A User-Centered, Interactive, Human-in-the-Loop Topic Modelling System

Zheng Fang[1], **Lama Alqazlan**[1], **Du Liu**[2], **Yulan He**[1,3,4] **and Rob Procter**[1,4]
[1]Department of Computer Science, University of Warwick, UK
[2]Bayes Business School, City, University of London, UK
[3]Department of Informatics, King's College London, UK
[4]The Alan Turing Institute, UK
{Z.Fang.4, Lama.Alqazlan, Rob.Procter}@warwick.ac.uk
yulan.he@kcl.ac.uk, Du.Liu@city.ac.uk

Abstract

Human-in-the-loop topic modelling incorporates users' knowledge into the modelling process, enabling them to refine the model iteratively. Recent research has demonstrated the value of user feedback, but there are still issues to consider, such as the difficulty in tracking changes, comparing different models and the lack of evaluation based on real-world examples of use. We developed a novel, interactive human-in-the-loop topic modeling system with a user-friendly interface that enables users compare and record every step they take, and a novel topic words suggestion feature to help users provide feedback that is faithful to the ground truth. Our system also supports not only what traditional topic models can do, i.e., learning the topics from the whole corpus, but also targeted topic modelling, i.e., learning topics for specific aspects of the corpus. In this article, we provide an overview of the system and present the results of a series of user studies designed to assess the value of the system in progressively more realistic applications of topic modelling.

1 Introduction

Huge amounts of unstructured, textual data are generated daily. As more data becomes available, it becomes more difficult to search, understand and discover the knowledge within it. Because of the human effort it requires, conventional qualitative approaches, such as Grounded Theory, (Glaser et al., 1968) are no longer feasible with such large volumes of data. Topic modelling is a potential solution that has received increasing attention in recent research (Heidenreich et al., 2019; Curiskis et al., 2020; Dantu et al., 2021; Goyal and Howlett, 2021) to help users organize, search, and understand large amounts of information. It is an unsupervised machine learning technique for identifying hidden topics in large, unstructured text corpora, in which a hidden topic is represented by a group of words describing a common theme. Users can easily identify the topics in each document and search for documents closely associated with a specific topic for a more in-depth study. However, the topics generated by conventional topic models are often incoherent and contain many unrelated words (Chang et al., 2009; Mimno et al., 2011; Boyd-Graber et al., 2014). Although these issues can be addressed by pre-processing the target data source, for example, by removing irrelevant words from the vocabulary list and adjusting the hyperparameters of the model, such as the number of topics, this requires familiarity with the algorithm. Hence, it is difficult for anyone who does not have some knowledge of topic modelling.

Human-in-the-loop topic modelling (HL-TM) incorporates human knowledge into the topic modelling process to address the aforementioned issues. It allows users who are not experts in topic modelling to refine the model through a set of refinement operations, such as adding words to a topic, removing words from a topic, splitting topics, or merging topics (Jagarlamudi et al., 2012; Wang et al., 2012; Choo et al., 2013; Hoque and Carenini, 2015; Lund et al., 2017; Smith et al., 2018). While most of these studies did not feed the refinement operations into an iterative retraining process, Smith et al. (2018) implemented a fully interactive, user-centered HL-TM system, and examined how the user experience is affected by issues arising in interactive systems, such as unpredictability, trust and lack of control. However, there are still limitations to their work. First, their system only allows users to refine the model sequentially, meaning that once a user updates the model, a new model overrides the previous model. This prevents users from comparing the effects of applying different refinement operations to the same model, making it difficult to find the most appropriate ones. Furthermore, because the previous model is no longer accessible, users may find it challenging to decide whether

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 505–522

the new version is really an improvement. Second, their system does not allow users to retract the changes they made after the underlying model was updated. This becomes a problem when users make inappropriate changes that lead to unexpected results. It is especially frustrating when a user has spent a lot of time refining the model and the whole effect is ruined by one inappropriate change. Third, they use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as their underlying model.

Since LDA is a full-analysis model that can only learn topics from the whole corpus, its application is limited to when users have no *apriori* assumptions about the topics of the corpus. While users can refine the model by adding prior knowledge in an attempt to turn an unrelated topic into one that focuses on the aspect of interest, the phenomenon of higher order co-occurrence in LDA (Heinrich, 2009) may prevent any infrequent words related to the aspect of interest being assigned to the unrelated topic. For instance, given a set of posts about health, researchers may wish specifically to analyze the impact of food on health. Researchers would add food-related words such as “food”, “eat” to an unrelated topic. If these words have a relatively low frequency of occurrence in the posts, then the system may not turn the unrelated topic into a food-related topic. Fourth, an important question that has not yet been explored is how to signal to users when their assumptions do not match reality (Kumar et al., 2019), which may bias the refinement process. For instance, a user may think that a technology-related topic should contain words like “apple, google", while the input corpus has the word “apple" related to fruit. By adding “apple" to the topic, the topic will be contaminated with fruit-related words. Such incorrect refinements would result in poor results.

To address these issues, we implemented a novel HL-TM system that supports six refinement operations, including *add words, remove words, swap word order, remove documents, merge topics*, and *split topic* (Lund et al., 2017; Smith et al., 2018). Unlike Smith et al. (2018), where users can only refine the model sequentially and cannot retract the changes they made, in our system, users can make different attempts at refinement to the same model node and compare the resulting models. A complete refinement history is also presented, allowing users to track their changes from the first step. These also ensure that users can revert to the previous model node when an inappropriate refinement operation is applied. Instead of using LDA as the underlying model, our system uses the query-driven topic model (QD-TM) from (Fang et al., 2021) as the underlying model. The advantage is that the QD-TM not only supports the full-analysis capabilities from LDA, i.e., learning the overall topics from the whole corpus, but also supports users in performing targeted analysis, i.e., learning topics focused on specific aspects of the corpus by mitigating the effects of higher order co-occurrence phenomenon in LDA (Fang et al., 2021).

Moreover, we also implemented a novel automatic topic words suggestion feature to guide users in adding appropriate words to the selected topic. This feature extracts a list of candidate words related to the topic from which users can select words to add to the topic. Our evaluation results demonstrate the usefulness of this feature, where the suggested words are closely related to the selected topic and align better with the ground truth. We also conducted a series of user studies designed to assess the value of the system on real world application scenarios.

This work makes the following contributions: (1) a novel interactive HL-TM system with an advanced user interface that allows users to train different models from the same model node and retract inappropriate changes applied; (2) the use of QD-TM as the underlying model to support both the full-analysis and targeted-analysis topic modelling capabilities; (3) a novel and efficient topic words suggestion feature to guild users add appropriate words to the selected topics; and (4) a small scale user study and two detailed studies designed to replicate real-world application scenarios.

2 Background

Query-driven topic model is a semi-supervised topic modelling algorithm developed by Fang et al. (2021). It allows users to specify a simple query in words or phrases and return query-related topics. The original model involves a two-stage modelling process; in the first stage, the model infers one topic for each concept as well as other unrelated topics, while in the second stage, the model extends the topic of each concept into a set of subtopics. In our work, we are only interested in the first stage of the model. If there is no query input, the model works as a conventional topic model.

The model is based on a variant of a Hierarchical

Dirichlet Process (HDP) (Teh et al., 2006), which is a nonparametric Bayesian model that assumes that a restaurant (i.e., a document) has a set of tables and serves dishes (i.e., topics) from a global menu. A single dish is only served at a single table for all customers (i.e., words) who sit at that table. We developed our version of the algorithm based on the Gibbs sampling technique. For a word w_{ji} at document j and position i, the probability for sampling an existing table t is:

$$p(t_{ji} = t \mid t^{-ji}, k) \propto \mathbb{1}(w_{ji}, k_{jt}) n_{jt}^{-ji} f_{k_{jt}}^{-w_{ji}}(w_{ji}) \quad (1)$$

and the probability for sampling a new table t^{new} is:

$$p(t_{ji} = t^{new} \mid t^{-ji}, k) \propto \alpha p(w_{ji} \mid t^{-ji}, t^{new}, \mathbf{k}) \quad (2)$$

Here, t^{-ji} are the table assignments of all other words. k_{jt} is the topic assignment of table t at document j. n_{jt}^{-ji} is the number of words in document j at table t and $f_{k_{jt}}^{-w_{ji}}(w_{ji})$ is the probability of w_{ji} assigned to topic k_{jt}. $\mathbb{1}(w_{ji}, k_{jt})$ is an indicator function, which takes on value 0 if w_{ji} is a pre-defined word associated to a topic z that $z \neq k_{jt}$, and 1 otherwise. α is a hyperparamenter of the model. The probability for sampling a topic $k_{jt^{new}}$ for the new table is:

$$p(k_{jt^{new}} \mid t, k^{-jt^{new}}) \propto \mathbb{1}(w_{ji}, k_{jt}) m_k f_k^{-w_{ji}}(w_{ji}) \quad (3)$$

Human-in-the-Loop topic modelling has received a lot of attention in recent years. Boyer et al. (2017) created a Human–Machine methodology for identifying Systems Thinking topics in a large corpus of text. Users are required to subjectively identify and provide seed documents describing the topic to guide the topic model's training. The methodology, however, did not incorporate any prior knowledge into the modelling process, instead it simply modified the training corpora. Various refinement operations, such as adding words, removing words, adding documents, removing documents, creating topics, merging topics, or removing topics, were implemented to better utilize human knowledge in the topic modelling process (Hoque and Carenini, 2015; Lund et al., 2017; Smith et al., 2018). During the model sampling process, the refinements change the prior knowledge of the underlying model. To understand the usefulness of different refinement operations, Lee et al. (2017) conducted a user-centered approach to find out how non-expert users interpret topic models and what refinement operations they want most, but they only implemented a basic system without full interaction, so the refinement operations they applied did not update the underlying topic model.

Smith et al. (2018) took one step forward from Lee's work by implementing a fully interactive, user-centered HL-TM system and further examined how common interactive machine learning challenges, such as unpredictability, latency and trust, affect the user experience. Kumar et al. (2019) were the first to comparatively evaluate different refinement operations, as well as two feedback injection frameworks, namely informed priors and constraints. They not only suggested that future research should test the system with end users, since their experiments only used simulated user behavior, but also mentioned that it's important to signal to users when their assumptions do not match reality. Though other HL-TM systems exist such as UTOPIAN (Choo et al., 2013) and ConVisIT (Hoque and Carenini, 2015), they use alternative approaches such as non-negative matrix factorization and fragment quotation, and do not provide the complete set of refinement operations that users need (Lee et al., 2017).

Our work can be seen as an extension of the work of Smith et al. (2018). Compared to these studies, it not only provides a fully interactive HL-TM system incorporating various refinement operations but also provides a more user-centered design, such as a complete refinement history and a topic words suggestion feature to enhance the user experience.

3 Proposed System

We introduce our implementation and the interface design in this section.

3.1 Refinement Implementations

Our system uses Gibbs sampling as the inference technique and adopts the constraint method described in Kumar et al. (2019) to inject new information. Every time a user provides a feedback to the system, it first forgets table-word assignment t_{ji} and then injects new information into the system using a potential function $f(k, w, j)$ (Yang et al., 2015) of the hidden topic k of word w in document j. The equation (1) then becomes:

$$p(t_{ji} = t \mid t^{-ji}, k) \propto \mathbb{1}(w_{ji}, k_{jt}) n_{jt}^{-ji} f_{k_{jt}}^{-w_{ji}}(w_{ji}) f(k_{jt}, w_{ji}, j) \quad (4)$$

Prior work (Lee et al., 2017; Smith et al., 2018) discovered that users typically prefer simple refine-

ment operations, therefore, we implemented the following six refinement operations that are commonly used by users in previous studies:

Add word x to topic z. We update the potential function $f(k, w, j)$ such that $f(k, w, j) = 1$ if $k = z$ and $w = x$, otherwise it is assigned a value of 0.

Remove word x from topic z. We update the potential function $f(k, w, j)$ such that $f(k, w, j) = 0$ if $k = z$ and $w = x$, otherwise it is assigned a value of 1.

Swap word order of w_1 and w_2 in topic z so that w_2 has higher order than w_1. We first compute the ratio r between the difference $n_{w_1,z} - n_{w_2,z}$ and n_{w_2}, where $n_{w_1,z}$ and $n_{w_2,z}$ are the counts of w_1 and w_2 in topic z, respectively, and n_{w_2} is the counts of w_2 in all topics except z. We then update the potential function $f(k, w, j)$, such that $f(k, w, j) = 1$ if $k = z$ and $w = w_2$, otherwise it is assigned δ, where $\delta = 0$ if $r > 1$, otherwise $\delta = 1.0 - r$.

Remove document d from topic z. We update the potential function $f(k, w, j)$ such that $f(k, w, j) = 0$ if $k = z$ and $j = d$, otherwise it is assigned a value of 1.

Merge topic t_2 into t_1. In the next Gibbs sampling iteration, we only sample topics for words assigned to t_2 in the previous Gibbs sampling iteration. We update the potential function $f(k, w, j)$ such that $f(k, w, j) = 1$ if $k = t_1$ and $w \in t_2$, otherwise it is assigned a value of 0.

Split topic t into two topics using seed words s, i.e., s need to be moved from t to a new topic t_n. To do this, we first create a new topic using the nonparametric model. Then, we apply the *add word* operation to add all the words in s to t_n.

3.2 Topic words suggestion

We developed a novel topic words suggestion feature in QD-TM to let users decide which words should appear or not appear in the topic words. Using only subjective input may result in topic words that are spurious, but combining the suggested words can help users to steer the model in the right direction. The topic words suggestion feature is integrated into the Gibbs samplings process of the model. It has two stages. First, it samples an indicator of whether a document is relevant to a topic or not. Second, it updates the suggested words from the relevant documents.

Document relevance sampling For each topic k, if a document contains any of the suggested words and is sampled to be relevant to the topic in the previous Gibbs sampling iteration, then it is assumed to be relevant to the topic in the current Gibbs sampling iteration. Otherwise, we use the following equation modified from Wang et al. (2016) to decide if the document is relevant to the topic:

$$p(r_k = c \mid \mathbf{r}, \pi, \beta, \gamma_{\mathbf{r}}, \gamma_{\mathbf{ir}}) \propto \begin{cases} (C_c^{R(-m)} + \gamma_r) \times \frac{\prod_v^V \Gamma(n_{k,v}^{-m} + \beta)}{\Gamma(\sum_v^V (n_{k,v}^{-m} + f_{c,m,v}) + V\beta)} \\ \text{if } c = 1 \\ (C_c^{R(-m)} + \gamma_{ir}) \times \frac{\prod_v^V \Gamma(n_{\tilde{k},v}^{-m} + \beta)}{\Gamma(\sum_v^V (n_{\tilde{k},v}^{-m} + f_{c,m,v}) + V\beta)} \\ \text{if } c = 0 \end{cases} \tag{5}$$

where $C_c^{R(-m)}$ is the number of documents under relevance status c excluding the current document m, $n_{k,v}^{(-m)}$ is the counts of term v in target topic k excluding the words from the current document m, $f_{c,m,v}$ is the frequency of term v in document m under relevance status c, and π indicates the Bernoulli distribution over relevance status. β, γ_r, and γ_{ir} are hyperparameters, and V is the vocabulary size of the dataset.

Automatic keywords expansion To get suggested words for a topic in each Gibbs sampling iteration, we first calculate the score of term v in the relevant documents of the topic:

$$Score(v) = P_R(v) log \frac{P_R(v)}{P_C(v)}, \tag{6}$$

where $P_R(v)$ is the probability of term v in the relevant documents and $P_C(v)$ is the probability of term v in the whole corpus. We extract terms with high scores as the candidate terms. We then calculate the cosine similarity between each candidate term and the embedding of the topic, and only add terms to the suggested words list when the similarity is greater than 0.5. The embedding of a topic k can be obtained by:

$$emb(k) = \sum_m^M P^i(k, m) emb(m), \tag{7}$$

where $p^i(k, m)$ is the probability of m^{th} word of topic k in i^{th} iteration, M is the total number of representative words in the topic, which is usually set to 10, and $emb(m)$ is the pretrained word embedding of the m^{th} word.

3.3 Interface Design

The user interface of the system consists of two windows (Appendix A). In the first window (Figure A1), users can define the hyper-parameters of a topic model, such as the initial number of topics. If users are interested in topics describing specific concepts or aspects of the corpus, they can also define the prior knowledge of a topic model in this window. If no prior knowledge is provided, then the model behaves as a conventional topic model without any prior knowledge.

The second window (Figure A2) displays the detailed information of the trained model. Users can view the model as a list of topics. Users can also refine the model using the refinement operations implemented and track model change histories. Different from previous systems where users can only refine a model sequentially and cannot retract the changes they made, we include a model history panel with a novel model tree structure so users can make different attempts at refinement to the same model node and compare the resulting models. To help users interpret each topic, we also show a set of topic labels from the automatic topic labeling algorithm (Mei et al., 2007), where a topic label is a phrase that summarises the main idea of the topic. A detailed description of the interface can be found in Appendix A.

4 Evaluation

The evaluation is divided into two parts. In the first part, we evaluated the performance of the topic words suggestion feature in a controlled, laboratory experiment, where we compared two versions of our system. In the second part, we evaluated the HT-LM system by applying it in two realistic topic modelling use cases.

4.1 Topic words suggestion evaluation

Datasets Three commonly used datasets for topic modelling were chosen: the 20newsgroup [1] dataset containing 18k news posts from 20 categories; the TagMyNews [2] dataset containing 32k short English news from 7 categories; and the SearchSnippets (Xu et al., 2017) dataset containing 12k short English news from 8 categories.

Baselines Our HL-TM system uses QD-TM as the underlying model. To test whether the *topic words suggestion* feature can improve the original model, we used only this feature to refine the QD-TM and compared the refined model with the original QD-TM.

Parameterisation We focused on the targeted topic modelling capabilities of QD-TM. To make fair comparisons, we adopted the same experimental settings described in Fang et al. (2021). We used query phrases to represent the main concept of each category in a dataset, following the same setup as in Fang et al. (2021). Categories that do not have meaningful names were removed from the datasets, e.g., `talk.politics.misc` in the 20Newsgroup dataset. The query phrases were integrated into the model using the first window of our HL-TM interface.

We used our HL-TM system to refine the original QD-TM and only used the *add word* refinement operation. In each Gibbs sampling iteration, for each topic, we added all suggested words to the topic. For the hyper-parameters in equation (5), β is set to 0.5, and both γ_r and γ_{ir} are set to 1. We ran 2000 Gibbs sampling iterations.

Results We evaluated the quality of the final models in terms of topic coherence and document retrieval performance as in Fang et al. (2021). Better precision@K scores indicate that the learnt topics are more discriminative and representative. Higher coherence scores indicate better topic interpretability. Table 1 shows the performance of adding suggested topic words to the model. We observed that the refined model achieves higher coherence scores compared with the original model. It indicates that the suggested words are semantically related to each topic and can help produce more coherent topics. The better precision@K scores also show that the suggested topic words are highly relevant to the predefined categories of the dataset. This is expected as the word suggestion feature in our system tends to identify more important words related to the category of the corresponding topic. Instead of blindly adding the suggested words to the model, our system allows users to decide which suggested words should be added. This allows them to use their domain knowledge to further filter out noise.

4.2 Laboratory evaluation

Description To compare our system with the state-of-the-art human-in-the-loop topic modelling system (Smith et al., 2018), we recruited 20 participants via a university campus email list to run

[1] http://qwone.com/ jason/20Newsgroups/
[2] http://acube.di.unipi.it/tmn-dataset/

Model	20news		TagMyNews		SearchSnippets	
	Coherence	Precision@K	Coherence	Precision@K	Coherence	Precision@K
QD-TM	0.445	0.612	0.413	0.736	0.475	0.741
QD-TM + words suggestion	**0.482**	**0.634**	**0.477**	**0.763**	**0.481**	**0.825**

Table 1: Average results for adding suggested topic words for model refinement. We ran each model five times. We used the topic coherence metric C_V from Röder et al. (2015)

a small-scale user study, with one subject group working with an equivalent of Smith et al. (2018) – our system with no model history, no words suggestion, and no topic labeling (old system) – and one group working with our full-featured system (new system). In a factorial design with two independent variables, we used two corpora of Reddit posts focusing on online teaching platforms (corpus A) and Twitter tweets discussing the 2021 United Nations Climate Change Conference (corpus B). Participants were randomly allocated into two groups. The first group did corpus A with the old system and corpus B with the new system, and the second group did corpus B with the old system and corpus A with the new system. The task was to conduct a qualitative analysis of the datasets.

Participants randomly started with either the old system or the new system to eliminate the influence of training effects. We then evaluated the average topic quality of the two system conditions. In addition to comparing topic quality across systems, we also asked participants questions about how much they like the new features. All participants were fluent English speakers and non-experts in topic modelling. Each participant received a £20 Amazon gift card as payment for the experiment.

Dataset and Topic Model We used two datasets for the experiments. The Reddit dataset contains 9,651 posts focusing on online teaching platforms. These posts were randomly sampled from the original dataset taken from Alqazlan et al. (2021). The Twitter dataset contains 8,990 tweets related to the 2021 United Nations Climate Change Conference. We used keywords related to the conference to search for tweets between 31 October 2021 and 17 November 2021. These keywords (see Table B1) were checked and provided by social scientists researching COP26 climate change tweets through a few rounds of manual inspection of the tweets. We used QD-TM as the underlying model, and used a standard stop words list and 2000 Gibbs sampling iterations to initialize the topics. We set the initial topic number to 10 and 13 respectively for the Reddit and Twitter datasets, respectively, based on our previous experience with these datasets. For each subsequent update during the task, 10 Gibbs sampling iterations were run.

Procedure Sessions took around 60 minutes on average, and they were conducted face-to-face. We began by introducing participants to topic modelling and how to use the tools to refine topics. For both systems, we asked them to first read the top five posts of each topic to interpret the underlying theme and then read the corresponding top 10 topic words and use the provided refinement operations to refine the topic until the top 10 topic words and the top 5 posts are consistent with the interpreted theme. For the new system, participants could also access the new features – model history, word suggestions, and topic labelling to assist their in refinements. We asked participants to click the *apply refinements* button to update the underlying model after they have added all the necessary refinements to a topic. They were allowed to undo any operations that have not been applied to the underlying model. After they finished the refining tasks, we asked them to rate how much they are satisfied with the resulting topics and how much they like the new features. Due to time limitations, we only asked participants to refine the first five topics of each system. The first five topics of the starting model for each corpus can be found in Table B2 and Table B3.

Findings We recorded user interactions with the tools. Users performed a total of 1,284 refinements using the two systems. Among the top four most used operations were the *delete words* operation (used 554 times), the *swap words* operation (used 512 times), the *add words* operation (used 171 times), and the *delete document* operation (used 81 times). This is consistent with the findings of Smith et al. (2018), who also observed that these operations were the four most used operations (excluding the add to stop words operation since our system didn't implement this). For the new system,

the *add words* operation was applied 115 times in total, among which 56 times were adding suggested words, which indicates that the suggested words operation had high usage. We report the usage of refinements for each subject group in Table B4.

To evaluate whether using the new system can result in better topics than the old system, participants were asked to rate their satisfaction with the final topics of the two systems on a five point scale (1 not very satisfied, 5 very satisfied). For the Reddit dataset, the average satisfaction score for the new system was 4.2 (SD=0.67), while the score for the old system was 3.89 (SD=0.78). Although a Mann-Whitney U test (U=30.5, z=0.83887, p=.200) showed this difference was not statistically significant, 8 out of 10 participants were satisfied with final topics of the new system. For the Twitter dataset, the score for the new system was 4.0 (SD=0.71), while the score for the old system was 3.2 (SD=0.67). A Mann-Whitney U test (U=18.5, z=1.8985, p=.029) showed that the difference was statistically significant, suggesting that the new system can help improve topic quality.

To evaluate whether the model tree feature can help users track changes, we asked participants to rate their agreement with the statement, "I was able to remember what the model looked like before my updates", on a five point scale (1 strongly disagree satisfied, 5 strongly agree), for both the old and new systems. The average agreement was 3.1 (SD=1.07) for the old system and 3.9 (SD=0.85) for the new system. A Mann-Whitney U test (U=116, z=2.25868, p=.012) showed the difference to be statistically significant. Participants were also asked to rate their agreement with the statement, "the model tree feature of the new system can help me track my updates". The average agreements for the statement was 4.35 (SD=1.04). This strongly suggests that the model tree feature of the new system can help users track their changes. To evaluate the usefulness of the words suggestion feature, we asked participants to rate their agreement with the statement, "the suggested words feature of the new system can help me identify relevant words of the topics". The average agreements for the statement was 4.1 (SD=0.64). 17 out of 20 participants stated that the feature is very useful, suggesting that the feature is a good supplement for adding words to topics. The finding is consistent with the observation of user interactions with the new system that around 50% (56 out of 115) of added words were from the suggested words list. For the topic labelling feature, participants were asked if they agreed with the statement, "the topic labels from the full-featured system can help me interpret the meaning of the topics". The average score for the statement was 3.25, and 11 out of 20 mentioned that the provided labels for some topics are totally irrelevant. This suggests that the algorithm (Mei et al., 2007) doesn't fit the datasets well and suggests the needs for a more accurate topic labelling algorithm in the future.

Suggestions Though the user studies show promising results, participants also had suggestions for improving the new system. Six mentioned that the swap words operation didn't fit their needs well. Instead of swapping the order of the two selected words, they prefer to allow the selected words to be inserted in new positions, which provides more flexibility when ranking topic words. One participant suggested allowing the undo any of the previously added operations directly, rather than starting with the most recently operation.

4.3 Use case one: Tutors' experiences in the Gig Economy

Description To evaluate our system on real-world document analysis tasks, we invited a researcher who is investigating the Gig Economy to use our model. In previous work, Alqazlan et al. (2021) applied LDA to identify topics that are related to tutors' experiences in the Gig Economy from a Reddit dataset where tutors posted about their experiences of working on online teaching platforms. We asked the researcher to use our system to identify the relevant topics in the dataset and assess whether using our system can produce better results than LDA. As Alqazlan et al. (2021) has previously found that a 17-topic LDA extracted the most number of relevant topics, for a fair comparison, the initial number of topics for our system was also set to 17. The researcher was encouraged to use any refinement operations provided by the system until she felt that the top 5 relevant posts of each detected topic revolved around one main theme.

Results A total of 16 models were trained and two branches were created during the refinement process, with models 11, 12 and 13 constituting one branch and models 14, 15 and 16 constituting the other branch. Both of these branches were extended from model 10, which was trained sequentially

from the initial model. By using the model tree panel to compare these models, it was found that Model 16 produced the most satisfactory results. Of all the refinement operations, *remove document* was used the most and *swap word order* was used the least. The researcher commented that the *swap word order* could be more useful if it only adjusts the position of a word in the topic, rather than swapping a word with another one.

We compare LDA with Model 16 and present the qualitative results here. For both models, the researcher was asked to manually assign a label to each topic. For the LDA model, 11 topics were found to be relevant to tutors' experience in the Gig Economy, two of which (Table C1) were missed by Model 16, while for Model 16, 16 topics were found to be relevant, of which seven topics (Table C2) were not identified by the LDA. This indicates that our HL-TM system is able to assist the researcher to identify more relevant topics.

To determine whether using our system can help produce better quality topics, we measured the topic coherence score for the 9 relevant topics identified by both LDA and Model 16 (Table C3). We used the best performing topic coherence measure CV based on the external corpus (Wikipedia) (Röder et al., 2015). We found that our system produced better topic coherence scores for six topics, while the LDA produced better scores for only three topics. This finding is consistent with the researcher's view that only two topics from the LDA were of better quality than Model 16. We present the relevant topics from both LDA and Model 16 in Appendix C.

4.4 Use case two: Patenting strategies for pre-determined patent value categories

Description We invited a second researcher, who is working on patenting strategies, to evaluate our system. In a previous study, Ribeiro and Shapira (2020) has identified a set of patent value categorises (Table D1) for synthetic biology patents. However, their work was based entirely on human evaluation and was, therefore, time-consuming, and biased evaluations could occur, so only 102 patent documents were analysed. In this use case, the researcher aimed to use our system to categorise a larger patent sample to the pre-defined categories and reduce human selection bias. 2607 related patents in the United States Patent and Trademark Office (USPTO) dataset were used. The pre-defined categorises are presented in Appendix D.

To identify documents in their predefined categories in the dataset, the researcher used the first window of the interface to incorporate prior knowledge. A list of final concept words of each category is presented in Table D2 in Appendix D. The initial number of topics is set to 20.

Results By using the first window of the user interface to incorporate priori knowledge, the initial model was able to roughly infer topics for the predefined categories, although the topics were not yet of high quality. We then asked the researcher to examine the top 10 relevant documents of these topics to assess the categorisation quality. The average precision is 0.3. The researcher then refined these topics using the system until she felt that it best categorised the top 5 relevant documents into each category and was satisfied with the resulting topic words. In total 23 models were trained and five branches were created, with Model 23 being the best one. After the refinements, we asked the researcher to further examine the top 10 relevant documents of the refined topics to compare with the initial model. The average precision for the refined topics is 0.96, which is much higher than the result from the initial model. This verifies that the use of the system can help identify a larger sample of documents in pre-defined categories than the previous manual evaluation method. In total, the *add word* operation was applied 62 times, among which 41 times were related to adding suggested words from the system. This shows that our topic words suggestion feature can indeed identify words that are highly relevant to specific topics.

We also compared the topic coherence scores for the focused topics from the initial model (Table D2) and from Model 23 (Table D3). The scores are 0.404 and 0.434, respectively. It shows that using the system can help produce better quality topics.

5 Conclusions and Future work

We have developed a novel user-centered, interactive, HL-TM system to address the limitations of the prior work (Smith et al., 2018). An advanced user interface with a model history panel is presented to allow users compare different models and retract inappropriate changes. The use of the QD-TM as the underlying model supports both the full-analysis and targeted-analysis capabilities. A novel topic words suggestion feature is also integrated into the system and the evaluation shows

that the feature is promising for suggesting reasonable and coherent words for topics. A small scale experimental user study and two detailed use cases further verified the usefulness of the system on real-world tasks. From the use cases, we observed that both researchers trained many different models with multiple branches created (use case one has 16 models trained and two branches created, while use case two has 23 models trained and five branches created). This shows that researchers are very likely to change their mind when refining a topic model, so allowing researchers to compare different models and retract their changes is helpful. In the future, we plan to improve the system based on the results from the user evaluations and conduct more extensive evaluations to gather feedback on deploying the system in a wide variety of applications.

Limitations

We used two small datasets in the user evaluations, so participants were not affected by latency issues, where the user would have to wait while the algorithm performs updates. Since QD-TM is more complex than LDA, and the words suggestion feature further increased the complexity of the computation, a large dataset could lead to latency problems. The time complexity of our system is O(DLK + KV), where D is the number of documents used, L is the average document length, K is the number of topics and V is the vocabulary size. We tested the system with two different dataset sizes (20,000 tweets vs. 8,990 tweets). By using 20,000 tweets, the average wait time for model updates during user interaction (10 Gibbs sampling iterations) was 34 seconds, compared to 14 seconds using 8,990 tweets.

According to Smith et al. (2018), longer wait times can negatively affect users when using interactive systems, suggesting that latency could be an issue in our system as the size of the dataset increases. However, as is well known, these effects can be mitigated through the provision of time affordances (Conn, 1995). We will address this in the future.

In addition to the latency issue, the laboratory evaluation we conducted also has limitations. Not all the participants were familiar with qualitative analysis. It is likely that participants with extensive experience in qualitative analysis identify more valuable refinements to the models, resulting in better quality topics.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council (grant no. EP/V048597/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (EP/V020579/1). ZF receives the PhD studentship jointly funded by the University of Warwick and China Scholarship Council.

References

Lama Alqazlan, Rob Procter, and Michael Castelle. 2021. *Workshop on Natural Language Processing for Digital Humanities, 18th International Conference on Natural Language Processing*.

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

Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 225255.

Ryan C Boyer, William T Scherer, Cody H Fleming, Casey D Connors, and N Peter Whitehead. 2017. A human–machine methodology for investigating systems thinking in a complex corpus. *IEEE Systems Journal*, 12(3):2937–2948.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19(12):1992–2001.

Alex Paul Conn. 1995. Time affordances: the time factor in diagnostic usability heuristics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 186–193.

Stephan A Curiskis, Barry Drake, Thomas R Osborn, and Paul J Kennedy. 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034.

Ramakrishna Dantu, Indika Dissanayake, and Sridhar Nerur. 2021. Exploratory analysis of internet of things (iot) in healthcare: a topic modelling & co-citation approaches. *Information Systems Management*, 38(1):62–78.

Zheng Fang, Yulan He, and Rob Procter. 2021. A query-driven topic model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1764–1777, Online. Association for Computational Linguistics.

Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. 1968. The discovery of grounded theory; strategies for qualitative research. *Nursing research*, 17(4):364.

Nihit Goyal and Michael Howlett. 2021. "measuring the mix" of policy responses to covid-19: Comparative policy analysis using topic modelling. *Journal of Comparative Policy Analysis: Research and Practice*, 23(2):250–261.

Tobias Heidenreich, Fabienne Lind, Jakob-Moritz Eberl, and Hajo G Boomgaarden. 2019. Media framing dynamics of the 'european refugee crisis': A comparative topic modelling approach. *Journal of Refugee Studies*, 32(Special_Issue_1):i172–i182.

Gregor Heinrich. 2009. A generic approach to topic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 517–532. Springer.

Enamul Hoque and Giuseppe Carenini. 2015. Convisit: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 169–180.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

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. *arXiv preprint arXiv:1905.09864*.

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*, 105:28–42.

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 896–905.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Barbara Ribeiro and Philip Shapira. 2020. Private and public values of innovation: A patent analysis of synthetic biology. *Research policy*, 49(1):103875.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Carson Sievert and Kenneth Shirley. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

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*, pages 293–304.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581.

Di Wang, Marcus Thint, and Ahmad Al-Rubaie. 2012. Semi-supervised latent dirichlet allocation and its application for document classification. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 306–310. IEEE.

Shuai Wang, Zhiyuan Chen, Geli Fei, Bing Liu, and Sherry Emery. 2016. Targeted topic modeling for focused analysis. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1235–1244.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31.

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*, pages 308–317.

Appendix A: Interface Design

The user interface of the system consists of two windows. If users are interested in topics describing specific concepts or aspects of the corpus, they can define the prior knowledge of a topic model in the first window (Figure A1). They can use a query phrase to define the concept of interest. The input query phrase is expanded to a set of candidate concept words using the concept words extractor from Fang et al. (2021). Users can use their knowledge to determine which words should be included in a topic and click the "+" buttons next to them to add to the final concept words list of a topic. They can also add words that are not in the candidate words list to the final list based on their knowledge. The words in the final list can be removed by clicking the corresponding "-" buttons. It also supports viewing the top relevant documents of the input query by changing the viewing options from "By keywords" to "By documents".

If users have no preferred interest in the topics of the corpus, they can leave the final concept words list of each topic empty, and the model behaves as a conventional topic model without any prior knowledge. Users can click the "apply" button to apply the settings to the underlying model. Users can view the resulting topics by clicking the "View" button in the bottom, which takes them to the second window of the user interface.

The second window of the interface (Figure A2) has three panels: *model history* panel (left panel), *model detail* panel (middle panel), and *inter-topic distance map* (right panel). The *model history* panel consists of two subpanels: the *model tree* panel (top panel), which displays the refined models in a tree structure, and the *refinements history* panel (bottom panel). Every time a user refines a selected model, a new model is added to the tree. As shown in Figure A2, users can also create new branches from the same model node. For example, model 5, model 6 and model 7 are all refined from model 4. The *refinements history* panel allows users to view the refinements history between two connected models by clicking on the edge between them. Users can also view the pending refinements of the selected model as shown in Figure A2.

The pending refinements will not be applied to the underlying model until users click the "Apply Refinement" button. We also include an "undo" button to allow users to undo previously added refinements. The *model history* panel allows users to compare previously refined models and keep track of their previous changes to further assist the refinement process. By clicking the "Download models" button, users can download the entire model tree to their local machine, and by clicking the "Load models" button, users can load a previously downloaded model tree to continue the refinements.

The *model detail* panel, shown in the middle of Figure A2, provides an overview of the selected model's topics, as well as the top words in each one. Weight represents the prevalence of the topic in the whole corpus. Users can rename a topic by typing a new name in the "Topic" column. Users can also merge or split topics in this panel. The right side of Figure A2 is the *inter-topic distance map*. Each circle represents a topic, with its size representing the topic weight in the corpus. The *inter-topic distance map* is an intuitive way to reveal the quality of a topic model where a larger distance between topics indicates a better model (Sievert and Shirley, 2014). By clicking the "view" button on the window, the selected topic's specifics will then be presented as shown in Figure A3.

The top left side of Figure A3 shows a set of topic labels from the automatic topic labeling algorithm (Mei et al., 2007), where a topic label is a phrase that summarises the main idea of the topic. The aim of the topic label is to help users interpret the topic. We also present the top words of the topic with corresponding topic-word weights. Weights indicate the prevalence of the words in the topic. Users can apply *add, delete* or *swap word order* refinement operations here. The right side of Figure A3 displays the top documents associated with the topic and ranks them based on their weight, representing the document's prevalence of the topic. Users can apply the *remove document* refinement operation here. By clicking the "view" button, users can view the details of the selected document.

When users click the "add" button in Figure A3, an "add words" sub-panel (Figure A4) appears. A list of words suggested by the topic words suggestion feature is displayed in the panel. Users can select from the suggested words or from their own knowledge to add words to the selected topic.

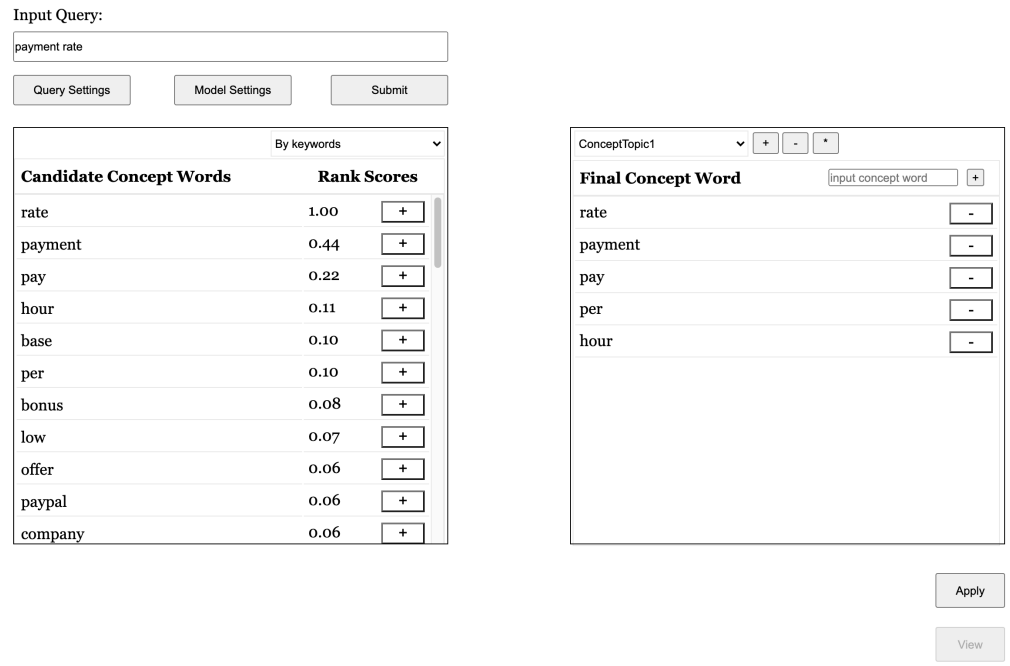

Figure A1: The first window of the user interface. Users can define the prior knowledge of a topic model here.

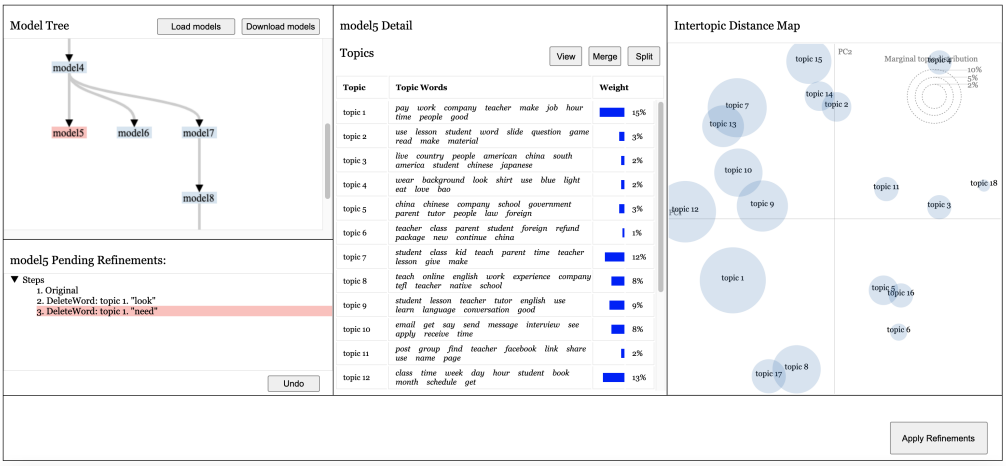

Figure A2: The second window of the user interface, where users can view the details of the selected model and apply merge or split topics refinement operations. Users can also compare different models that have been previously refined and keep track of their previous changes.

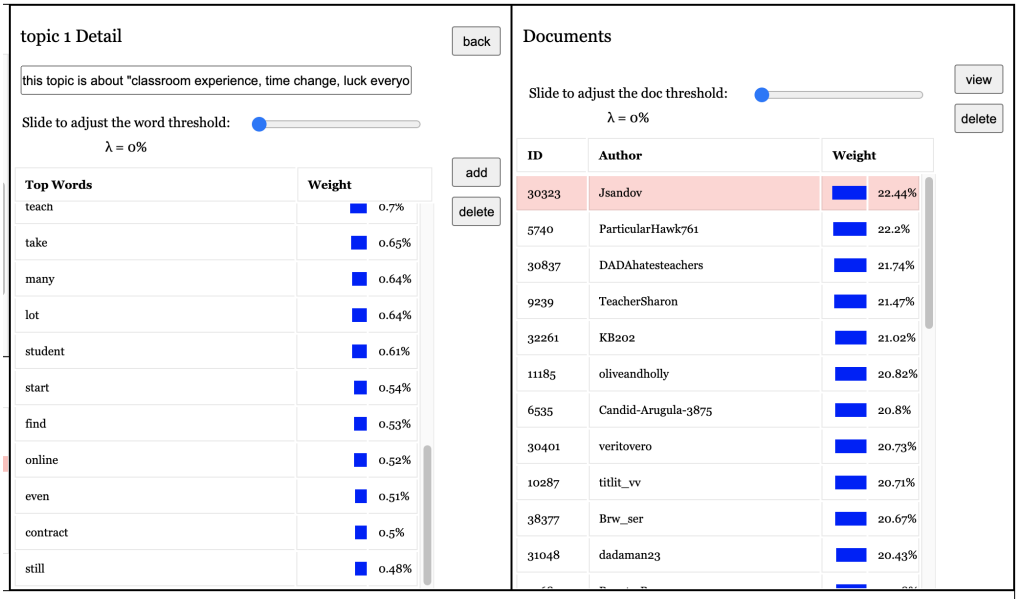

Figure A3: The specifics of a selected topic. Users can apply add, remove or swap word order refinement operations here.

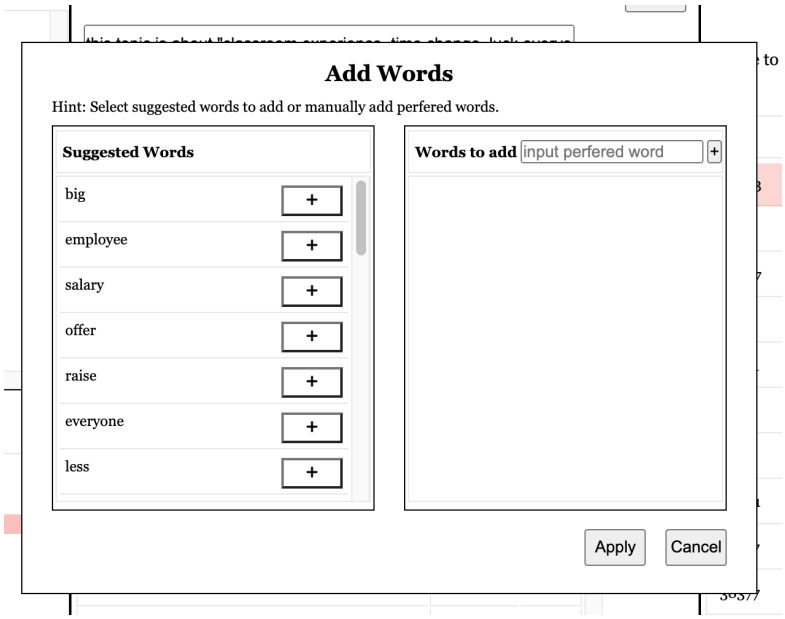

Figure A4: The "add words" sub-panel.

Appendix B: Tables for the small-scale user study (section 4.2)

keywords
police liaison, Police Liaison Officer, PLO, blue bib, #peacefulprotest, #right-toprotest, Police, Policing, Anarchy, Anarchism, Violence, #ClimateJustice, #ClimateCrisis, #ClimateAction, cop26, #GlobalDayofAction, #GlobalDayfor-ClimateJustice, #greenwash, Pigs, Fascist, Stormtrooper, Nazi, crowd, march, rally, mob, extremist, class, privilege, eco-zealot, vandal, nutter, lunatic, eco-fascist, hypocrite, far-left, chaos, stunt, marxist, terror, rabble, anarchist, filth, rozzer, clown, pc plod, old bill, polis, wankers, woke, eco-warrior, campaign, radical, extremist, zealot, authoritarian, tyranny, tyrant, numpty, numpties, scum, disgusting, awful, unbelievable, evil, frightening, selfish, vicious, violent, thug, brutal, vicious, fight, carnage, blood, injury, hostility, aggression, force, assault, invasion, offensive, friction, stress, strain, damage, hurt, harm, block, kettle, contain, arrest, imprison, charge, thankyou, thanks, carnival, festival, fun, enjoy, party, tension, tense, disrupt, soft

Table B1: Keywords used to retrieve tweets related to the 2021 United Nations Climate Change Conference.

Topic	Top 10 words
topic 1	hour class week pay month work time day start make
topic 2	email apply work company good check send interview hire process
topic 3	teach english online native tefl degree experience company teacher certificate
topic 4	student class give rating lesson feedback time parent level bad
topic 5	tutor account student lesson video group profile share bank paypal

Table B2: The first five topics of the starting model for the Reddit dataset. Both the new and old systems used the same starting model.

Topic	Top 10 words
topic 1	climate action change cop people glasgow world today day amp
topic 2	party vote tory labour johnson boris government mps paterson sleaze
topic 3	amp emissions countries climate carbon energy gas global coal fuel
topic 4	charge pay people covid money case work claim give email
topic 5	game great play team today enjoy amp fun win time

Table B3: The first five topics of the starting model for the Twitter dataset. Both the new and old systems used the same starting model.

Operation	New system/Reddit	New system/Twitter	Old system/Reddit	Old system/Twitter	Total
Delete words	74	205	198	77	554
Reorder words	118	152	139	103	512
Add own words	14	45	32	24	115
Add suggested words	7	49	/	/	56
Delete document	12	37	20	12	81
Split topic	0	3	0	4	7
Merge topics	0	1	0	1	2

Table B4: The usage of refinements for each subject group. "New system/Reddit" indicates the use of new system with the Reddit dataset.

Appendix C: Tables for the use case one (section 4.3)

Label	Top 10 words
Reasons to join or leave a platform	company, teacher, job, work, pay, people, make, money, online, think
Miscommunication with platform management	know, anyone, video, tutor, help, ask, let, please, apply, interview

Table C1: Relevant topics identified by LDA, but missed by Model 16.

Label	Top 10 words
Minimum and living wages	pay, work, company, teacher, make, hour, rate, wage, job, time
Nationalities of students	china, people, country, live, chinese, student, american, america, world, government
A restriction on what tutors can wear while tutoring	wear, background, shirt, blue, eat, light, use, bao, love, dino
Chinese new law on out-of-school tutoring	teacher, china, parent, chinese, company, foreign, school, class, new, student
Differences between a professional tutor and a community tutor	student, lesson, teacher, tutor, use, english, learn, language, make, conversation
Discussions about Facebook groups created by platforms' management	post, group, find, facebook, link, see, teacher, use, share, name
Tutors express anger and dissatisfaction with the platform	people, make, good, fuck, post, shit, go, feel, na, sorry

Table C2: Relevant topics identified by Model 16, but missed by LDA.

Label	Top 10 words (LDA)	Top 10 words (Model 16)
Bookings and working hours	week, hour, day, time, book, class, slot, schedule, open, month	class, week, time, day, hour, book, student, schedule, month, slot
Rating system	student, teacher, work, class, give, rating, really, think, month, company	student, rating, call, minute, reservation, time, hour, ph, tutor, week
Teaching materials and methods	student, use, lesson, question, word, ask, say, make, learn, conversation	word, use, game, read, student, slide, play, question, lesson, sentence
Technical issues	class, minute, student, call, time, show, start, late, happen, reservation	use, issue, work, app, internet, problem, try, phone, computer, test
Payment	lesson, pay, time, student, hour, teacher, rate, tutor, base, minute	class, hour, pay, per, bonus, teach, minute, base, lesson, student
Experiences with kids in class	kid, teach, level, well, old, year, think, really, feel, say	student, class, kid, teach, parent, time, lesson, give, teacher, make
Hiring process	email, send, message, say, get, group, reply, back, try, see	interview, apply, video, demo, hire, good, application, new, company, process
Job requirements	english, native, live, country, speaker, language, china, work, american, non	teach, online, english, native, tefl, company, work, experience, teacher, school
Bank transfers and transaction fees	rating, account, pay, demo, use, bank, test, paypal, payment, internet	account, bank, pay, paypal, use, payment, transfer, payoneer, fee, money

Table C3: Relevant topics identified by both LDA and Model 16

Appendix D: Tables for the use case two (section 4.4)

Category	Definition
Market and industrial opportunities	The potential of the invention to enter existing markets
Cost and efficiency	Reduction of production costs associated with more efficient processes
Increasing compound yields	Improvements in compound productivity based on novel processes
Upscaling production	Taking production to the commercial level
Scientific advancements	Contribution to knowledge production
Environmental sustainability	Contribution to environmental quality and preservation
Human health	Improvements in the quality of human health
Food security	Avoiding competition with human food sources
Animal health	Improvements in the quality of animal health

Table D1: categories table from (Ribeiro and Shapira, 2020).

Category	Final Concept Words
Market and industrial opportunities	market, commercial, value, exists
Cost and efficiency	Reduction of production costs associated with more efficient processes
Increasing compound yields	compound, yield, increasing, plant
Upscaling production	production, improve, scale, large
Scientific advancements	advancement, benefit, filed
Environmental sustainability	renewable, sustainable, energy
Human health	health, human, patient, cancer, disease
Food security	food, security, supply, preparation, chain
Animal health	animal, health, disease

Table D2: Final concept words for each pre-defined category.

Category	Topic (Top 10 words)
Market and industrial opportunities	healthy, improved, serum, level, variant, commercial, value, concentration, woman, sample
Cost and efficiency	efficiency, increase, cost, increasing, efficient, enhance, reaction, amplification, low, target
Increasing compound yields	plant, tolerance, improved, compound, yield, marker, increasing, herbicide, soybean, resistance
Upscaling production	improve, improved, production, activity, antibody, enzyme, property, polypeptide, stability, expression
Scientific advancements	benefit, fold, effective, greater, composition, field, size, skin, provide, material
Environmental sustainability	renewable, energy, product, source, diesel, produced, carbon, lipid, sustainable, fuel
Human health	disease, patient, effective, cancer, amount, antibody, human, treatment, administering, subject
Food security	chain, healthcare, ge, mm, food, preparation, light, healthy, column, region
Animal health	health, animal, disease, effective, national, institute, dose, determined, administration, amount

Table D3: Focused topics identified by the initial model.

Category	Topic (Top 10 words)
Market and industrial opportunities	value, commercial, market, industrial, exists, business, global, adding, hepcidin, year
Cost and efficiency	efficiency, improve, increase, increasing, cost, enhance, reaction, efficient, amplification, low
Increasing compound yields	plant, efficiency, compound, yield, increasing, transformation, improved, gene, resistance, increased
Upscaling production	improve, production, improved, improvement, improving, higher, activity, greater, expression, enzyme
Scientific advancements	benefit, provide, field, improved, skin, cleaning, enzyme, surface, cellulase, material
Environmental sustainability	renewable, energy, reducing, biofuels, source, diesel, biofuel, carbon, lipid, product
Human health	disease, patient, cancer, diagnosis, effective, human, liver, lung, clinical, antibody
Food security	chain, sample, food, preparation, assay, light, supply, improved, donor, culture
Animal health	health, animal, medical, disease, care, healthy, nutrition, population, clinical, bethesda

Table D4: Focused topics identified by Model 23.