pandemics. Thus, community resilience planning occurs in the context of multiple, potentially competing objectives, such as adaptation, sustainability, and public health. However, much of the existing guidance on community resilience planning is formulated independently of other objectives that compete for a community's limited resources.

Clavin et al. [13] examine the literature on the often technical and administratively complicated planning processes undertaken to address community objectives, focusing on planning for resilience, climate adaptation, and sustainability (RAS). [2] Their review is focused on developing a methodology for assessing the quality of resulting plans. In the process, the authors highlight areas of overlap, including similarities and differences in terminology, as well as opportunities for synergy across planning processes.

Clavin et al. [17] take this a step further by conducting a rigorous content analysis of RAS planning *guidance*, which are typically reports produced by government or non-governmental organization with the purpose of providing communities with a "recipe" for conducting the planning process. Figure 8 illustrates the steps in the RAS annotation workflow for conducting content analysis.

[2] Resilience is the ability to prepare for anticipated hazards, adapt to changing conditions, and withstand and recover rapidly from disruptions.[14] Adaptation is the process of adjustment to actual or expected climate and its effects to moderate harm or exploit beneficial opportunities.[15] Sustainability refers to "development that meets the needs of the present without compromising the ability of future generations to meet their own needs.[16]

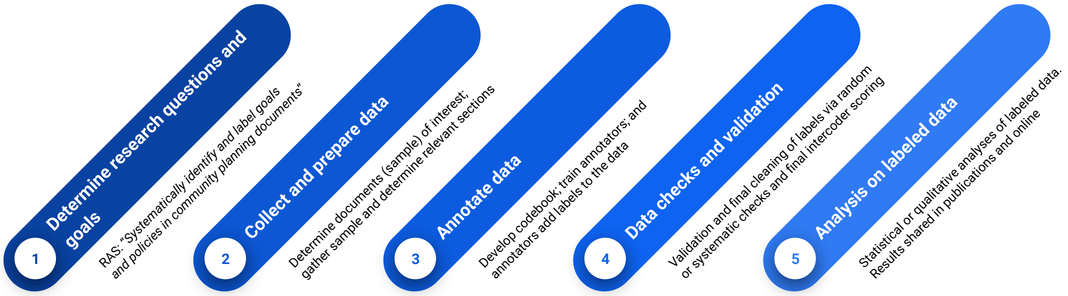

1
Determine research questions and goals
RAS: *"Systematically identify and label goals and policies in community planning documents"*
2
Collect and prepare data
Determine documents (sample) of interest; gather sample and determine relevant sections
3
Annotate data
Develop codebook; train annotators; and annotators add labels to the data
4
Data checks and validation
Validation and final cleaning of labels via random or systematic checks and final intercoder scoring
5
Analysis on labeled data
Statistical or qualitative analyses of labeled data. Results shared in publications and online

Fig. 1. Example content analysis for RAS

As shown in Fig. 1, the primary research goal is to "systematically identify and label goals and policies in community plans." In particular, one of the research questions is to what extent planning *guidance* can provide insight into the themes generally found in actual community plans. The RAS content analysis allows the authors to quantify, in terms of term frequencies, areas of similarities and differences across the different objectives that might be addressed by resilience, adaptation, and sustainability planning. In the process, the authors carefully developed and iterated on a codebook that was used to annotate the planning guidance documents chosen for the study. The result is a rigorously validated, hand-annotated corpus specific to the domain of RAS planning guidance.

Of course, this requires significant human labor, and in our RAS example, estimated in total to be around five months of annotation, review, and validating labels through intercoder checks. In particular, Fig. 2 shows that Step 3 in Fig. 1 is not a linear process. "Annotate data" is an iterative process that includes: development as well as updating of the codebook used for annotation, based on periodic check-ins with annotators, in addition to training and potential re-training of annotators.

Moreover, data checking and validation in Step 4 is almost as time consuming as the initial annotation process itself: four annotators spent three months adding labels to the documents, and roughly two additional months on validation. The final, complete corpus consists of 2625 labeled documents, representing roughly 25 source planning guidance reports.

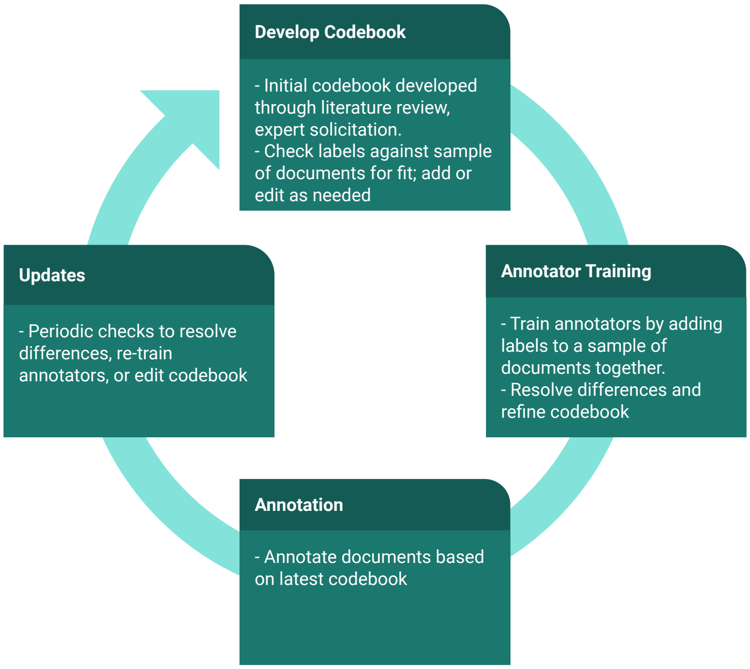

Develop Codebook
- Initial codebook developed through literature review, expert solicitation.
- Check labels against sample of documents for fit; add or edit as needed
Annotator Training
- Train annotators by adding labels to a sample of documents together.
- Resolve differences and refine codebook
Annotation
- Annotate documents based on latest codebook
Updates
- Periodic checks to resolve differences, re-train annotators, or edit codebook

Fig. 2. Annotation itself is an iterative process.

2.1. Scaling?

Recall that the hand-annotated corpus corresponds to planning *guidance*. It would be straightforward, though labor-intensive, to hand annotate additional passages of text from similar source guidance.

But what about community plans themselves? Ultimately, the goal of these studies is to extract actionable intelligence for communities planning for RAS. Specifically, if we look at actual community plans for existing approaches to RAS objectives, what can communities learn in the future for their own planning processes? For instance, for streamlining planning across resilience, adaptation, and sustainability.

Moreover, consider that communities conduct planning processes for many other objectives. In addition to resilience, sustainability, and climate-adaptation plans, communities often develop other plans. Table 1 presents examples of community plan types. These range from very high level plans such as comprehensive plans that almost all communities develop from time to time, to more specialized plans such as small area or neighborhood plans that tend to be tailored to the community.

Table 1. Examples of plan types and their purpose in a community. Source: "Plan Integration for Resilience Scorecard Guidebook."[1]

Plan Type	Purpose
Comprehensive/General Plan	Main community planning document
Hazard Mitigation Plan	Reduce long-term risk to human life and infrastructure
Disaster Recovery Plan	Address disaster recovery related needs
Area Plans: • Downtown (Redevelopment) • Small Area/Neighborhood/District • Waterfront • Corridor Plan	Address planning issues pertaining to a portion of the community
Functional or Sector-specific Plans: • Transportation/Transit • Parks/Open Space • Economic Development • Environmental Management • Resilience • Sustainability • Climate Adaptation and/or Mitigation • Housing (Consolidated/Strategic) • Wildlife Management • Wildfire Protection	Focus on individual or related functions or sectors in need of specialized planning

Scaling the content analysis workflow in Fig. 1 for each of these scenarios introduces increasing complexity, in addition to requiring significant time and resources.

2.2. Summary of the problem

The RAS content analysis is an example of the more generic problem of transforming text found "in the wild" into structured data for analysis, as described in Sec. 1. Community plans, or even planning guidance reports, do not adhere to a particular format, so there is a great deal of variation in the amount of information as well as in how that information is presented across plans—even within a single community.

Annotating source text is one way to organize unstructured data into a meaningful representation that is suitable for analysis. Moreover, human expertise is critical to the organization, as the source text is often technical and domain-specific. If we want a large quantity of text to analyze, we need more efficient tools than manual human annotation.

As this example highlights, and the literature review in Sec. 3 shows, the ultimate goal is not to fully automate such tasks but rather to incorporate machine learning methods in a process that remains human-centered in order to leverage technical expertise.

The question we address in this report is whether we can use machine learning to assist humans in research tasks such as content analysis. In particular, we wish to reduce the time required to annotate text while maintaining confidence in the resulting labels.

3. Literature

Recall the motivating question we posed in Sec. 1: to what extent can machine learning assist a human with traditional text analysis, such as content analysis or grounded theory? In particular, to what extent can machine learning assist humans in efficiently and effectively reviewing, labeling, categorizing, and analyzing a large sample of technical documents.

We first review the literature on computational methods in social science, which tackles such tasks. We then review applications of machine learning specific to the domains of our motivating example (Sec. 2): resilience, climate adaptation, and sustainability (RAS) planning.

3.1. Computational methods in social science

As we alluded to in Sec. 1, the analysis of text is a typical problem encountered in the social sciences. For instance, political scientists that want to compare political candidate ideologies, based on speeches they give; psychologists that are interested in mapping patient intake forms to latent personality types; and communications researchers examining trends in public discourse before and after a disruptive event through the analysis of social media posts. The common thread is that there is a wealth of data embedded in a broad variety of texts and the ultimate problem is how to operationalize that text into data suited to answering any number of research questions. The literature offers several potential options for integrating computational methods into social science research.

DiMaggio [18] compares approaches to text analysis in social science and computer science, while noting the differences between the fields is not massive. One key difference is what DiMaggio [18] terms “social scientists’ unhealthy obsession” with model-based inference, as opposed to computer science’s focus on model utility for specified tasks.[3] The result is a relatively higher prioritization of corpus curation in social science. On the other hand, DiMaggio points out that computer scientists tend to place more trust in human judgment as the “gold standard” for benchmarking machine learning algorithms. In contrast, social scientists are much more aware of and comfortable with the fact that humans are prone to systematic biases and logical fallacies. One thing both fields have in common, DiMaggio argues, is that both humans and algorithms are equally bad at nuance. These similarities and differences suggest potential common ground for collaboration on cognitively similar tasks such as document annotation. Moreover, as DiMaggio points out: “topic-modeling programs require lots of decisions that most social scientists are ill-equipped to make,” which argues for an interdisciplinary, human-in-the-loop approach as we propose in this report.

[3] This contrast was famously observed more generally between “model-based” statistics and “algorithmic” machine learning by Leo Breiman in 2001 [19].

Within political science, Grimmer and Stewart [7] may be the definitive introductory review of computational methods for content analysis.[4] Grimmer and Stewart [7] begin with an overview of computational methods for automated text analysis, covering supervised and unsupervised learning, scaling methods, and validation (semantic, including predictive, and convergent validity). In the process, Grimmer and Stewart [7] address common misconceptions, errors, and potential pitfalls (e.g., the "best model" is typically application-specific). Given the breadth of methods and potential for misuse, the authors provide four guiding principles as a framework for automatic content analysis, the most relevant of which is that computational methods for "text amplify resources and augment humans."[5] Finally, it is worth noting that much has changed since 2013: while Grimmer and Stewart [7] conclude by urging political scientists to contribute rather just consume algorithms, this has already happened in political science specifically as well as across many fields in the social sciences more broadly.[6]

Lucas et al. [23] review supervised, unsupervised, and scaling methods for analyzing text in multiple languages within comparative politics, for the analysis of blogs, newspaper articles, speeches, deliberations, and political-party manifestos.[7] Multilingual text analysis presents unique challenges, most importantly validating the alignment of topics across languages, and the authors propose can be addressed by using Structured Topic Models (STM) to incorporate metadata as document-level covariates to pool information and compare across documents.[8] Lucas et al. [23] illustrate the use of STMs with a case study of Arabic and Chinese text. The key takeaway is that "automated text analysis... amplifies human effort," similarly to Grimmer and Stewart.[7]

Wilkerson and Casas [24] discuss the use of text as data in political science research, delineating four areas of advancement and opportunity for research: classification, scaling, text reuse (which focuses on "discovering instances of similar language usage"), and using syntax and semantics to understand political relationships. Given the extensive use of topic models in political science, Wilkerson and Casas [24] discuss the particular problem of topic instability (where multiple runs of the same topic model may result in topics with little overlap) and the challenge of validation in the absence of gold labels across many applications. Moreover, typical validation is often focused on a single model, with varying numbers of topics, rather than exploring results from multiple models. The methodol-

[4] Based on 3733 citations on Google Scholar as of 2024-04-15: https://scholar.google.com/scholar?lookup=0&q=grimmer+stewart+2013&hl=en&as_sdt=0,21

[5] The first is a restatement of George Box's famous quip that "all models are wrong, but some are useful." The third stresses that the best model is context and application specific, whereas social scientists tend to think in terms of "global" best. The final principle highlights the importance of validation.

[6] See for instance the use of Exploratory Graph Analysis (EGA) in psychology [20, 21] and the development of metrics to measure partisanship in congressional speeches in economics [22].

[7] In particular, the Manifesto Project, which supports "policy preferences covers over 1000 parties from 1945 until today in over 50 countries on five continents:" https://manifesto-project.wzb.eu/

[8] In addition, multilingual analysis presents preprocessing challenges, from integrating many potential encodings of the source text to translating the raw text to a single language.

ogy we present in this report approaches text analysis problems agnostic to the model, with the goal of allowing users to “plug and play“ into a modular system. Moreover, the user study we present in Sec. 4.3 is focused specifically on validation across different topic models. Finally, a human-in-the-loop system as we present in this report can support an additional level of quality control by allowing users to hand-annotate a small set of documents and then validate the output with machine assistance against such gold labels. We demonstrate this in an additional validation study we conduct using the gold labels from our motivating example in Sec. 2.

In the domain of communications, Zamith and Lewis [25] compare traditional content analysis (which they call the “human coder”) and computational approaches (which they call the “algorithmic coder”). Much of what Zamith and Lewis [25] discuss involves leveraging computational methods to curate or parse large data sets down to something a human can easily and quickly process/validate on a second pass. They propose a “hybrid” approach to blend the best of both human and algorithmic coders: in other words, to leverage human expertise “contextual sensitivity and validity” and the scalability and reliability of computational approaches. The hybrid approach of Zamith and Lewis [25] sounds a lot like the human-in-the-loop approach we propose in this report and their paper provides some justification for using human-in-the-loop for domain-specific, technical documents.

Similarly, Boumans and Trilling [26] review the literature on automated content analysis in journalism and other domains, as well as methods used for digital journalism. The recommendations for conducting machine-assisted content analysis include: interdisciplinary teams; code literacy; developing new methods and tools (echoing Grimmer and Stewart [7]; and, importantly, sharing code.

Baumer et al. [8] compare grounded theory and topic modeling from the lens of communications and information science. Using a motivating case study of social media behavior, they explore similarities and differences in analyzing survey responses, highlight tradeoffs in choosing between methods, and make recommendations for combining the two approaches (a “mixed-methods” approach). The main similarity (“convergence”) is a correspondence between grounded theory themes and topic model results (i.e., the topics). As Baumer et al. [8] observe, both facilitate meaningful discourse on the data and while terminology may differ, both follow a similar analytical process that requires a significant amount of expert judgment. On the other hand, the main difference (“divergence”) is that a topic model can align pretty well with grounded theory themes even without the use of contextual information (as the authors state, the “model does not need to understand”). More obviously, qualitative methods such as grounded theory require much greater time commitment.

Building on these convergences and divergences, Baumer et al. [8] provide a framework for human-in-the-loop text analysis. The authors see computational methods as complementing human-driven qualitative research. Thus, topic models should be used to facilitate document organization and prioritization by thematic coherence, which is the approach we propose in this report (see Sec. 4). Muller et al. [27] further explore convergence in approaches (both are iterative and require human input) and potential hybrid approaches. An open question posed is what metrics are best suited for "detecting concepts, themes, or constructs grounded in text."

In light of increasing interest and use of Natural Language Processing (NLP) in cognitive science, Lake and Murphy [9] compare human and algorithmic approaches to psychological semantics (i.e., how humans represent the meaning of words and how humans interpret language based on those mental models). While NLP shows great promise as a theory of psychological semantics, Lake and Murphy [9] observe a few key shortcomings that must be addressed. In particular, they find that model performance is often strongly tied to existing corpora. This works well for tasks such as word similarity but less so for domain-specific meaning, e.g., psychological judgement. The authors argue for the development of more complex word representations, potentially through multimodal models such as visual-language models that receive human feedback.

Kjell et al. [28] explore the use of NLP to measure psychological constructs based on open-ended survey responses. Psychological constructs (such as attitudes and emotions) are typically measured based on close-ended responses, often on a Likert scale. However, as Kjell et al. [28] observe, respondents can encode deep psychological meaning in qualitative responses. Traditional NLP methods do not work "off the shelf" for such a technical text. Thus, the authors propose the use of semantic measures to "statistically measure, differentiate and describe subjective psychological states." The authors benchmark this approach with traditional methods and find comparable validity. While not explicitly human-in-the-loop, developing semantic measures requires domain expertise to conduct latent semantic analysis (that is, it requires significant human expertise).

3.2. Domain-specific applications: RAS

We now turn to the domain of our motivating example: resilience, climate adaptation, and sustainability (RAS). While the literature in these domains is nascent and promising, many of the applications are "mechanical" uses of computational methods (e.g., off the shelf topic models and word embeddings, as well as classification). The preceding literature review offers a potential roadmap for future research in RAS that embraces a more human-centric approach as we propose in this report.

Biesbroek et al. [29] train a neural network to classify text passages (or "blocks") from policy documents in order to predict if a passage is about 'Adaptation', 'Mitigation', or 'Non-climate.' The authors collected and scraped text from documents tagged "policy papers" in the United Kingdom's collection of 'gov.uk' Ministerial websites, hand-selecting a small subset pertaining to each category to annotate (10 each for Adaptation and Mitigation, 20 for Non-climate). Each block roughly corresponded to a paragraph, with some additional pre-processing and a maximum block length of 200 words. The result was a corpus consisting of nearly 14 000 blocks of text, with about 3800 corresponding to Adaptation and 5500 to Mitigation. The ultimate trained model performed fairly well at classification, based on the usual metrics (precision and recall) as well as expert review, and the authors demonstrate potential use to scale the size of the labeled corpus.

Berrang-Ford et al. [30] conducted a machine-assisted literature review of the academic climate adaptation literature. Specifically, the authors systematically collected over 48 000 articles through bibliometric search to learn what adaptation responses are observed in practice. The authors adopt a process that combines machine learning with human expertise in order to sift through a large amount of technical text. First, the authors developed predetermined inclusion criteria to screen articles for further study. Second, the authors used a classifier to assist with screening with respect to the inclusion criteria. This step required a team of four researchers to manually screen and annotate a small subset of articles based on the criteria. The hand-annotated articles were then used to train the classifier to predict whether an article is relevant to the study. The result was about 2000 articles selected for further study by teams of humans to confirm whether the article satisfied the inclusion criteria as well as to draw further insights on adaptation practices. The final corpus consisted of almost 1700 articles that allowed the authors to not only assess the “global stocktake” of implemented adaptation but to also identify priorities for future research.

Brinkley and Stahmer [31] use topic modeling on a corpus of 461 California planning documents. The goal is to identify common themes across communities through “automated” document annotation and thus the task is closer to the one we are concerned with in this report. Unsurprisingly, while the authors talk about “automating” content analysis, the approach implemented in the paper requires significant human input. In that sense, Brikley and Stahmer [31] implicitly leverage a human-in-the-loop approach to provide domain expertise. Interestingly, the authors use an “off-the-shelf” package for LDA to cluster documents by themes, but use the underlying plan’s geographic information to plot variation in themes across communities.

Finally, Saheb et al. [32] review the academic literature on sustainable artificial intelligence, AI, which they define as the use of AI in support of sustainability and sustainable energy goals. In particular, “the term sustainable AI describes how AI can be used to enable societies to achieve their sustainability goals.” The authors combine clustering, LDA, and a contextual topic model (based on BERT) with a cluster-based content analysis to identify eight themes in the field, which highlight research gaps and lead to recommendations for further research (ie, based on expert human judgment of the topics). The authors are able to explore each of the discovered topics, associated topics, and evolution of topics over time. The ‘mixed method” approach taken by Saheb et al. [32] is a hybrid computational approach that fundamentally relies on human expertise for the content analysis.

4. Methodology

In this section, we present the methodology we developed for the generic problem of extracting knowledge from technical documents. Specifically, we describe:

1. Sec. 4.1: Machine learning methods for assisting humans in the document labeling problem described in the motivating example (Sec. 2).
2. Sec. 4.2: A human-in-the-loop system developed to implement machine-assistance for the use of document labeling
3. Sec. 4.3: User study results to validate the use of different models for machine-assisted document labeling.

4.1. Machine Learning Techniques

First, we describe the machine learning methods we consider for the task of document annotation. In particular, we summarize theoretical, simulation, and user study results comparing the use of various topic models in a system that provides active learning as a baseline.[33]

The methods combine supervised and unsupervised learning as follows:

- Unsupervised: Topic models automatically group documents thematically and present potential labels to the user
- Supervised: Active learning takes user labels as feedback and provides recommendations on similar documents for labeling

The value of using active learning with topic models has been shown by Poursabzi et al. [34], in a system they call Active Learning with Topic Overviews (ALTO). Thus, we focus on comparing different classes of topic models under the assumption that active learning is valuable for the user's task.

4.1.1. Topic Modeling

As the literature review in Sec. 3 shows, topic modeling has been a popular machine technique for conducting text analysis in the social sciences for instance, for semi-automated content analysis or machine-assisted grounded theory. The classical and most popular topic model technique is Latent Dirichlet Allocation (LDA), which has over 50 000 citations.[35] Figure 3 shows a simple example of automatic document clustering using topic models, which can allow a social scientist to quickly grasp the most popular topics within a massive number of documents.

When conducting content analysis, a researcher can develop a codebook of themes (and consequently labels) for investigation within a corpus.[5] Given a large set of unstructured documents, a topic model can automatically discover the latent semantic themes for the

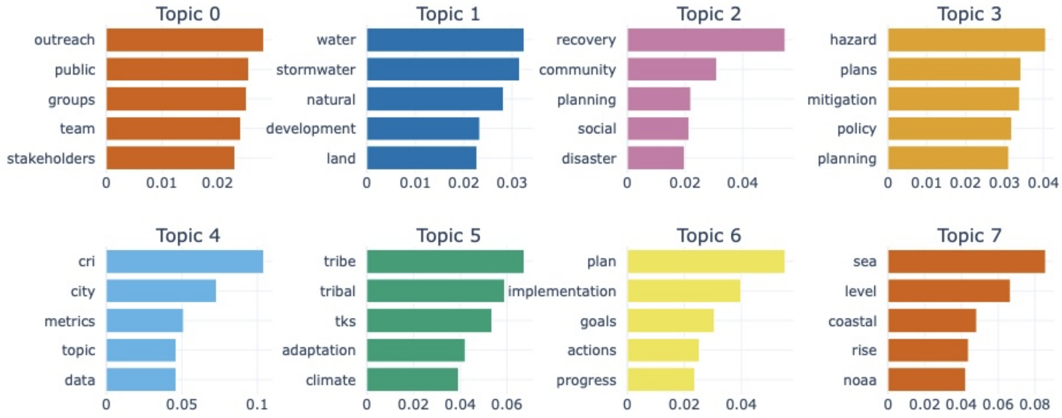


Fig. 3. Example illustrating how a topic model can automatically group a set of community resilience documents with a pre-defined number of topics. Topic models automatically group similar documents together and suggest a common theme for group of documents.

documents. The researcher can then explore the topics for alignment with the hypotheses (the initial codebook developed before looking at the data) and potentially use the emergent themes from the topic model to update the codebook in an iterative process.

Grounded theory treats annotation as an iterative and complicated process where human annotators have to wander around a massive number of unstructured documents and assign them to human created categories.[8] Topic models can assist with the initial annotation process by providing annotators a global overview of potent categories in a set of documents.

4.1.2. Topic Model Variations

In this section, we introduce various topic models and compare their differences and advantages. Different topic models have quite different ways of automatically clustering documents, and the generated topics are quite diverse for different topic modeling methods.

The most popular topic model is classical LDA, first introduced by Blei, Ng, and Jordan [35] in 2003. The model treats each document as a distribution of words, and uses probabilistic Bayesian inference to automatically cluster documents based on their word distributions. Classical LDA is a simple algorithm that requires less computational resources to run relative to more modern topic models.

A variation of LDA is supervised LDA (sLDA).[36] Unlike classical LDA, sLDA requires a set of known response variables in advance prior to automatic clustering. Suppose we have a set of movie reviews with star ratings, and we want to analyze what people talk most about for each level of star rating. In this case, LDA cannot directly relate the generated topics with star ratings, and it can only generate very general topics and review clustering where people talk most about for all star levels. However, sLDA can generate topics that correlate with the star levels, where the generated topics are more representative to a particular star level and the reviews are clustered more based on similar star ratings. Thus, sLDA can effectively incorporate response variables to generate topics that are more representative of the corpus as a whole. In the case of our motivating example, the response variables are the annotator inputs.

With the resurgence of neural networks, neural topic models (NTM) have become a quite popular research direction for computer science researchers. NTMs take advantage of pre-trained neural networks, where the network is trained on a large amount of text corpora and have already learned mass word relationships and associations from the corpora. NTMs take the pre-trained network's embedding that contains rich contextual information to encode a set of documents and perform automatic topic model clustering on them. Different NTMs have quite different pipelines to cluster documents and generate topics, and the quality of generated topics are quite diverse across different NTMs. Current popular NTMs include the contextualized topic model (CTM),[37] embedded topic model (ETM,)[38] and Bertopic.[39]

4.1.3. Active Learning

Merely having a set of topics and themes is not sufficient for a researcher to conduct thorough data analysis on a corpus of technical text. For instance, social scientists using grounded theory are interested in exploring on the contents of specific documents to come up with labels for the documents. To support more efficient exploration of topic model outputs, we use active learning. Active learning is a machine learning technique to continuously guide users' attention to individual documents based on users' previously input labels.[34]

To illustrate how active learning can support data analysis, consider the following idealized document annotation process. Suppose we have a set of documents without any categories, titles, or labels. Mimno et al. [40] define the annotation (or "coding") process as attaching meaning to the set of documents by assigning documents into groups or clusters, where each cluster is assigned a particular meaning, also referred to as a 'label' (or 'code').

In this idealized annotation process, traditional approaches to assigning labels are labor-intensive and complicated. Researchers or annotators usually assign labels to documents by manually going through each document in the corpus, either in order or randomly, as illustrated in our motivating RAS example in Sec. 2. In addition, due to the complexity and diversity of annotator-selected documents, and the limitations of human working memory, annotators can quickly lose track of the labels they have previously created as they annotate more and more documents. There is a danger of creating redundant or repetitive labels that may be covered by the existing label set or of creating labels that are much too detailed. For example, if an annotator creates 100 labels for 2000 documents, and the annotator reads a new document and creates a label 'water infrastructure and natural development' but there is an existing label 'public infrastructure and natural resource development' that is more general and can cover the newly created label, then the annotator inadvertently creates duplicate or over-specific labels during the annotation process.

Thus, a process of refinement of labels is often needed to merge or replace labels with very similar labels, which is itself a long and tedious process. Another way to combat annotators creating over-specific labels is for the annotator to explore the list of existing labels they created and try to pick a suitable label for each document. If no suitable label exists to represent the summary meanings of the current document, the annotator creates a new label and adds it to the list of existing labels. However, as the number of existing labels grows, this process requires much more mental effort from the annotator.

Active learning solves the challenges presented by this idealized annotation process, as illustrated in Fig. 4:

1. It can continuously redirect the user by recommending the most beneficial document for the annotator to read next, based on their existing inputs.

2. It can provide a ranked list of existing labels based on the probability a label belongs to the new document, reducing the cognitive burden of going through existing label list to create or assign labels.

3. It can reduce the number of documents needed to label by predicting labels for the set of documents, based on existing labels and labeled documents. As the annotator labels more documents, active learning can usually predict reasonable labels for the remaining documents.

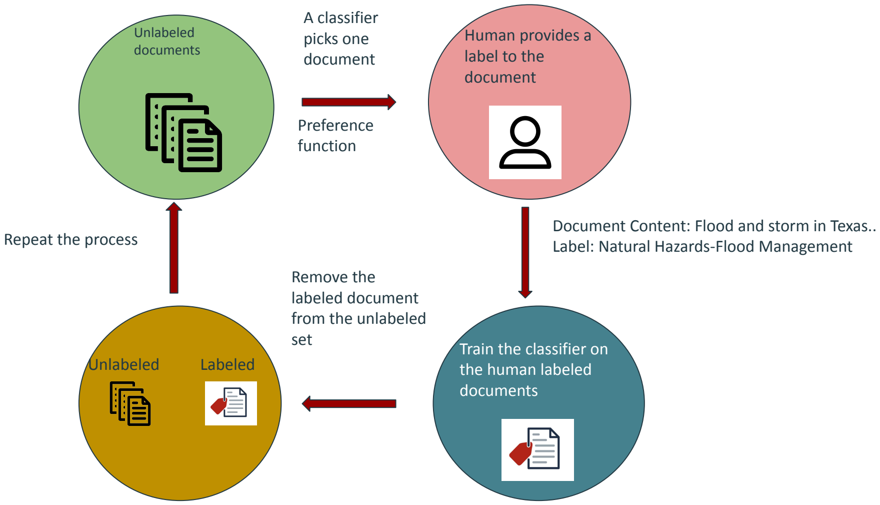

Unlabeled documents
A classifier picks one document
Preference function
Human provides a label to the document
Document Content: Flood and storm in Texas..
Label: Natural Hazards-Flood Management
Repeat the process
Remove the labeled document from the unlabeled set
Unlabeled
Labeled
Train the classifier on the human labeled documents

Fig. 4. By repeating the process of active learning to manually label at least 30 documents, the classifier starts predicting accurate labels for the rest of the documents.

As we discuss in Sec. 4.3, we conduct a user study to approximate a plausible threshold for the number of documents needed to be labeled within an hour for active learning to predict labels for all the documents, where the quality of user created labels and active learning predictions for all the documents in an hour is comparable to the ground-truth labels created by human annotators and through an inter-annotator agreement process that can be time-consuming to complete.

4.1.4. Combination of Topic Modeling and Active Learning

Active Learning provides local direction to annotators by guiding them towards documents to annotate next based on their previous inputs. However, merely focusing on individual documents to come up with a label that can not only represent the meaning of a specific document, but can also take a broader view of possible themes in the set of documents is challenging. With active learning alone, annotators do not have access to the set of potential topics in a document corpus.

Topic models can automatically cluster the set of documents into various topics at the beginning of the analysis, while providing a set of popular keywords for each topic. When using topic modeling in the process of content analysis and document annotation, annotators can take the broad view of the topics into account to create more general topics for individual documents and are more likely to create better label groups to categorize documents. In addition, for each document, a topic model generates a list of probabilities the document belongs to each topic. The list of probabilities also acts as part of inputs of active learning, where active learning can:

1. Constantly pick diverse documents from various topics in the process to maximize the label categories annotators create with a minimal number of documents needed to be labeled, thus reducing the chance annotators label documents with the same theme consecutively, which can be inefficient.
2. Improve the active learning classifier's ability to accurately predict labels for the set of documents while requiring fewer annotator label inputs.

Finally, for analysis of technical text such as content analysis, it is important for a human to be able to provide context and guide the topic model toward meaningful rather than spurious topics. This is the captured in the third requirement: human-in-the-loop.

4.2. Human-in-the-loop implementation

Next, we describe the human-in-the-loop system we developed to evaluate the use of various topic models and active learning to assist humans with document annotation.

Our requirements for the system are as follows:

- Focus on human-in-the-loop and user-input;

- Modular architecture that supports “plugging in” machine learning models;
- Deployable for validation through general public user study and expert pilot study.

The overall methodology decouples the machine learning methods from the implementation for computer-human interaction and provides a roadmap for interdisciplinary collaborations.

As previously discussed, we do not seek to automate human tasks but rather to assist humans in performing tasks more effectively. In the case of social science research, for instance, human input is crucial to the analysis of technical text and the literature on computational social science is increasingly recognizing the need for human-in-the-loop workflows that allow domain experts to guide and provide feedback to machine learning models.

Modularity is critical and has become more so with the explosion in large language models (LLMs), and it will be useful to be able to compare LLMs to more “traditional” methods such as topic models in the future. As such, the system is independent of the actual machine learning methods used for text analysis tasks.

To achieve modularity, a standard data structure is agreed upon. The data structure defines the communication between the models and the deployed system. The defined data structure grants the system the ability to work with different models and allows comparison of results across models.

This modularity allows us the flexibility to test the use of different models, as we discuss below in Sec. 4.3. Not only can we create different experimental conditions for testing the models, we can easily ingest different data sets. This allows us to validate the models through both large scale user studies, in which we crowdsource participants without any particular domain expertise, as well as more tailored validation studies with domain experts, as we describe in Sec. 5.

4.2.1. Overall Process

Figure 8 illustrates the document-annotation process in our human-in-the-loop implementation. Additional details are provided in Appendix A.[9]

The process features are grouped as follows:

- Log-in and Instructions: Once a user is selected to take part in the user study, a unique ID is issued. An unauthorized ID will be denied access. After successfully logging in, the set of instructions for the study are displayed for the user’s perusal.

[9] The source code for the human-in-the-loop system, which we call annotune (Annotation with the Text Understanding and Navigation Engine), can be found at https://github.com/usnistgov/annotune.

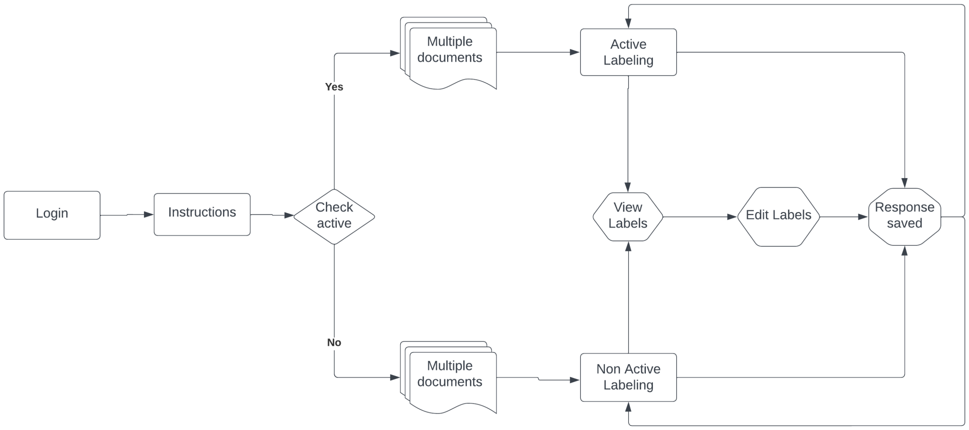


Fig. 5. The document annotation process in a human-in-the-loop system.

- Documents List: After proceeding from the instructions, a list of documents is presented to the user to annotate. The list provides a preview (i.e., the first few characters) of each document. The first document in the list is the document recommended by the current model assigned to the user in their experimental condition. Users have the freedom to select which document to annotate and can select any document in the list.
- Document Annotation: This feature allows users to read the entire document and annotate as deemed fit. Some models have keyword clusters displayed that can be highlighted to provide contextual meaning of the words in the documents, based on the topic model. Users can also view previously labeled documents and relabel. Older labels are saved and can be reused.
- Other Miscellaneous Features: The annotation app shows the user the elapsed time while logging the times spent on each page. Each session is saved so the user can stop and return to the task as needed.

4.3. User Study Validation

Given our interest in active learning and topic modeling to support document annotation, we form the following set of hypotheses:

1. Active learning combined with topic model can help annotators more efficiently create labels for documents.
2. Neural topic models combined with active learning can help annotators create higher quality labels than classic LDA combined with active learning.

3. Supervised LDA can update and generate new topics and clusters that correlate with annotator inputs, which can improve annotators' overall experience on reading the updated generated topics.

We then incorporate active learning and topic modeling in our modular human-in-the-loop implementation to conduct a user study to validate our hypotheses. We conduct a 60 user user-study on the Congressional Bills dataset.[10] The user study is important as it provides a systematic approach to evaluating the usefulness of these methods with actual humans in the loop.

4.4. Experimental Design

We separate our user study into four groups based on the following experimental conditions used to test our hypotheses:

1. Active Learning Only (NONE)
2. Contextualized Topic Model with Active Learning (CTM)
3. Latent Dirichlet Allocation with Active Learning (LDA)
4. Supervised Latent Dirichlet Allocation with Active Learning (sLDA)

Building on Poursabzi et al. [34], our baseline condition assumes active learning is valuable for document annotation. Note that the Contextualized Topic Model (CTM) is the specific neural topic model tested in the user study. Further simulation results for alternative neural topic models are provided in Li et al.[33] As the combination of neural topic models with active learning builds on ALTO,[34] we call this combination Topic-Enabled Neural Organization and Recommendations, or TENOR.[33]

Prior to the user study, we conduct a power analysis to determine the minimum number of annotators needed per group in order to detect statistical differences across the four experimental conditions. As a result, we assign 15 annotators to each group.

In addition, we conducted several pilot sessions to test the interface, collecting feedback on user experience and debugging to ensure successful deployment in the full user study.

4.4.1. Annotator Label Quality Measurement

Measuring the quality of labels created by different annotators is an impossible task,[41] where there are no right or wrong labels and each annotator can come up with their own definition of labels. To combat this issue, we introduce three machine learning clustering metrics to evaluate the quality of annotator labels from our user study. The three metrics reflect a comparison of annotator clustering against the official expert clustering in the Congressional Bills dataset.

[10] http://www.congressionalbills.org/codebooks.html

Purity: Purity evaluates how *pure* an induced cluster is. It measures what proportion of documents are placed in a cluster that is not commingled with documents with another gold label.[42] Purity can easily fail if each document is placed into a unique cluster that has no other documents.

Adjusted Normalized Mutual Information (ANMI): Normalized Mutual Information [43] assesses clustering quality by measuring the interdependence between true and predicted labels. ANMI [44], an enhancement of NMI, corrects for the chance alignment of clusters.

Adjusted Rand Index (ARI): ARI [45] measures the similarity between two clusters by evaluating the agreement between them while also accounting for random cluster agreement.

All of the above metrics range from 0 to 1, where a higher value of the metric is "better." Any one of the above clustering metrics is not perfect on its own, but if all of three are higher for one experimental condition than another, we can say that we feel more confident that one group creates labels more similar to the official expert labels than the other group.

4.5. User Study Ethics

Our user study details including the process, experimental design, payment, maximum number of users, potential risks, and privacy were submitted and approved by the University of Maryland Institutional Review Board. To best of our knowledge, our study did not contain any risks to the users. To ensure privacy, we assigned a unique identifier to each user to protect anonymity. We did not collect any personal information from the users. Users were allowed to withdraw from the study at any time if they did not agree with our user study requirements.

4.6. User Study Results

We solicit users from the crowdworking site Prolific,[11] where each user is paid $\$20$ an hour to complete our one-hour study. Each annotator is required to sit for one hour to complete the study. Users are instructed to read the task instructions and to get familiar with the user interface page first, then they proceed to read documents and create labels for them. After their time is up, they will complete a set of feedback questions on their experience, where the questions measure their mental effort, reliance on the topic model, and reliance on active learning to label documents.

Given that our user study is sampled from a general population, rather than domain experts from a specific discipline, we use a more user-friendly (though still somewhat technical) corpus: the Congressional Bills data set. One advantage of using this corpus for our user study is that it has been used extensively coded by trained annotators with years of refinement of the labels. The corpus has hierarchical labels and includes popular topics

[11] https://www.prolific.com/

about general public policies that users of any background can understand. We keep track of the evaluation metrics for every minute passed in the study for all the users. Within each group of 15 users, we take the median from the 15 users' purity, ARI, ANMI metrics per minute as our group-level evaluation metric. In addition, we also take the median of the metric values for each document the annotators labeled, for each of the four groups. We present the results of the evaluations for all the groups in Fig. 6.

Users create labels more similar to the official labels when using contextualized topic model. Figure 6 shows that the CTM group users have the highest gain of purity, ARI, and ANMI as both a function of minutes elapsed and number of documents labeled. A higher score for all three metrics indicates that the user-created labels are more similar to the Congressional Bills ground truth assigned categories. Although users in the study have different backgrounds, the CTM group still makes higher quality labels than users in the other three groups. In the next section, we describe a follow-up user study we conducted to validate the generalization of our conclusion that CTM is better than LDA in a setting where users have similar backgrounds (in fact, users are domain experts) on the RAS guidance document corpus.[17]

Discussion and implications. Within the context of advancing computational methods for text analysis, and in particular for social science research tasks such as content analysis, our results address both and model selection and evaluation.

With respect to model selection, the results show that using CTM with an active learning classifier can help annotators produce high quality label sets quickly, according to cluster metrics and human ratings, suggesting that the right choice of NTMs can be better than LDA for content analysis. However, simulation results show that classical topic models (i.e., LDA) remain competitive with NTMs. Thus, the question is not about NTM versus LDA but rather the right choice of NTM for specific tasks.

Moreover, the results show that current automated metrics do not provide a complete picture of topic modeling capabilities. Our study demonstrates how human-in-the-loop task-based evaluation can be used to compare neural, supervised, and classical topic models for content analysis and label set creation.

In the next section, we explore how these results may generalize when annotators have domain expertise with the text.

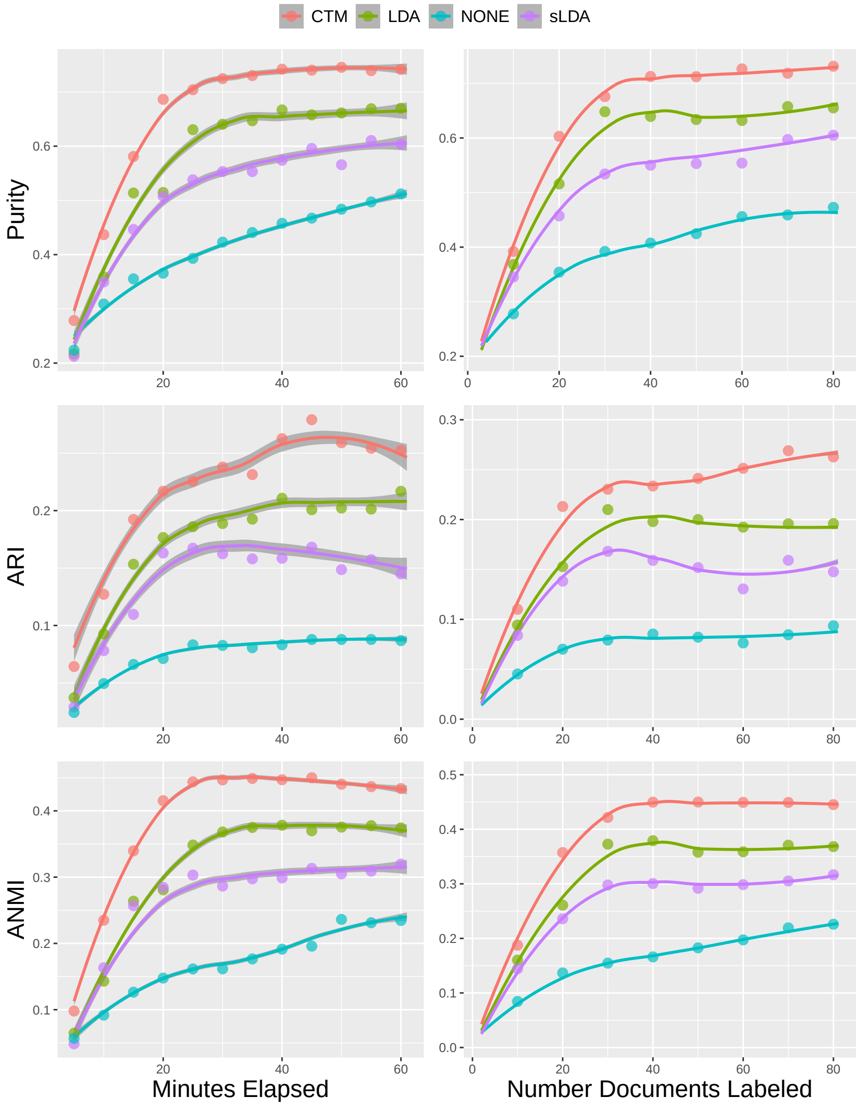


Fig. 6. User study evaluation metrics gain for each group. The left shows the median evaluation metric gain, from minute 5 to minute 60, in the user study. The right shows the metric gain, from 2 documents labeled to 80 documents labeled, for each group. CTM combined with active learning helps annotators create the labels with the best quality, on average, compared to the other groups. The bands, which represent group variance around the median, suggest consistency in evaluation metric values across users.

5. Expert Validation: Resilience, Adaptation, and Sustainability (RAS) Planning

In this section, we return to our motivating example from Sec. 2. Recall that Clavin et al. [17] created a hand-annotated corpus based on planning guidance for resilience, climate adaptation, and sustainability (RAS) community planning. The documents were hand-annotated through a rigorous content analysis, but the process was a time and labor intensive effort.

Given the user study results presented in Sec. 4.3, a natural question is whether the results generalize to data found "in the wild." The goal is to determine to what extent the combination of active learning with topic models (and in particular the Contextualized Topic Model, or CTM), is useful to domain experts for real-world document-annotation tasks. We explore the tradeoff between efficiency gain, in terms of time spent annotating documents, and label quality, in terms of the evaluation metrics from Sec. 4.3 (Purity, ANMI, and ARI).

Leveraging the Clavin et al. [17] hand-annotated RAS guidance corpus for ground truth, we conducted a validation study with domain experts at NIST. Specifically, we recruited eight staff members within the NIST Community Resilience Program (CRP) to use our human-in-the-loop implementation to annotate documents from the RAS corpus. The participants who volunteered have expertise ranging across the domains of resilience, adaptation, and sustainability and included both early career and senior staff.

5.1. Expert Validation: Purpose

For the validation study with RAS domain experts, the three main questions we ask are:

1. Generalizability: To what extent do evaluation metrics from domain experts correlate with user study results?
2. User experience: What do domain experts think of using the human-in-the-loop implementation for document-annotation tasks?
3. Value: What is the tradeoff between efficiency gain (speed) and label quality (reliability)?

For the first question, we focus on comparing LDA with CTM, both with active learning, splitting the participants into two groups (four per group). The volunteers were asked to annotate documents for up to one hour.

For the second question, we assess responses from the post-study questionnaire we used for the user study. In addition, we obtain informal feedback from our colleagues on their overall experience.

For the third question, we extrapolate how long it would take, on average, for humans to annotate the entire RAS guidance corpus (roughly 2600 documents) and balance that against the potential loss in label quality.

Finally, it is worth noting the estimated number of labeled documents per hour is a lower bound since (a) experts in the validation study did not have access to the codebook, as it would have taken too long to train the volunteers; and (b) potentially, human learning and active learning both reduce time to annotate as more documents are annotated. In principle, a codebook should reduce time to annotate documents, though in practice it may be a lot more interactive and iterative. Our assumption therefore is that, as the number of labeled documents increases, using a codebook should not decrease the rate of document annotation.

5.2. Expert Validation: Results

We had six expert volunteers to validate our results on the RAS guidance document corpus.[12] We pick the top two winning conditions from the last user study– CTM and LDA— and assign three users to each group. Aside from the domain-specific data, the expert validation study setup is identical to the user study we described in Sec. 4.3. Figure 7 shows results from the domain expert validation study.

5.2.1. Generalizability

CTM users create higher quality labels than LDA users, when users have similar domain-specific expertise on the dataset. The results are largely consistent with our user study results: as the number of labeled documents increases, CTM performs better than LDA on all three evaluation metrics. This set of results from our expert validation study are encouraging, as they suggest that the results from the user study generalize to "real" data used for content analysis, providing additional confidence in using our proposed human-in-the-loop approach for applications that require the analysis of technical text.

5.2.2. User experience

At the end of the annotation session, we collected qualitative user feedback on the experience with our human-in-the-loop system. Specifically, we see more positive and more detailed feedback from CTM users, relative to LDA users, about improving the usability and experience of the human-in-the-loop interface.

Overall, LDA users did not pay attention to the keywords generated by LDA. Upon observing the topic keywords, we see that LDA generates too broad and general keywords that do not help users in interpreting and understanding the contents of a specific document. On

[12] While we had eight total volunteers, two of the volunteers did not spend the required time annotating documents.

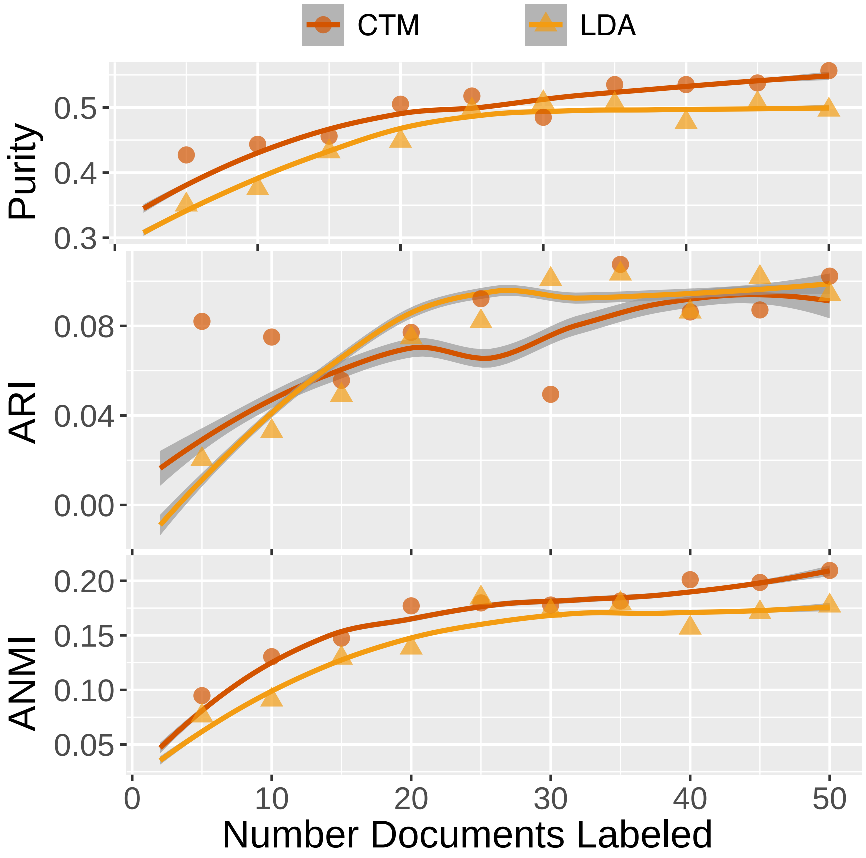


Fig. 7. We track the expert users' purity, adjusted Rand Index, and adjusted normalized mutual information on the number of documents labeled with 3 experts in each group. CTM (contextualized neural topic model) does better than LDA overall on the three metrics. The bands, which represent group variance around the median, suggest consistency in evaluation metric values across users.

the other hand, CTM generates compact and specific keywords that might serve as labels for specific documents.

One user in the CTM group thinks the tool has a bit of learning curve, “but once you get used to it is easy to use.” With regard to the guidance document corpus, the lack of a codebook did introduce additional challenges to the users: “I do think it is so hard to label things without knowing the research question or why I am labeling items since I could be looking for very different things depending on the purpose and most text could be labeled in lots of different ways.”

Finally, a summary of the recommendations from the CTM users include:

1. Incorporating a method to evaluate the performance of user and providing feedback on how they are doing (maybe after 15 minutes). This can be a training step.
2. It would be nice if the tool provides an estimated number of passages that should be labeled within one hour.
3. It would be useful if the tool provides an estimated time that needs to be spent on each passage to assign a label.

5.2.3. Value

Based on the results of the expert validation study, we extrapolate how long it would take, on average, for humans to annotate the entire RAS guidance corpus.

- Assuming 50 documents per hour implies 1.2 minutes per document, for a total of 52.5 hours to annotate 2625 documents.
- We can round that up to 60 hours of labor required per annotator.
- Let’s assume the annotator spends 4 hours a day, for 3 days a week, actually annotating (in the case of the RAS content analysis, the annotators had more focused time on annotation but did not spend all of their time on annotation).
- At these assumed rates, an annotator will require 5 weeks to annotate all 2625 documents.

That’s a bit over a month, compared to the three months that were actually required in the original, purely human RAS content analysis. In other words, machine assisted annotation improves productivity by about (at minimum) 58%.

How do we value that productivity gain? Suppose our annotators are paid the prevailing minimum wage ($15 per hour in Maryland as of January 1, 2024).[13] At the new rate of productivity, we only need 5 weeks to produce the same amount of output (which was initially worth $15 × 12 weeks = $7200 in total wages, assuming 40 hours of work per

[13] https://www.montgomerycountymd.gov/humanrights/min-wage.html. Accessed: 2024-02-09.

week). The value of time savings is \$15 × (12 - 5) = \$4200. Moreover, the same annotator's effective wage is now \$7200 / 5 weeks = \$36 per hour, which is an indirect way of obtaining higher skill labor at a lower cost. In our RAS example, this is a lower bound. The real value arguably comes from the time savings to the senior researchers who participated in the content analysis, who would either spend less time on annotating text on average or altogether delegate annotation fully to the junior researchers while focusing on higher-level tasks such as coordination and planning.

An important caveat, as stated above, is that the volunteers in our expert validation study did not have access to the codebook. If we assume annotators use a codebook, rather than building a label set from scratch, the time savings should be much greater.

Of course, there is no free lunch! The efficiency gain comes with a potential loss in label quality. As Fig. 7 shows, none of the three metrics appear close to 1 and this loss in label quality could represent substantial risk to the research objectives. How does a researcher evaluate that risk against the reward of higher efficiency (which could translate the same total amount of labor into significantly more labeled data)? This is not straightforward and depends on the subjective value of label quality (rather than just the wage rate), assuming one could place a dollar value on label quality, as well as what that loss in label quality translates to in practice.

In an actual content analysis, the time saved to annotate documents can be used for more frequent intercoder checks, codebook updates, and data validation steps, which could improve label quality over the baseline. Moreover, a researcher would need to track the metrics beyond the hour shown in Fig. 7 to see if label quality increases over time or stabilizes, whether the observed loss in label quality is acceptable (is the researcher concerned about type I or type II errors), and whether additional human input can improve the evaluation metrics (and weigh that against how much effort is required to do so). We leave these as open questions for future research.

5.3. Implications

We began with the premise that machine assistance can improve the efficiency and productivity of experts in conducting labor-intensive tasks such as document annotation for content analysis. Our user studies support this premise. However, as we highlight in our discussion on the value of machine assistance, the potentially larger productivity gains come from freeing experts from tedious tasks (such as manually annotating text) while "upskilling" non-experts (e.g., students are potentially able to lead and conduct a content analysis that typically requires substantial domain expertise).[46, 47] This is an example of how artificial intelligence has the potential to improve labor market outcomes for the middle class by extending the reach of expertise, as hypothesized by Autor: "For workers with foundational training and experience, AI can help to leverage expertise so they can do higher-value work."[48]

6. Conclusion

While the user study and expert validation study are encouraging, much work remains in order to have a process that can be integrated into a text analysis pipeline such as content analysis. As we described in Sec. 4, the machine-assisted process we propose is not specific to any one topic model. Though our user study results suggest that the contextualized topic model (CTM) may be preferable to LDA for enhancing the document annotation workflow, the user study did not compare alternative neural topic models. Moreover, as large language models (LLMs) have come to dominate natural language processing, we are exploring how LLMs compare with traditional topic models for human-in-the-loop document annotation.

Additionally, while our human-in-the-loop implementation was largely developed to enable user studies, it also serves as a proof of concept for a modular system for conducting research. Though we are sharing the source code for other researchers and for transparency,[14] we are working on refining the system so that in principle it can be deployed for actual research.

Finally, we have only focused so far on the “annotation” part of the pipeline but as Fig. 1 in Sec. 2 illustrates, the text analysis workflow has several time-consuming steps that could be bottlenecks in efficiently analyzing a large corpus. For instance, we assume that the documents to be analyzed are in already in an easily ingestible, plain-text format. However, the RAS content analysis Clavin et al. [17] conducted required

- Data curation: Gathering source text through a combination of careful internet search and knowledge of the domain; and
- Data pre-processing: Converting the source text, most often large PDF files, into plain-text “snippets” (i.e., the documents) for analysis.

It remains to be seen to what extent machine learning could assist in these and other steps in the pipeline, and which steps would still require significant human labor.

The key message for future applications from is that collaborative, truly interdisciplinary research is needed to advance the use of machine learning to assist humans with the analysis of technical text.

[14] https://github.com/usnistgov/annotune

References

[1] Malecha M, Masterson JH, Yu S, Berke P (2021) Plan Integration for Resilience Scorecard Guidebook: Spatially evaluating networks of plans to reduce hazard vulnerability - Version 2.0 (Institute for Sustainable Communities, College of Architecture, Texas A&M University, College Station, Texas), Available at https://planintegration.com/wp-content/uploads/2023/03/PIRS-Guidebook-v2.0_2021.09.pdf.

[2] Brundage MP, Sexton R, Hodkiewicz M, Dima A, Lukens S (2021) Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters* 27:42–46. https://doi.org/10.1016/j.mfglet.2020.11.001. Available at https://www.sciencedirect.com/science/article/pii/S2213846320301668

[3] Castano S, Ferrara A, Furiosi E, Montanelli S, Picascia S, Riva D, Stefanetti C (2024) Enforcing legal information extraction through context-aware techniques: The ASKE approach. *Computer Law & Security Review* 52:105903. https://doi.org/10.1016/j.cl sr.2023.105903. Available at https://www.sciencedirect.com/science/article/pii/S0 267364923001139

[4] Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B, Rand DG (2014) Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58(4):1064–1082. https://doi.org/10.1111/ajps.12103. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12103 Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12103

[5] Bengtsson M (2016) How to plan and perform a qualitative study using content analysis. *NursingPlus Open* 2:8–14. https://doi.org/10.1016/j.npls.2016.01.001. Available at https://www.sciencedirect.com/science/article/pii/S2352900816000029

[6] Cho JY, Lee EH (2014) Reducing Confusion about Grounded Theory and Qualitative Content Analysis: Similarities and Differences. *The Qualitative Report* 19(64):1–20. Available at http://www.nova.edu/ssss/QR/QR19/cho64.pdf.

[7] Grimmer J, Stewart BM (2013) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3):267–297. https://doi.org/10.1093/pan/mps028. Publisher: Cambridge University Press Available at https://www.cambridge.org/core/journals/political-analysis/article/text-as-data-the-promise-and-pitfalls-of-automatic-content-analysis-methods-for-political-texts/F7AAC8B2909441603FEB25C156448F20

[8] Baumer EPS, Mimno D, Guha S, Quan E, Gay GK (2017) Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68(6):1397–1410. https://doi.org/10.1002/asi.23786. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23786 Available at https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23786

[9] Lake BM, Murphy GL (2023) Word meaning in minds and machines. *Psychological Review* 130(2):401–431. https://doi.org/10.1037/rev0000297. Place: US Publisher: American Psychological Association

[10] Wang ZJ, Choi D, Xu S, Yang D (2021) Putting Humans in the Natural Language Processing Loop: A Survey. https://doi.org/10.48550/arXiv.2103.04044. arXiv:2103.04044 [cs] Available at http://arxiv.org/abs/2103.04044

[11] Clavin CT, Dabreau AM, Walpole EH (2020) Resilience, Adaptation, and Sustainability Plan Assessment Methodology: An Annotated Bibliography. *NIST* Last Modified: 2020-09-28T09:09-04:00 Publisher: Christopher T. Clavin, Avery M. Dabreau, Emily H. Walpole Available at https://www.nist.gov/publications/resilience-adaptation-and-sustainability-plan-assessment-methodology-annotated.

[12] Silge J, Robinson D (2016) tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *Journal of Open Source Software* 1(3):37. https://doi.org/10.21105/joss.00037. Available at https://joss.theoj.org/papers/10.21105/joss.00037

[13] Clavin CT, Dabreau AM, Walpole EH (2020) Resilience, Adaptation, and Sustainability Plan Assessment Methodology: An Annotated Bibliography. *NIST* Last Modified: 2020-09-28T09:09-04:00 Publisher: Christopher T. Clavin, Avery M. Dabreau, Emily H. Walpole Available at https://www.nist.gov/publications/resilience-adaptation-and-sustainability-plan-assessment-methodology-annotated.

[14] Community Resilience, https://www.nist.gov/community-resilience. Accessed: 2024-03-01.

[15] USGCRP (2023) *Fifth National Climate Assessment* (U.S. Global Change Research Program, Washington, DC, USA). https://doi.org/10.7930/NCA5.2023

[16] Roostaie S, Nawari N, Kibert CJ (2019) Sustainability and resilience: A review of definitions, relationships, and their integration into a combined building assessment framework. *Building and Environment* 154:132–144. https://doi.org/10.1016/j.buildenv.2019.02.042. Available at https://www.sciencedirect.com/science/article/pii/S0360132319301532

[17] Clavin CT, Walpole EH, D'Abreau A, Wong S (2021) All Roads Lead to Resilience? A Structured Comparison of Community-Oriented Guidance for Resilience, Climate Adaptation, and Sustainability Planning. In Review.

[18] DiMaggio P (2015) Adapting computational text analysis to social science (and vice versa). *Big Data & Society* 2(2):2053951715602908. https://doi.org/10.1177/2053951715602908. Publisher: SAGE Publications Ltd Available at https://doi.org/10.1177/2053951715602908

[19] Breiman L (2001) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* 16(3):199–231. https://doi.org/10.1214/ss/1009213726. Publisher: Institute of Mathematical Statistics Available at https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full

[20] Golino H, Shi D, Christensen AP, Garrido LE, Nieto MD, Sadana R, Thiyagarajan JA, Martinez-Molina A (2020) Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods* 25(3):292–320. https://doi.org/10.1037/met0000255. Place: US Publisher: American Psychological Association

[21] Golino H, Moulder R, Shi D, Christensen AP, Garrido LE, Nieto MD, Nesselroade J, Sadana R, Thiyagarajan JA, Boker SM (2021) Entropy Fit Indices: New Fit Measures for Assessing the Structure and Dimensionality of Multiple Latent Variables. *Multivariate Behavioral Research* 56(6):874–902. https://doi.org/10.1080/00273171.2020.1779642. Publisher: Routledge _eprint: https://doi.org/10.1080/00273171.2020.1779642 Available at https://doi.org/10.1080/00273171.2020.1779642

[22] Gentzkow M, Shapiro JM, Taddy M (2019) Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica* 87(4):1307–1340. https://doi.org/10.3982/ECTA16566. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16566 Available at https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16566

[23] Lucas C, Nielsen RA, Roberts ME, Stewart BM, Storer A, Tingley D (2015) Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis* 23(2):254–277. https://doi.org/10.1093/pan/mpu019. Publisher: Cambridge University Press Available at https://www.cambridge.org/core/journals/political-analysis/article/computerassisted-text-analysis-for-comparative-politics/CC8B2CF63A8CC36FE00A13F9839F92BB

[24] Wilkerson J, Casas A (2017) Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science* 20(1):529–544. https://doi.org/10.1146/annurev-polisci-052615-025542. _eprint: https://doi.org/10.1146/annurev-polisci-052615-025542 Available at https://doi.org/10.1146/annurev-polisci-052615-025542

[25] Zamith R, Lewis SC (2015) Content Analysis and the Algorithmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis. *The ANNALS of the American Academy of Political and Social Science* 659(1):307–318. https://doi.org/10.1177/0002716215570576. Publisher: SAGE Publications Inc Available at https://doi.org/10.1177/0002716215570576

[26] Boumans JW, Trilling D (2016) Taking Stock of the Toolkit. *Digital Journalism* 4(1):8–23. https://doi.org/10.1080/21670811.2015.1096598. Publisher: Routledge _eprint: https://doi.org/10.1080/21670811.2015.1096598 Available at https://doi.org/10.1080/21670811.2015.1096598

[27] Muller M, Guha S, Baumer EP, Mimno D, Shami NS (2016) Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. *Proceedings of the 19th International Conference on Supporting Group Work* (ACM, Sanibel Island Florida USA), pp 3–8. https://doi.org/10.1145/2957276.2957280. Available at https://dl.acm.org/doi/10.1145/2957276.2957280

[28] Kjell ONE, Kjell K, Garcia D, Sikström S (2019) Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods* 24(1):92–115. https://doi.org/10.1037/met0000191. Place: US Publisher: American Psychological Association

[29] Biesbroek R, Badloe S, Athanasiadis IN (2020) Machine learning for research on climate change adaptation policy integration: an exploratory UK case study. *Regional Environmental Change* 20(3):85. https://doi.org/10.1007/s10113-020-01677-8. Available at https://doi.org/10.1007/s10113-020-01677-8

[30] Berrang-Ford L, et al. (2021) A systematic global stocktake of evidence on human adaptation to climate change. *Nature Climate Change* 11(11):989–1000. https://doi.org/10.1038/s41558-021-01170-y. Number: 11 Publisher: Nature Publishing Group Available at https://www.nature.com/articles/s41558-021-01170-y

[31] Brinkley C, Stahmer C (2021) What Is in a Plan? Using Natural Language Processing to Read 461 California City General Plans. *Journal of Planning Education and Research* :0739456X21995890https://doi.org/10.1177/0739456X21995890. Publisher: SAGE Publications Inc Available at https://doi.org/10.1177/0739456X21995890

[32] Saheb T, Dehghani M, Saheb T (2022) Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis. *Sustainable Computing: Informatics and Systems* 35:100699. https://doi.org/10.1016/j.suscom.2022.100699. Available at https://www.sciencedirect.com/science/article/pii/S2210537922000415

[33] Li Z, Mao A, Stephens D, Goel P, Walpole E, Dima A, Fung J, Boyd-Graber J (2024) Improving the tenor of labeling: Re-evaluating topic models for content analysis. 2401.16348.

[34] Poursabzi-Sangdeh F, Boyd-Graber J, Findlater L, Seppi K (2016) ALTO: Active Learning with Topic Overviews for Speeding Label Induction and Document Labeling. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Berlin, Germany), pp 1158–1169. https://doi.org/10.18653/v1/P16-1110. Available at https://aclanthology.org/P16-1110

[35] Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3(null):993–1022.

[36] Blei DM, McAuliffe JD (2010) Supervised topic models. 1003.0783.

[37] Bianchi F, Terragni S, Hovy D (2021) Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *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)* (Association for Computational Linguistics, Online), pp 759–766. https://doi.org/10.18653/v1/2021.acl-short.96. Available at https://aclanthology.org/2021.acl-short.96

[38] Dieng AB, Ruiz FJR, Blei DM (2019) Topic modeling in embedding spaces. 1907.04907.

[39] Grootendorst M (2022) Bertopic: Neural topic modeling with a class-based tf-idf procedure. 2203.05794.

[40] Baumer EPS, Mimno D, Guha S, Quan E, Gay GK (2017) Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *J Assoc Inf Sci Technol* 68(6):1397–1410. https://doi.org/10.1002/asi.23786. Available at https://doi.org/10.1002/asi.23786

[41] Kleinberg J (2002) An impossibility theorem for clustering. *Advances in Neural Information Processing Systems*, eds Becker S, Thrun S, Obermayer K (MIT Press), Vol. 15. Available at https://proceedings.neurips.cc/paper_files/paper/2002/file/43e4e6a6f341e00671e123714de019a8-Paper.pdf.

[42] Zhao Y (2005) *Criterion Functions for Document Clustering*. Ph.D. thesis., USA. AAI3180039.

[43] Strehl A, Ghosh J (2003) Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* 3(null):583–617. https://doi.org/10.1162/153244303321897735. Available at https://doi.org/10.1162/153244303321897735

[44] Amelio A, Pizzuti C (2016) Correction for closeness: Adjusting normalized mutual information measure for clustering comparison: Correction for closeness: Adjusting nmi. *Computational Intelligence* 33. https://doi.org/10.1111/coin.12100

[45] Sundqvist M, Chiquet J, Rigaill G (2022) Adjusting the adjusted rand index: A multinomial story. *Comput Stat* 38(1):327–347. https://doi.org/10.1007/s00180-022-01230-7. Available at https://doi.org/10.1007/s00180-022-01230-7

[46] Brynjolfsson E, Li D, Raymond L Generative AI at Work. https://doi.org/10.48550/arXiv.2304.11771. 2304.11771 Available at http://arxiv.org/abs/2304.11771

[47] Dell'Acqua F, McFowland III E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. https://doi.org/10.2139/ssrn.4573321. Available at https://papers.ssrn.com/abstract=4573321

[48] Autor D AI Could Actually Help Rebuild The Middle Class Available at https://www.noemamag.com/how-ai-could-help-rebuild-the-middle-class.

Appendix A. Implementing a Modular System for Testing Human-in-the-loop Document Annotation

This appendix provides details on the implementation of the modular system we developed to conduct user studies of human-in-the-loop document annotation using different methods, which we call annotune (Annotation with the Text Understanding and Navigation Engine).[15]

Appendix A.1. Overall Process

The aim of the app is to aid in annotation of topics, track all saved labels, and track the times used to label each documents for the user study. Figure 1 shows the process.

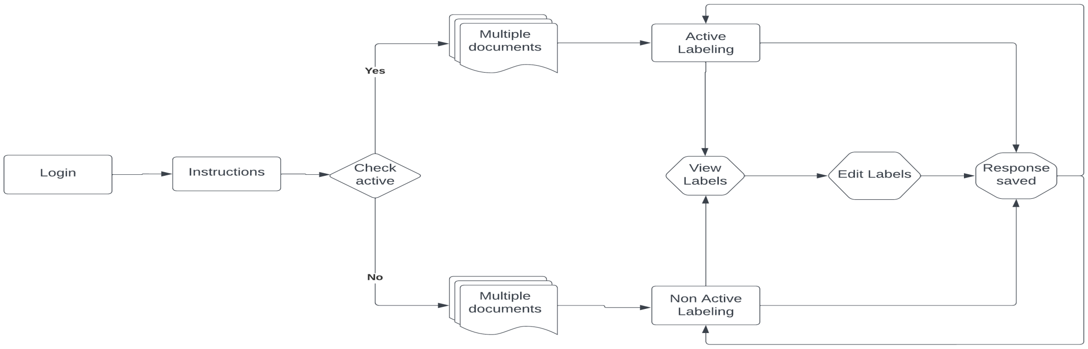


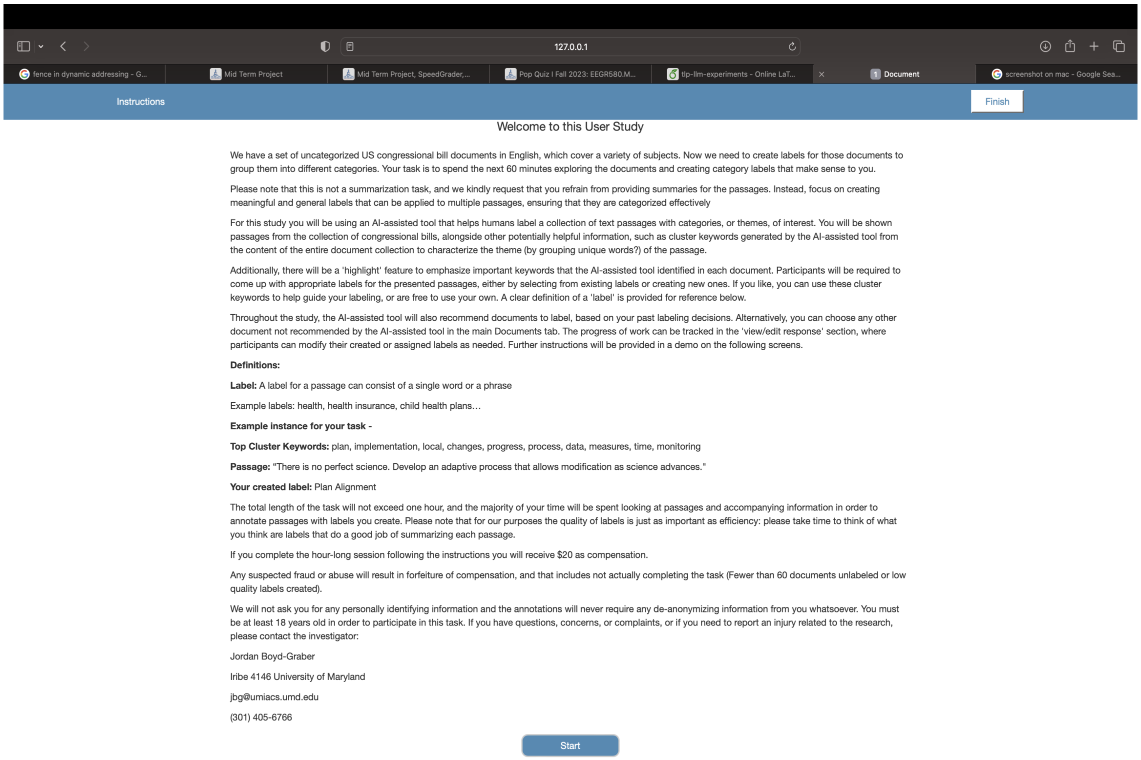
Fig. 8. The annotation process encoded within the interface.

Appendix A.1.1. Login

The first page the user sees is the login page where users are required to enter their unique user ID (UUID). When the user enters the UUID and submits the login form via a post request, the submitted name is extracted from the form and validated to see if the name exists in the dictionary of allowed user names. Once it is validated, session variables are set for the particular user. The session variables are username, labels, `user_id`, and labeled_document.

- `username`: the UUID of the user
- `Labels`: All past labels created by the user
- `user_id`: The id assigned to this user by the model
- `labeled_document`: The ids of documents that have already been labeled.

[15] See source code at: https://github.com/usnistgov/annotune.


127.0.0.1
fence in dynamic addressing - G...
Mid Term Project
Mid Term Project, SpeedGrader,...
Pop Quiz I Fall 2023: EEGR580.M...
tlp-llm-experiments - Online LaT...
Document
screenshot on mac - Google Sea...
Instructions
Finish

Welcome to this User Study

We have a set of uncategorized US congressional bill documents in English, which cover a variety of subjects. Now we need to create labels for those documents to group them into different categories. Your task is to spend the next 60 minutes exploring the documents and creating category labels that make sense to you.

Please note that this is not a summarization task, and we kindly request that you refrain from providing summaries for the passages. Instead, focus on creating meaningful and general labels that can be applied to multiple passages, ensuring that they are categorized effectively

For this study you will be using an AI-assisted tool that helps humans label a collection of text passages with categories, or themes, of interest. You will be shown passages from the collection of congressional bills, alongside other potentially helpful information, such as cluster keywords generated by the AI-assisted tool from the content of the entire document collection to characterize the theme (by grouping unique words?) of the passage.

Additionally, there will be a 'highlight' feature to emphasize important keywords that the AI-assisted tool identified in each document. Participants will be required to come up with appropriate labels for the presented passages, either by selecting from existing labels or creating new ones. If you like, you can use these cluster keywords to help guide your labeling, or are free to use your own. A clear definition of a 'label' is provided for reference below.

Throughout the study, the AI-assisted tool will also recommend documents to label, based on your past labeling decisions. Alternatively, you can choose any other document not recommended by the AI-assisted tool in the main Documents tab. The progress of work can be tracked in the 'view/edit response' section, where participants can modify their created or assigned labels as needed. Further instructions will be provided in a demo on the following screens.

Definitions:

Label: A label for a passage can consist of a single word or a phrase

Example labels: health, health insurance, child health plans…

Example instance for your task -

Top Cluster Keywords: plan, implementation, local, changes, progress, process, data, measures, time, monitoring

Passage: “There is no perfect science. Develop an adaptive process that allows modification as science advances."

Your created label: Plan Alignment

The total length of the task will not exceed one hour, and the majority of your time will be spent looking at passages and accompanying information in order to annotate passages with labels you create. Please note that for our purposes the quality of labels is just as important as efficiency: please take time to think of what you think are labels that do a good job of summarizing each passage.

If you complete the hour-long session following the instructions you will receive $20 as compensation.

Any suspected fraud or abuse will result in forfeiture of compensation, and that includes not actually completing the task (Fewer than 60 documents unlabeled or low quality labels created).

We will not ask you for any personally identifying information and the annotations will never require any de-anonymizing information from you whatsoever. You must be at least 18 years old in order to participate in this task. If you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator:

Jordan Boyd-Graber

Iribe 4146 University of Maryland

jbg@umiacs.umd.edu

(301) 405-6766

Start

Fig. 9. The instructions page provides detailed instructions on the user study and navigating the interface.

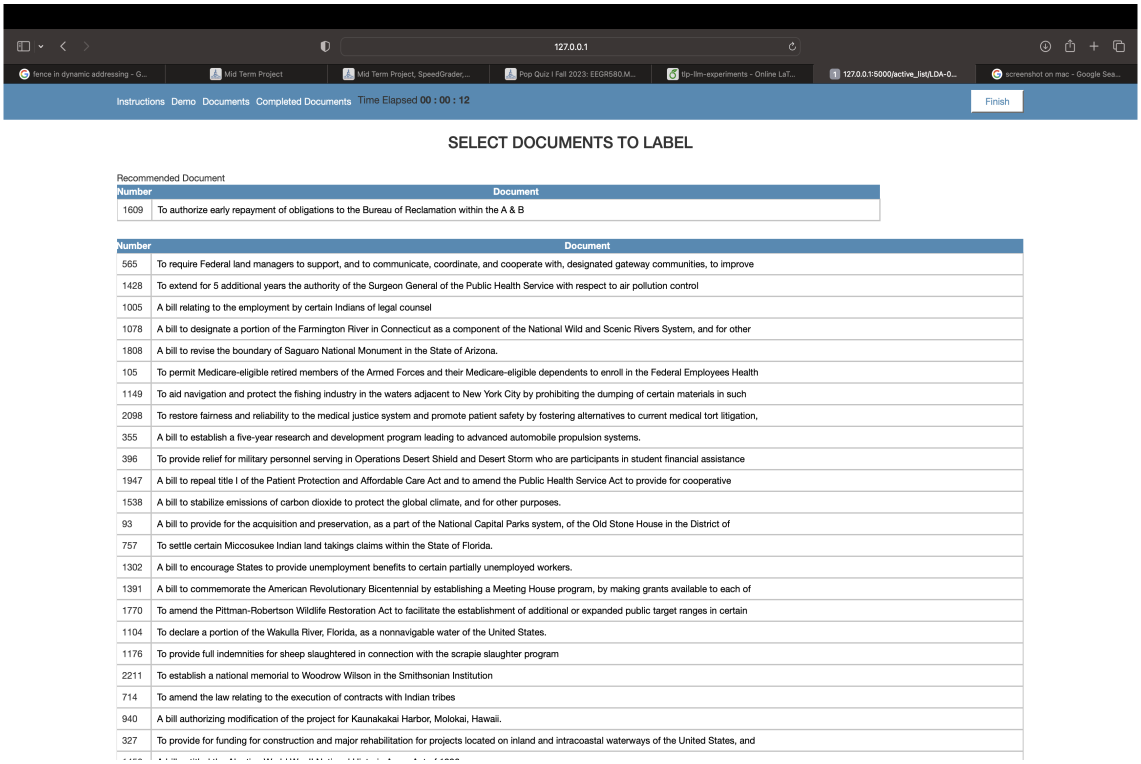
If a user is a new user or starting their session for the first time, a post request is sent to the model to create a new user. If user is valid, this information is then saved in a `users.json` file in the `static/users/` directory. If the user is a returning user, the session data are extracted from the `users.json` and they may return to the where they left off. Usernames that are not validated are redirected back to the login page

Appendix A.1.2. Instructions

Upon the login, instructions for the user study are shown to the user, as seen in Fig. 9. After users read the instructions, they may click on start and the user study session begins. When the start button is selected, a post request is sent to the model to get the topic list. There are different sets of models, some with active learning and others without active learning. When the start button is selected, a get request is sent to the model and a list of documents, their keywords, and topic clusters are requested. The check active function checks whether the model assigned to this user has active learning or not. It does so by checking the number of topic clusters that the model sends. If it is just one, it is active learning, else, its not. Once the decision is made, the appropriate route is selected and routed to.

Appendix A.1.3. Document List with active learning

Before annotating, the app offers two options that depend on the type of model being used. When the model includes active learning, there is no clustering and all the documents are listed together. A model-recommended document is selected and presented at top of the list. When a document is selected, the user is redirected to the `/get_label/` route to label. Fig. 10 shows how the active list page is displayed by the app.


127.0.0.1
fence in dynamic addressing - G...
Mid Term Project
Mid Term Project, SpeedGrader,...
Pop Quiz I Fall 2023: EEGR580.M...
tlp-llm-experiments - Online LaT...
127.0.0.1:5000/active_list/LDA-0...
screenshot on mac - Google Sea...
Instructions
Demo
Documents
Completed Documents
Finish

Time Elapsed **00 : 00 : 12**

SELECT DOCUMENTS TO LABEL

Recommended Document

Number	Document
1609	To authorize early repayment of obligations to the Bureau of Reclamation within the A & B

Number	Document
565	To require Federal land managers to support, and to communicate, coordinate, and cooperate with, designated gateway communities, to improve
1428	To extend for 5 additional years the authority of the Surgeon General of the Public Health Service with respect to air pollution control
1005	A bill relating to the employment by certain Indians of legal counsel
1078	A bill to designate a portion of the Farmington River in Connecticut as a component of the National Wild and Scenic Rivers System, and for other
1808	A bill to revise the boundary of Saguaro National Monument in the State of Arizona.
105	To permit Medicare-eligible retired members of the Armed Forces and their Medicare-eligible dependents to enroll in the Federal Employees Health
1149	To aid navigation and protect the fishing industry in the waters adjacent to New York City by prohibiting the dumping of certain materials in such
2098	To restore fairness and reliability to the medical justice system and promote patient safety by fostering alternatives to current medical tort litigation,
355	A bill to establish a five-year research and development program leading to advanced automobile propulsion systems.
396	To provide relief for military personnel serving in Operations Desert Shield and Desert Storm who are participants in student financial assistance
1947	A bill to repeal title I of the Patient Protection and Affordable Care Act and to amend the Public Health Service Act to provide for cooperative
1538	A bill to stabilize emissions of carbon dioxide to protect the global climate, and for other purposes.
93	A bill to provide for the acquisition and preservation, as a part of the National Capital Parks system, of the Old Stone House in the District of
757	To settle certain Miccosukee Indian land takings claims within the State of Florida.
1302	A bill to encourage States to provide unemployment benefits to certain partially unemployed workers.
1391	A bill to commemorate the American Revolutionary Bicentennial by establishing a Meeting House program, by making grants available to each of
1770	To amend the Pittman-Robertson Wildlife Restoration Act to facilitate the establishment of additional or expanded public target ranges in certain
1104	To declare a portion of the Wakulla River, Florida, as a nonnavigable water of the United States.
1176	To provide full indemnities for sheep slaughtered in connection with the scrapie slaughter program
2211	To establish a national memorial to Woodrow Wilson in the Smithsonian Institution
714	To amend the law relating to the execution of contracts with Indian tribes
940	A bill authorizing modification of the project for Kaunakakai Harbor, Molokai, Hawaii.
327	To provide for funding for construction and major rehabilitation for projects located on inland and intracoastal waterways of the United States, and

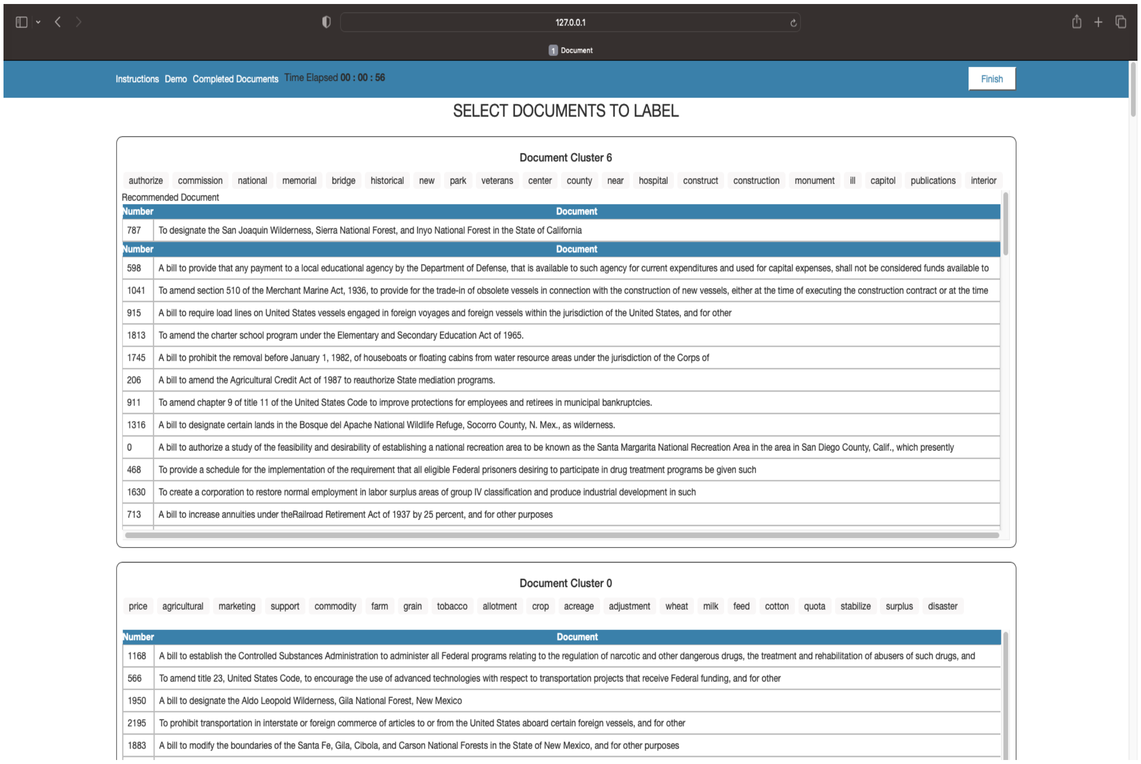
Fig. 10. The active learning document list shows a list of recommended documents for annotation, with the current recommendation at the top.

Appendix A.1.4. Document List without active learning

When an active learning model is not being used, the model generates a list of document clusters. Keywords are associated with each cluster to enhance the user's understanding of the document's context. The user interface is designed to facilitate the user experience by isolating the recommended topic on the top and displaying six documents within each cluster.

Appendix A.1.5. Labeling with active learning

The `active` function in the annotation app manages the active list functionality for users engaged in the active learning process. The function takes two parameters: `name`, which



Time Elapsed 00 : 00 : 56

SELECT DOCUMENTS TO LABEL

Document Cluster 6

authorize commission national memorial bridge historical new park veterans center county near hospital construct construction monument ill capitol publications interior

Recommended Document

Number	Document
787	To designate the San Joaquin Wilderness, Sierra National Forest, and Inyo National Forest in the State of California

Number	Document
598	A bill to provide that any payment to a local educational agency by the Department of Defense, that is available to such agency for current expenditures and used for capital expenses, shall not be considered funds available to
1041	To amend section 510 of the Merchant Marine Act, 1936, to provide for the trade-in of obsolete vessels in connection with the construction of new vessels, either at the time of executing the construction contract or at the time
915	A bill to require load lines on United States vessels engaged in foreign voyages and foreign vessels within the jurisdiction of the United States, and for other
1813	To amend the charter school program under the Elementary and Secondary Education Act of 1965.
1745	A bill to prohibit the removal before January 1, 1982, of houseboats or floating cabins from water resource areas under the jurisdiction of the Corps of
206	A bill to amend the Agricultural Credit Act of 1987 to reauthorize State mediation programs.
911	To amend chapter 9 of title 11 of the United States Code to improve protections for employees and retirees in municipal bankruptcies.
1316	A bill to designate certain lands in the Bosque del Apache National Wildlife Refuge, Socorro County, N. Mex., as wilderness.
0	A bill to authorize a study of the feasibility and desirability of establishing a national recreation area to be known as the Santa Margarita National Recreation Area in the area in San Diego County, Calif., which presently
468	To provide a schedule for the implementation of the requirement that all eligible Federal prisoners desiring to participate in drug treatment programs be given such
1630	To create a corporation to restore normal employment in labor surplus areas of group IV classification and produce industrial development in such
713	A bill to increase annuities under theRailroad Retirement Act of 1937 by 25 percent, and for other purposes

Document Cluster 0

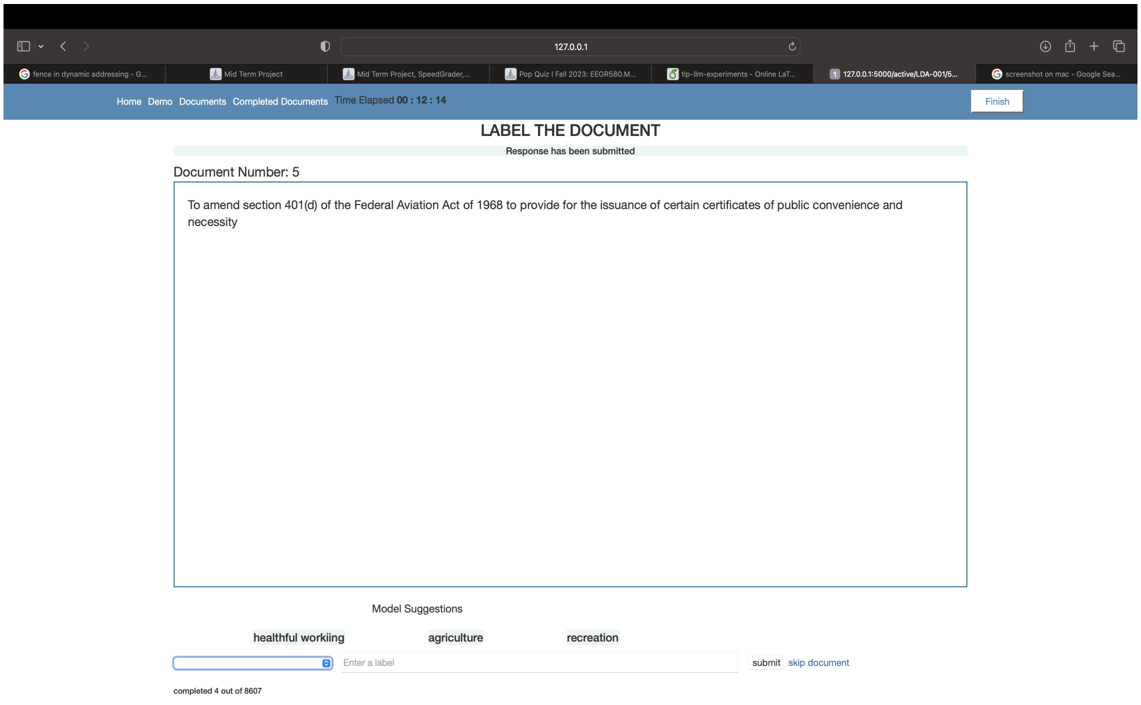
price agricultural marketing support commodity farm grain tobacco allotment crop acreage adjustment wheat milk feed cotton quota stabilize surplus disaster

Number	Document
1168	A bill to establish the Controlled Substances Administration to administer all Federal programs relating to the regulation of narcotic and other dangerous drugs, the treatment and rehabilitation of abusers of such drugs, and
566	To amend title 23, United States Code, to encourage the use of advanced technologies with respect to transportation projects that receive Federal funding, and for other
1950	A bill to designate the Aldo Leopold Wilderness, Gila National Forest, New Mexico
2195	To prohibit transportation in interstate or foreign commerce of articles to or from the United States aboard certain foreign vessels, and for other
1883	A bill to modify the boundaries of the Santa Fe, Gila, Cibola, and Carson National Forests in the State of New Mexico, and for other purposes

Fig. 11. The non-active learning document list shows documents within each document cluster.

represents the user's name extracted from the URL, and `document_id`, which represents the ID of the document being labeled.

The document annotation page with active learning is shown in Fig. 12. Relevant information about the document, such as its text and previous labels, is fetched from the `all_texts` dictionary based on the provided `document_id`. The function also selects a random document ID from the remaining recommended block to serve as a skip option. In case of a POST request (when the user submits labels), the function records the user's response time and saves the labeled data. It then sends a request to the model to recommend the next document for annotation. The user's session data, including labeled documents and labels, are updated, and the user is redirected to label the next document. The function uses these parameters and data to dynamically generate the active learning interface (`activelearning.html`) where users can interact with the document, provide labels, and navigate through the active learning process. Flash messages are used to notify users when their response has been successfully submitted. It is important to note that the function relies on the model and helper functions (`save_time`, `save_response`, `save_labels`, and `all_texts`) for complete functionality.


127.0.0.1
fence in dynamic addressing - G...
Mid Term Project
Mid Term Project, SpeedGrader,...
Pop Quiz I Fall 2023: EEGR580.M...
tlp-llm-experiments - Online LaT...
127.0.0.1:5000/active/LDA-001/5...
screenshot on mac - Google Sea...
Home
Demo
Documents
Completed Documents
Finish
Model Suggestions
healthful workiing
agriculture
recreation
Enter a label
submit
skip document

Time Elapsed 00 : 12 : 14

LABEL THE DOCUMENT

Response has been submitted

Document Number: 5

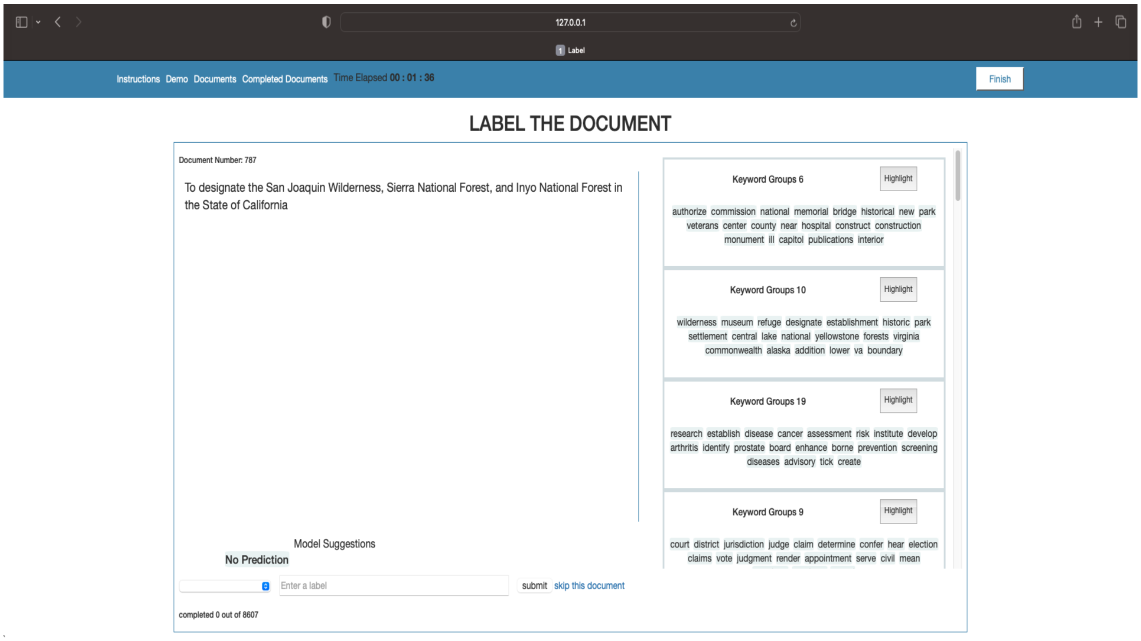
To amend section 401(d) of the Federal Aviation Act of 1968 to provide for the issuance of certain certificates of public convenience and necessity

completed 4 out of 8607

Fig. 12. The active learning page for annotating (labeling) documents showing document ID and text, with suggested labels at the bottom.

Appendix A.1.6. Labeling without active learning (Non-active label)

The `non_active_label` function in the Flask application manages the annotation process for documents shown in an environment without active learning. It handles both GET and POST requests, allowing users to view and annotate specific documents. It displays keywords according to a rank giving by the model. These keywords help a user to get the context of the topic by highlighting the keywords present in the text. The user interface presents the document's text, along with suggestions for annotation. Users can select a label from the provided suggestions or enter a custom label. Additionally, users can skip the current document and proceed to the next one. When the user submits a label, the function processes the input, records the response time, and communicates with the model recommend the next document for labeling. The function then redirects the to the non-active label page with updated information about the next document to annotate. Throughout the process, the function keeps track of various metrics, such as elapsed time, the number of labeled documents, and the total number of documents to be annotated. These metrics are displayed on the user interface, providing users with real-time feedback on their progress.



Time Elapsed 00 : 01 : 36

LABEL THE DOCUMENT

Document Number: 787

To designate the San Joaquin Wilderness, Sierra National Forest, and Inyo National Forest in the State of California

Keyword Groups 6

authorize commission national memorial bridge historical new park veterans center county near hospital construct construction monument ill capitol publications interior

Keyword Groups 10

wilderness museum refuge designate establishment historic park settlement central lake national yellowstone forests virginia commonwealth alaska addition lower va boundary

Keyword Groups 19

research establish disease cancer assessment risk institute develop arthritis identify prostate board enhance borne prevention screening diseases advisory tick create

Keyword Groups 9

court district jurisdiction judge claim determine confer hear election claims vote judgment render appointment serve civil mean

Model Suggestions

No Prediction

Enter a label

completed 0 out of 8607

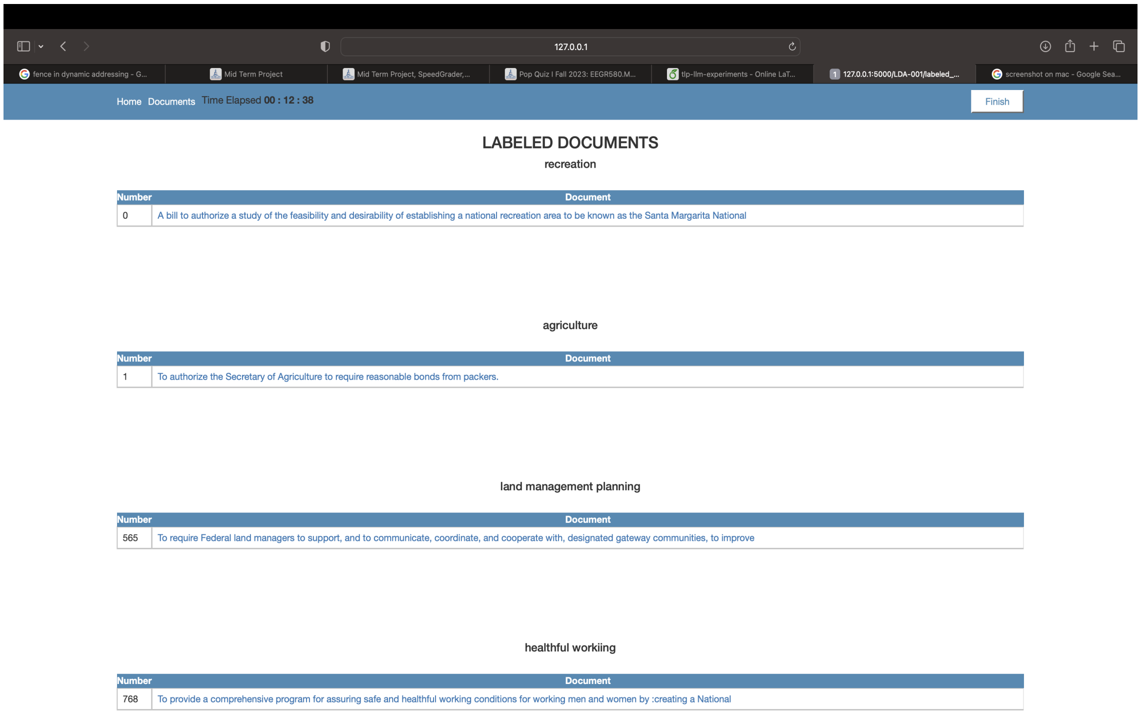
Fig. 13. The document annotation page in an environment without active learning, including top topics and associated keyword suggestions for labels on the right.

Appendix A.1.7. Document re-labelling

Users have the flexibility to edit labels for previously annotated documents through a straightforward process. The system presents a user-friendly interface where labeled documents are grouped by assigned labels. When users choose to edit responses, they are seamlessly guided to the relevant annotation page designed for the selected document. Users can then select the document to view the whole document text and its associated label. If a user wants to modify the labels, there is an edit response option that can be selected. Upon selecting "Edit response," users are then redirected to the appropriate document label page to assign a new label to the document.

Appendix A.1.8. Authorization

The authorization process in the annotation application is designed to secure user access and ensure that only authenticated individuals can interact with the system. The process begins when a user attempts to log in by providing their credentials, typically a `username`. This credential is then verified against stored user data, which includes information such as usernames. If the provided credential matches an entry in the system, the user is successfully authenticated, and the authorization process proceeds. Upon successful authen-


127.0.0.1
fence in dynamic addressing - G...
Mid Term Project
Mid Term Project, SpeedGrader,...
Pop Quiz I Fall 2023: EEGR580.M...
tlp-llm-experiments - Online LaT...
127.0.0.1:5000/LDA-001/labeled_...
screenshot on mac - Google Sea...
Home
Documents
Finish

Time Elapsed **00 : 12 : 38**

LABELED DOCUMENTS

recreation

Number	Document
0	A bill to authorize a study of the feasibility and desirability of establishing a national recreation area to be known as the Santa Margarita National

agriculture

Number	Document
1	To authorize the Secretary of Agriculture to require reasonable bonds from packers.

land management planning

Number	Document
565	To require Federal land managers to support, and to communicate, coordinate, and cooperate with, designated gateway communities, to improve

healthful workiing

Number	Document
768	To provide a comprehensive program for assuring safe and healthful working conditions for working men and women by :creating a National

Fig. 14. This page displays documents that have already been labeled and allows the user to reassign labels to previously annotated documents by clicking on the document to re-enter the relevant document-annotation page.

tication, the application initializes a user session. A session is a way to track and manage the state of interactions by a user with the application during a specific period. During this session initiation, critical information about the user is stored in the session variables. This information may include the user's name, a unique user ID, labels associated with the user, and a record of previously labeled documents. The user's identity and session information are then stored securely to maintain the user's context and permissions as they navigate through different parts of the application.

Appendix A.1.9. Time Tracking

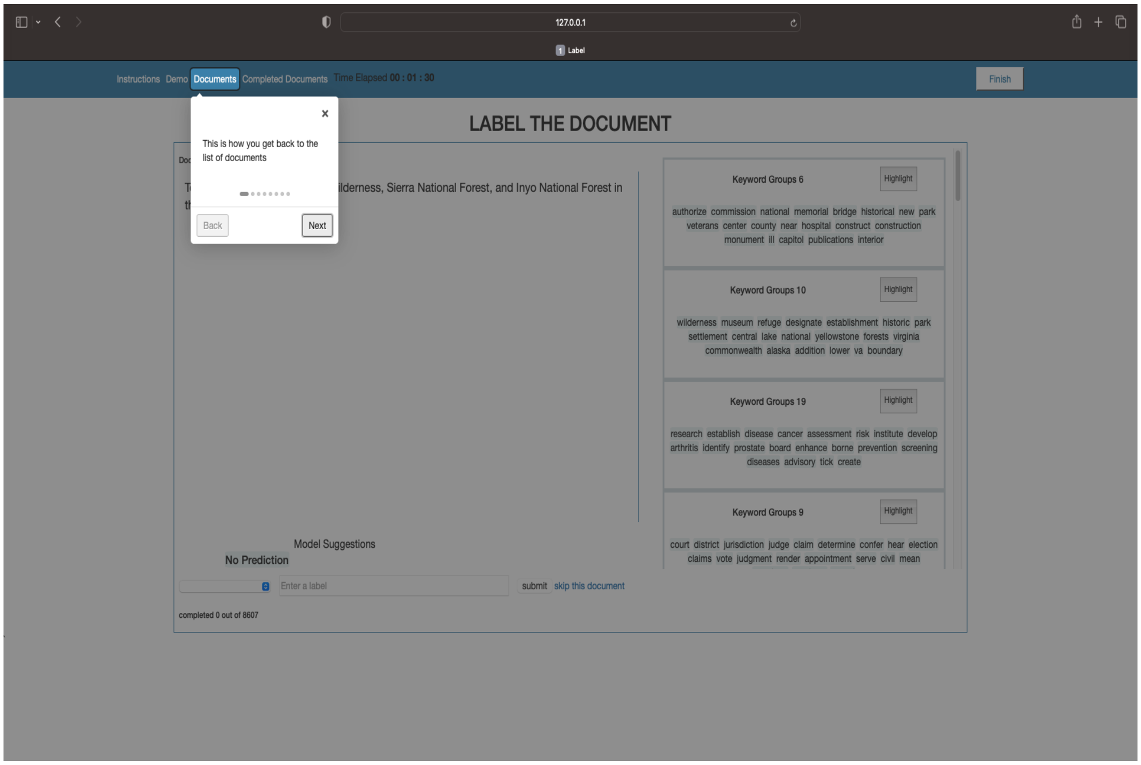
Time tracking in the annotation app plays a crucial role in monitoring and recording the user's interactions and activities throughout their session. It involves capturing and storing timestamps at various points during the user's engagement with the application, providing valuable insights into the time spent on specific tasks. The tasks may be selecting documents to annotate, finding the right label for a document, and editing labels.

The following are the key aspects of time tracking in the document annotation app:

- Session Start: The time tracking process begins when a user initiates a session by logging into the application. The moment of login is recorded as the starting point for tracking the user's activity.
- Task Duration Measurement: During the session, the app records the time spent on different tasks, such as viewing documents, labeling, and navigating between pages. This helps researchers understand how much time is dedicated to each activity. The system records the time taken to annotate a document, view a topic, or complete an annotation task. This data assists in evaluating efficiency and productivity under specific experimental conditions in our user studies.
- Elapsed Time Calculation: The application calculates the elapsed time by comparing the current timestamp with the session start time. This information is often displayed to users, providing a real-time view of their session duration.

Appendix A.1.10. App tutorial

For first-time users logging into the annotation application, the on-boarding process is enhanced through the integration of 'intro.js'. This feature serves as a guided tour, systematically introducing users to the various functionalities and buttons within the application. By providing informative highlights and tool tips, 'intro.js' ensures a user-friendly and navigable experience, offering valuable insights into the diverse actions that can be performed within the application. This on-boarding mechanism significantly contributes to user engagement and understanding, empowering users to make the most of the app's features right from the start.



Time Elapsed 00 : 01 : 30

LABEL THE DOCUMENT

This is how you get back to the list of documents

Doc

T ilderness, Sierra National Forest, and Inyo National Forest in th

Keyword Groups 6

authorize commission national memorial bridge historical new park veterans center county near hospital construct construction monument ill capitol publications interior

Keyword Groups 10

wilderness museum refuge designate establishment historic park settlement central lake national yellowstone forests virginia commonwealth alaska addition lower va boundary

Keyword Groups 19

research establish disease cancer assessment risk institute develop arthritis identify prostate board enhance borne prevention screening diseases advisory tick create

Keyword Groups 9

court district jurisdiction judge claim determine confer hear election claims vote judgment render appointment serve civil mean

Model Suggestions

No Prediction

Enter a label

completed 0 out of 8607

Fig. 15. At any time, users can click through an optional tutorial to learn how to navigate and use the interface. The tutorial is shown upon entering the application and exiting the instructions page.

Appendix B. List of Symbols, Abbreviations, and Acronyms

The following acronyms are used in this report.

ALTO Active Learning with Topic Overviews.

ANMI Adjusted Normalized Mutual Information.

ARI Adjusted Rand Index.

CTM Contextualized Topic Model.

HCI Human-Computer Interaction.

LDA Latent Dirichlet Allocation.

LLM Large Language Model.

NLP Natural Language Processing.

NTM Neural Topic Model.

RAS Resilience, Adaptation, and Sustainability.

sLDA Supervised LDA.

TENOR Topic-Enabled Neural Organization and Recommendations.