Making MultiModalTopicExplorer User-Adaptive

Felipe González-Pizarro, and Harshinee Sriram

{felipegp,hsriram}@cs.ubc.ca

Department of Computer Science, University of British Columbia

1 Introduction

The constant increase in the volume of textual data has led to the development of various algorithms intended to summarize and understand unstructured textual data (Peter et al., 2015). A promising solution to this problem is topic modeling, a robust statistical approach for extracting core themes or *topics* from large text corpora. When a topic modeling algorithm is applied to a corpus of documents (e.g., a collection of news articles), the resulting topic model includes a list of topics. Each topic is a large set of terms ranked according to their relevance to the topic and its distribution over the corpus documents (El-Assady et al., 2018).

Although powerful, the results of these algorithms do not provide a descriptive name for each identified topic; therefore, humans must be involved in assigning a meaning to each one, for instance "politics", "economy", or "sports" (Chang et al., 2009; Dou et al., 2011). Visual text analytic researchers have designed algorithms and visual representations to support topic interpretation, attempting to make topic models more human-readable and informative (Yang et al., 2017; Dou et al., 2013). Topic modeling visualization tools help understand the meaning of topics and their quality (Kherwa and Bansal, 2019). However, designing effective mechanisms to enable people to interact with visual representations of topic models is still an open challenge (Jelodar et al., 2019).

First, no visualization tools provide explicit functionalities to analyze the quality of the generated topics. Second, current visual representations lack support for multi-modal corpora (image and text) and often provide limited support to examine the evolution of topics over time. For the courses, CPSC503 AND CPSC547, Gonzalez-Pizarro and Alavi (2021) designed and developed MultiModalTopicExplorer, a web-based interactive visualization of topic modeling algorithm results to address these limitations. This tool allows users to identify the most relevant keywords, most relevant images, and the most relevant documents of the twenty most frequent topics. To validate the functionalities of their tool, the authors designed and conducted a user study with a small group of computer scientists. They asked them to evaluate two popular topic models using the functionalities of MultiModalTopicExplorer. Their results showed that participants felt successful in accomplishing the tasks, however it required high effort and mental demand (Gonzalez-Pizarro and Alavi, 2021).

Topic models might contain hundreds or even thousands of topics. Unfortunately, current topic modeling visualization tools such as MultiModalTopicExplorer do not allow users to explore the topics that are most relevant to them easily. To address this problem, in this work, we extend the functionalities of MultiModalTopicExplorer to allow users to identify topics that are more relevant to them and that are more aligned with their understanding of the domain and current information needs. In the final version of MultiModalTopicExplorer, users can express their preferences by selecting a set of terms of their interest. Then, the system will consider that preferences and recommend a set of topics to match current users' information needs.

We conducted a within-subject user study to test the potential of these extended functionalities. We ask users to inspect and rate the quality of topics from a large-scale Antisemitic and Islamophobic dataset (González-Pizarro and Zannettou, 2022) by using the non-user adaptive (baseline) and user-adaptive (experimental) versions of MultiModalTopicExplorer. Our results show that when participants used the experimental system, they reported lesser physical and mental demands and higher task success rates while evaluating topics. To the best of our knowledge, this is the first attempt to make a topic modeling visualization tool user-adaptive.

The manuscript is organized as follows. Section 2 summarizes related work on visual representations of topic models, discussing limitations and positioning our expected contribution. Section 3 introduces MultiModalTopicExplorer, describing the modifications to make it user-adaptive. Section 4 presents our user study, and Section 6 summarizes its results. Section 6 offers discussions of the results, their limitations, and lines of future work. Finally, Section 7 presents our conclusions.

Ethical considerations. We emphasize that we rely entirely on publicly available and anonymous data shared on 4chan's /pol/. We follow standard ethical guidelines (Rivers and Lewis, 2014), like reporting our results on aggregate and not attempting to deanonymize participants.

Disclaimer. This manuscript contains Antisemitic and Islamophobic textual and graphic elements that are offensive and are likely to disturb the reader.

2 Topic modeling visualization tools

Designing visual representations of topic models requires making decisions about a series of aspects. In this section, we revise prior work about one critical issue to our work: how to allow users to explore the set of topics generated by applying a topic modeling algorithm to a corpus.

Topic model visualization tools offer different layouts to help users get a global view of the corpus. iVisClustering (Lee et al., 2012) uses a graph-based layout, where nodes correspond to documents and the length of an edge between nodes represents the cosine similarity between a pair of documents. Documents that are associated with a topic are closer to each other and share the same node color. For each topic, a color-bounded rectangle shows a list of its most salient terms (see Figure 1 (a)).

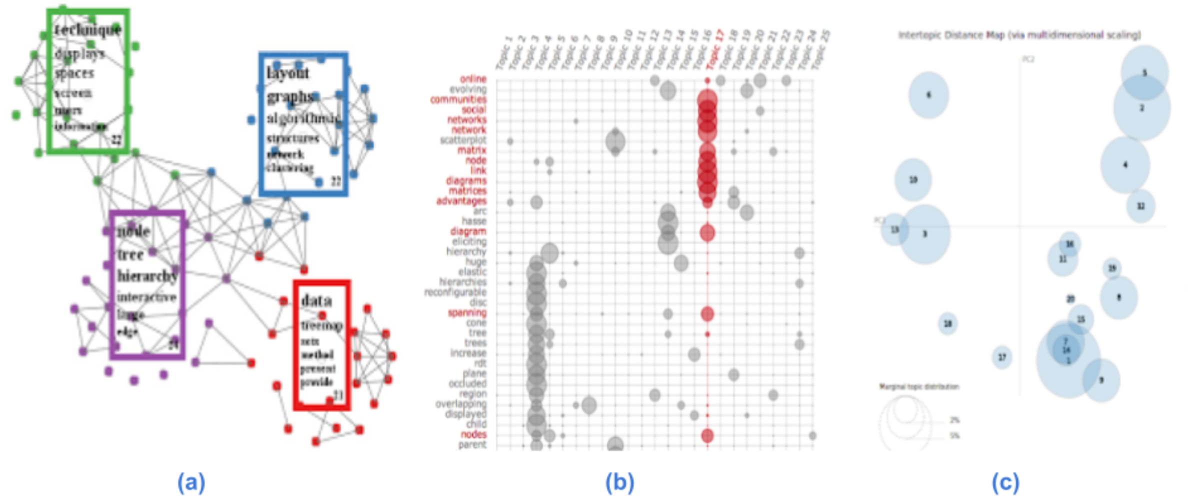


Figure 1: Layout of global view of topics in: (a) iVisClustering (b) Termite (c) LDAvis

.

Termite (Chuang et al., 2012) displays term-topic distributions using a matrix layout. Rows correspond to terms and columns to topics. Termite uses circles to encode term probabilities. Circles with larger areas represent the most frequent terms (see Figure 1 (b)).

LDAvis (Sievert and Shirley, 2014) projects topics into a two-dimensional space. Topics are represented by circles and their positions are determined by the distance between topics. The circle size indicates a topic's prevalence in the corpus (see Figure 1 (c)).

Other tools enable users to visualize and compare global views of topic models of different corpora. TopicPanorama (Liu et al., 2014) uses graphs to represent the topic models of each corpus. A graph matching method and a density-based graph layout on top of these graphs allow displaying them in a single and coherent visualization (see Figure 2 (a)). Alternatively, TopicFlow (Malik et al., 2013) focuses on depicting the corpora dynamics. It allows users to visualize the evolution of topics over time. TopicFlow displays topics as boxes and uses paths between them to represent topic similarity. The box size depends on the number of documents associated with the topic. The most prevalent topics are shown at the top of the chart. Topic flow uses colours to encode topic stages: emerging (green), ending (red), continuing (blue), or standalone (orange), see Figure 2 (b).

The reviewed topic modeling visualization tools aimed to allow users to get a global view of topics, and some of them allow users to compare topics from two corpora. However, while they are helpful, they do not explicitly have functionalities to deal with many topics, such as hundreds or even thousands (Sievert and Shirley, 2014). When that is the case, users must inspect each topic individually to find those relevant to their current information needs. Given that it is challenging to visualize many topics

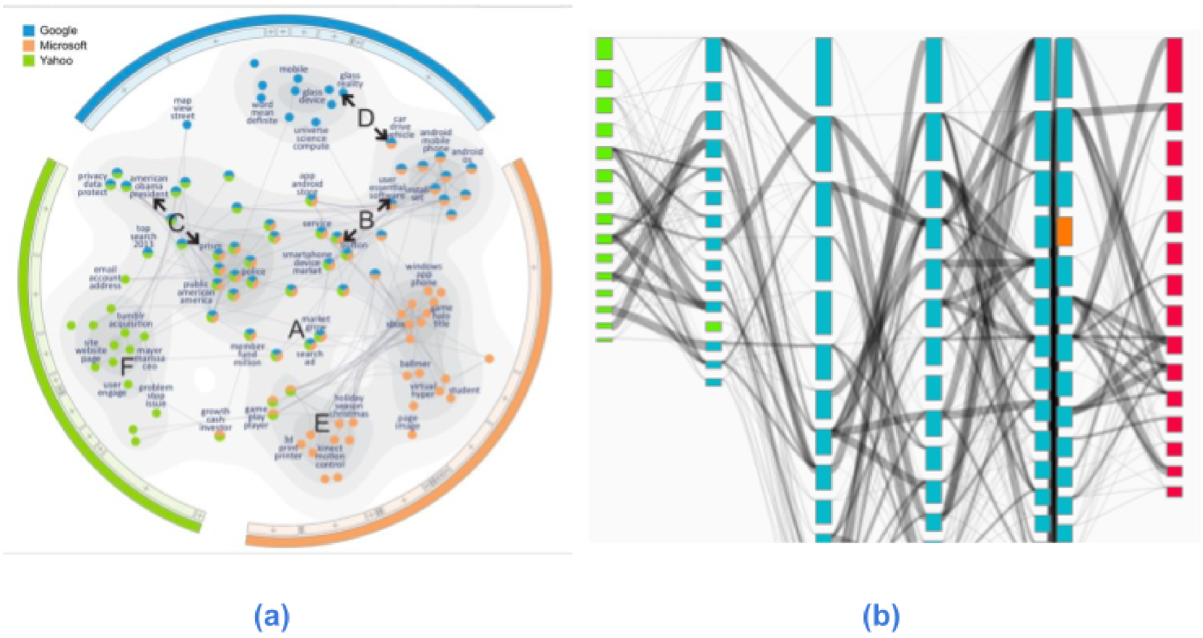


Figure 2: Multi-corpora comparison layout in: (a) TopicPanorama (b) TopicFlow

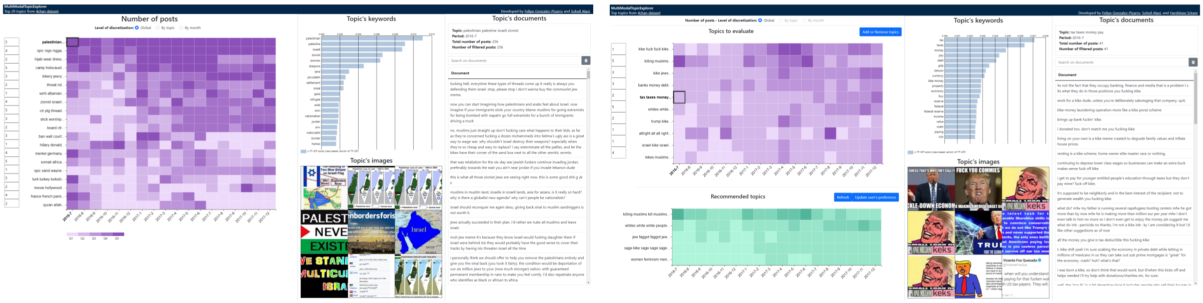
.

compactly, it is necessary to develop tools that can adapt to users, allowing them to quickly find the topics of their interest, saving them time and workload while evaluating the quality of topics.

3 User Adaptive MultiModalTopicExplorer

This manuscript proposes a user-adaptive system for MultiModalTopicExplorer to allow users to find topics relevant to them. This section describes the functionalities of the two versions of MultiModalTopicExplorer: the non-user adaptive (our Baseline system) and the user-adaptive (our final system). Both versions support users in evaluating topic model algorithms results and allow them to answer the following questions: (1) What are the most prevalent topics of the corpus?; (2) How do these topics evolve (e.g., during which historical incidents which topics become trending), and (3) What is the meaning of each topic?.

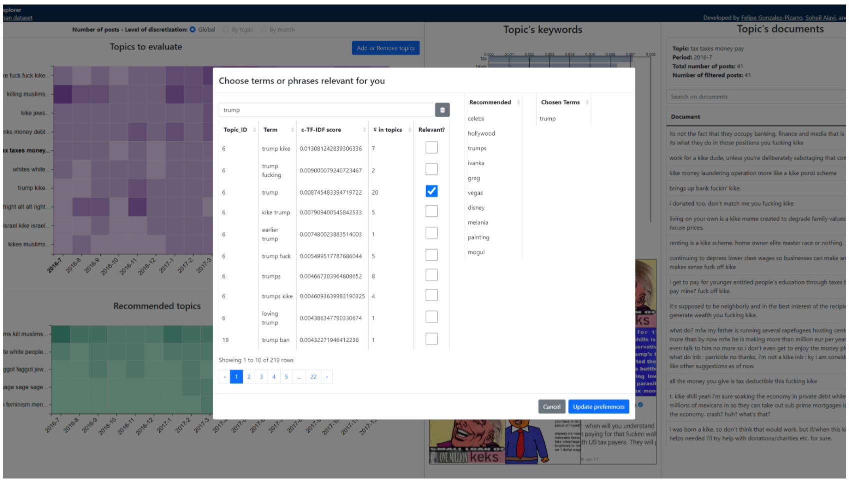
Figure 3 (a) shows the initial version of MultiModalTopicExplorer. Each row from the heatmap represents a topic, and each column represents a timespan. Gonzalez-Pizarro and Alavi (2021) used luminance to encode quantiles calculated on the number of posts on different discretization levels: global (divided by the total number of posts), by topic (divided by the total number of posts for that topic), or by month (divided by the total number of posts in that month). Users can click on any heatmap's cells to visualize information from topics on periods. Users can visualize the top 20 relevant keywords and six most relevant images associated with that topic by selecting a cell. They can also visualize the documents associated with that topic in that specific period. Figure 3 (a) shows the most relevant keywords, most relevant images, and documents from the topic “palestinian”, “palestine”, and “israeli”, in July 2016. Participants can report the quality of each topic on the left of the heatmap.



(a) (b)

Figure 3: MultiModalTopicExplorer layouts: (a) Non user adaptive (b) User-adaptive

We made several modifications to make the system user-adaptive and allow users to find and evaluate topics of their interest. First, we created a modal view to allow users to indicate their preferences (see Figure 4). This modal displays a table with the top 20 keywords for each topic. From this table, users must select the keywords that are relevant to them. After implementing this functionality, we realized that the universe of keywords that users might choose is increasingly large. A topic model must contain hundreds or even thousands of topics, making it difficult for users to find important terms. To mitigate this problem, we implemented a keywords recommendation system. The system suggests terms to users based on their previous selected terms (see details in Section 3.1). Users must click “Update preferences” when they finish indicating their preferences.



Choose terms or phrases relevant for you

trump

Topic_ID	Term	c-TF-IDF score	# in topics	Relevant?
6	trump kike	0.013081242839306336	7	□
6	trump fucking	0.009000079240723467	2	□
6	trump	0.008745483394719722	20	☑
6	kike trump	0.007909400545842533	5	□
6	earlier trump	0.007480023883514003	1	□
6	trump fuck	0.005499517787686044	5	□
6	trumps	0.004667303964808652	8	□
6	trumps kike	0.004609363998190325	4	□
6	loving trump	0.004386347790330674	1	□
19	trump ban	0.00432271946412236	1	□

Showing 1 to 10 of 219 rows

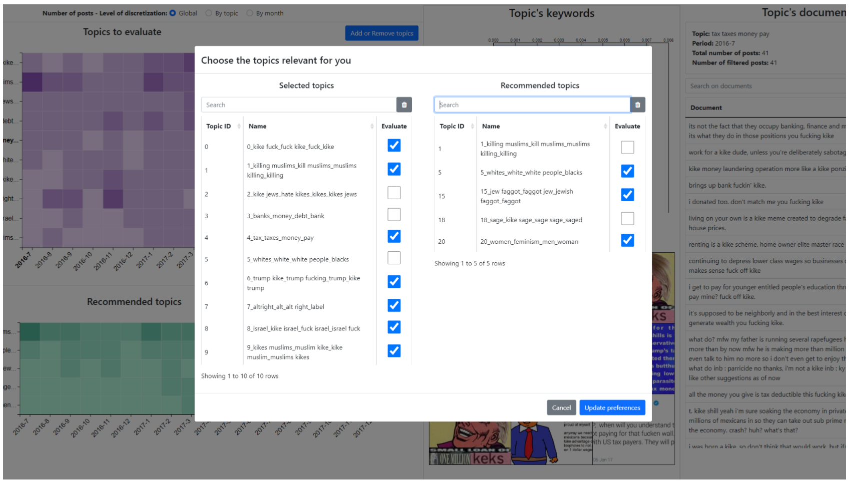
Recommended
celebs
hollywood
trumps
ivanka
greg
vegas
disney
melania
painting
mogul

Chosen Terms
trump

Figure 4: On this modal, users need to select keywords that are relevant to them to express their preferences. To help them during this process, MultiModalTopicExplorer recommends keywords based on previously chosen terms.

After indicating the terms users are interested in, the system retrieves the ten topics that best match users' preferences. Users can visualize these topics in the main heatmap (the purple one, see Figure 3 (b)). This is the visual component that allows users to report topics' quality. The system also recommends five additional topics based on users' preferences. These topics are displayed in the secondary heatmap (the green one, see Figure 3 (b)). Users can click on any of these cells to visualize topic-related information, such as their most relevant keywords, most relevant images, and associated documents. They also can click on the "Refresh" button (see Figure 3 (b)) to get new five additional recommendations. Users can modify their preferences at any time by clicking the button "Update user's preferences" (see Figure 3 (b)).

The user-adaptive version of MultiModalTopicExplorer allows users to report the quality of ten topics, those that appear in the main heatmap. Users can add and remove topics from that heatmap by clicking the "Add or Remove topics" button. Here, a modal view is displayed (see Figure 5) with a list of the current topics in the main heatmap and with a list of the recommended topics, those that appear in the secondary heatmap (green one). This modal allows users to select the topics they wish to maintain in the main heatmap for further evaluation. When users finish their selection, they must click the "Update preferences" button.



Choose the topics relevant for you

Selected topics

Search

Topic ID	Name	Evaluate
0	0_kike fuck_fuck kike_fuck_kike	☑
1	1_killing muslims_kill muslims_muslims killing_killing	☑
2	2_kike jews_hate kikes_kikes_kikes jews	□
3	3_banks_money_debt_bank	□
4	4_tax_taxes_money_pay	☑
5	5_whites_white_white people_blacks	□
6	6_trump kike_trump fucking_trump_kike trump	☑
7	7_altright_alt_alt right_label	☑
8	8_israel_kike israel_fuck israel_israel fuck	☑
9	9_kikes muslims_muslim kike_kike muslim_muslims kikes	☑

Showing 1 to 10 of 10 rows

Recommended topics

Search

Topic ID	Name	Evaluate
1	1_killing muslims_kill muslims_muslims killing_killing	□
5	5_whites_white_white people_blacks	☑
15	15_jew faggot_faggot jew_jewish faggot_faggot	☑
18	18_sage_kike sage_sage sage_saged	□
20	20_women_feminism_men_woman	☑

Showing 1 to 5 of 5 rows

Figure 5: This modal allows users to add and remove topics from the main heatmap. MultiModalTopicExplorer shows a list of the current topics in the main heatmap. It also shows a list of the recommended topics.

3.1 Keywords recommendation system

As we described above, users can indicate their preferences by selecting the keywords of their interest. Given that the number of possible options is extensive, MultiModalTopicExplorer recommends keywords that might be worthy of including. To do so, we follow several steps. First, we employ word embeddings to get a vector representation of the corpus terms. Based on the co-occurrence of terms, word embeddings create a reduced multi-dimensional representation of a corpus (Mikolov et al., 2013). Such representation can identify the semantic proximity among the corpus terms and expose the semantic context in which they are used (Rho et al., 2018; González-Pizarro et al., 2022).

Every time that the user selects a new keyword in the modal view (see Figure 4), the system creates a vector that represents the chosen keywords based on their word embedding representation. Then, we use cosine similarity to measure the similitude between the generated vector and the vector representation of the corpus terms. We always provide users with a list of ten keywords. Seven of them have the highest similarity to the generated vector, and three have the lowest similarity. We took this decision to reduce the "Filter bubble" (Nguyen et al., 2014) and provide users with a higher diversity of content.

Word embeddings are one of the most efficient ways to represent words. They occupy less space than the traditional one-hot encoded vectors, and they maintain the semantic representation of words. In this project, we use Glove (Pennington et al., 2014), an unsupervised learning algorithm for obtaining vector representation of terms. We used the pre-trained model trained on Wikipedia and Gigaword, which includes 6 billion tokens and 400,000 keywords (University, 2014).

3.2 Topics recommendation system

The user-adaptive version of MultiModalTopicExplorer recommends topics based on users' preferences. To do so, we create a vector representation of topics based on their 20 most relevant keywords. We only considered the terms that appear in our pre-trained Glove word embeddings model. When keywords are not present in the word embedding, we look for their stem generated by the PorterStemming algorithm (Willett, 2006). For instance, if the keyword "happiness" is not present in Glove, we look for its stem "happy". If that is the case, that is the word representation that we use.

In this system, a topic "a" $\overrightarrow{T_a}$ will be represented by the set of eligible keywords, $e^{a}_{d^{th}}$, where d^{th} indicates the position in the list of keywords for the Topic $\overrightarrow{T_a}$. Thus, let $\overrightarrow{U}$ the vector representation of the keywords chosen by the users. The similarity between the user-chosen keywords $\overrightarrow{U}$ and a topic "a" $\overrightarrow{T_a}$ will be indicated by the following equation:

$$Similarity(\overrightarrow{U}, \overrightarrow{T_a}) = \frac{\overrightarrow{U} \cdot \overrightarrow{T_a}}{\|\overrightarrow{U}\| \|\overrightarrow{T_a}\|} = \frac{\sum_{i=1}^{n} \overrightarrow{U}_i \overrightarrow{T}_{ai}}{\sqrt{\sum_{i=1}^{n} (\overrightarrow{U}_i)^2} \sqrt{\sum_{i=1}^{n} (\overrightarrow{T}_{ai})^2}} \quad (1)$$

where n indicates the number of selected terms.

MultiModalTopicExplorer always recommends a set of five topics. Three of them with the highest similarity with the chosen keywords, and two of them with the lowest similarity. Thus, we mitigate the "filter bubble' (Nguyen et al., 2014) and provide a rich set of options to users.

4 User study

We conducted a within-subjects user study to understand how the extended functionalities of MultiModalTopicExplorer support users during the exploration and evaluation of topic models. In our study, we asked participants to identify the quality of topics by using the non-adaptive (baseline system) and the user-adaptive (experimental system) versions of MultiModalTopicExplorer. We expected that users took less time and reported a lower workload when analyzing topics with the user-adaptive MultiModalTopicExplorer system. We provide details about the study design, user tasks, and technical implementation below.

4.1 Apparatus and materials

We used a desktop computer available in the research lab to run both versions of MultiModalTopicExplorer. We also used an eye-tracker to gain insights into user behaviour while analyzing topics. We attached a Tobii Pro Fusion eye-tracker to the computer screen, and we used the Tobii Pro Lab software to collect eye-tracker data.

A counterbalancing sheet helped with assigning each participant their IDs and the order in which they would interact with both topic modeling systems. After using the system, each participant had to answer questionnaires delivered to them via Qualtrics[1]. An additional laptop was placed on the side for the participants to answer those questions.

We used a modified version of the NASA Task Load Index (NASA-TLX) (NASA, 2022; Hart and Staveland, 1988) to assess the user experience (see Table 1). This widely used instrument measures the overall perceived workload when executing tasks. The questionnaire identified six dimensions: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration level. We removed from the original questionnaire the question associated with temporal demand because we captured the time users spent using the system with the eye-tracker. We also added a new dimension, “Relevance of topics”, to allow users to express if the evaluated topics match their preferences. We asked participants to answer this questionnaire twice. After using the baseline system, and after using the experimental system.

A third questionnaire was created to determine which version of MultiModalTopicExplorer they prefer (see Table 2). As can be observed, the terms “baseline-system” and “participant-adaptive system” were avoided when creating the questions. This was done to mitigate any possible bias that could have been induced due to the nomenclature of these systems.

[1] https://ubc.qualtrics.com/

Table 1: User experience questionnaire

Dependent variable	Question
Physical demand	How physically demanding was it to rate topics using this system?
Mental demand	How mentally demanding was it to rate topics using the system?
Task success	How successful do you think you were in rating topics?
Task performance	Using this system, how hard did you have to work to achieve your perceived level of performance?
Effort	How confident are you with your task performance using the system?
Relevance of topics	Do the shown topics match your interest?

Table 2: Questionnaire to assess participants' preferences.

Which system showed you interesting topics to rate?
Which system showed you topics based on your preferences?
Which system helped you discover new kinds of topics?
Which system are you most likely to choose for rating topics in the future?
Using which system are you most satisfied with for the task of rating topics?
Overall, which system do you prefer?

4.2 User study Design

The experiment used a 1X2 within-subjects design. The independent variable was the type of topic modeling system shown to the participant, which had two levels: the baseline (non-adaptive) system and the experimental (user-adaptive system). The task for the experiment was to rate the topics presented in the system from 1-5 based on their coherence/quality. 1 meaning that the topic is incoherent and 5 meaning that it is highly coherent. A topic is considered coherent if there appears to be some logical connection between the images, keywords, and documents that form the topic. The seven dependent variables were task completion time, level of physical demand, level of mental demand, task success, task performance, effort, and relevance of topics. The task completion time was determined from the timestamp when the participant began rating the topics to the timestamp when they ended rating all topics. The other variables were a part of the user experience questionnaire and were measured in the form of Likert scale ratings ranging from 1 to 5 (see Table 1). Here, 1 means that the dependent variable had a low influence on the participant's experience and 5 means that the dependent variable had a strong influence on the participant's experience. Additionally, participant preferences were considered with a third questionnaire (see Table 2).

4.3 Hypotheses

We modify MultiModalTopicExplorer to allow users to find and evaluate topics of their interest. When participants use the user-adaptive system, we expect that they report lower workloads and take less time in completing the assigned task. Thus, in the user study, we test the following six hypotheses:

- **H1**: Participants will take less time to complete the task with the user-adaptive system than with the non-user-adaptive system.
- **H2**: Participants will undergo less mental demand when completing the task with the user-adaptive system than with the non-user-adaptive system.
- **H3**: Participants will have higher task success with the user-adaptive system than with the non-user-adaptive system.
- **H4**: Participants will have higher confidence in task performance with the user-adaptive system than with the non-user-adaptive system.
- **H5**: Participants will require less effort when completing the task with the user-adaptive system than with the non-user-adaptive system.
- **H6**: Participants will find the topics in the user-adaptive system more relevant than those in the non-user-adaptive system.

4.4 Dataset

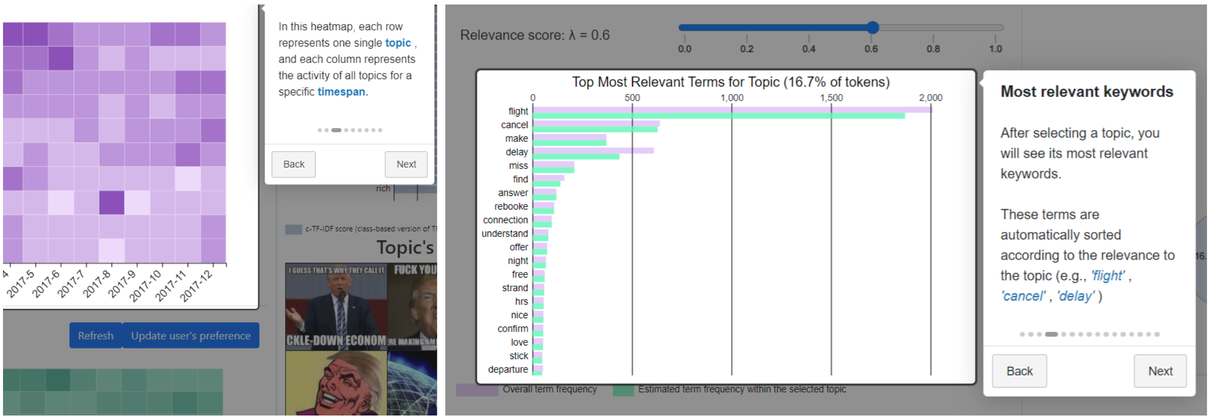
We asked participants to analyze topics from 4chan, particularly the Politically Incorrect board (/pol/). This is the main board for discussing world events and politics and is infamous for spreading conspiracy theories (Zannettou et al., 2017; Tuters et al., 2018) and racist/hateful content (Hine et al., 2017; Zannettou et al., 2020). We used the publicly available dataset released by (González-Pizarro and Zannettou, 2022); this dataset includes 573,513 Antisemitic/Islamophobic multimodal posts shared on 4chan in the period between July 1, 2016, and December 31, 2017. The topic model was generated by using BERTopic (Grootendorst, 2022). This is a new topic modeling technique that leverages Transformers to create dense clusters allowing for easily interpretable topics while keeping essential words in the topic descriptions.

Why this dataset? Social media sites such as Twitter, Facebook, and 4chan allow participants to share their ideas and opinions instantly. However, there are several ill consequences, such as online harassment, trolling, cyber-bullying, fake news, and hate speech. Exploring these conversations could help us understand how these communities interact on these platforms. Moreover, it is the first step before creating automated hate speech detection and content moderation systems.

4.5 Participants recruitment, training, and evaluation procedure

We recruited 6 participants (1 female, ages 23-26), most of whom were students from our research lab. To ensure some level of diversity, the participant population comprised 5 CS students and 1 non-CS student. At the end of their participation, participants were remunerated with food.

As the experiment was a within-subjects design, all participants interacted with both topic modeling systems. The order in which each participant was shown both systems was counterbalanced with a Latin square design. Before the experiment began, each participant was given a quick tutorial on both systems (see Figure 6). The tutorial included a brief description of the topic modeling systems, the type of information presented, and how it should be interpreted. Then, the participant was introduced to the concept of topic quality/coherence, which led to a description of the task they were required to perform using both systems. The participant was then shown some sample topics from the dataset to provide them with more information about the kind of topics they would be rating. Due to the sensitive nature of these topics, the participant was informed that they might discontinue their participation at any point in time if they were uncomfortable. Next, the participant was introduced to the eye-tracker attached to the desktop monitor. They were informed that the eye-tracker would first include a short calibration session, followed by the real monitoring beginning when they use the systems. With this, the tutorial ended.



In this heatmap, each row represents one single **topic** , and each column represents the activity of all topics for a specific **timespan.**

Most relevant keywords

After selecting a topic, you will see its most relevant keywords.

These terms are automatically sorted according to the relevance to the topic (e.g., *'flight'* , *'cancel'* , *'delay'*)

Figure 6: Snapshots of the interactive tutorial explaining the representation of the topics (left) and the chart with the most relevant terms (right)

Once the eye-tracker started recording the participant, they underwent a short calibration session followed by a full view of the first system assigned to them. An experimenter was present for all questions that the participant may have. After the participant finished rating all topics presented by the system, the experimenter asked them for their consent to store their eye-tracking data. Then, the participant answered a questionnaire about the system they had just interacted with (see Table 1). This process was then repeated for the second system that was shown to the participant. Once they finished independently evaluating both systems, they were provided with a third questionnaire asking them to compare the two versions of MultiModalTopicExplorer (see Table 2).

5 Results

We present our results in three steps. First, we display the results obtained for all dependent variables to determine if our hypotheses were supported or not. Second, we examine the results from the third (comparative) questionnaire to make inferences on overall participant preferences. Third, we present our results from the eye-tracking data obtained from all participants.

5.1 Participant data for all dependent variables

As mentioned in section 4.2, the seven dependent variables were task completion time, level of physical demand, level of mental demand, task success, task performance, effort, and relevance of topics.

5.1.1 Task completion times

To measure the task completion time, we only consider the time spent in rating the quality of the topics. We took this decision because the initial interaction with the user-adaptive system involves the additional step of selecting keywords and topics that interest the participant. Table 3 presents the task completion times for each participant when interacting with the baseline system (non-user-adaptive) and experimental system (user-adaptive system).

We perform two Shapiro-Wilk tests for the task completion times corresponding to each topic modeling system to check for the normality of the data. Based on the tests, both sets of task completion times do not depart significantly from normality (non-participant adaptive system: W(6) = .918, p = .580; participant adaptive system: W(6) = .958, p = .926). Hence, to test for significance, we perform a right-tailed paired t-test. No statistically significant difference was observed between the baseline and the experimental systems, t(5) =0.3, p=.629. We also determine the effect size with Cohen's d (Lakens, 2013), which is 0.08, implying that a minimal effect size is observed. Due to this minimal observed effect and the low power of our test, we did not find support for **H1:** *Participants will take less time to complete the task with the User Adaptive system than with the Non-User Adaptive system.*

Table 3: Time (in minutes) spent in rating topics in the baseline system (non-user-adaptive) and experimental system (user-adaptive)

	Baseline system	Experimental system
Participant #1	10.33	11.28
Participant #2	7.08	6.13
Participant #3	17.43	14.10
Participant #4	19.46	18.16
Participant #5	15.26	13.55
Participant #6	8.29	12.41
Mean	12.98	12.61
Standard Deviation	5.12	3.94
Standard Error Measure (SEM)	2.09	1.61

5.1.2 Dependent variables from the system-specific questionnaires

This section analyzes the remaining six dependent variables: physical demand, mental demand, task success, task performance, effort, and relevance of topics. As these dependent variables are in the form of Likert scale ratings (ranging from 1-5), we standardize the scores with their means and then normalize them so that all scores fall in the range of [0, 1]. Then, we perform five Shapiro-Wilk tests, one for each dependent variable, to check if there is a significant departure from normality in the obtained scores. For all five tests, the received scores depart significantly from normality. Hence, to check for a significant difference between the scores for the baseline system and the participant-adaptive system for all five variables, we perform five Mann-Whitney U tests. The results of these tests are in Table 4.

From Table 4, we observe that none of the differences are deemed statistically significant. However, there is a possibility that this insignificance is a result of the small sample size. So, we calculate the effect size (r) for each of the six dependent variables. We observe a small effect size of task success and no effect size for the other variables. This indicates the possibility that our tests lack power. Ultimately, there is not enough evidence that a significant difference is observed for these dependent variables. Hence, we reject hypotheses **H2**, **H3**, **H4**, **H5**, and **H6**.

Table 5 shows the average of the resulting scores per dimension. Participants generally note that the user-adaptive system requires lesser physical and mental demand and has a higher task success rate. However, we also observe that participants typically have lesser confidence in their task performance and require more effort to rate topics when interacting with the user-adaptive system. Additionally, participants note that neither topic modeling system gives them relevant topics.

Table 4: Results of the Mann-Whitney U Tests performed on each dependent variable

Dependent variable	Mann-Whitney U Test with Z scores	Effect size r $(z/\sqrt{N})$
Physical demand	U = 14, $N_{bas} = N_{exp} = 6$, p=3 Z=-0.091, p=0.93	0.04
Mental demand	U = 15, $N_{bas} = N_{exp} = 6$, p=5 Z=0.40, p=0.69	0.16
Task success	U = 14, $N_{bas} = N_{exp} = 6$, p=5 Z=-0.56, p=0.58	0.23
Task performance	U = 17.5, $N_{bas} = N_{exp} = 6$, p=5 Z=0, p=1	0
Effort	U = 17.5, $N_{bas} = N_{exp} = 6$, p=5 Z=0, p=1	0
Relevance of topics	U = 17.5, $N_{bas} = N_{exp} = 6$, p=5 Z=0, p=1	0

Table 5: Average Standardized and normalized Likert scale responses

Dependent variable	Baseline system	Experimental system
Physical demand	0.14	0.08
Mental demand	0.64	0.53
Task success	0.84	0.91
Task performance	0.86	0.80
Effort	0.43	0.38
Relevance of topics	0.05	0.06

5.2 Results from the comparative system questionnaire

Table 6 shows the results of the comparative questionnaire. First, we note the system that most participants for each criterion selected. Based on this, we perform six one-proportion z-tests for each of the six criteria to compare if the observed preference is significant or not. The results of these tests are in Table 7. We notice that the preference for the user-adaptive topic modeling system is significantly higher than the other options when it comes to the criteria of (1) which system helps the participant discover more topics, (2) which system will the participant most likely choose for rating topics and (3) which system does the participant prefer overall. This suggests that the user-adaptive topic modeling system has certain merits when it comes to user experience, especially when the task is related to topics discovery and topics rating. For some of the other criteria, we observe some surprising results. For example, most participants mention that neither system showed them more interesting topics and that they achieve the highest amount of task satisfaction with the non-adaptive system. However, as shown in Table 7, these results are not statistically significant.

Table 6: Results from the comparative questionnaire.

Criteria	Baseline	Experimental	Neither
Which system showed you more interesting topics?	0	2	4
Which system showed you topics based on your preferences?	2	3	1
Which system helped you discover more topics?	1	5	0
Which system are you most likely to choose for rating topics?	0	5	1
Which system are you most satisfied with for the task?	4	2	0
Overall, which system do you prefer?	1	5	0

Table 7: Results of the one-proportion z-tests (ns=not supported, * = p<0.05, ** = p<0.01)

Criteria	Preferred system	Z value	CI
Which system showed you more interesting topics?	Neither (4)	1.76 (ns)	[0.29, 1.04]
Which system showed you topics based on your preferences?	Experimental (3)	0.89 (ns)	[0.10, 0.90]
Which system helped you discover more topics?	Experimental (5)	2.62 (**)	[0.53, 1.13]
Which system are you most likely to choose for rating topics?	Experimental (5)	2.62 (**)	[0.53, 1.13]
Which system are you most satisfied with for the task?	Baseline (4)	1.76 (ns)	[0.29, 1.04]
Overall, which system do you prefer?	Experimental (5)	2.62 (**)	[0.53, 1.13]

5.3 Analysing participant behaviour with a eye-tracker

Eye-tracking data comprises two features: fixations and saccades (see figure 7). Fixations are periods when the eye is kept aligned with the target for a particular duration, processing the image details. On the other hand, Saccades are the type of eye movement used to move the fovea rapidly from one point of interest to another. Hence, fixations help analyze if the eye is trying to process a particular set of information, whereas saccades monitor the shift in focus from one fixation to another.

Our study focuses on the number of fixations generated by the participant when they are interacting with the topic modeling systems. Before performing our analysis, we define specific areas of interest (AOIs) in both topic modeling systems. These AOIs have been highlighted in figure 8 . Each rectangle with a dark border is an AOI. For both topic modeling systems, there are four specified AOIs. To analyze the eye-tracking data, we consider the number of fixations within these AOIs for both topic modeling systems. This is done to check if there is an observed difference in participant behaviour when interacting

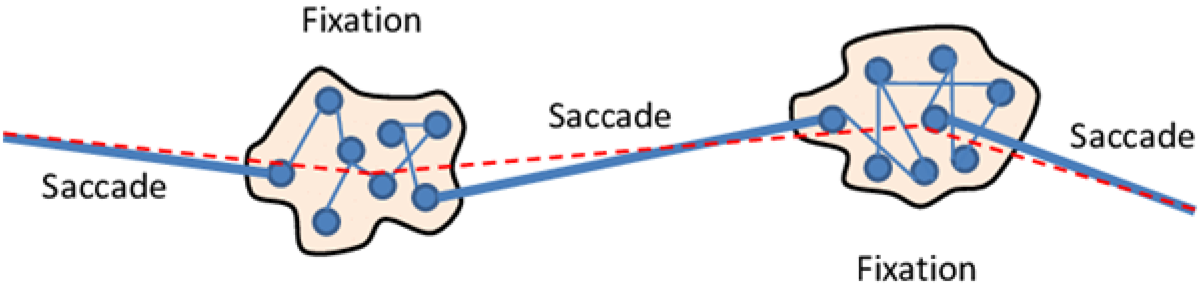
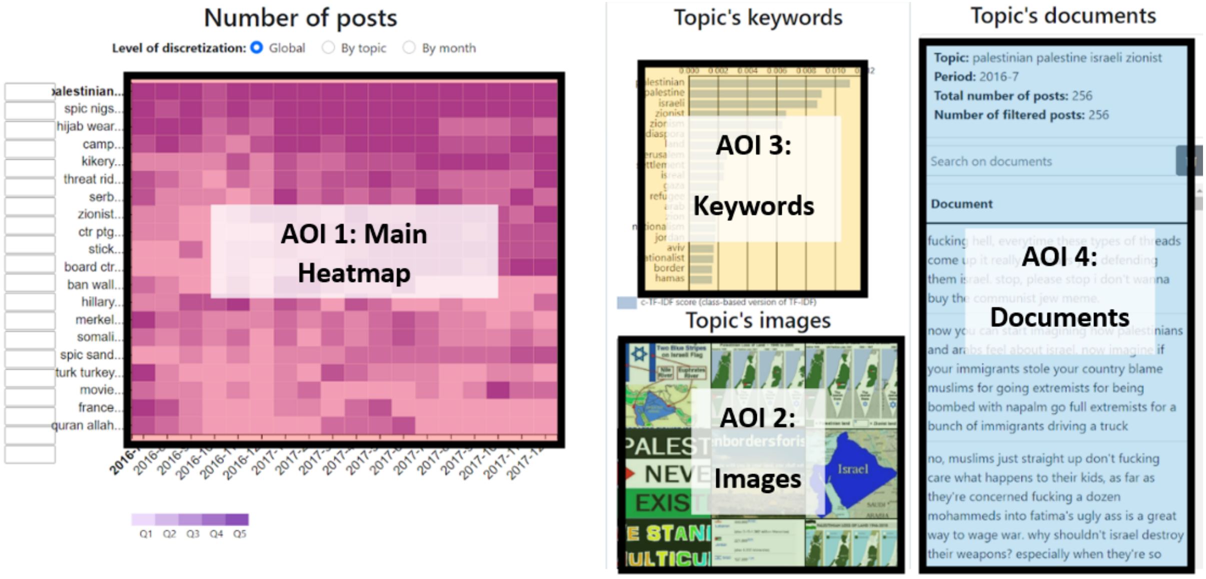


Figure 7: Two types of data measured with the eye-tracker



Topic: palestinian palestine israeli zionist
Period: 2016-7
Total number of posts: 256
Number of filtered posts: 256

Search on documents

Document

fucking hell, everytime these types of threads come up it really [illegible] defending them israel. stop, please stop i don't wanna buy the communist jew meme.

now you can start imagining how palestinians and arabs feel about israel. now imagine if your immigrants stole your country blame muslims for going extremists for being bombed with napalm go full extremists for a bunch of immigrants driving a truck

no, muslims just straight up don't fucking care what happens to their kids, as far as they're concerned fucking a dozen mohammeds into fatima's ugly ass is a great way to wage war. why shouldn't israel destroy their weapons? especially when they're so

Figure 8: The specified Areas of Interest (AOIs) for both topic modeling systems

with topics provided to them versus when they are interacting with topics personally selected by them.

Four Shapiro-Wilk tests, each for the number of fixations per participant with respect to one AOI, indicate that all data sets deviate significantly from normality. Hence, we perform four Mann-Whitney U tests with the number of fixations corresponding to each AOI to check for significance. Table 8 shows the results of the four tests. Our results show that the differences in the number of fixations for each topic modeling system are not statistically significant for any AOI. However, we do observe a small effect size of documents, hinting at the low power of this test. Overall, we do not observe any notable difference.

6 Discussion

This work explores the potential of making MultiModalTopicExplorer user-adaptive to allow users to inspect and evaluate topics relevant to them. We conducted a user study to identify the potential of the added functionalities. Our results show no statistically significant difference for any of the six dependent variables of the user-experience questionnaire. Upon further examination, we believe that these results

Table 8: Results of the Mann-Whitney U Tests performed on number of fixations corresponding to each topic modeling system for every area of interest

Area of Interest (AOI)	Mann-Whitney U Test with Z values	Effect size
Main heatmap	U = 12, $N_{bas} = N_{exp} = 6$, p=2 Z=0, p=1	0
Images	U = 10, $N_{bas} = N_{exp} = 6$, p=2 Z=-0.42, p=0.67	0.17
Keywords	U = 11, $N_{bas} = N_{exp} = 6$, p=2 Z=-0.21, p=0.83	0.09
Documents	U = 8, $N_{bas} = N_{exp} = 6$, p=2 Z=-0.84, p=0.40	0.34

might be explained by the nature of the selected dataset. The topics are skewed towards a particular kind of content (see Section 4.4) because the dataset mainly includes Antisemitic and Islamophobic content. As a result, even though the actual topics shown to users with both systems were different, these topics were from the same category. Hence, in this case, user-adaptivity was hardly noticeable and, in some cases, futile. In this section, however, we consider the actual averages of the normalized and standardized Likert scale scores and discuss them qualitatively.

Regarding the average scores for the user-experience questionnaires, we believe that task success score is higher in the user-adaptive system because participants can select how many topics they want to rate and swap existing topics for other topics if they need to. On the other hand, the confidence in task performance is possibly lower because many topics hardly have any documents, which makes it difficult to rate their quality. Most participants face a dilemma: Are coherent the topics that show a high correlation between their most relevant keywords, most relevant images, and most relevant documents, even in the cases when they have a short number of documents associated with them? The standard definition of coherence does not consider the number of documents (Röder et al., 2015); thus, such topics should be rated with a high coherence score. However, we noticed that user study participants were still confused about this matter.

We also observe that the user-adaptive system seems to require more effort to use. We believe that the two-fold recommendation system, one for keywords and another for topics, possibly leads to too many variables for the participant to control. Thus, it seems that much effort has to be put into initializing the system, and the focus of the users is taken out of the primary task (rating topics). Additionally, when browsing through the recommendations, participants are subjected to an increased amount of offensive content because of the very nature of the dataset. Hence, they have to put more effort into actively avoiding offensive content as much as possible while still browsing through the recommendations to select keywords and topics.

Moreover, we observe that participants generally note that neither topic modeling system provides relevant topics. The answer to this might lies in the nature of the dataset. Most topics fall into broad clusters containing the same kind of information (Antisemitic/Islamophobic content). As a result, even if the participant selects a keyword that they think is from a different category, this keyword is likely a part of a topic related to Antisemitism and Islamophobia.

Finally, when using both topic modeling systems, we noticed that participants were unsure about how topics are formed. They did not understand why multiple topics are created with similar keywords and why some topics have a low number of documents. This could explain why participants report a higher task satisfaction when using the non-user-adaptive system. The baseline system always show the most popular topics. As these topics are the most popular, they usually have a sufficient amount of information for the purposes of rating.

6.0.1 Limitations, future work, and lessons learned

As with any study, our research has limitations. The chosen dataset was not broad enough to cater to users with different interests. Most topics were highly triggering and contained a similar type of information and references. In future studies, we plan to use large datasets with a more variety of topics, such as the traditional 20 newsgroup dataset (Albishre et al., 2015).

We also noticed a high amount of poor-quality topics. Those topics contained only a few documents and contained a high amount of redundant keywords. This is an interesting finding, considering that the recent paper of BERTopic was recently published promising high-quality topics (Grootendorst, 2022). Given the poor performance of this algorithm, we plan in the future consider other options, such as the traditional LDA (Blei et al., 2003) or the recently proposed Contextualized Topic model algorithm (Bianchi et al., 2021).

We also plan to improve the performance of our recommendation system. Topics usually are represented by a list of terms. We believe that by considering the order of the keywords, we can improve the representation of topics. We also plan to consider the activity in each topic into account, such as the number of documents associated with them. Thus, avoid users getting topics with a scarce amount of information.

Although powerful, topic modeling algorithms sometimes generate topics with incoherent or loosely connected terms (Smith et al., 2018; Wang et al., 2019; Bianchi et al., 2021). Part of this problem is because most topic modeling approaches focus on the co-occurrence of terms as the primary signal to detect the semantic relations among them (Harrando and Troncy, 2021). As a result, these algorithms do not capture semantic and lexical relations between words that are not present in the corpus (Harrando and Troncy, 2021; Song et al., 2020; **?**). Prior work has suggested using external knowledge to overcome this drawback (**?**), and common sense knowledge is one promising alternative (Harrando and Troncy, 2021).

In this light, we plan to design a visual common sense based component to allow users to identify the quality of topics quicker, especially when the relationship between the topics' terms is not easy to identify. We did several initial experiments to visualize the relationship between the most relevant terms of topics. In these experiments, we used ConceptNet (Speer et al., 2017). This is a large-scale concept-centric knowledge base (Chakrabarty et al., 2021) that models lexical and semantic relationships (e.g., "party" *like* "flu"; "flu" *not desires* "person"). We choose ConceptNet given its vast number of concepts (approximately 1.5M nodes) and types of relations (34 in total).

Then, we used the NetworkX Python library (developers, 2014) to identify the shortest path between the pair of terms, and then we created a compact representation of the resulting graph (see an example in Figure 9)

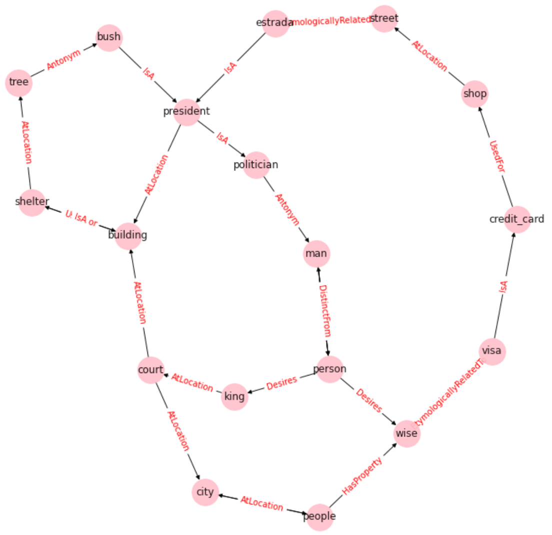


Figure 9: Visual representation of the relations between the keywords "visa", "court", and "president" based on a common sense knowledge base

.

After implementing these visualizations, we noted several challenges. First, some relations in ConceptNet do not make sense. For instance, the knowledge base indicates that "president" is the opposite of "man" (see Figure 9). Therefore, we need to implement mechanisms to identify spurious relations. Also, we noticed that visualizations are complicated to read when the number of keywords increases. To mitigate this problem, we plan to experiment with different visualization techniques such as matrix views.

We did not evaluate this visual common sense based component during the user study because we spent most of our time working on the front-end and back-end of the user-adaptive version of MultiModalTopicExplorer. On the front-end, we made several changes to improve the user experience.

We also made several modifications to allow the system to visualize more than 2,000 topics. This was a great success, given that the initial version of MultiModalTopicExplorer does not support more than twenty topics. Also, we made several changes to reduce the execution time of our recommendation systems.

7 Conclusions

In this manuscript, we presented an interactive visualization system to address some limitations of previous topic modeling visualization tools related to exploring a vast number of topics. The new functionalities of MultiModalTopicExplorer aim to support users while finding topics relevant to them. In addition, we conducted a within-subject user study to test the potential of the added functionalities of our tool. We asked participants to identify the quality of topics from a large-scale real-world dataset. Our results show that participants when using the user-adaptive system, reported lesser physical demand, lesser mental demand, and a higher task success rate while accomplishing the assigned task.

References

K. Albishre, M. Albathan, and Y. Li. Effective 20 newsgroups dataset cleaning. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 98–101. IEEE, 2015.

F. Bianchi, S. Terragni, and D. Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.96. URL `https://aclanthology.org/2021.acl-short.96`.

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

T. Chakrabarty, Y. Choi, and V. Shwartz. It's not rocket science: Interpreting figurative language in narratives. *arXiv preprint arXiv:2109.00087*, 2021.

J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces*, pages 74–77. ACM, 2012.

N. developers. Network analysis in python. `https://networkx.org/`, 2014.

W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pages 231–240. IEEE, 2011.

W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19 (12):2002–2011, 2013.

M. El-Assady, F. Sperrle, R. Sevastjanova, M. Sedlmair, and D. Keim. Ltma: Layered topic matching for the comparative exploration, evaluation, and refinement of topic modeling results. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, pages 1–10. IEEE, 2018.

F. Gonzalez-Pizarro and S. Alavi. Multimodaltopicexplorer: Topic modeling for exploring multi-modal data from asynchronous online conversations. Technical report, 2021.

F. González-Pizarro and S. Zannettou. Understanding and detecting hateful content using contrastive learning. *arXiv preprint arXiv:2201.08387*, 2022.

F. González-Pizarro, A. Figueroa, C. López, and C. Aragon. Regional differences in information privacy concerns after the facebook-cambridge analytica data scandal. *Computer Supported Cooperative Work (CSCW)*, 31(1):33–77, 2022.

M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

I. Harrando and R. Troncy. Discovering interpretable topics by leveraging common sense knowledge. In *Proceedings of the 11th on Knowledge Capture Conference*, pages 265–268, 2021.

S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

G. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *ICWSM*, 2017.

H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11): 15169–15211, 2019.

P. Kherwa and P. Bansal. Topic Modeling: A Comprehensive Review. 2019.

D. Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, page 863, 2013.

H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. ivisclustering: An interactive visual document clustering via topic modeling. In *Computer graphics forum*, volume 31, pages 1155–1164. Wiley Online Library, 2012.

S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. Topicpanorama: A full picture of relevant topics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 183–192. IEEE, 2014.

S. Malik, A. Smith, T. Hawes, P. Papadatos, J. Li, C. Dunne, and B. Shneiderman. Topicflow: visualizing topic alignment of twitter data over time. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 720–726. ACM, 2013.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

NASA. Nasa task load index (nasa-tlx). `https://human-factors.arc.nasa.gov/groups/TLX/downloads/HFES\_2006\_Paper.pdf`, 2022.

T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

J. Peter, S. Szigeti, A. Jofre, and S. Diamond. Topicks: Visualizing complex topic models for user comprehension. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 207–208. IEEE, 2015.

E. H. R. Rho, G. Mark, and M. Mazmanian. Fostering civil discourse online: Linguistic behavior in comments of# metoo articles across political perspectives. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–28, 2018.

C. M. Rivers and B. L. Lewis. Ethical research standards in a world of big data. *F1000Research*, 3(38): 38, 2014.

M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, pages 399–408. ACM, 2015.

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

A. Smith, V. Kumar, J. Boyd-Graber, K. Seppi, and L. Findlater. 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. ACM, 2018.

D. Song, J. Gao, J. Pang, L. Liao, and L. Qin. Knowledge base enhanced topic modeling. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 380–387. IEEE, 2020.

R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

M. Tuters, E. Jokubauskaitė, and D. Bach. Post-truth protest: how 4chan cooked up the pizzagate bullshit. *M/c Journal*, 21(3), 2018.

N. L. S. University. Glove: Global vectors for word representation. `https://nlp.stanford.edu/projects/glove/`, 2014.

J. Wang, C. Zhao, J. Xiang, and K. Uchino. Interactive topic model with enhanced interpretability. In *IUI Workshops*, 2019.

P. Willett. The porter stemming algorithm: then and now. *Program*, 2006.

Y. Yang, Q. Yao, and H. Qu. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics*, 1(1):40–47, 2017.

S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *IMC*, 2017.

S. Zannettou, J. Finkelstein, B. Bradlyn, and J. Blackburn. A quantitative approach to understanding online antisemitism. In *ICWSM*, 2020.