# Evaluating Interactive Topic Models in Applied Settings

Sally Gao
sally.gao@thomsonreuters.com
Thomson Reuters Labs
Boston, MA, USA

Milda Norkute
milda.norkute@thomsonreuters.com
Thomson Reuters Labs
Zug, Switzerland

Abhinav Agrawal
abhinav.agrawal@thomsonreuters.com
Thomson Reuters Labs
Bengaluru, Karnataka, India

## ABSTRACT

Topic modeling is a text analysis technique for automatically discovering common themes in a collection of documents. "Human-in-the-loop" topic modeling (HLTM) allows domain experts to steer and adjust the creation of topic models. In this case study, we use a custom-built HLTM interface to assess the impact of human refinement on model interpretability and predictive performance in collaboration with an analytics team within our organization. Using a small dataset ($\approx 12k$ documents) of responses drawn from an organizational employee satisfaction survey, we compare the pre- and post-refinement models using both human judgments and automated metrics. We find that human refinement can enhance interpretability and predictive performance in some cases, but may lead to overfitting on the training data, which negatively impacts model quality. Furthermore, we observe that existing evaluation methods don't sufficiently and clearly capture topic model quality in applied settings, and propose guidance for further HLTM tool development.

## CCS CONCEPTS

• **Human-centered computing** → **Systems and tools for interaction design**; • **Information systems** → **Document topic models**.

## KEYWORDS

Interactive topic model, Interactive machine learning, Interpretability, Explainability, Topic modeling, Human in the loop

**ACM Reference Format:**
Sally Gao, Milda Norkute, and Abhinav Agrawal. 2024. Evaluating Interactive Topic Models in Applied Settings. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3613905.3637133

## 1 INTRODUCTION

Topic modeling is an unsupervised statistical modeling technique used to discover latent "topics" within a collection of documents. The most widely-used approach is Latent Dirichlet Allocation (LDA), which models a topic as a distribution over a vocabulary, and each document as a distribution over the topics [3, 10]. A user can get a sense of the themes present in the corpus by examining the most prominent words and topics outputted by the model [6, 19]. Additionally, topic models can be used as predictive models on unseen documents [4, 22].

A good topic model consists of meaningful and well-differentiated topics that, taken together, accurately represent the document collection as a whole. However, due to the complexity of the task, topic models often return imperfect results, or generate topics that do not fully align with intended modeling goals [7, 13, 14]. For instance, topics can be incoherent, missing, duplicated, or contain multiple themes instead one. Human-in-the-loop (HLTM) topic modeling, also known as interactive topic modeling, aims to address this problem [1, 8, 11, 13, 16, 23, 25, 26]. It allows end users who aren't experts in topic modeling algorithms to encode their knowledge into the topic model through direct iterative refinement.

This is a promising approach to creating high-quality topic models in applied settings. Incorporating human feedback resolves user concerns about mismatches between anticipated and resultant topics. To facilitate interaction, the topics are visualized in a user interface (UI), emphasizing explainability and building trust in results. HLTM promotes the use of topic models in diverse industry applications such as sentiment analysis for market research, identifying common issues in customer support data, and tracking topics on social media, to name just a few examples.

However, existing research on HLTM has thus far only occurred under experimental settings, using students or crowdsourced workers as test subjects [8, 13, 16, 23, 24]. To be truly useful, it must be shown that HLTM systems perform well in real-world applied settings, where it is important that a topic model is able to perform well on previously unseen data, and also be meaningful and interpretable to end users.

In this case study, we compare two human-refined models with a pre-refinement baseline model. Our purpose is to investigate the following questions:

(1) Does human refinement via HLTM result in topics that are interpretable and thematically distinct?
(2) Do topic models refined via HLTM perform well on new data?
(3) How do users interact with HLTM in applied contexts?

To assess model interpretability, we use word intrusion, a well-established human evaluation task in the topic modeling literature [7]. In addition, drawing on recent ideas that link topic interpretability with the ability to label topics [9, 18], we ask users to label every topic in their refined models, and measure label concurrence as part of the human evaluation.

To assess model performance on downstream tasks, we use an automated metric, Normalized Pointwise Mutual Information (NPMI) [15], as well as a human evaluation task called topic identification, a modified version of the topic intrusion task developed by Chang et al. [7] (see details in Section 5.2). In both cases, we use a held-out


*CHI EA '24, May 11–16, 2024, Honolulu, HI, USA*

ACM ISBN 979-8-4007-0331-7/24/05.
https://doi.org/10.1145/3613905.3637133

test set (a reserved set of documents not used in the model creation and refinement process) to assess model quality.

Based on the findings of this limited-scope case study, HLTM refinement can sometimes improve predictive performance on new data and results in topics that have high label agreement, suggesting high interpretability. However, we also find that human refinement may result in overfitting on the existing data, resulting in poor classification performance on the held-out set. In addition, user interviews revealed the need for greater emphasis on document-topic assignments rather than topic-word lists, with users using expressing the desire to see which words within a document are associated with a certain topic. When it comes to evaluation methods, we advocate for greater emphasis on held-out documents, rather than using topic-word lists. Lastly, we report on user strategies and challenges faced during the refinement process, and uncover some potential directions for future HLTM tool development.

## 2 USE CASE

For our study, we collaborated with four members of an analytics team within our organization that provides workforce metrics, reports, and data visualizations on employee sentiment across the organization. As part of their work, this team periodically collects data through employee surveys. Analysis of textual survey responses involves manual compilation of keywords by the analytics team, which is laborious and time-consuming. As these surveys are conducted on a regular periodic basis, there is a need for a topic model to perform well on future data. In addition, the team has found existing third-party topic modeling solutions unflexible and unsatisfactory, making this an ideal use case for applied HLTM.

To investigate whether a custom-built HLTM solution would better serve the team's needs, we used an anonymized dataset of free-form text responses to the question: "How likely are you to recommend Thomson Reuters to a friend or family member?"

We split the dataset into a train set ($N \approx 10k$), which we use to create the baseline (pre-refinement) model, and a held-out test set ($N \approx 2k$), which we use for both human and automatic model evaluation. The train set comprises survey responses collected between March 2020 through February 2022, while the test set spans March and April 2022. The average survey response was 35 words long.

We recruited four study particiants, all members of the aforementioned analytics team. Two of them were tasked with refining the baseline model, while the other two were tasked with evaluation. All participants were aged between 25-40 and had worked as analysts at Thomson Reuters for 3-8 years. All had some previous familiarity with statistics, but not topic modeling.

## 3 HLTM TOOL

### 3.1 Interface

We present an example screenshot of our UI and its associated functionalities in Fig. 1. Topics initially appear with a generic label (e.g. "Topic 1"), which users can edit. Selecting a topic allows a user to see the words and documents associated with a topic in greater detail.[1] Each word is followed by a number in parentheses, which denotes the count of the word in the topic. Users can paginate through both words and documents, and undo up to 10 previous actions.

[1]Our vocabulary also includes two- and three-word terms, also known as bigrams and trigrams, but for consistency we refer to all terms as "words" throughout this paper.

### 3.2 HLTM Implementation

We implement seven refinement operations commonly found in the HLTM literature: **Promote Word**, **Demote Word**, **Add to Stopwords**, **Demote Document**, **Merge Topics**, **Split Topic**, and **Create Topic**. We deviate from previous work [14, 23, 24] by using the words "promote" and "demote" instead of "add" and "delete". This was done to reduce confusion and to reflect the fact that the membership of a word or a topic within a model is probabilistic rather than binary.

*3.2.1 Gibbs Sampling.* We use the collapsed Gibbs sampling algorithm proposed by Griffiths and Steyvers [10], which is favored by HLTM developers due to its low latency [24]. For document collection $D$, we have $T$ topics, each represented by a multinomial distribution over the vocabulary $V$. Under the hood, the topic model comprises two matrices: $\theta$, a $D * T$ matrix where $\theta_d$ represents the topic distribution for document $d$, and $\phi$, a $T * V$ matrix where $\phi_t$ represents the word distribution of topic $t$. The probability of a topic assignment $z = t$ given observed token $w$ in document $d$ is:

$$P(z = t|z\_, w) \propto (n_{d,t} + \alpha)\frac{n_{w,t} + \beta}{n_t + V\beta}$$

where $z\_$ are the topic assignments of all the other tokens, $n_{d,t}$ is the count of tokens in $d$ assigned to $t$, $n_{w,t}$ is the count of $w$ assigned to $t$, and $n_t$ is the count of all tokens assigned to $t$.

*3.2.2 Refinement Implementation.* Unlike previous HLTM implementations [14, 23], we maintain symmetric priors and directly modify the topic-word count $n_{w,t}$ or the document-topic count $n_{d,t}$. Then, we run the Gibbs sampler on a restricted set of words to obtain new topic assignments, and update $\phi$ or $\theta$ accordingly.

The advantage of this approach is that the model always aligns with the underlying data. Compared to the informed prior method, our implementation favors modeling the input corpus faithfully over satisfying user expectations, resulting in lower *user control* [14].

Our procedures for each refinement are described below:

**Promote word $w$ in topic $t$** Temporarily set $n_{w,t}$ to the highest topic-word count in $t$. For every instance of $w \notin t$, sample a new topic $z$. Update $\phi_{t,w}$ to reflect topic $t$'s new word distribution.

**Demote word $w$ in topic $t$** Temporarily set $n_{w,t}$ to zero and sample a new topic $z$ for every instance of $w \in t$. Update $\phi_{t,w}$ as above.

**Add to stopwords** Set $\phi_{t,w}$ to zero for every $w$ and remove every instance of $w$ from every document $d$, along with their corresponding topic assignments.

**Demote document** Set $n_{d,t}$ to zero. For every word $w \in t$, sample a new topic, repeating until there are no more words in $d$ assigned to $t$.

**Merge topics** We follow the procedure described by Smith et al. [23], and minimize the number of re-assignments necessary by always deleting the smaller of the two topics.

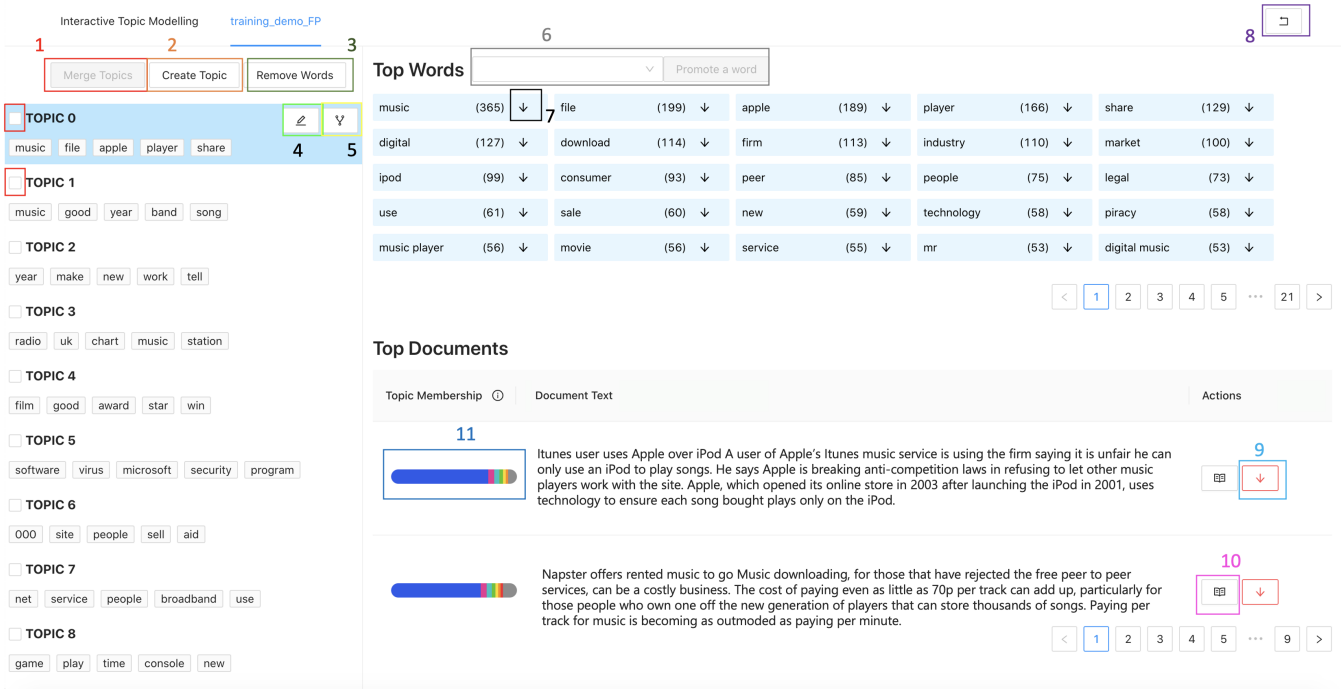



**Figure 1: Screenshot of the user interface of our HLTM tool. The numbered boxes correspond to the following functionalities: 1 - Merge Topics; 2 - Create Topic; 3 - Remove Word (Add to Stopwords); 4 - Rename Topic; 5 - Split Topic; 6 - Promote Word; 7 - Demote Word; 8 - Undo; 9 - Demote Document; 10 - View Full Document. 11 - Topic Membership (Users can see which other topics are present in a topic by hovering over this bar.)**

**Split topic $t$ based on list of seed words $s$** Create a second list $\hat{s}$, which comprises every word among the top 30 words in $t$ not in $s$. Move every word $w \in s$ to the new topic, $t_n$. Every word $w \in \hat{s}$ must remain in the original topic $t$, maximizing the model's ability to return two well-differentiated topics. Run the Gibbs sampler for several iterations over every remaining word $w \notin (s \cup \hat{s})$ in $t$, forcing it to only sample from $t$ or $t_n$.[2]

**Create topic $t_n$ from list of seed words $s$** Move every word $w \in s$ to $t_n$. Run the Gibbs sampler for several iterations, allowing the model to add new (related) words to $t_n$. Stop early if $n_{w,t_n}$ exceeds twice the number of initial words obtained from $s$.

## 3.3 Post-Interaction Messages (PIM)

Due to the challenge of striking a balance between applying user changes and preserving the statistical validity of the model, HLTM can sometimes result in what users perceive as unpredictable or unexpected behavior [14, 23, 24]. To mitigate this, we introduce a set of post-interaction messages (PIM) to the UI to alert the user whenever an interaction results in a major change in another topic. The goal is to help the user understand the secondary impacts of their most recent action, in addition to whatever local changes they intended. We implement PIM for Promote Word, Demote Word and Create Topic. We chose these interactions as they are the most likely to have moderate-to-large impacts other topics. Example PIM messages are shown in Fig.2.

To determine whether a given refinement to topic $t$ has caused a potentially significant change in another topic $t'$, we compare the most frequently-occurring words in each topic $t'$ before and after refinement. For each interaction, we display a PIM if the following conditions are met:

**Promote Word** If $w$ is within the top 20 words for any $t'$ prior to the interaction and its rank in $t'$ drops as a result of the interaction.

**Demote Word** If $w$ is within the top 20 words for any $t'$ prior to the interaction and its rank in $t'$ increases as a result of the interaction.

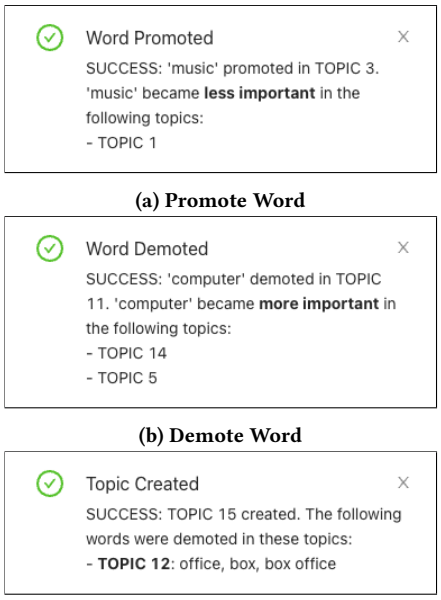
**Create Topic** If any $w$ in any $t'$ drops out of the top 20 words for $t'$ as a result of the interaction.

# 4 USER STUDY

## 4.1 Procedure

We obtain our baseline model by running Latent Dirichlet Allocation [3] on the train set with multiple values for $T$, and validate the vocabulary and initial number of topics with members of the analytics team described in Section 2. Then, using the model's document-topic matrix and topic-word matrix to obtain $n_{d,t}$ and $n_{w,t}$, we use the Gibbs sampling algorithm over the entire corpus

[2]Unlike the approach proposed by Pleple [21], which requires the user to provide two sets of seed words, this approach maximizes the model's ability to return two well-differentiated topics given a only single set of seed words.




**(a) Promote Word**

**(b) Demote Word**

**(c) Create Topic**

**Figure 2: Examples of post-interaction messages displayed to the user.**

for three iterations (omitting the update step) to obtain a word-level, count-based representation of the model.

The two refiners were introduced to topic modeling concepts and trained to use the tool before the refinement task. We phrased our refinement instruction as follows: *"You will be presented with a machine-generated topic model. Based on your knowledge of the data, please refine the topics so that the topics are: coherent, distinctive and aligned with your understanding of themes present in the data."* In addition, we required them to give each topic a custom label.

The participants were told to refine the models individually and in their own time, and to inform the researchers once they were done. Overall, the users reported spending between 3 and 6 hours on the task. We then scheduled individual follow-up interviews about their experiences with the tool. A single researcher conducted, transcribed and analyzed the interviews. Using the transcriptions, the researcher used thematic analysis [5] to create a coding schematic, then clustered the user comments using affinity diagrams. This revealed themes which were presented and discussed with the other researchers before the final themes were agreed on.

## 4.2 User Feedback

*4.2.1 Strategy.* Both users reported using similar refinement strategies. First, they went over the list of topics, looking only at top keywords, labeling them and deciding what the topic would be about. We found that the labeling feature helped users stay organized during the refinement process: *"It made sense to label topics first, seeing what the themes were already in the topics that were auto-generated"*. Next, they performed bigger actions like merging or splitting topics, followed by promoting or demoting words. This strategy was partially shaped by a desire to bring down the number of topics in the baseline model: *"My main criteria was to get the topics down to a number of manageable topics, so I wasn't going the other direction at this point in making more topics."* We found that both users preferred to handle a smaller number of topics than the 20 present in the baseline model.

*4.2.2 Need for collaboration.* Both users said they were done refining the model once they felt that they had created something usable. However, both believed their models could still be further improved by making further changes to the model in collaboration with their colleagues: *"I could probably further improve it if I had consulted with [a team member] who might think I missed some important topic."*

*4.2.3 Model explainability.* The users were sometimes confused by why certain words or documents were associated with certain topics. They saw value in being able to surface documents within a topic based on a word, or being able to see which words were associated with a topic within a document. Both users described wanting more document-level information, but handled their uncertainty differently. One user said: *"The context of that particular document didn't fit with my theme, but I couldn't find the word that was pulling it up to the front. I ended up just demoting the document so it didn't appear in the top ten."* Faced with a similar situation, the other user to hesitated over a refinement they wanted to perform, but ultimately did not: *"I found myself wanting to click on that word and have those three instances show up, because I didn't really understand how it was related to [my topic]. Without knowing exactly how that term was being used in the documents, I didn't want to demote it."*

*4.2.4 Control.* Despite the lack of explainability, the undo and post-interaction messages (PIM) features helped users feel in control. One user reported *" I got [...] I don't know if it's happiness or closure, when one of those [PIM] messages popped up. 'Because you did this, now that word is showing up more in this other topic.' That was a nice validation, because that's what I wanted to happen., I knew that if I demoted a word I could promote it back if I wanted to."* In addition, despite using undo infrequently, users felt reassured by the presence of the undo feature: *"I knew that if I demoted a word I could promote it back if I wanted to, or if I accidentally removed a word I could bring it back some way, so it didn't feel like I had to be like super careful. The undo button was very valuable."*

*4.2.5 Trust.* During their interviews, both users said that they trusted the tool, and also the underlying model to some extent. They viewed it as a useful partner that could make suggestions about which words should or shouldn't belong to certain topics: *"the machine might be better at this than I was, so maybe I would have never picked this word for this topic, but maybe I would have been wrong"*.

*4.2.6 Feedback on PIM..* The post-interaction messages were perceived very positively. In some instances, they piqued user curiosity and encouraged them to explore the effects of their actions on other topics: *"I remember either promoting or demoting one word had an impact to [another] topic. That got me interested and I wanted to see what actually happened. So I ended up checking the other topic to make sure, because I think when you promote a word it sometimes demotes it in another topic."*

# 5 MODEL EVALUATION

Throughout this section, we refer to the baseline model as $m_{initial}$ where the number of topics k=20, and the two user-refined models as $m_{R1}$ (k=12) and $m_{R2}$ (k=13).

## 5.1 Automatic Evaluation

Normalized pointwise mutual information (NPMI) is a widely-used automatic coherence metric that has been shown to correlate with topic interpretability [15, 20]. A topic has a high NPMI if its top N words tend to co-occur more than we would expect from random chance, based on joint word probabilities obtained from a reference corpus. Intuitively, NPMI can also be a signal of downstream predictive performance when a held-out test set is used as the reference corpus. We calculate NPMI using both the train and tests set as reference corpus, using the top 20 words from each topic and a window size of 10.[3]

As NPMI score ranges vary depending on the reference corpus, we present the difference in NPMI between the refined models and the pre-refinement baseline rather than absolute scores in Fig. 3. An difference greater than zero indicates that the refined model outperforms the baseline with respect to the reference corpus. Our results show this is the case for both of the refined models on both the train and test sets, indicating that the human refiners successfully used their domain expertise to improve the baseline model. We observe that while the train set slightly favors $m_{R1}$, this does not translate to better performance on the test set, where $m_{R2}$ is slightly superior. This suggests that $m_{R1}$ is somewhat overfitted, that is, overly tailored to the training set at the expense of general performance.

## 5.2 Human Evaluation

In this section, we describe three human evaluation tasks that we used to compare the pre- and post-refinement models. To assess whether the refiners' models are interpretable to their colleagues, we used different members of the same analytics team as evaluators.

*5.2.1 Word Intrusion.* This task evaluates thematic coherence by measuring how well a human evaluator can associate a topic label with its top words [7]. The evaluator is presented with five high-probability words from a topic, as well as a low-probability word that has a high probability in another topic. The goal of the task is to identify the word that does not belong. Model precision *MP* for a given topic is the fraction of evaluators agreeing with the model. This is averaged across the topics to obtain a model score.

*5.2.2 Topic Identification.* This task measures how well a model assigns topics to documents. We modify the topic intrusion task devised by Chang et al. [7], which requires at least three high-probability topics to be available for every document. Due to the brevity of the documents in our use case, most primarily belong to a single topic. We present a single high-probability topic and three randomly sampled low-probability topics, and ask the evaluator to identify the topic that is *most* representative of the document (hence "identification" instead of "intrusion"). To make the task more challenging and to measure the models' performance on downstream data, the documents were drawn from the held-out set.

We define topic log odds ($TLO_i$) for document $d$ and subject $s$ as the difference in the log-probability of the answer selected by $s$ and the log-probability of the "correct" answer according to the model. Like the original *TLO* for topic intrusion defined by Chang et al. 7, the upper bound of $TLO_i$ is still zero.[4]

*5.2.3 Label Concurrence.* In addition to the tasks described above, which are considered standard for human evaluation, we introduce an additional task to assess interpretability. This task measures how well a human evaluator can associate a topic label with its top words. Given a topic label produced by the model refiner, we present four topics, each represented by their top five highest-probability words. The evaluator is asked to match the label to its corresponding topic. Label Accuracy (*LA*) is simply the total number of correct answers divided by the number of questions.

In designing this task, we were motivated by recent ideas that link topic model interpretability with the ability to label topics and consensus on the topic labels [9, 18]. The task is also a signal of the distinctness of each topic, which is relevant because topic duplication is a frequent failing of traditional topic models [13], whereas a good topic collection contains unique topics.

*5.2.4 Procedure.* For each task, we create 10 questions per model and assign every question to both evaluators. For all tasks, we present the candidate answers in random order. To prevent the evaluators from learning the topic-word lists ahead of the word intrusion task, we instructed them to do in the tasks in the following order: Word Intrusion, Topic Identification, Label Concurrence.

| **Model** | *LA* | *MP* | $TLO_i$ |
|---|---|---|---|
| $m_{initial}$ | - | 0.5 (90%) | -3.452 (90%) |
| $m_{R1}$ | 0.85 (80%) | **0.7** (80%) | -8.056 (30%) |
| $m_{R2}$ | **0.95** (80%) | .45 (60%) | **-1.151** (90%) |
| (Avg.) | (**90**%) | (63%) | (76%) |

**Table 1: Mean Label Accuracy, Model Precision and Topic Log Odds scores across the three models. Annotator agreement percentages are displayed in parentheses.**

*5.2.5 Results.* We report the mean value of the associated metric for each task in Table 1, including statistics on annotator agreement for each task in parentheses.

On the label concurrence task, both refined models exhibit high accuracy and annotator agreement, showing that the refiners were able to produce topics whose themes were understandable to their colleagues.

Results for word intrusion and topic identification present a notable dichotomy. $m_{R1}$ scored highest on word intrusion, but performed worse than the unrefined baseline on topic identification. For $m_{R2}$, we observe the reverse: it achieves the best topic identification score, but scores just below the baseline model for word intrusion.

[3]Our vocabulary includes bigrams and trigrams, while NPMI is designed for unigrams. To account for inflated co-occurrence counts due to overlapping tokens in bigrams and trigrams, we eliminate any terms that overlap with unigram words.

[4]To avoid taking the log of 0, we treat doc-topic memberships of 0 as 1e-10. This imposes an artificial lower bound of approximately -23.025.

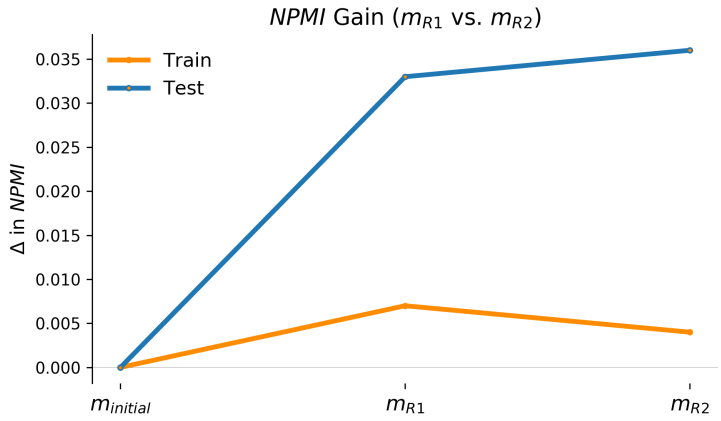



**Figure 3: NPMI for all models subtracted by the NPMI for the baseline model.**

While the NPMI results discussed in Section 5.1 suggested that $m_{R1}$ was only slightly better than $m_{R2}$, the topic identification task reveals that in fact, $m_{R1}$'s performance is severely limited compared to $m_{R2}$ when faced with new data. The low annotator agreement for $m_{R2}$ corroborates this, revealing that for each held-out document, none of the topics was a good fit.

Given that we find a reverse relationship between word intrusion and predictive performance, this calls into question whether word intrusion is a reliable task for topic model evaluation. Unlike topic identification, the word intrusion task is based solely on the word-topic list created by the model refiner, and consequently can result in a high score even for a severely overfitted model.

## 6 CONCLUSIONS

Throughout this study, our use case partners were consistently enthusiastic about the HLTM tool both inside and outside of formal interview sessions. We take this as a promising sign of the value of HLTM in applied settings. However, our study also revealed some key gaps and challenges for HLTM going forward.

A high-quality model should satisfy two requirements: it must be able to assign appropriate topics to future documents (i.e. be performant), and its topics must be adequately meaningful to end users (i.e. be interpretable). Our findings affirm the potential of HLTM when using domain experts as refiners in an applied setting and on a small dataset. Our users were able to create topics that were distinct and thematically well-defined, and the resulting models can perform well on unseen data, although not always.

The mixed results across our different evaluation approaches expose a key challenge when it comes to assessing topic models — namely, which evaluation metric to trust? Choosing the right evaluation approach is important because model selection can be especially challenging in applied contexts. In experimental settings, researchers typically rely on crowdsourcing to achieve enough statistical power to draw robust conclusions [7, 12, 17, 18]. However, this may not be not viable in business settings due to data privacy concerns and crowdworkers' lack of subject matter expertise. The sample size of our study is a more realistic scenario in a business context, emphasizing the need to select suitable evaluation tasks while rejecting unreliable ones.

We consider topic identification our most robust task, as it directly measures a model's ability to classify documents that were not seen in the training data, as opposed to indirectly capturing model quality by matching top words (as in word intrusion), or by comparing co-occurrence of top words (as in NPMI). We heavily question the efficacy of assessing topics solely based on top words. The idea that an interpretable top-words list does not necessarily lead to good document-topic assignments is supported by previous work [2, 17]. The problem is more severe for word intrusion, which only considers a topic's top five words and is unable to take unseen data into account. On the other hand, NPMI allows for a larger sample of top words and can be applied to a held-out set. When using the test set as reference corpus, NPMI scored the more performant model more highly, but nonetheless overstated the quality of the non-performant model in our case.

In light of these findings, we are drawn to the growing body of work that questions the prevailing trend of evaluating topic models based only on lists of top words, and instead propose focusing on evaluating document-topic assignments or even individual word assignments within documents to assess model quality [2, 9, 17, 18]. In addition, given the risk of overfitting during the refinement process, we emphasize the need to evaluate topic models based on held-out validation sets as opposed to directly on the input corpus, which is not uncommon according to Hoyle et al. [12].

Moving away from topic-word lists in favor of documents is supported by findings from our user interviews, which surfaced the need for more explainability at the document level. This provides an opportunity to mitigate excessive user trust. Although HLTM is frequently extolled as user-friendly and accessible, anecdotally we find that the statistical theory behind topic modeling is a challenge for adoption. Users sometimes deferred to machine judgments against their own intuition, hesitating over decisions like removing irrelevant terms from topics. Other times, they refined the models in unexpected ways. For instance, we were surprised to find that words like "really" were not demoted from a topic's top words,

as we do not consider such words thematically meaningful. Our findings in Section 4.2 suggest that this is due to excessive faith in the model. Our users reported high levels of trust in the system, even going so far as to say *"the machine might be better at this than I [am]"*, echoing the findings of Smith et al. [23].

Excessive trust presents a potential challenge in creating such tools for more independent use, and illustrates that more thought needs to be given to designing HLTM in ways that allow users to accurately judge when and when not to place trust in the system. Helping users understand *why* certain documents belong to certain topics is an essential step for future HLTM development. In addition to improving model quality as discussed above, a greater focus on documents would give users more agency and information. Users reported that they were sometimes confused by why a document was ranked highly inside a particular topic, and rather than just browsing topic words, they wanted to be able to see exactly which documents within a topic contained a particular word of interest. This could be achieved in the UI by allowing users to search and browse via words and documents, as well as displaying word-level topic assignments within documents. Such functionality is especially valuable if the model's intended use involves inference on new documents or other downstream tasks, as it would help users validate that the model is behaving according to their intentions during refinement.

## 7 FUTURE WORK

Because of the applied setting of our study, we were only able allocate two participants each to the refinement and evaluation tasks. As a result, the significance of our findings is limited by our small sample sizes. Further study and exploration is needed to support some of our conclusions.

Our findings unlock promising new directions for HLTM. In terms of HLTM tool design, we surfaced the need to help users understand *why* certain documents belong to certain topics. This could be achieved by allowing users to see the word-level topic assignments within a document, for instance. Additionally, as our post-interaction messages (PIM) resonated with users, another idea is to extend it by alerting users to the impact of their actions on documents, or exploring how different implementations of PIM affect model quality and user experience. In terms of model evaluation, in future we intend to focus on document-topic and document-word-topic assignments as discussed in Section 6, as well as extending label concurrence to documents. Finally, we see significant value in conducting user studies of HLTM in collaborative settings with multiple users working together to build a single topic model.

## ACKNOWLEDGMENTS

The authors would like to thank Nadja Herger, Rahul Jain, Sumit Das and Nikola Spasojevic for their assistance and support throughout the duration of this research.

## REFERENCES

[1] David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09* (Montreal, Quebec, Canada, 2009). ACM Press, Montreal, Quebec, Canada, 1–8. https://doi.org/10.1145/1553374.1553378

[2] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2017. An Automatic Approach for Document-level Topic Model Evaluation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (Vancouver, Canada, 2017). Association for Computational Linguistics, Vancouver, Canada, 206–215. https://doi.org/10.18653/v1/K17-1022

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (Mar 2003), 993–1022.

[4] Jordan L. Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of Topic Models. *Found. Trends Inf. Retr.* 11 (2017), 143–296. https://api.semanticscholar.org/CorpusID:67010161

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[6] Allison Chaney and David Blei. 2021. Visualizing Topic Models. *Proceedings of the International AAAI Conference on Web and Social Media* 6, 1 (Aug. 2021), 419–422. https://doi.org/10.1609/icwsm.v6i1.14321

[7] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'09)*. Curran Associates Inc., Red Hook, NY, USA, 288–296.

[8] Sanjoy Dasgupta, Stefanos Poulis, and Christopher Tosh. 2019. Interactive Topic Modeling with Anchor Words. *ArXiv* abs/1907.04919 (June 2019), 1–7. http://arxiv.org/abs/1907.04919 arXiv: 1907.04919.

[9] Caitlin Doogan and Wray Buntine. 2021. Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 3824–3848. https://doi.org/10.18653/v1/2021.naacl-main.300

[10] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101 (April 2004), 5228–5235. https://doi.org/10.1073/pnas.0307752101

[11] Enamul Hoque and Giuseppe Carenini. 2016. Interactive Topic Modeling for Exploring Asynchronous Online Conversations: Design and Evaluation of ConVisIT. *ACM Transactions on Interactive Intelligent Systems* 6, 1 (2016), 1–24. https://doi.org/10.1145/2854158

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

[13] Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive Topic Modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 248–257. https://aclanthology.org/P11-1026

[14] Varun Kumar, Alison Smith-Renner, Leah Findlater, Kevin Seppi, and Jordan Boyd-Graber. 2019. Why Didn't You Listen to Me? Comparing User Control of Human-in-the-Loop Topic Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 6323–6330. https://doi.org/10.18653/v1/P19-1637

[15] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Shuly Wintner, Sharon Goldwater, and Stefan Riezler (Eds.). Association for Computational Linguistics, Gothenburg, Sweden, 530–539. https://doi.org/10.3115/v1/E14-1056

[16] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* 105 (Sept. 2017), 28–42. https://doi.org/10.1016/j.ijhcs.2017.03.007

[17] Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic Evaluation of Local Topic Quality. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, 2019). Association for Computational Linguistics, Florence, Italy, 788–796. https://doi.org/10.18653/v1/P19-1076

[18] Fred Morstatter and Huan Liu. 2018. In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics. *Journal of Machine Learning Research* 18, 169 (2018), 1–32. http://jmlr.org/papers/v18/17-069.html

[19] David Newman, Timothy Baldwin, Lawrence Cavedon, Eric Huang, Sarvnaz Karimi, David Martinez, Falk Scholer, and Justin Zobel. 2010. Invited Paper: Visualizing Search Results and Document Collections Using Topic Maps. *Web Semant.* 8, 2–3 (jul 2010), 169–175. https://doi.org/10.1016/j.websem.2010.03.005

[20] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Los Angeles, California) *(HLT '10)*. Association for Computational

Linguistics, USA, 100–108.
[21] Quentin Pleple. 2013. *Interactive Topic Modeling*. Master's thesis. University of California, San Diego.
[22] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 248–256. https://aclanthology.org/D09-1026
[23] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo Japan, 293–304. https://doi.org/10.1145/3172944.3172965
[24] Alison Smith-Renner, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2020. Digging into user control: perceptions of adherence and instability in transparent models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 519–530. https://doi.org/10.1145/3377325.3377491
[25] Jun Wang, Changsheng Zhao, Junfu Xiang, and Kanji Uchino. 2019. Interactive Topic Model with Enhanced Interpretability. In *IUI Workshops*. ACM, Los Angeles, USA, 1–7.
[26] Yi Yang, Doug Downey, and Jordan Boyd-Graber. 2015. Efficient Methods for Incorporating Knowledge into Topic Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 308–317. https://doi.org/10.18653/v1/D15-1037