MultiModalTopicExplorer: A Visual Text Analytics System for Exploring a Collection of Multi-modal Online Conversations

Felipe González-Pizarro, and Soheil Alavi

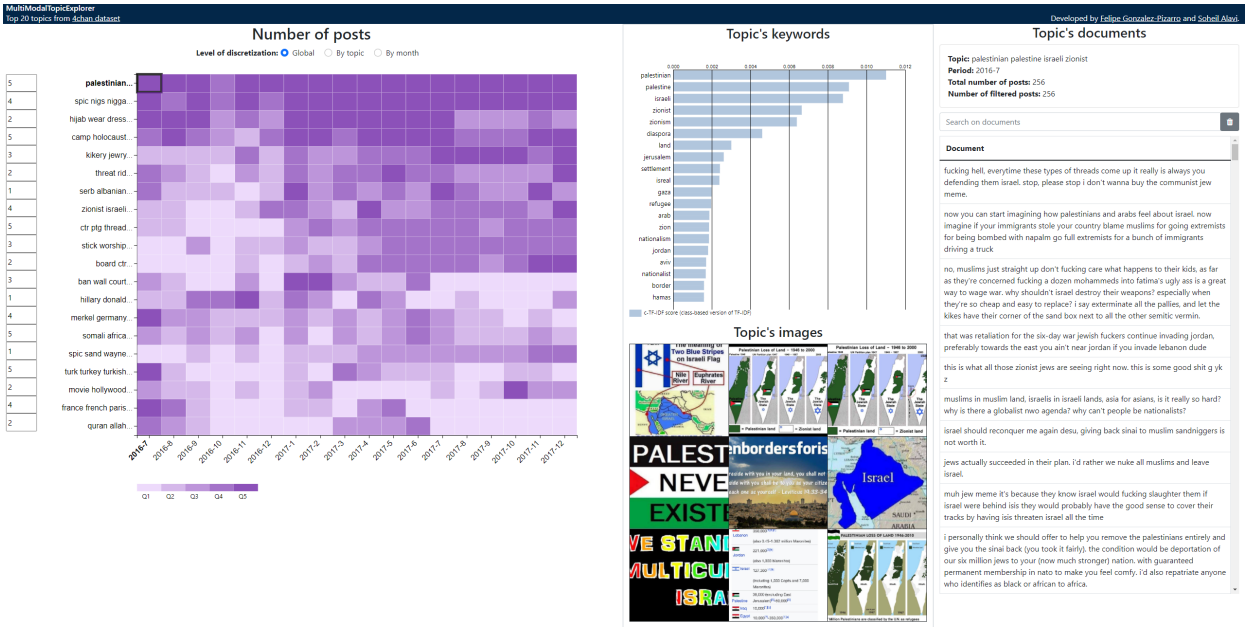

Fig. 1. MultiModalTopicExplorer layout. Each row from the matrix represents a topic, and each column represents a timespan. We use luminance to encode quantiles calculated on the number of posts on different discretization levels: global (divided by the total number of posts), by topic (divided by total number of posts for that topic), or by month (divided by total number of posts in that month). Users can click on matrix cells to visualize information from topics on periods. This example shows the most relevant keywords, most relevant images, and documents from the topic "palestinian", "palestine", and "israeli", in July 2016. Users can report the quality of each topic on the left of the matrix.

Abstract—
The constant increase in the volume of textual data has led to the development of various algorithms to summarize and understand this type of data. A promising solution is topic modeling, a statistical approach for extracting themes from large-scale datasets. Humans who directly interact with and interpret the output of topic modeling may rely on visualization tools to better interpret and evaluate the results. However, these tools still have limitations. First, they do not provide explicit functionalities to perform a qualitative analysis of the results. Second, current visual representations lack support for multi-modal conversations (image and text) and do not show the evolution of topics over time. We designed and developed MultiModalTopicExplorer, a web-based interactive visualization of topic modeling algorithm results to address these limitations. There are two key innovations in this work. First, MultiModalTopicExplorer allows users to report the quality of topics. Second, it shows the most relevant images for each topic to help users interpret topics. We designed and conducted a user study with four computer scientists. We asked them to evaluate two popular topic models using our tool. Our initial results show that participants felt successful in accomplishing the tasks, although it may have cost them some effort and mental demand.

Index Terms—topic modeling, multi-modal, bertopic, lda, clip, infovis, conversational texts

1 Introduction

The constant increase of the volume of textual data has led to the development of various algorithms intended to summarize and understand unstructured textual data [33]. A promising solution to this problem is topic modeling, a robust statistical approach for extracting core themes or *topics* from large text corpora. Thus, when a topic modeling algorithm is applied to a large corpus of documents, such as a collection of news articles, the results will include a list of topics, such as politics, economy, or sports. Each topic is defined by a set of descriptive words ranked according to their importance for the topic and by its distribution over the corpus documents [12].

Although powerful, topic models do not interpret themselves; therefore, humans must be involved [5, 9]. Visual text analytics researchers have designed algorithms and visual representations to support topic sense-making and interpretation, making probabilistic topic results legible and exploratory to a broader audience [10]. Topic modeling

- *Felipe González-Pizarro, PhD track Computer Science student at University of British Columbia. E-mail: felipegp@cs.ubc.ca.*
- *Soheil Alavi, PhD student at University of British Columbia. E-mail: salavis@cs.ubc.ca.*

visualization tools help in understanding topic models output and issues in modeling [18]; however, they still have limitations. Finding mechanisms to improve these visual representations is still an open challenge [16].

First, there is scarce support for qualitative analysis of topic models. Visual analytics systems might provide valuable insights about topic models' intrinsic properties and behaviors [6, 23]. However, current topic modeling visualization tools do not offer explicit functionalities to support users in evaluating topic modeling algorithm results.

Second, current topic modeling visualization tools lack support for multi-modal conversations. With the proliferation of web-based social media, there has been an exponential growth of asynchronous online conversations discussing a large variety of popular issues such as "US 2016 Election", or "Samsung watch release" [15]. Social media users post textual and image data to discuss these and other topics. Nevertheless, to the best of our knowledge, non of the current topic modeling visualization tools support image representation of topics.

Additionally, traditional topic modeling visualization tools do not consider how topics may change over time. The content discussed on social media websites is diverse, and users often react to current events as they happen, so the content is constantly evolving [39]. Therefore, topic modeling visualization end-users can be interested in analyzing the evolution of main discovered topics over time.

To address these limitations, we designed and developed MultiModalTopicExplorer, a web-based interactive visualization of topic modeling-generated topics. This tool aims to support users in evaluating topic model algorithms results. To do so, users are interested in answering the following questions:

- What are the most prevalent topics of the corpus?
- How do these topics evolve over time (e.g. during which historic incidents which topics become trending, etc.)?
- What is the meaning of each topic?

The remainder of the manuscript is organized as follows. Section 2 discusses two popular topic modeling algorithms. Section 3 summarizes related work about visual representations of topic models, discussing the current limitations and positioning this proposal. Section 4 describes our dataset. Section 5 describes tasks abstractions. Section 6 introduces our proposal. Section 7 includes the implementation details. Section 8 indicates the number of actual and estimated hours per milestone. Section 9 presents the user study method. Section 10 offers discussions, limitations, and future work. Finally, Section 11 includes our conclusions.

2 Topic Modeling Algorithms

This section briefly goes over two popular topic modeling algorithms we used in this project: LDA and BERTopic.

2.1 Latent Dirichlet Allocation

A large number of techniques have been proposed for the extraction and tracking of relevant topics over a large amount of text, where Latent Dirichlet Allocation (LDA) [1] is one of the most traditional and popular methods [27, 34]. The LDA model is based on the assumption that document collections have latent topics in the form of a multinomial distribution of words, which is typically presented to users via its *top-N* highest probability words [19]. LDA aims to discover topics from the corpus by finding the most optimal representation of two matrices: document-topic and topic-word. Figure 2 summarizes this process.

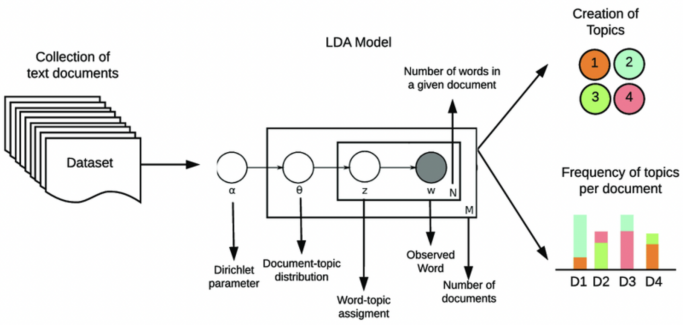

Fig. 2. LDA model: Alpha α controls per document topic distribution, M is the total documents in the corpus, N is the number of words in the document, w is the Word in a document, z is the latent topic assigned to a word, and theta (θ) is the topic distribution.

2.2 BERTopic

BERTopic [13] is a topic modeling technique that leverages Transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions. The algorithm can be split into three stages:

1. **Embed documents**: get document embeddings.
2. **Cluster embeddings**: cluster documents into semantically similar clusters.
3. **Extract representations for clusters**: create topic representations for clusters.

Figure 3 illustrates the internal architecture of BERTopic. BERTopic uses BERT [8] as the sentence embedding algorithm to get document representations during the first stage. Next, BERTopic uses UMAP [26] to reduce the dimensionality of the embeddings and cluster the resulting vectors with HDBSCAN [25]. In the third stage, BERTopic exploits from c-TF-IDF to generate representing topic keywords candidates. Finally, it calculates Maximal Marginal Relevance (MMR) [24] between candidate keywords and documents to improve the selection.

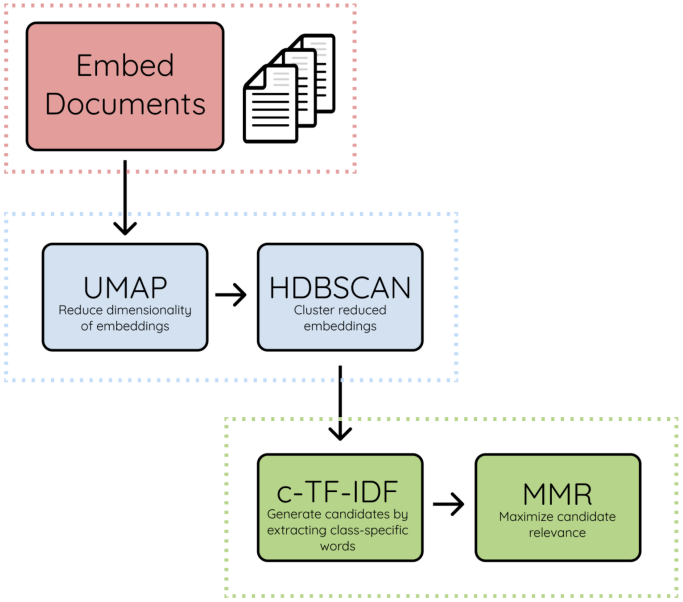

Fig. 3. BERTopic includes three stages: (1) Embed documents, (2) Cluster embeddings, and (3) Extract representation for clusters.

3 Topic Model Representation

The raw output of such topic modeling algorithms might be so complex that it can be difficult and time-consuming for non-expert users to understand it [4, 30, 37]. To address this need and add analytic value, previous work has explored different visual representation approaches to support

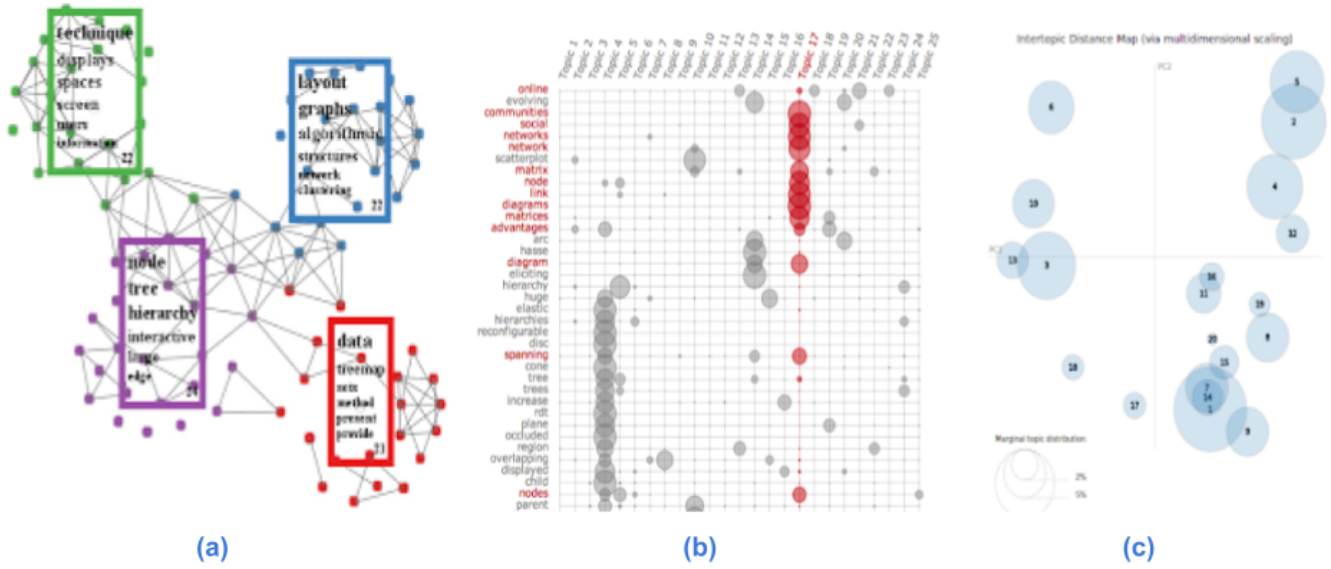

Fig. 4. Layout of global view of topics in: (a) iVisClustering (b) Termite (c) LDAvis

a human interpretation of topic models. Current topic modeling visualizations tools vary in topic keywords representations, documents view, and global views of topics.

3.1 Most relevant keywords

The most common output of topic modeling algorithms is the ranked list of the top terms of each particular topic [18]. They can be represented through different topic visualization techniques: (1) word lists; (2) word lists with bars; (3) word clouds; and (4) network graphs of terms [38]. Among these alternatives, simple visualizations such as word lists or word lists with bars allow users to understand topics quicker [38].

Usually, the top keywords are shown to users as a ranked list of the most frequent terms for each particular topic [4, 33]. In LDA, this is the same that ordering the terms by their topic-specific probability.

3.2 Most relevant documents

Sometimes the top keywords are not enough to identify the semantics of a topic [15]. That is the case when the most relevant terms are poorly connected, or when they include disparate [28] or generic terms (e.g., "yes", "like", "Mr", "maybe") [2, 21]. Due to that, it is better to include another level of information such as the most relevant documents to each topic to help end-users during topic interpretation [15, 40]. Indeed, previous research found that when topic modeling visualization tools display documents, users can read them to ensure topics' quality and verify if they satisfy their expectation [12].

There is no clear consensus regarding the best method to display documents associated with particular topics. For instance, visualizations that aim to support users in exploring asynchronous conversations position the most relevant documents of a topic according to their chronological ordering [14, 15]. Another method is to display the documents according to their contribution to the topic, as [39]. Thus, the most relevant documents always appear first. In LDA, this is the same as ordering the documents regarding the topic-document probability for each topic.

3.3 Global view of topics

Along with showing the most relevant keywords and the documents associated with topics, topic modeling visualization tools offer different layouts to help users get a global view of the topic model.

One alternative is to represent relevant keywords and documents from topics through a graph layout. That is the case of iVisClustering [20] where documents (graph nodes) from one topic are visualized as colored circles with the same color. The edges between nodes represent the similarity between documents based on cosine similarity. Controlling a slider makes edges with higher values than the slider value appear, and those with smaller values disappear. For each cluster, there is a color-bordered rectangle with the most representative keywords (see Figure 4 (a)).

A second approach consists in displaying the term-topic distributions through a matrix layout. In this approach, proposed in Termite [7], the rows correspond to terms and the columns to topics. It uses circular areas to encode term probabilities. Thus, the most frequent terms are represented by circles with a larger area (see Figure 4 (b)).

A third alternative consists in projecting the similarity between topics into a two-dimensional space. In this approach, proposed in LDAvis [36] (see Figure 4 (c)), the topics are represented as circles. Their centers are determined by computing the distance between topics and then using multidimensional scaling to project the inter-topic distances onto two dimensions. In this layout, each topic's overall prevalence is encoded using the areas of the circles, such that a more extensive area indicates a higher prevalence. This layout provides a global view of the topics, via their prevalences and similarities to each other, in a compact space.

3.4 Relevant commercial tools

As far as we know, there are not any relevant commercial tools or efforts from practitioners related to multi-modal topic modeling visualization.

4 Dataset

For this research effort, we focus on unstructured multimodal conversational data gathered from users' interactions on famous social media forums such as `4chan.org`.

4.1 4Chan dataset

4chan is an imageboard website with virtually no moderation. An *Original Poster* (OP) creates a thread by posting an image and a message. Content is organized in subcommunities, called boards with various topics of interest. Other users can post in the OP's thread with a message or an image. On 4chan, users do not need a registered account to post content. In this project, we focus on the politically incorrect board (pol), which has been shown to include a high volume of racist, xenophobic, and hateful content [29].

We used a random sample of the 4chan dataset [32], which contains threads of posts from the Politically incorrect board (/pol/). An example of a typical pol thread is given in Figure 5. Our base dataset contains more than 0.5 million posts over a period of 1.5 years (June 2016-Dec 2017). We took the following steps in our data preprocessing component:

1. Remove all html tags.
2. Lowercase the words.
3. Lemmatize the words.
4. Remove stopwords and punctuation.

Fig. 5. Example of a typical pol thread

During the training of our models, we tried different conditions in the data preprocessing: lemmatization, stemming, and removing stop words. With a preliminary qualitative analysis of the results, we concluded that lemmatized data with no stop words leads to a better performance of both models. Therefore, we used that combination as our base dataset.

4.2 Why this dataset?

Social media sites such as Twitter, Facebook, and 4chan allow users to instantly share their ideas and opinions. However, there are several ill consequences, such as online harassment, trolling, cyber-bullying, fake news, and hate speech. We believe that an exploration of these conversations could help understand how these communities interact on these platforms. Moreover, it is the first step before creating automated hate speech detection and mitigation systems.

4.3 Ethical considerations

We analyzed publicly available data, which are masked for privacy purposes. In addition, we followed standard ethical guidelines, not attempting to track users or deanonymize them.

5 Task Abstraction

MultiModalTopicExplorer aims to support users in evaluating topic models by answering questions such as: What are the most prevalent topics of the corpus? How do these topics evolve? What is the meaning of each topic? We expect users can be able to: :

1. Identify the most frequent topics from a corpus.
2. Identify the popular topics in a certain period of time.
3. Identify when a topic achieved its highest popularity.
4. Compare the frequency of the identified topics throughout the time.
5. Identify the most important words for each topic.
6. Identify the most relevant images for each topic.
7. Identify the relevant documents for each topic in a certain period of time.
8. Find documents that contain certain keywords.
9. Rate topics.

6 MultiModalTopicExplorer

Current topic modeling visualization tools do not provide explicit funtionalities to perform a qualitative analysis of the results. Also, there is a lack of support for multi-modal conversations (image and text) and do not show the evolution of topics over time. We designed and developed MultiModalTopicExplorer, a web-based interactive visualization of topic modeling algorithm results to address these limitations. The MultiModalTopicExplorer layout is illustrated in Figure 1. This section provides an overview of the visual interface features.

6.1 Identify and rate the most frequent topics

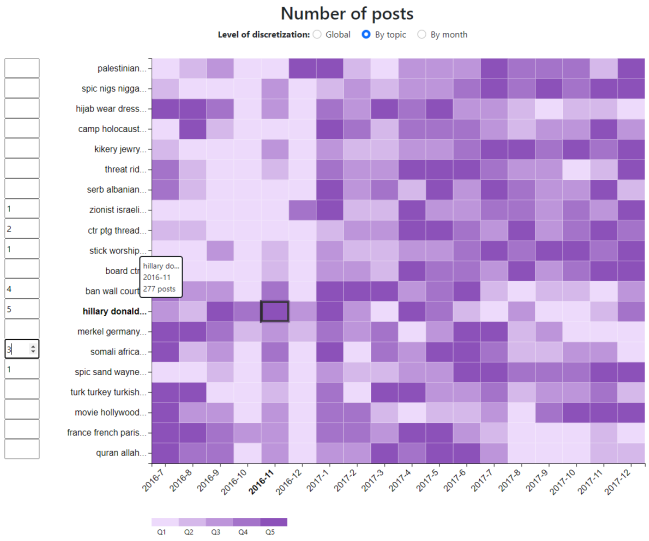

Fig. 6. MultiModalTopicExplorer allow users to identify and rate the most relevant topics from the corpus

In this project, we used LDA and BERTopic to identify the main themes from our dataset. While these algorithms can identify many topics (e.g., 100, 500, 1000 topics), we believe that users are especially interested in visualizing the most frequent ones.

We incorporated a matrix (see Figure 6) to allow users to perform *Task 1: Identify the most frequent topics from a corpus*. Each row represents a topic. The default name of each topic is defined by its most relevant keywords. The topics are sorted according to their total number of posts. Thus, the most popular topics appear first. A summary of what, why, and how analysis of the component is described in Table 1.

One of the goals of this proposed tool is to allow end-users to rate topics. We included a box on the left of the topics' names to allow users to *Task 9: Rate topics*. Users can type on those boxers a number between 1 and 5. Usually, a higher score indicates higher quality [19].

6.2 Evolution of topics

MultiModalTopicExplorer also allows users to identify the evolution of the most frequent topics over time. Figure 6 shows a matrix where each column represents a period (month), and each row represents a topic. We use luminance to encode quantiles calculated from different discretization levels: global, by topic, or by month. When "global" is selected, users can compare the popularity of two topics in any time

Table 1. What-Why-How analysis of most relevant topics component

What: Data	Table: one categorical key attribute (topic name), one quantitative value attribute (total number of posts associated to the topic).
What: Derived	Ordered key attribute total number of posts associated to the topic).
How: Encode	Express value attribute with vertical position.
Why: Task	Identify most frequent topics
Scale	Items: twenty.

period in the corpus. When "by topic" is selected, users can answer the question: When was this topic more popular? A darker color indicates a higher quantile. On the other hand, when "by month" is selected, users can answer: What were the most popular topics in each month?

Users can mouse over matrix cells to get the number of posts for that topic in that period. When users click a matrix cell, the most relevant keywords and most relevant images for that topic are displayed. Additionally, the conversation view on the right shows a list of posts for that topic in that specific period.

The matrix allow users to perform *Task 2: Identify the popular topics in a certain period of time*; *Task 3: Identify when a topic achieved its highest popularity*; and *Task 4: Compare the frequency of the identified topics throughout the time.* A summary of what, why, and how analysis of the component is described in Table 2.

Table 2. What-Why-How analysis of evolution of topics component

What: Data	Table: one categorical key attribute (topic name) and one ordered key attribute(month), one quantitative value attribute (number of posts).
What: Derived	Ordered attribute with five levels (we calculated quantiles)
How: Encode	2D matrix alignment of area marks, luminance-map.
Why: Task	Find clusters, outliers, summarize
Scale	Categorical attribute levels: dozens to hundreds (twenty rows, hundreds of columns). Quantitative attribute levels: 5 (luminance levels).

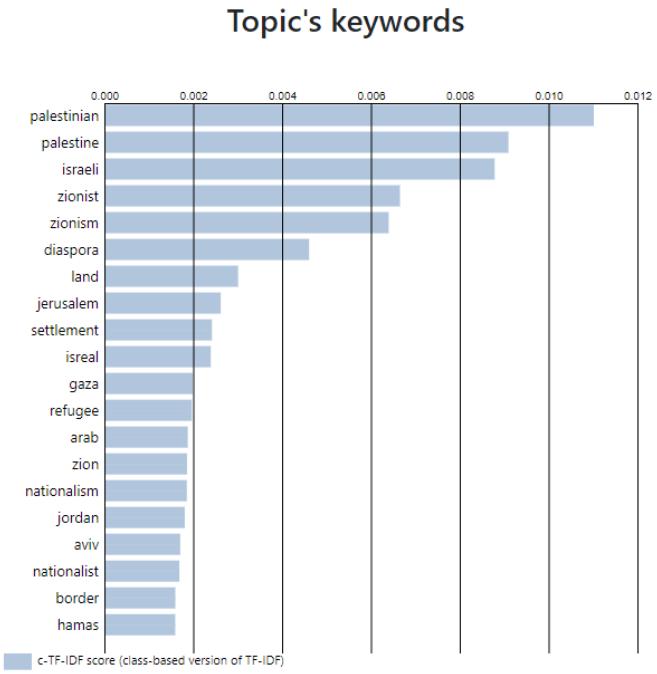

Fig. 7. MultiModalTopicExplorer allow users to visualize the most relevant keywords of the selected topic

Table 3. What-Why-How analysis of most relevant keywords component

What: Data	Table: one categorical key attribute (keyword), one quantitative value attributes (c-TF-IDF score)
What: Derived	Ordered key attribute (by c-TF-IDF score).
How: Encode	Line marks, express value attribute with aligned horizontal position, separate key attribute with vertical position.
How: Encode	Express value attribute (c-TF-IDF score) with vertical position
Why: Task	Lookup and compare values.
Scale	Key attribute: twenty levels.

6.3 Most relevant keywords

The most relevant keywords panel of MultiModalTopicExplorer (see Figure 7) depicts a horizontal bar chart for the most relevant terms to the selected topic. For each term, a bar is unfolded. This kind of linked selection allows users to examine a large number of topic-term relationships compactly and supports users in topic interpretation [36]. The most useful terms to a given topic are ranked according to their topic-term probability or c-TF-IDF score regarding if the results are from LDA or BERTopic, respectively.

This component allows users to perform *Task 5: Identify the most important words for each topic.* A summary of what, why, and how analysis of the component is described in Table 3.

6.4 Most relevant images

We believe that images can help users to have a complementary perspective on the meaning of each topic. Figure 8 shows the nine most relevant images of the topic. When users click on these images, they can see them full size. Also, they can see the original post. These images are retrieved from the 4chan dataset using CLIP [35], a neural network model, which recognizes a wide variety of visual concepts in images and associates them with names. We use this model to retrieve the most related images from the dataset to the top twenty keywords of each topic. To the best of our knowledge, this is the first topic modeling visualization tool that supports image content.

This component allows users to perform *Task 6: Identify the most relevant images for each topic.* A summary of what, why, and how analysis of the component is described in Table 4.

Table 4. What-Why-How analysis of most relevant images component

What: Data	Table: one categorical key attribute (image), one quantitative value attribute (relevance score to the selected topic)
How: Encode	Express value (relevance) with horizontal and vertical spatial position.
Why: Task	Identify and compare
Scale	Items: nine.

Fig. 8. MultiModalTopicExplorer allows users to identify the most relevant images for each topic.

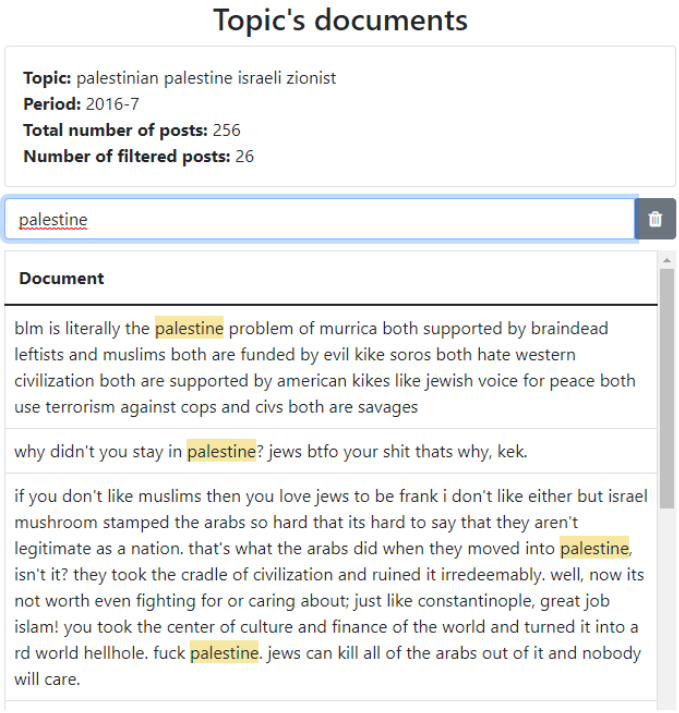

Fig. 9. MultiModalTopicExplorer allows to identify the relevant documents for each topic in a certain period of time. Also, it allows to find documents that contain certain keywords.

6.5 Conversations/documents view

The Topic's documents component (see Figure 9) allows users to perform *Task 7: Identify the relevant documents for each topic in a certain period of time*, and *Task 8: Find documents that contain certain keywords*. Users can search for specific keywords to understand how them are being used. We highlight the searched term in the documents. We also indicate the number of filtered posts. A summary of what, why, and how analysis of the component is described in Table 5.

Table 5. What-Why-How analysis of conversations view component

What: Data	Table: one categorical key attribute (document), one ordered key attribute (month)
How: Encode	Vertical position separate documents. Text marks. Color hue to highlights searched terms.
Why: Task	Part-to-whole relationship, Lookup documents and terms.
Scale	Items: hundreds to thousands

6.6 Scenario of use

In this section we present a usage scenario, showing how typical users would utilize different components in the MultiModalTopicExplorer to accomplish tasks we mentioned in section 5.

> John is an NLP researcher working on developing and analyzing topic modeling algorithms for conversational datasets gathered from social media such as 4chan. As part of his job, he has developed multiple different topic modeling algorithms for the 4chan website. He needs to do a qualitative analysis of the topics generated by the mentioned algorithms. 4chan dataset contains thousands of posts per month, and John wants to analyze the top most frequent topics in the corpus. John is also interested in seeing how these topics have evolved throughout the time and cross-checking them with the important events that have happened within that period to see if the topics match the mental model he has or not.

John starts to explore the top most frequent topics by checking out the topic-time matrix displayed on the left side of the screen in MultiModalTopicExplorer. Each row in this matrix represents a topic, and each column represents a month. Topics are sorted based on their overall frequency from top to bottom in the corpus. The luminance of matrix cells encode quantiles calculated from different discretization levels: global, by topic, or by month. While checking out the matrix in the global mode, John starts examining the topic named "hillary donald trump benie" (13th from the top). Investigating that row, he sees that the cell representing November 2016 is darker than other cells. By hovering his mouse over the cell, he finds that there were 277 posts for that topic in November 2016 (see Figure 10).

John decides to investigate topic number 13 further. He clicks on the proper cell representing the topic in November 2016 (Matrix["hillary donald trump bernie"]["November 2016"]). Immediately Topic's keywords view (see Figure 11) and the Topic's Images view (see Figure 12) get updated. Additionally, the Topic's documents view also gets updated (see Figure 13).

John wishes to grasp a better understanding of what are the contexts in which the keyword "hillary" has been used. He has two options.

First, he mouses over the keyword "hillary" on the barplot. That makes "hillary" bold and highlights all instances of it in the Topic's documents view. Then, by scrolling down the Topic's document view, John can see all the retrieved documents (whether they include "hillary" or not).

Second, John can search for the word "hillary in the Topic's documents view and use the filter option that MultiModalTopicExplorer provides to focus only on the documents that "hillary" has appeared in them (see Figure 14). It is worth mentioning that the number of filtered documents shown in the header of Topic's documents view gets updated after each search result. To start over with a different search,

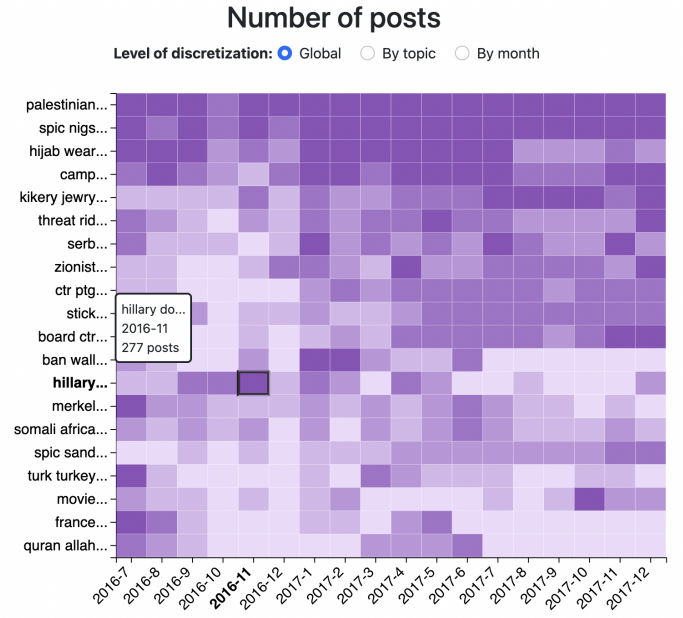

Fig. 10. The topic-time matrix in MultiModalTopicExplorer. The selected cell's luminance represent the quantile of the "hillary donald trump bernie" topic in 2016-11. Also by hovering on the cell a white window appears on the top left side of it, showing the following info: topic's name, time, number of posts in that period.

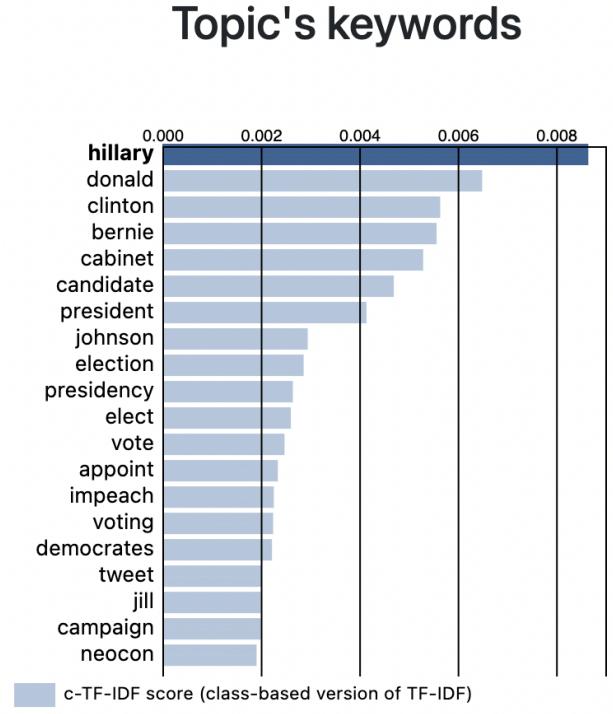

Fig. 11. The Topic's keywords view after selecting the "hillary donald trump bernie" topic from topics-time matrix. Hovering the mouse on "hillary" has made it bold.

Fig. 12. The Topic's Images view after selecting the "hillary donald trump bernie" topic from topics-time matrix. Top 9 relevant images are retrieved by CLIP. Clicking on each image opens a new window showing the full-size image and its original caption.

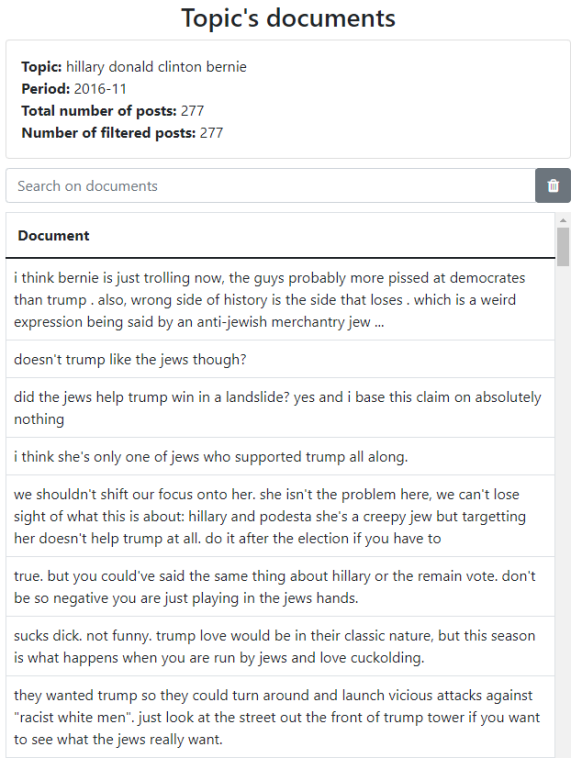

Fig. 13. The conversations/documents view after selecting the "hillary donald trump bernie" topic from topics-time matrix on November, 2016. The table shows a list of the documents belonging to this topic that has been posted during that period.

clicking on the garbage icon next to the search field would suffice to reset the results.

Examining the documents, John figures that this topic is about the US election in 2016. To gain a better understanding of the "hillary donald trump bernie" topic and further investigate his hypothesis, John checks out some of the relevant images that the Topic's images view has provided him. John finds face photos from the candidates Hillary Clinton and Donald Trump. He finds those images relevant and clicks on them, opening them in full size, and revealing their original caption.

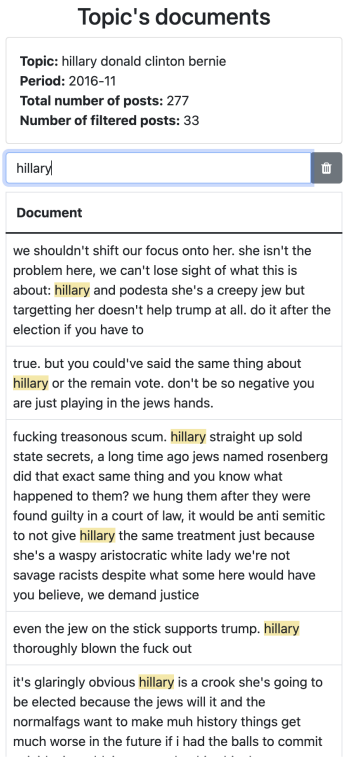

Fig. 14. The Topic's documents view after selecting the "hillary donald trump bernie" topic from topics-time matrix. After searching for "hillary" in the search field, all 33 relevant documents that contain "hillary" in them are retrieved while "hillary" is highlighted in them.

The images also confirm his hypothesis that the topic is about the US election matching his memories of the events in November 2016 during which the US election happened. "Hillary Clinton" and "Donald Trump" were the lead candidates, and it made sense to see their names as the most relevant keywords for this topic.

Finally, John rates the quality of this topic by selecting a number between 1 and 5. A higher score indicates a higher quality of the topic. He chooses five because all the information displayed is related to the same event. Then, he goes on to evaluate the next topic. He can, later on, compare the aggregated rated qualities of different topic models that he has implemented to determine which topic model is superior.

7 Implementation Details

The front-end of MultiModalTopicExplorer was implemented using: D3.js, Jquery, Bootstrap, Bootstrap-table, Highlight.js, HTML, and CSS. In our backend, we used FastAPI, Nginx, uvicorn, and gunicorn. The full code is available online at `https://github.com/gonzalezf/MultiModalTopicExplorer`. We also released a demo of our tool, which is accessible at `http://multimodaltopicexplorer.ml`. We recommend using Google Chrome or Microsoft Edge web browsers.

7.1 BERTopic

The input to BERTopic is the base dataset described in section 4. One of the main differences between BERTopic and LDA is that, unlike LDA, we do not need to generate bigrams and trigrams for BERTopic to make it possible to get phrases of length two or more as keywords. Instead, BERTopic handles that during its *extract representations for clusters* step automatically.

For this research effort, we exploited from open-source GitHub repository BERTopic [13]. As the first stage of BERTopic (getting BERT representations) was parallelizable, we used four GeForce GTX 1080 Ti GPUs with a frequency of 33MHz to speed up the training process. With the mentioned resources training the model for 0.5 million samples took around 2 hours.

One of the constraints of the current BERTopic version is it cannot fit the model to more than 130000 samples. Hence, based on the developer's recommendation, we split the data into randomly selected chunks of 130000 samples, fit the model to the first chunk, and transformed the other chunks to get the topics (predicted the appropriate topics for the rest).

Another difference between BERTopic and LDA is that BERTopic can automatically tune the topic number hyperparameter during the training. After the training step, we ended up with 815 topics.

7.2 LDA

In section 4, we explained how we preprocess the data by applying lemmatization and removing the stop words from the original text. In order to make LDA capable of finding phrases of more than one word as the keywords, we also calculated the bigrams and trigrams for the base corpus. After removing the stop words from bigrams and trigrams, we added them to the base dataset and created the preprocessed data for LDA.

For this research effort, we used the gensim LDAmulticore library, which can take advantage of multiprocessing on the CPU to speed up the training process. One of the important hyperparameters in training the LDA model is the number of topics, k, that we are interested in finding in the corpus. From our experience in training BERTopic, we guesstimated that the ballpark of optimal k would be somewhere between 600-1000 topics. Hence, given the limited time and resources, we trained LDA with six different values for k. The total training took 6 hours on a CPU machine with 9 cores. Next, we calculated C_v, the *Coherence Score*, for each of these models. The concept of topic coherence combines a number of measures into a framework to evaluate the coherence between topics inferred by a model. C_v measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and cosine similarity.

The results in Figure 15 show that the model trained with $k = 600$ has the highest Coherence Score. Hence, we used the outputs of this model for the rest of this paper.

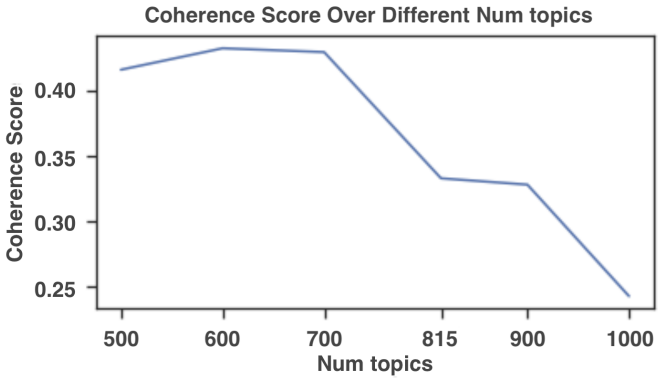

Fig. 15. Calculated coherence measure for the following list of topic numbers: 600, 700, 815, 900, 1000

7.3 Data Post Processing and Aggregation

After training the LDA and BERTopic models, the next step is extracting aggregated data necessary for the tasks described in Section 5. Our developed scripts extract the necessary information in the following steps:

1. Sort topics based on their frequency and get top 20 most frequent topics.

2. For documents (doc) associated with top 20 frequent topics create records of: <original doc, preprocessed doc, topic_id, topic_name, year, month>

3. Group the output of step 2 by date and topic to get the frequency of most frequent topics per month.

4. Retrieve the top 20 most relevant topic keys per topic. For LDA based on probability of words within the topic. For BERTopic based on c-TF-IDF score.

7.4 CLIP

We used Contrastive Language-Image Pre-Training (CLIP) [35] to identify the most relevant images from each topic. CLIP is a neural network trained on various (image, text) pairs. This model learns the relationship between a whole sentence and the image it describes; in a sense that when the model is trained, given an input sentence, it will be able to retrieve the most related images corresponding to that sentence.

To accomplish our goal, we followed several steps. First, we used CLIP to obtain a latent space embedding for each image from our dataset. Then, for each topic, we retrieved their 20 most relevant terms. We used those terms as "queries". For each topic, we used cosine similarity to obtain the nine most similar images to the most relevant keywords.

8 Milestones

Table 6 shows the milestones of this project and the estimated and actual number of hours required to complete it.

9 User Study Design

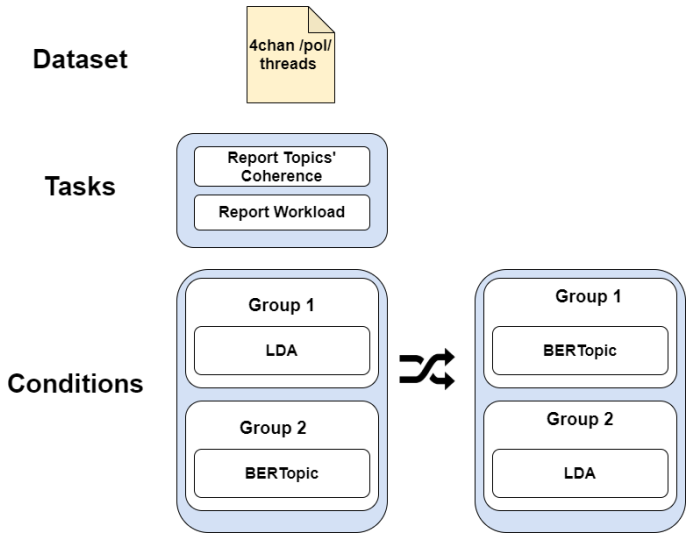

Fig. 16. Dataset, tasks, and conditions in our user study

This manuscript introduces MultiModalTopicExplorer, an interactive topic modeling visualization tool that aims to help NLP-researchers evaluate topic models' quality. One method to identify the quality of automatic generated topics is by measuring their coherence [31], which can be automatically calculated or reported by users [11, 19]. Topics are coherent when there are evident semantic relationships among their constituent components (e.g., keywords, documents, images) [11, 17, 19]. Considering this, we propose to test the following null hypothesis:

H_0: *There are no differences in the coherence of topics emerged from BERTopic and LDA.*

To investigate this hypothesis, we conducted a within-subjects user study to evaluate the performance of two popular topic modeling algorithms: BERTopic and LDA. We asked participants to interpret and report the coherence of topics using our proposed tool. They rated each topic on a 5-point scale. A higher value indicates a higher coherence.

In addition, we asked participants to report the perceived workload after completing the tasks. We used the NASA Task Load Index (NASA-TLX) to allow users to self-report the workload perceived on a scale from 0 to 100. This questionnaire identifies six dimensions: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration level. It is the most common method to evaluate and report the overall workload level perceived during the task [3]. Figure 16 summarizes the user study setup.

For our user study, we recruited computer scientists who understand the English language. Thus, they can read the top keywords and posts for each topic from the selected dataset. The user study was conducted online because of COVID-19 pandemic restrictions.

For the dependent variable, its normal distribution was analyzed by Shapiro-Wilk's test. To compare topics' coherence scores from each user study condition, We used the Mann-Whitney U test because our data do not follow a normal distribution. The statistical procedures were performed with a cut-off for significance at 0.05 using Python.

Figure 17 shows the distribution of the topics' coherence scores by conditions. We found significant differences between conditions (U = 4811.0 $N_{bertopic}$=80, N_{lda} = 84, $p < .001$), thus we can reject our null hypothesis: *There are no differences in the coherence of topics emerged from BERTopic and LDA*. Topics emerged from BERTopic have higher quality ($M = 3.9$, $SD = 1.21$) than topics emerged from LDA ($M = 2.92$, $SD = 1.09$) according to user study participants.

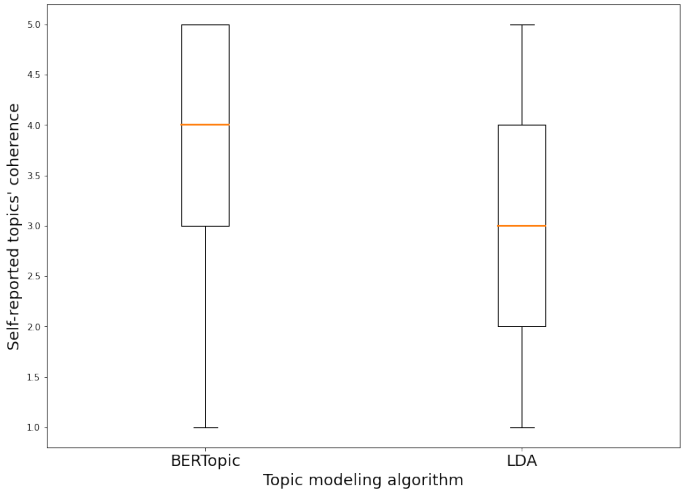

Fig. 17. Distribution of coherence scores in topics emerged from BERTopic and LDA. A higher score indicates a higher coherence

We asked participants to report the perceived workload after evaluating the topic models. Figure 18 shows the distribution of participants responses to the questionnaire.

10 Discussion and Future Work

We designed and developed a novel tool to support users in evaluating topic models. This section offers discussions regarding the implication of our results, limitations, and future work.

The results from our within-subject user study point out that people felt successful after evaluating two topic models using our tool, which hints that MultiModalTopicExplorer is helpful for the given tasks.

Our results from our user study also show that participants required effort and mental demand to accomplish the tasks. However, it is not clear if that is because our tool might be complex or challenging to use or because of the nature of the user study tasks. Our proposal contains

Table 6. Estimated (Est.) and actual (Act.) number of hours per milestone. We also report if the milestone is associated with the course CPSC 503, CPSC 547, or both.

Task	Est. Total	Est. Felipe	Est. Soheil	Act. Total	Act. Felipe	Act. Soheil	Courses
Literature review	6	3	3	6	3	3	Both
Brainstorming and mock-up design	8	4	4	8	4	4	Both
Write proposal	32	16	16	32	16	16	Both
Investigate previous similar implementations	10	5	5	10	5	5	Both
Learn D3.js	16	8	8	24	8	16	547
Data preprocessing	20	10	10	20	5	15	Both
BERTopic implementation and training	30	15	15	30	10	20	Both
(BERTopic) Integrate the model with the visualizer	10	5	5	20	10	10	Both
(BERTopic) Identify the most important topics and visualize them (Figure 1 (a))	10	5	5	15	7.5	7.5	Both
(BERTopic) Aggregate BERTopic topics for each month and visualize them (Figure 1 (b))	30	15	15	30	15	15	Both
(BERTopic) Find the most important words for each topic and Visualize the term C_TF_IDF (Figure 1 (c))	15	7.5	7.5	15	7.5	7.5	Both
(BERTopic) Retrieve sample of conversations for each month	10	5	5	3		3	Both
LDA implementation and training	30	15	15	30	10	20	503
(LDA) Integrate the model with the visualizer	10	5	5	20	10	10	503
(LDA) Identify the most important topics and visualize them (Figure 1 (a))	10	5	5	15	7.5	7.5	503
(LDA) Aggregate LDA topics for each month and visualize them (Figure 1 (b))	15	7.5	7.5	20	10	10	503
(LDA) Find the most important words for each topic and Visualize the term Probabilities (Figure 1 (c))	10	5	5	20	10	10	503
(LDA) Retrieve sample of conversations for each month	5	2.5	2.5	3		3	503
Visualize sample of conversations for each month (Figure 1 (e))	5	2.5	2.5	5	5		547
Preprocess and retrieve the most related images for each topic	5	2.5	2.5	15	15		Both
Mouse hover text highlight feature in the VisTool	5	5		5	5		547
Search for keywords feature	5	5		5	5		547
Mouse hover on the matrix feature	5	5		5	5		547
User study and analysis	10	5	5	15	10	5	Both
Release a demo	4	2	2	6	6		547
Presentation	10	5	5		5	5	547
Final Report	20	10	10		10	10	547
Total	300	150	150	407	204.5	202.5	

several components to support users during the evaluation of topic models. In future studies, we plan to compare MultiModalTopicExplorer with similar tools further to examine the advantages and disadvantages of our approach.

Although the number of study participants is small, the results are promising. Future user studies should include a more significant number of users to make the results more robust. We must also mention that we recruited general computer scientists and no natural language processing researchers, which might impact the results. Ideally, we would have preferred to have participants with NLP backgrounds.

We focused on the 4chan dataset in this project. However, our current implementation also supports multi-modal datasets from other social media sources. NLP experts might be interested in evaluating their topic model algorithms with different datasets to get insights into the limitations of their approach.

While the end goal of this tool is to support users in evaluating topic modeling algorithm results, we believe that NLP experts are not the only potential users of MultiModalTopicExplorer. We believe that researchers from other domains (e.g., social science, social computing) might be interested in identifying the main topics from their datasets and how these topics evolve.

After implementing both topic modeling algorithms, we discovered that the 20 most frequent topics retrieved by LDA cover 18% (92,525 out of 500,000 posts) of the base dataset. For BERTopic, the results are different. The top 20 topics retrieved by BERTopic cover 70% (356,124 posts) of the dataset. While these results are not decisive regarding which model is superior to the other, it shows that our design choice to focus only on the 20 most frequent topics was a reasonable decision, as these cover a decent portion of the documents.

Matrix views can be helpful to find clusters, outliers and summarize data. Our current design consider 20 rows (topics) and 18 columns (months). We recognized that users could be interested in visualizing a large number of topics and months. Our current implementation support that case. When larger data is available, the height and width of matrix cells will be adjusted automatically to fit all the information on users' screens. Our implementation has some limitations. The size of the matrix cells will be extremely tiny when the number of rows and cells is more than hundreds. In future work, we plan to incorporate a scroll bar in the x-axis and y-axis and use a focus + context approach

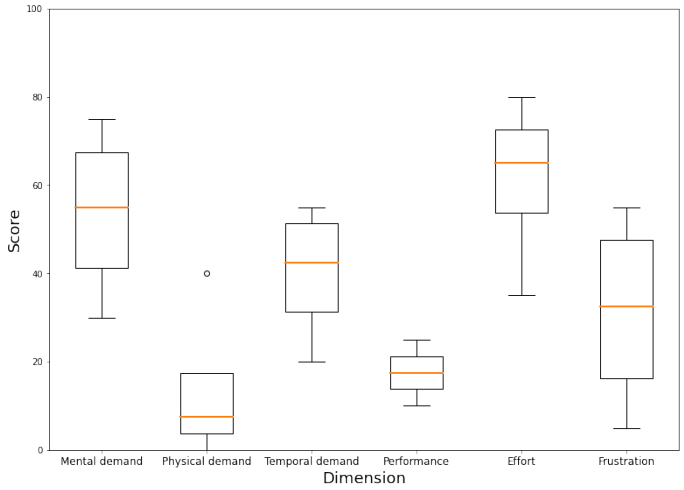

Fig. 18. Distribution of participants responses to the NASA TLX questionnaire. A lower score indicates a better result

[22]. Seek a solution to compactly visualize a larger number of topics is an exciting and open research question [36].

To the best of our knowledge, this is the first topic modeling visualization tool that provides images of topics. We believe that this component can be helpful, especially when the most relevant keywords are generic or do not provide enough information [14]. In future research, we should conduct user studies to evaluate how this component improve the interpretation of topics. Moreover, our current implementation allows users to visualize each topic's nine most relevant images. A future version of MultiModalTopicExplorer should allow users to visualize a larger number of images (e.g., by including a vertical scroll bar). Future user studies should also identify how many images are helpful during topic interpretation.

We use a barplot to show the most relevant keywords associated with each topic. Currently, there is no consensus regarding how many keywords help identify the meaning of each topic. In a future version of MultiModalTopicExplorer we plan to allow users to adjust the number of keywords displayed.

In conversations with user study participants, we learned that they used the conversation view to identify the context of specific keywords. According to them, that functionality was beneficial when they did not know the term's meaning. However, some participants felt overwhelmed after visualizing the entire list of documents associated with a topic in a specific month. It could be possible that not all the documents are helpful for users. Future research could be conducted to identify the helpful number of documents to visualize.

One method to identify the quality of topics is by measuring their coherence, which can be automatically calculated or reported by users [19]. Our current implementation allows users to report the quality of each topic. However, we also believe that we should also provide the automatic coherence score for the entire model in a future version.

11 Conclusion

In this manuscript, we presented an interactive visualization system to address some limitations of previous topic modeling visualization tools related to the scarce support for a qualitative analysis of the results, scarce support of multi-modal data, and lack of visualizations to allow users to visualize the evolution of topics.

In addition, we conducted a within-subject user study to test the functionalities of our tool. We asked participants to evaluate the performance of two popular topic modeling algorithms: LDA and BERTopic. Our results suggest that topics generated from BERTopic are significantly better according to human judgment. Also, the results hint that users felt well about their performance while evaluating the topic models using our tool.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, Mar. 2003.

[2] J. Boyd-Graber, D. Mimno, and D. Newman. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 225255, 2014.

[3] A. Cao, K. K. Chintamani, A. K. Pandya, and R. D. Ellis. NASA TLX: Software for assessing subjective mental workload. *Behavior research methods*, 41(1):113–117, 2009.

[4] A. J. Chaney and D. M. Blei. Visualizing Topic Models. In *ICWSM*, 2012.

[5] 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*, pp. 288–296, 2009.

[6] N.-C. Chen, M. Drouhard, R. Kocielnik, J. Suh, and C. R. Aragon. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–20, 2018.

[7] 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*, pp. 74–77. ACM, 2012.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018.

[9] 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)*, pp. 231–240. IEEE, 2011.

[10] 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.

[11] M. Efron, P. Organisciak, and K. Fenlon. Building topic models in a federated digital library through selective document exclusion. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011.

[12] 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)*, pp. 1–10. IEEE, 2018.

[13] M. Grootendorst. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics., 2020. doi: 10.5281/zenodo.4381785

[14] E. Hoque and G. Carenini. ConVis: A visual text analytic system for exploring blog conversations. In *Computer Graphics Forum*, vol. 33, pp. 221–230. Wiley Online Library, 2014.

[15] E. Hoque and G. Carenini. ConVisIT: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pp. 169–180. ACM, 2015.

[16] 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.

[17] R. M. Kaplan, J. Burstein, M. Harper, and G. Penn. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

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

[19] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539, 2014.

[20] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An interactive visual document clustering via topic modeling. In *Computer graphics forum*, vol. 31, pp. 1155–1164. Wiley Online Library, 2012.

[21] T. Y. Lee, A. Smith, K. Seppi, N. Elmqvist, J. Boyd-Graber, and L. Findlater. The human touch: How non-expert users perceive, interpret, and

fix topic models. *International Journal of Human-Computer Studies*, 105:28–42, 2017.

[22] R. Li, E. Hoque, G. Carenini, R. Lester, and R. Chau. ConVIScope: Visual Analytics for Exploring Patient Conversations. In *2021 IEEE Visualization Conference (VIS)*, pp. 151–155. IEEE, 2021.

[23] R. Li, W. Xiao, L. Wang, H. Jang, and G. Carenini. T3-Vis: a visual analytic framework for Training and fine-Tuning Transformers in NLP. *arXiv preprint arXiv:2108.13587*, 2021.

[24] M. Masala, S. Ruseti, M. Dascalu, and C. Dobre. Extracting and Clustering Main Ideas from Student Feedback Using Language Models. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova, eds., *Artificial Intelligence in Education*, pp. 282–292. Springer International Publishing, Cham, 2021.

[25] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[26] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[27] E. Meeks and S. B. Weingart. The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1):1–6, 2012.

[28] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 262–272, 2011.

[29] A. Mittos, S. Zannettou, J. Blackburn, and E. De Cristofaro. "And We Will Fight for Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 452–463, 2020.

[30] J. Murdock and C. Allen. Visualization techniques for topic model checking. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[31] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 100–108, 2010.

[32] A. Papasavva, S. Zannettou, E. De Cristofaro, G. Stringhini, and J. Blackburn. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. *14th International AAAI Conference On Web And Social Media (ICWSM) 2020*, 2020.

[33] 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)*, pp. 207–208. IEEE, 2015.

[34] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[36] 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*, pp. 63–70, 2014.

[37] A. Smith, T. Hawes, and M. Myers. Hiearchie: Visualization for hierarchical topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 71–78, 2014.

[38] A. Smith, T. Y. Lee, F. Poursabzi-Sangdeh, J. Boyd-Graber, N. Elmqvist, and L. Findlater. Evaluating visual representations for topic understanding and their effects on manually generated topic labels. *Transactions of the Association for Computational Linguistics*, 5:1–16, 2017.

[39] A. Smith, S. Malik, and B. Shneiderman. Visual analysis of topical evolution in unstructured text: Design and evaluation of topicflow. In *Applications of Social Media and Social Network Analysis*, pp. 159–175. Springer, 2015.

[40] 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.