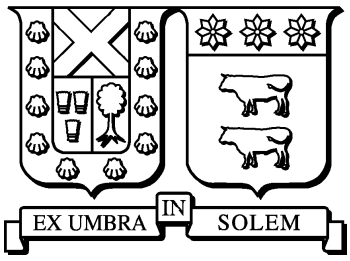
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



"TOPICVISEXPLORER: SUPPORTING MULTI-CORPORA COMPARISON THROUGH VISUAL EXPLORATION OF TOPIC MODELING"

FELIPE ANDRÉS GONZÁLEZ PIZARRO

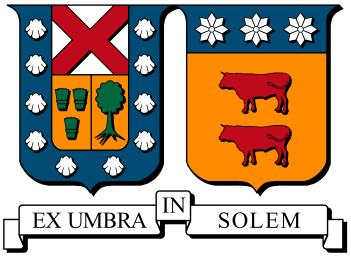
Tesis para optar al grado de

MAGÍSTER EN CIENCIAS DE LA INGENIERÍA INFORMÁTICA

PROFESOR GUÍA: CLAUDIA LÓPEZ

SEPTIEMBRE 2021

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"TOPICVISEXPLORER: SUPPORTING MULTI-CORPORA COMPARISON THROUGH VISUAL EXPLORATION OF TOPIC MODELING"

Tesis de Grado presentada por
FELIPE ANDRÉS GONZÁLEZ PIZARRO

como requisito parcial para optar al grado de
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA INFORMÁTICA

PROFESOR GUÍA: CLAUDIA LÓPEZ
PROFESOR CORREFERENTE: MARCELO MENDOZA
PROFESOR EXTERNO: EVANGELOS MILIOS (DALHOUSIE UNIVERSITY)

SEPTIEMBRE 2021

Acknowledgements

During my four years pursuing my MSc degree, many people have helped me to reach my goals. First, I want to thank my parents because their love has accompanied me all the way. Also, I want to thank my sister, who had illuminated my life. I am so happy that you are studying an area that you feel so passionate about. I wish you success in your soon-to-be thesis defense. I also owed a debt of gratitude to my aunt and grandparents, who taught me the importance of studying hard and getting a degree since I was a kid. I want to thank the Palma family, who are part of my family too. We have celebrated the good news and supported each other during challenging times.

Special gratitude to Ignacio Tampe, my common-law partner, for your tireless support and love. The last years have been difficult, but you have always given me the strength and help I needed. Also, thank you so much for listening to my research ideas and for expressing your sincere feedback. I love when we have passionate conversations about computer science.

I also thank all my friends. Together we have overcome difficult times and have celebrated like never before. I love you very much. I hope you continue to accompany me throughout my life.

Special gratitude to the professors: Claudia López, Evangelos Milios, Fernando Paulovich, Marcelo Mendoza, Cecilia Aragón, and Savvas Zannettou, who during the last years have helped me so much to improve my research skills and grow professionally. All of you have helped me discover the research areas that I love. Because of you, nowadays, I can continue my studies pursuing a Ph.D. in Computer Science.

To all of them, thank you very much.

Resumen

El constante aumento en el volumen de datos de tipo texto ha llevado al desarrollo de varios algoritmos destinados a resumir y comprender este tipo de datos. Una solución prometedora este problema es el modelado de temas (en inglés conocido como *topic modeling*), un enfoque estadístico para extraer temas de alto volúmenes de datos. Humanos que interactúan e interpretan directamente el resultado de estos algoritmos pueden usar herramientas de visualización para interpretar mejor los resultados, sin embargo, estas herramientas todavía tienen una limitación significativa. Las representaciones visuales actuales permiten refinar y comparar temas basados solo en sus palabras claves, lo que genera un rendimiento deficiente cuando estas son demasiado genéricas, están mal conectadas o no proporcionan suficiente información. Para abordar este problema, propongo TopicVisExplorer, un conjunto de visualizaciones interactivas que soporta *Latent Dirichlet Allocation (LDA)*. Esta propuesta tiene por objetivo ayudar a los usuarios durante el refinamiento y comparación de temas. Tres innovaciones claves de este trabajo buscan apoyar refinamiento del modelo de tema e identificar temas similares de uno o dos corpus. (1) Propongo un algoritmo de fusión de temas que considera tanto términos como documentos de los tópicos, (2) un nuevo algoritmo de división de temas basado en sus documentos, (3) y una métrica que estima la similitud entre temas en base a sus palabras y documentos más relevantes. Realicé un estudio de usuarios con 95 usuarios no expertos para evaluar las funcionalidades de TopicVisExplorer. Los resultados muestran que los participantes pudieron identificar los temas que necesitan mejorar su calidad. Aproximadamente la mitad de los participantes mejoraron la coherencia de su modelo después de aplicar operaciones de división y fusión de temas. Además, los participantes pudieron identificar temas similares entre dos corpus. Aquellos que utilizaron la métrica de similitud propuesta cometieron menos errores que aquellos que usaron una métrica base.

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 high volumes of data. Humans who directly interact with and interpret the output of topic modeling may rely on visualization tools to better interpret the results. However, these tools still have a significant limitation. Current visual representations allow to refine and compare topics based only on their most relevant keywords, leading to poor performance when these terms are too generic, poorly connected, and do not provide enough information. To address this problem, I propose *TopicVisExplorer*, a set of web-based interactive visualizations of topics estimated using Latent Dirichlet Allocation (LDA). These visualizations aim to support users during topic refinement and comparison. There are three key innovations in this work. I propose (1) a topic merging algorithm that considers both terms and documents of two independent topics, (2) a new document-based topic splitting algorithm, and (3) a topic similarity metric that estimates the similarity between topics regarding their most relevant keywords and most relevant documents. I conducted a user study with 95 non-expert users to evaluate TopicVisExplorer functionalities for refining and comparing topics from a large-scale real-world Twitter dataset. The results show that participants were able to identify topics that need further refinement to improve their quality. About half of the participants improved the topic model coherence after applying topic splitting and topic merging operations. Moreover, they were able to identify similar topics between the two corpora. Those who used the proposed topic similarity metric made significantly fewer erroneous matches than those who used a current state-of-the-art topic similarity metric.

Content Index

List of Tables

List of Figures

Glossary

- **LDA:** Latent Dirichlet Allocation (LDA) is a topic modeling algorithm. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.
- **PMI:** Pointwise Mutual Information (PMI) is a measure of association used in information theory and statistics. The automatic coherence metric based on PMI considers a sliding window and the pointwise mutual information of all word pairs of the given top words.
- **NPMI:** Pointwise mutual information can be normalized between [-1,1], resulting in -1 (in the limit) for never occurring together, 0 for the independence of the variables, and +1 for complete co-occurrence.
- **Word embedding:** Based on the co-occurrence of terms, word embeddings create a reduced multi-dimensional representation of a corpus. Such representation can be used to analyze the semantic proximity among the corpus terms.
- **Topic coherence:** A set of statements or facts is said to be coherent if they support each other. Thus, in coherence topics, there is a semantic similarity between high-scoring words.

Chapter 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 [54]. 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 [20].

Although powerful, topic models do not interpret themselves; therefore, humans must be involved [10, 15]. 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. [16]. Topic modeling visualization tools help in understanding topic models output and issues in modeling [32]; however, they still have limitations. Finding mechanisms to improve these visual representations is still an open challenge [30].

First, in some cases, automatically generated topics can be noisy; or may not align well with user's needs [27, 34]. Previous work has shown that incorporating human knowledge into the topic modeling output is a promising approach to create high-quality topic models [40].

Some topic modeling visualization tools allow users to add or remove words in topics, split generic topics, or merge similar topics [40]. The last two operations are considered the most relevant ones [27]. Indeed, users prefer a topic model interface that allows merging and splitting topics over those not supporting topic refinement operations. [27]. However, a limited number of topic modeling visualization tools support topic merging and topic splitting (see [39, 12, 27, 19]).

Topic modeling visualization tools that enable users to merge two independent topics share a significant shortcoming [27, 12, 39]. During the merge operation, they do not consider the relevance of all the keywords and documents associated with those topics, which might impact the quality of the results [2]. Moreover, not all of them provide user evaluations, making it difficult to understand how users apply this operation in real-world scenarios [40].

Topic modeling visualization tools that support topic splitting operation share two limitations. First, they do not support Latent Dirichlet Allocation (LDA), which is a widely-used topic modeling technique even today [76, 39, 75, 20, 34]. Moreover, they split topics based only on the topics' top keywords, which according to users, sometimes are way too generic [27] and do not provide enough information to understand the topic [40]. This impacts the performance of the topic splitting operation, which in some cases can not accurately split the topics into meaningful sub-topics, making end-users feel frustrated about the results [27].

Secondly, there is a low number of tools supporting multi-corpora comparison, and those that exist (see [41, 63]) do not support LDA. Comparison of text datasets can support different tasks, for instance: (1) assessing whether people discuss the same topics in Twitter and Instagram during a controversial event, (2) analyzing whether the conversation about a theme is similar or no in two periods of time, (3) and examining if a theme is discussed in similar ways in different languages.

Moreover, all topic modeling visualization tools use topic similarity metrics to provide a global view of the topics and help users identify how topics differ. While current topic similarity metrics are helpful [68, 65], they only model the topics as a ranked list of terms. As a result, they do not perform well when topics consist of noisy or poorly connected sets of terms [40, 7].

To address these limitations, I designed, developed, and evaluated *TopicVisExplorer*, a set of web-based interactive visualizations of LDA-generated topics. This tool supports topic refinement and multi-corpora comparison through a interactive visual exploration of topic models. It enables users to adjust topics by topic splitting and topic merging operations. Moreover, it includes a newly proposed topic similarity metric that evaluates the similarity between topics considering their relevant keywords and documents.

I conducted a user study to evaluate the potential of TopicVisExplorer in helping non-expert users interpret the results of LDA topic models for a large-scale real-world dataset. The results suggest that participants could identify topics that need further refinement to improve their quality. About half of the user study participants improved the automatic coherence of the resulting topic model after applying topic merging and topic splitting. Moreover, after comparing the proposed topic similarity metric with a baseline, the results suggest that the proposed metric can reduce the number of erroneous matches when comparing two datasets.

In summary, the primary contributions of this work are fourfold:

- A topic merging operation that considers both terms and documents of the two independent topics
- A new document-based topic splitting operation that allows users to split a generic topic into two subtopics after indicating the most relevant documents for each new item.
- A word embedding-based topic similarity metric, which evaluates the similarity between the topics' most relevant keywords and most relevant documents.
- Provide evidence of how non-expert users can improve topics from one corpus and compare topics from two corpora through a visual exploration of topic modeling results.

The remainder of the manuscript is organized as follows. Chapter 2, summarizes prior work about visual representations of topic modeling algorithm outputs, discussing the current limitations and positioning this proposal. Chapter 3 introduces the proposed topic modeling visualization system, including its topic merging operation, its document-based topic splitting

mechanism, and the new topic similarity metric. Chapter 4 presents the user study method. Chapter 5 summarizes its results. Finally, Chapter 6 offers discussions and conclusions of the results, their limitations, and future work.

Chapter 2

Related Work

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) [6] is one of the most traditional and popular methods [45, 55]. 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 [37]. 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 [60, 9, 50]. To address this need and add analytic value, previous work has explored different visual representation approaches to support a human interpretation of topic models. These approaches vary in topic model representation, techniques of topic modeling refinement, and support for inter-topic comparison.

2.1. Topic model representation

The most common output of topic modeling algorithms is the ranked list of the top terms of each particular topic [32]. 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 [62]. Among these alternatives, simple visualizations such as word lists or word

lists with bars allow users to understand topics quicker [62].

Usually, the top keywords are shown to users as a ranked list of the most frequent terms for each particular topic [9, 54]. In LDA, this is the same that ordering the terms by their topic-specific probability. The problem with representing topics this way is that frequent common terms in the corpus often appear near the top of such lists for multiple topics, making it hard to users to find the differences between them [59].

To mitigate this problem, an intrinsic measure to rank terms within topics was proposed. It is called *Lift* [64], and it is defined as the ratio of a term's probability within a topic to its marginal probability across the corpus. Thus, let $\phi_{k\omega}$ denote the probability of the term $\omega \in 1, .., V$ occurring in topic $k \in 1, ..., K$, where V denotes the number of terms in the vocabulary and K the number of topics. Let p_ω denote the marginal probability of the term ω in the corpus. The ordering of keywords by *lift* is given by :

$$lift(\omega, k) = (\frac{\phi_{k\omega}}{p_\omega}) \tag{2.1}$$

This measure generally decreases the rankings of globally frequent terms, which can be helpful for topic interpretation. Nevertheless, it can be noisy in some cases by giving high rankings to very rare terms that occur in only a single topic. While such terms may contain useful topical content, if they are very infrequent, the topic may remain challenging to interpret [59].

Another intrinsic measure was proposed to mitigate *lift* limitations. It is called *relevance* [59], and it is based on both term's frequency as well as its *exclusivity*, the degree to which its occurrences are limited to only a few topics. Thus, the *relevance* of term ω to topic k given a weight parameter λ (where $0 \leq \lambda \leq 1$) is defined as:

$$r(\omega, k|\lambda) = \lambda log(\phi_{k\omega}) + (1 - \lambda) log(\frac{\phi_{k\omega}}{p_\omega}) \tag{2.2}$$

where λ determines the weight given to the probability of term ω under topic k relative to its *lift* (measuring both on the log scale). Setting $\lambda = 1$ results in the ranking of terms in

decreasing order of their topic-specific probability, and setting $\lambda = 0$ ranks terms solely by their *lift*. A user study found that the optimal value of λ for topic interpretation is 0.6 [59].

Sometimes the top keywords are not enough to identify the semantics of a topic [27]. That is the case when the most relevant terms are poorly connected, or when they include disparate [48] or generic terms (e.g., “yes”, ”like”, ”Mr”, ”maybe”) [40, 7]. 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 [27, 73]. 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 [20].

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 [26, 27]. Another method is to display the documents according to their contribution to the topic, as [63]. 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.

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 [39] 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 2.1 (a)).

A second approach consists in displaying the term-topic distributions through a matrix layout. In this approach, proposed in Termite [13], the rows correspond to terms and the columns to topics. It uses circular areas to encode term probabilities. Thus, the most frequent

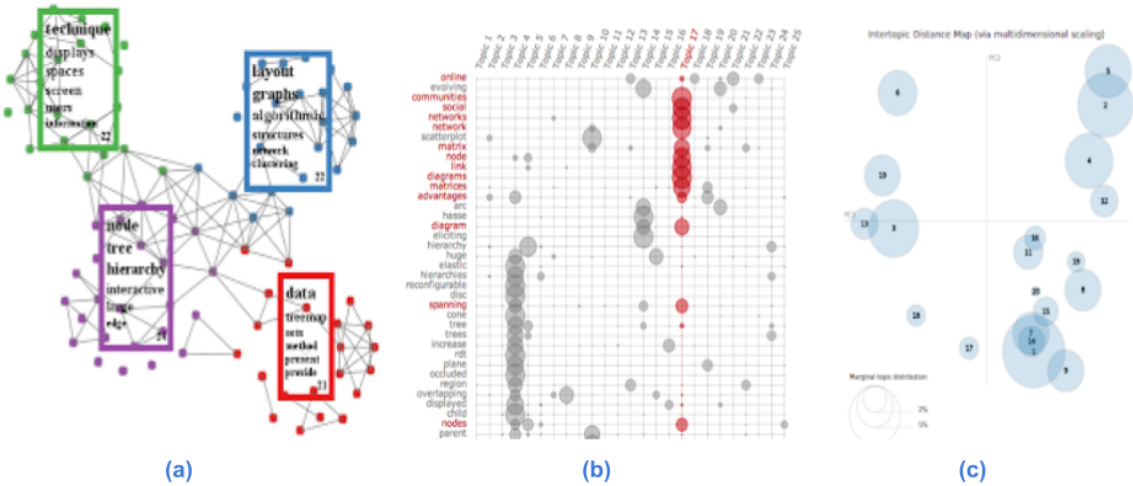


Figure 2.1: Layout of global view of topics in: (a) iVisClustering [39] (b) Termite [13] (c) LDAvis [59].

terms are represented by circles with a larger area (see Figure 2.1 (b)).

A third alternative consists in projecting the similarity between topics into a two-dimensional space. In this approach, proposed in LDAvis [59] (see Figure 2.1 (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.

The previous approaches allow users to get a global view of one topic model, but they do not support multi-corpora comparison. Considering that in some cases, users may be interested in performing analysis across multiple sources to compare topics of their interest [41] or identify changes in topics across time [43], some visual representations have been designed to address these needs.

An approach to get a global view of multiple topic models consists of modeling each corpus as a topic graph. In TopicPanorama [41], a graph matching method and a density-based graph layout over these topic graphs allow displaying common and distinctive topics among

multiple datasets (see Figure 2.2 (a)).

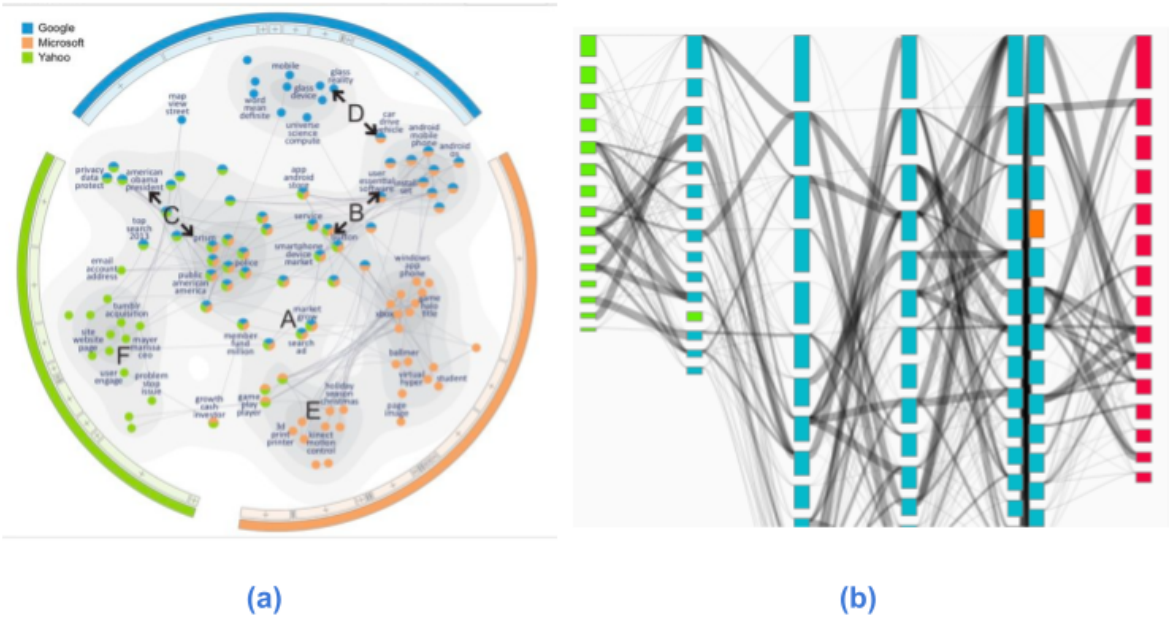


Figure 2.2: Multi-corpora comparison layout in: (a) TopicPanorama [41] (b) TopicFlow [43]

.

Sankey diagrams [52] are another approach to display the similarity between topics from different corpora. In TopicFlow [43] they are used to allow users to visualize the evolution of topics over time. It represents topics as boxes, and the path between them symbolizes their similarity. The box' sizes depend on the number of documents attributed to the topic, and they are ordered horizontally from the top by decreasing size. Therefore, the most prevalent topics are at the top of the chart, and the user can quickly see how the frequency of a topic evolves. The paths are weighted by topics similarities. Thus, topics that are more similar are connected with a wider link. Colour is used in the chart to distinguish topics by their evolution state: emerging (green), ending (red), continuing (blue), or standalone (orange) (see Figure 2.2 (b)).

2.2. Topic modeling refinement

Generated topic models are not perfect as they can include poor quality topics: (1) they may contain incoherent or loosely connected topics [61, 67]; (2) some topics can be misaligned with a domain expert's understanding of the corpus [61], and (3) topics can be noisy and may not match the users' current information needs [27, 67]. Rather than reconfiguring and rebuilding a model when users are unsatisfied with its quality, they can improve it through interactive operations such as adding or removing words in topics, merging similar topics, and splitting generic topics [40]. Among these operations, merging and splitting topics are the methods with the highest relevance level by end-users of visualization topic modeling tools [27]. Nonetheless, just a few current topic modeling visualization tools allow users to perform these operations.

There are some strategies to support users in aggregating multiple similar topics into a single one. For instance, in topic-graph representation, the new merged topic is created by union the nodes and edges of the two initial topics [27]. Another alternative is to perform a document-based merge operation. This is the case of UTOPIAN [12], which automatically selects a fraction of the most relevant documents of the initial topics and adds their topic contribution. For instance, suppose two documents where the contributions for three topics are represented as (0.7, 0.2, 0.1) and (0.3,0.5,0.2), respectively. When merging topics 1 and 2, the final contribution of these documents must be set to (0.7+0.2, 0.1) and (0.3+0.5, 0.2) for the merged topic and topic 3.

While the latter approach seems promising, there are some limitations. First, it only considers a subset of documents associated with each topic, which might impact the quality of the resulting topic [2]. Also, it was designed only for the Non-Negative Matrix Factorization topic modeling algorithm [12]. Moreover, the authors do not provide user evaluations, so it is unclear how users would apply the refinement operation in real datasets [40].

There are also strategies to support users exploring subtopics from a generic one. In cases when topics are represented as topic clusters, systems are able to split a topic by extracting a sub-graph of the initial topic. This process can be automatic, as in ConVisIT [27]. Here, the system splits the chosen topical cluster into n sub-clusters applying a k-way min-cut graph

partitioning algorithm with a normalized cut (n-cut) criteria. The optical number of subtopics (n) is automatically determined, but because of time constraint imposed by the nature of the operation, the maximum number of possible subtopics is five.

Another approach consists in performing a keyword-based topic splitting, as is done in UTOPIAN [12]. In this case, the system assumes that the user expects to split a generic topic into only two subtopics. For this purpose, the system creates two reference vectors, v_a and v_b, for the expected resulting sub-topics. Users manipulate these two topic vectors via topic keyword refinement interaction, an operation that enables to change the keyword's weight in a topic. For instance, users might increase the weight of a particular keyword in the first one while decreasing/removing the weight of the same keyword in the second one.

There are two main limitations of the reviewed approaches. First, users who used ConVisIT become frustrated when the automatically generated sub-topics were not accurately separated according to the user needs [27]. Second, although UTOPIAN [12] provides the capability to split a topic into two in a user-driven way through topic keyword refinement, this operation is only based on the most relevant keywords of the selected topic. As a result, users will follow a complex and frustrating process when the top keywords are too generic, disparate, or do not provide meaningful and precise information to allow them to recognize well-defined subtopics [27, 40].

2.3. Inter-topic comparison

Topic similarity metrics evaluate how topics from one topic model differ from each other [59]. These metrics are being used during multi-corpora comparisons as well. Several metrics have been introduced in state-of-the-art to estimate the similarity between topics [65], and there is no consensus on which is the best [68].

There are six commonly used topic similarity metrics to detect a temporal organization of similar topics [33]: (1) Jaccard's coefficient, (2) Kendall's coefficient, (3) Discounted cumulative gain, (4) Cosine similarity, (5) Kullback-Leibler Divergence, and (6) Jensen-Shannon Divergence. An evaluation of these metrics on LDA topic models created from a dataset of

nine months of Korean Web news shown that Jensen-Shannon divergence generates inter-topic similarities better aligned to the corpus [33]. This evaluation compared the negative log-likelihood for the six metrics to measure how well the model explains the corpus..

In topic similarity networks: graphs in which nodes represent latent topics in text collections, and links represent similarity among topics, Hellinger distance-based metric, another quantifier of the similarity between a pair of probability distributions, is an alternative metric and might be a better approach to compare topics from different corpora than traditional methods such as Kullback-Leibler divergence or their symmetric variation Jensen-Shannon Divergence [42].

Based on these claims, previous work had compared the performances of the Jensen-Shannon divergence, Hellinger distance-based metric, and other two topic similarity metrics to examine how they align with human judgments [68]. These experiments included the cosine similarity metric, which measures the similarity between two vectors/distributions by finding the cosine of the angle between them. It also included a newly proposed similarity metric called by them as *Word-Embedding based metric*. This approach computes the word semantic similarity of the top keywords of a topic using word embeddings [46]. Each topic is represented by its top n words ranked by its words' posterior topic probabilities. Let W_i be the set of top n words for topic i and Vec_p the vector of word p in the word embedding model, the similarity of two topics i and j is given by:

$$WES(\theta_i, \theta_j) = \sum_{p \in W_i} \min_{\forall q \in W_j} cosine(Vec_p, Vec_q) \tag{2.3}$$

The results show that cosine similarity and word embedding-based metrics perform better and appear to be complementary. Cosine similarity estimates similarity better when the topics share the exact high-frequency words. On the other hand, the word embedding-based metric can outperform cosine similarity when topics share words lexicographically different but with similar meanings [65]. For instance, if the most relevant terms of two topics are: {"vote", "president"}, and {"election", "votes"}, respectively, the word embedding-based metric will capture the semantic relationships between these two topics and indicate them as a match. Cosine similarity metric will not do the same given that it can not capture semantic

relationships between terms.

While current state-of-the-art topic similarity metrics have potential, they still have limitations. A shortcoming shared by all these revised metrics is related to how they model the topics. They consider topics as a multinomial distribution over the vocabulary or as a ranked list of words [33, 71, 68, 18, 65]. The problem with representing topics only in this way is that these metrics are sensitive to the high dimensionality of the vocabulary [65] and can assign high similarity to topics that contain ambiguous words, thus generating solutions that do not strongly correlate with human judgments [1, 65].

Overall, there is a critical shortcoming among topic modeling visualization tools that facilitate users to refine and compare topics. Its topic refinement operations and their topic similarity metric consider only the topics' relevant keywords, leading to poor performance when these terms are generic, poorly connected, and do not provide enough information by themselves [40, 7]. In those cases, the top keywords are not enough to identify the semantics of a topic [27]. Therefore, it is necessary to design, develop, and evaluate topic modeling visualization tools that include another level of information, such as the most relevant documents to each topic [73, 66], to improve the quality of the results.

Chapter 3

TopicVisExplorer

I propose an interactive visualization system called *TopicVisExplorer* to address some limitations of previous topic modeling visualization tools. This approach supports users in refining topics of one topic model and evaluating the similarity between topics from one or two topic models. Different visual components are designed for this purpose, some of which are original, while others are borrowed from existing tools [59, 63]. This section provides an overview of the visual interface features.

3.1. Layout #1: Topic modeling refinement

The first layout of TopicVisExplorer is illustrated in Figure 3.1. It has four primary visual components to allow users to get a sense of the global view of topics and visualize specific information of topics such as its most relevant keywords and documents. It also provides components to allow users to apply topic merging and topic splitting.

3.1.1. Global view of topics

The central panel of the layout (see Figure 3.1 (a)) presents a global view of the topics and aims to answer the question: "How do topics relate to each other?" In this layout, I plot the

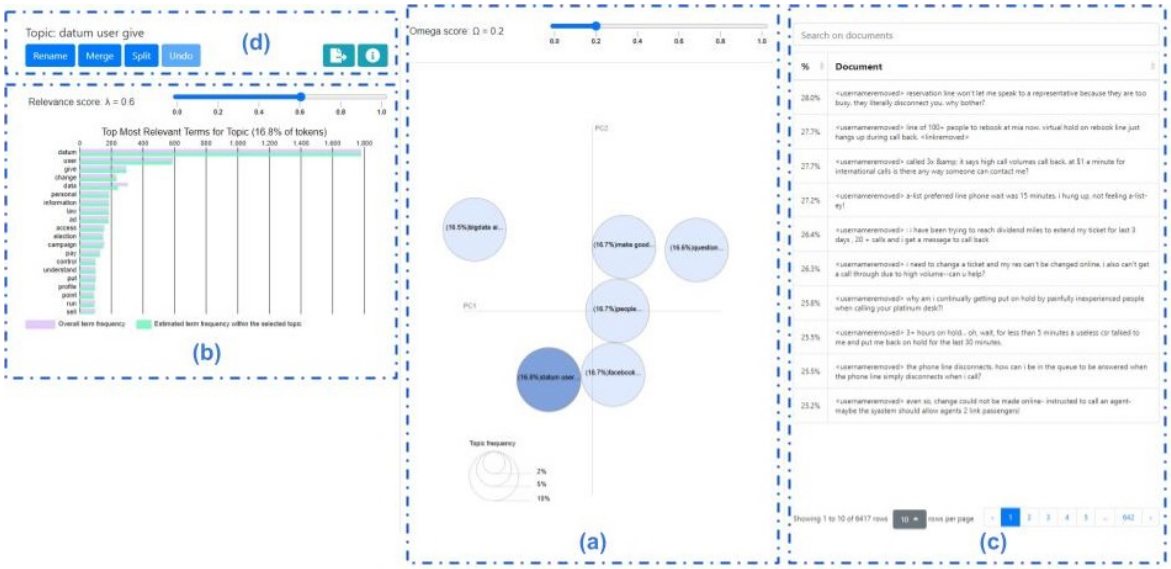


Figure 3.1: Topic modeling refinement scenario. (a) Global view of topics, (b) Topic's most relevant keywords. (c) Topic's most relevant documents.

topics as circles in a two-dimensional space whose centers are determined by computing the distance between topics. I use Principal Coordinate Analysis (PCoA) [31] to transform an inter-topic similarity matrix onto two components, as is done in [59]. Thus, similar topics appear closer, while distinct topics appear more distant from each other. The inter-topic similarity matrix is a *n*x*n* matrix, where *n* is the number of topics. The PCoA output is a *n*x*2* matrix.

The central panel (see Figure 3.1 (a)) also supports users to answer the question: "How prevalent is each topic?". For this purpose, I encode each topic's frequency using the circle area, as in [59]. Thus, prevalent topics appear with a larger area. Additionally, users visualize the name of each topic inside each circle area. The default name for each topic corresponds to its three most relevant keywords, but users can change it by clicking the *rename* button (see Figure 3.1 (d)). In cases when the circles' area is tiny, and topics' names can not be displayed, a text label with the topic's name will appear when users mouse over a circle.

Users can visualize the most relevant keywords and the most relevant documents of a topic by clicking on the circle that represents the topic (see Figure 3.1 (a)).

3.1.2. Topic's keywords and topics' documents

The left panel of the layout (see Figure 3.1 (b)) depicts a horizontal bar chart for the most relevant terms to the selected topic. For each term, two bars are unfolded. Violet bars represent the corpus-wide frequency of a given term, and the green bar represents the topic-specific frequency of such term. This kind of linked selection allows users to examine a large number of topic-term relationships compactly and supports users in topic interpretation [59]. The most useful terms to a given topic are ranked according to the *relevance* score, allowing users to flexibly rank terms in order of usefulness for topic interpretation [59]. A higher relevance score designates greater importance to the frequency of terms within the selected topic (green bar). However, at the same time, it reduces the importance of their exclusivity. In other words, how rare these words are on other topics. A slider allows users to alter the rankings of terms to aid topic interpretation. By default, the slider value is set to 0.6, as is suggested by a prior user study [59].

This proposal includes a new component to visualize the most relevant documents associated with the selected topic. In the layout on the right panel (see Figure 3.1 (c)), the documents are sorted according to their contribution to the topic; thus, the most relevant documents appear first. In this panel, users can also search for documents that contain specific terms. Doing so will highlight the searched term in the documents (see Figure 3.2).

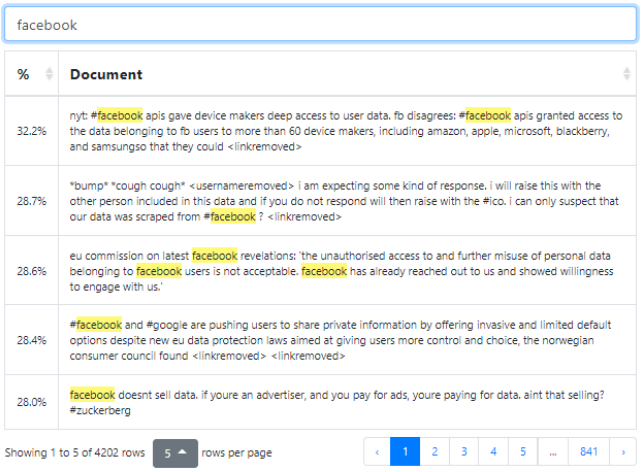


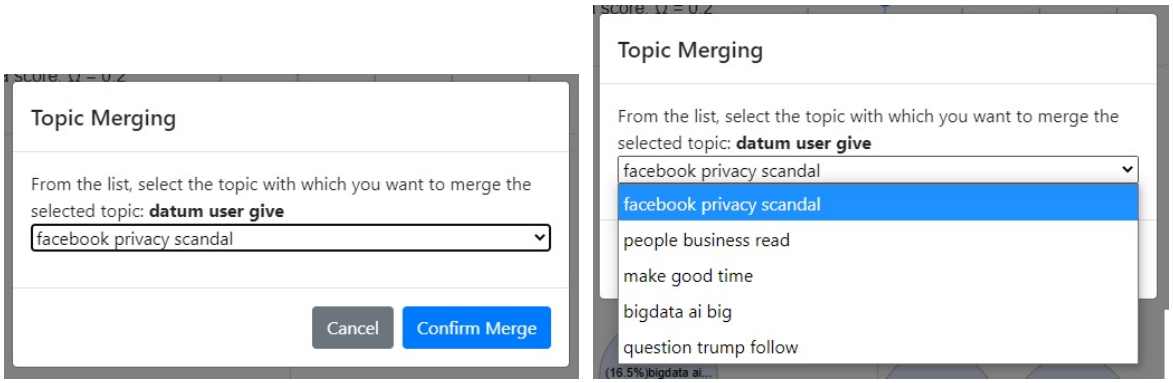
Figure 3.2: Most relevant documents to the selected topic. Results are filtered by the keyword: "facebook"

3.1.3. Topic modeling refinement operations

The system enables users to refine a topic model. Users can find buttons to (1) merge two topics; and (2) split a topic into two new subtopics (see Figure 3.1 (d)).

3.1.3.1. Merging two topics into a single one

In some cases, users might be interested in joining two highly similar topics into a single one to avoid redundancy [27]. On TopicVisExplorer, when users wish to merge two similar topics, they must select one of them and then click the *merge button* (see Figure 3.1 (d)). That action will deploy a modal view with a dropdown list with all the possible topics that the user may select to complete the merge operation (see Figure 3.3). Topics in this dropdown list are sorted according to their similarity to the selected one. Thus similar topics to the current one will appear between the first options. After choosing a topic from that list, the new merged topic will emerge, and the visualization will be updated accordingly.



(a) (b)

Figure 3.3: (a) Modal view to make a merge operation over the topic “datum user give”. (b) Drop-down list of topics that the user may choose to finish the merge operation.

To merge two independent topics, the system adds together the probability distribution of words of these two independent topics. It also adds together the probability distribution of these topics in all the documents. As a result, the visualization shows the new most relevant keywords and most relevant documents for the new merged topic. The system also calculates

the frequency of the merged topic, which corresponds to the sum of the topics' frequency of the two original topics. Moreover, it also estimates the similarity between the merged topic and the other topics.

3.1.3.2. Splitting one topic into two subtopics

In some cases, after exploring a topic, users may find that it is too generic and that splitting it into two subtopics might be more suitable for topic interpretation [40]. TopicVisExplorer users can perform this operation. Users who wish to split a currently selected topic on the central panel must click the *split button* (see Figure 3.1 (d)). As a result, a modal view will appear on the screen (see Fig. 3.4), with the most relevant documents associated with the selected topic. The user must select the documents that are mostly related to the new subtopics. There is no maximum number of documents possible to choose for each subtopic, and not every document needs to be categorized. However, at least one document should be indicated as a seed for the new subtopic. After confirming this operation, the visualization will be updated accordingly to show the new two subtopics.

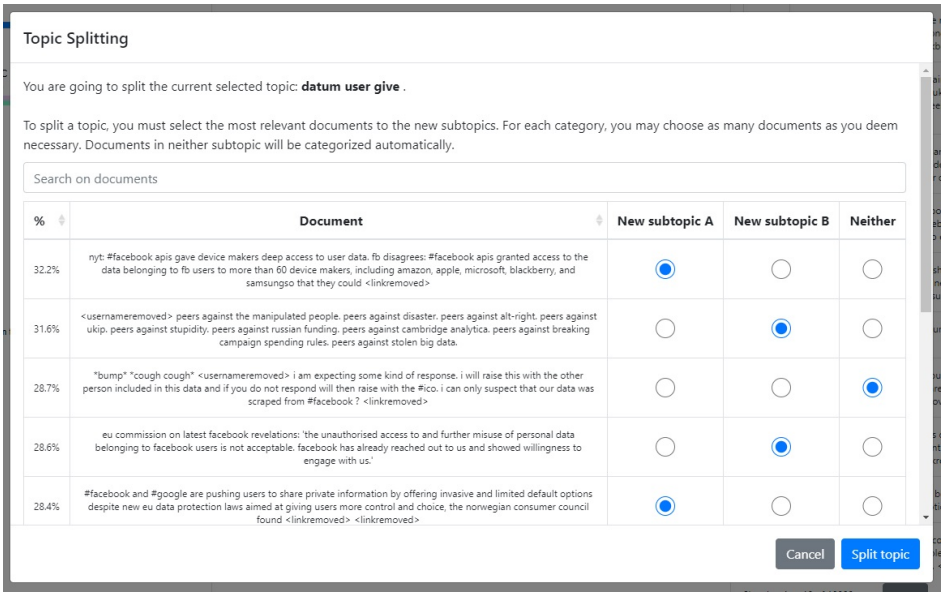


Figure 3.4: Modal view of the topic splitting operation for a topic

I assume that the user expects to split a generic topic into only two subtopics, as [12]. These new subtopics must share a common semantic meaning at a high level but with minor differences in their details. To do so, I propose a new document-based topic splitting algorithm.

This operation includes several steps. First, the system creates a document vector for each corpus document (see Section 3.3). Then, for each new subtopic, it initializes a subtopic vector calculating the average of the document-vectors associated with the subset of documents that are mostly related to each subtopic according to the user's criteria. Regarding the documents not categorized by the user, the system calculates the similarity between these documents and the recently created subtopic-vectors using cosine similarity. Documents not categorized by the user will be assigned to the most similar subtopic.

3.2. Layout #2: Topic model comparison

Figure 3.5: Topic models comparison layout. (a) Overview of the relationship between topics. Top keywords and top documents of each corpus are displayed in (b) and (c)

There have been many tools in recent years designed to expose a single model to the user. However, these tools rarely facilitate direct comparison between models [2]. TopicVis-Explorer also includes a visual representation to compare two topic models (see Figure 3.5). Here, the central panel (see Figure 3.5 (a)) contains a Sankey diagram[1] to provide an overview of the relationships among the topics from different corpora, as in [43]. The topics are represented as boxes and are colored according to which dataset they belong to. The path between the boxes indicate the similarity between topics. They have a width proportional to their similarity, thus topics that are more similar are connected with a wider link.

Cluttering can make the figure challenging to interpret. To mitigate this problem, boxes' position are automatically determined to minimize the number of crossings between links. Additionally, at the top of the Sankey diagram, users can use a slider to visualize only links between topics with a similarity score higher than a given value (see examples in Figure 3.6).

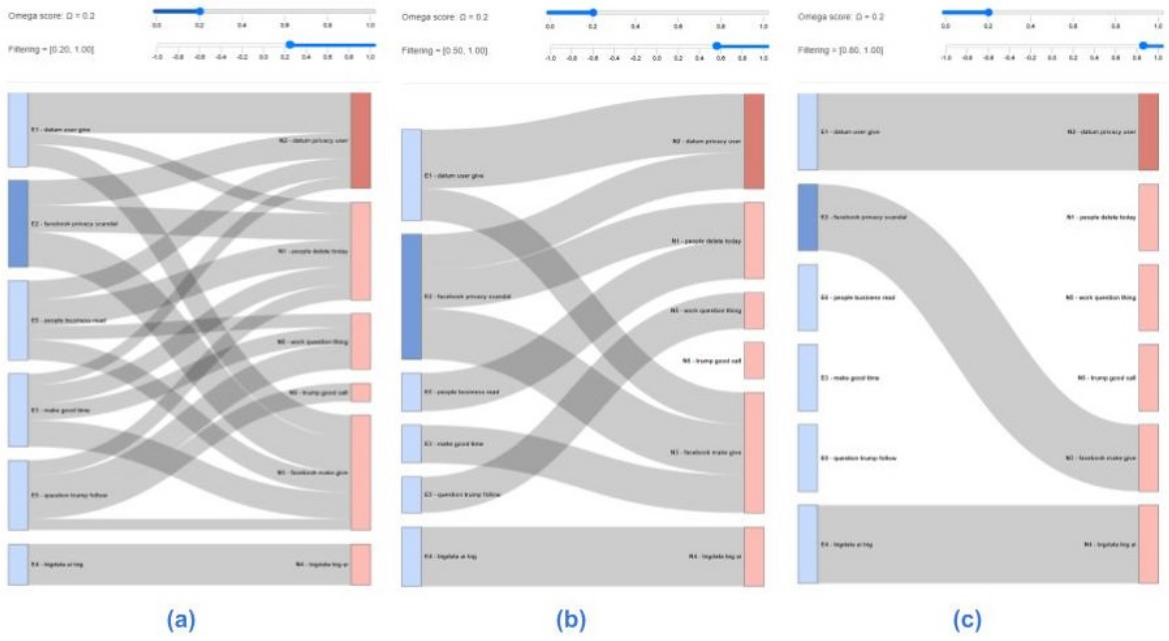


Figure 3.6: Representation of the similarity between topics on three different filtering values

Similar to the previous scenario, the topic default name is given by their three most relevant keywords, calculated after setting up the relevance score into 0.6. Users can change that name by clicking the *rename* button.

[1]Based on code from: https://github.com/d3/d3-sankey

When users select a topic from the left column on the central panel by clicking on its respective box (see Figure 3.5 (a)), its most relevant keywords and most relevant documents are displayed in the left panel (see Figure 3.5 (b)). In the same way, similar information is shown in the right panel for topics of the right column of the central panel (see Figure 3.5 (c)). In this layout, for each topic, users can also explore different ordering of terms changing the relevance score, and search for specific terms between its most relevant documents.

3.3. Inter-topic similarity

TopicVisExplorer introduces a new topic similarity metric that exploits the nature of word embedding and considers the topics' most relevant keywords and documents.

Based on the co-occurrence of terms, word embeddings create a reduced multi-dimensional representation of a corpus [46]. Such representation can identify the semantic proximity among the corpus terms and expose the semantic context in which they are used [57, 22]. This multi-dimensional representation can also capture conceptual relationships between the most relevant terms and the most relevant documents of topics. Thus, it overcomes the absence of semantics in the traditional similarity measures available in state of the art [65].

The proposed topic similarity metric includes several components; as described bellow.

Let $\vec{\omega}$ be the vector that represents the term ω from the corpus as a word embedding. The vector that represents the contribution of the top n words of a topic k weighted in order of their position r after sorting them by their relevance score for the topic k, is defined as:

$$\vec{X_k} = \sum_{i=1}^{n} \frac{\vec{\omega_i}}{r(\omega_i, k|0.6)} \tag{3.1}$$

Let $\vec{d_{sk}}$ be a document-vector of the document $s \in 1, ..., S_k$, where S_k denotes the number of

documents associated to a topic k:

$$\vec{d_{sk}} = \sum_{j=1}^{T_s} \frac{\vec{\omega_j}}{r(\omega_i, k|0.6)} \tag{3.2}$$

with $j \in 1, ..., T_s$, where T_s denotes the number of words on the document s. The vector that represents the top n documents, weighted according to their contribution δ to the topic k is defined as:

$$\vec{Y_k} = \sum_{i=1}^{n} \delta_{ik} \vec{d_{ik}} \tag{3.3}$$

For each topic k, a topic vector $\vec{Z_k}$ is generated according to Eq.3.4 (where $0 \leq \Omega \leq 1$):

$$\vec{Z_k} = (1 - \Omega)\vec{X_k} + (\Omega)\vec{Y_k} \tag{3.4}$$

where Ω *(omega)* determines the balance between the contribution of the keywords and documents for the vectorial representation. Notice that TopicVisExplorer users can explore different *omega* (Ω) values.

Having computed the topic vectors $\vec{Z_k}$, the similarity of two topics a and b can be computed by the pairwise word semantic similarity shown in Eq.3.5:

$$WES(\theta_a, \theta_b) = cosine(\vec{Z_a}, \vec{Z_b}) \tag{3.5}$$

3.3.1. Inter-topic similarity in TopicVisExplorer

The proposed topic similarity metric compares topics considering their most relevant documents and keywords. The relationship between these two entities is given by the parameter *omega score*. When calculating the similarity between two topics, a higher *omega score* implies higher importance of their most relevant documents but a lower significance of their most relevant keywords. Users can modify this parameter in both TopicVisExplorer layouts.

In the topic modeling refinement layout, a slider on the top of the central panel (see Figure 3.1 (a)) allows users to change the value of the *omega score*. Changing this parameter will

immediately update the inter-topic distances on the visualization. I use Procrustes analysis [23, 35] to align the different inter-topic distances for each *omega score* value. Procrustes analysis is a method that can find the optimal translation, rotation, and scaling between datasets in order to move them into a common frame of reference (see Figure 3.7). This allows the user to distinguish the topics that vary between two different *omega score* values.

Figure 3.7: Inter-topic similarity in the topic modeling refinement layout. The results for two omega scores are displayed: (a) 0.05 and (b) 1.0

For topic model comparison, users can also interact with the omega score slider to modify the topic similarity metric settings (see Figure 3.8). In this scenario, users can mouse over the links to get more precise information about the similarity between topics. This action highlights the selected link and displays the topic similarity score.

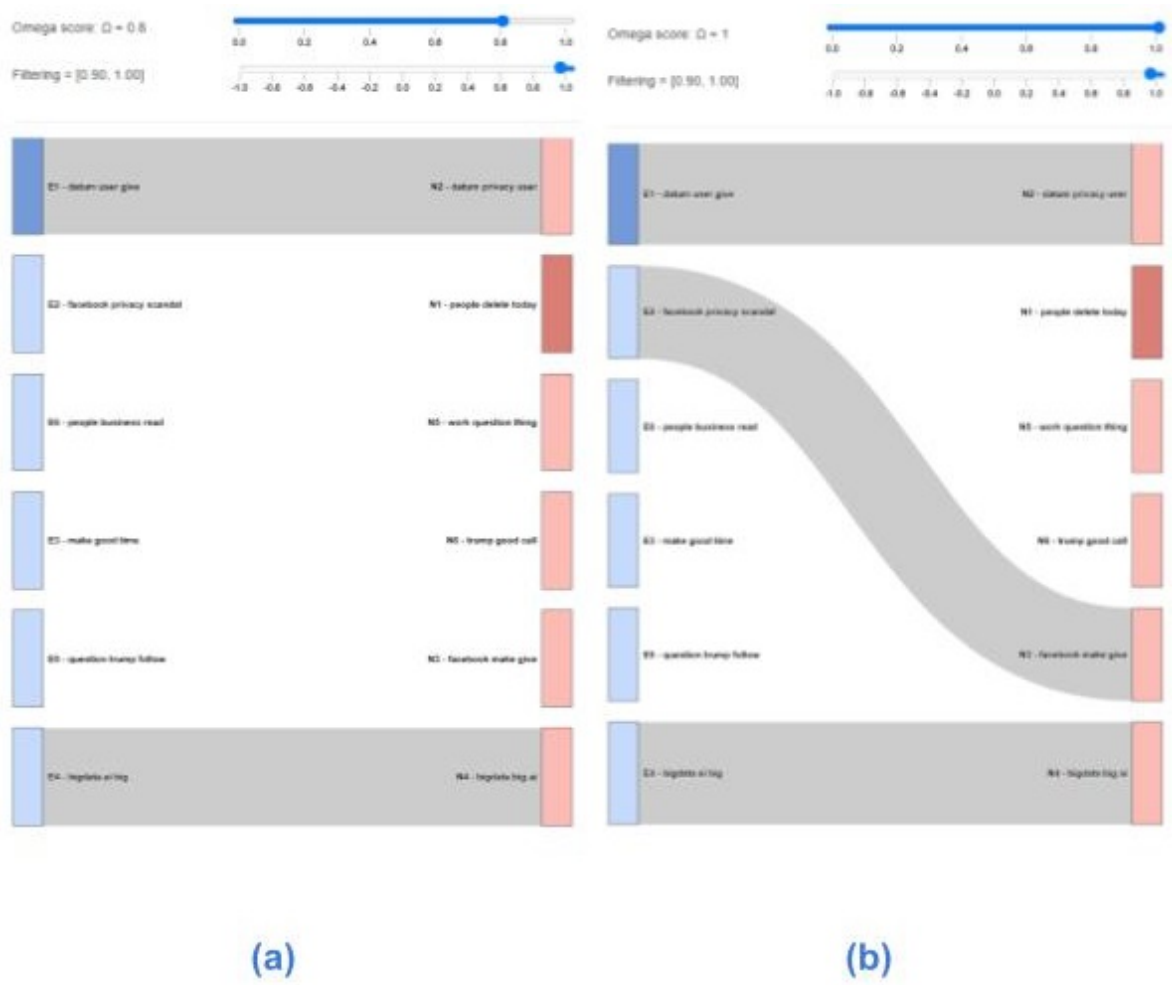


Figure 3.8: Inter-topic similarity in the topic models comparison layout. The results for two omega scores are displayed: (a) 0.8 and (b) 1.0

3.4. TopicVisExplorer in the hands of non-expert users

Compared to tools that visualize static topic models, TopicVisExplorer provides mechanisms to allow end-users to refine these models. While there are a few topic modeling visualization tools that allow topic model refinement, most of them have not been tested with non-expert users on real-world tasks [61, 15]. Therefore, the implications of topic modeling refinement functionalities regarding user experience are not yet well understood. For example, it is unclear if the results obtained after refining a topic model improves topics' coherence.

Moreover, only a few topic modeling visualization tools allow users to compare simultaneously two topic models. In the case of TopicFlow [43], a usability study was conducted with 18 participants. The study results showed positive feedback regarding some of TopicFlow's functionalities, for instance, a low level of users' perceived effort and frustration while finding the two most similar topics. However, there is no evidence about the users' performance regarding precision and recall while finding more similar topics among two datasets.

Additionally, while there is a need to incorporate topic similarity metrics based on word-embeddings in topic modeling visualization tools [65], it has not been done yet. There is no information about how these metrics impact the user experience of end-users of topic modeling visualization tools in realistic scenarios.

To the best of my knowledge, TopicVisExplorer is the only topic modeling visualization tool that allows users to simultaneously (1) inspect in detail one topic model and apply topic modeling refinement operations, and (2) evaluate the similarity between two topic models. Current visualization tools focus only on one of these scenarios [12, 59, 43, 27]. The results of a user study might reveal how non-expert users refine, and compare topics. These results can be useful for other researchers interested in designing topic modeling visualization tools using a user-centered approach.

Chapter 4

User study design

This manuscript introduces *TopicVisExplorer*, an interactive topic modeling visualization tool that aims to help users refine and compare topic models.

Topic model refinement operations such as topic merging and splitting can improve the quality of topics models and allow users to adjust them to their needs [61]. However, only limited research has focused on allowing users to perform these operations [61]. While prior work has proposed strategies to support users merging similar topics and splitting generic ones [39, 12, 27], they still have, in some cases, poor performance and make users feel frustrated when the results are not what they expected [27]. The performance of current state-of-the-art topic merging and topic splitting operations is poor when topics are not well defined [3]. In these cases, its most representative keywords are too generic, disparate, or poorly connected [40, 7].

TopicVisExplorer includes topic merging and topic splitting operations in order to mitigate the current limitations of previous approaches. The merge operation adds together the probability distribution of words of two independent topics. It also adds together the probability distribution of these topics in all the documents. It is in line with prior research that indicates that topics should not be modeled as a ranked list of words but as a set of both term distribution and document distributions [3]. Moreover, the topic splitting operation aims to help users divide a topic into two subtopics. Instead of acting based on only a set of keywords,

users can perform this operation by categorizing documents associated with a topic into two new subtopics. This extra level of information might enable users to identify the context in which terms are being used, which can be helpful when these keywords do not provide enough information by themselves [66].

I expect that both functionalities allow users to improve the generated topic model and adjust it to their needs [61].

One method to identify the quality of topics is by measuring their *coherence* [51], which can be automatically calculated or reported by users [17, 37]. Topics are coherent when there are evident semantic relationships among their constituent components (e.g., keywords, documents) [17, 37, 51]. Considering this, I propose to test the following *null* hypothesis:

H_{0a}: *There are no differences in the coherence of topics between people who apply TopicVisExplorer's mechanisms for topic splitting and topic merging and people who do not.*

Topic modeling visualization tools make use of topic similarity metrics to evaluate how topics differ. While current state-of-the-art topic similarity metrics are powerful, they still share a limitation: they are keyword-based. As a result, when the quality of topics is poor, they establish matches between topics with frequent ambiguous terms, generating solutions that do not strongly correlate with human judgments [65].

Good quality topics are modeled by a skewed distribution toward a small set of words from the complete dictionary [3]. When this is not the case, noisy topics consist of general words, commonly used across a broad of documents within the corpus [11]. For domain experts, the content of these topics is insignificant and often meaningless [3]. Furthermore, well-defined topics are also inclined to appear heavily in a small subset of documents [3]. Thus, if a topic is estimated to generate words in an extensive range of documents, it is far from having a defined and authentic identity [3].

Considering the characteristics of good quality topics, previous work has indicated that topic

similarity metrics can benefit from considering the relative importance of topics in documents [3, 2, 66]. Investigating the distribution of topics over documents might allow distinguishing good quality topics from noise ones. The additional knowledge about the relationships between words in documents can expose the level of coherence of the discovered topic [72, 66]. Indeed, when both term-distribution and document-distribution are combined, they provide a better judgment about the topic significance when compared to each one of them individually [3].

Thus, I expect a topic similarity metric that considers both the most relevant keywords and the most relevant documents to achieve better performance than topic similarity metrics that do not consider both levels of information. In this light, I also propose to test the following null hypothesis:

H_{0b}: *There are no differences in the performance and error rate when comparing topics between people who use a topic similarity metric based on keywords and documents and those who use a keyword-based similarity metric.*

To investigate these hypotheses, I conducted a between-subjects user study to understand how functionalities of TopicVisExplorer may help non-expert users obtain coherent topics and achieve better performance during the comparison of topics. The study consisted of two scenarios (see Figure 4.1). In each of them, participants were randomly assigned into the *experimental* or *baseline* group. In the first scenario, participants interpreted and reported the coherence of topics using the first TopicVisExplorer layout. Participants from the experimental group were able to complete this task using topic merging and topic splitting. These functionalities were not available for the baseline group. In the second scenario, participants identified similar topics of two datasets using the second layout of TopicVisExplorer. The experimental group completed the task using the proposed topic similarity metric. Individuals from the baseline group performed the task using a baseline topic similarity metric [70]. After finishing each scenario, participants reported the perceived workload.

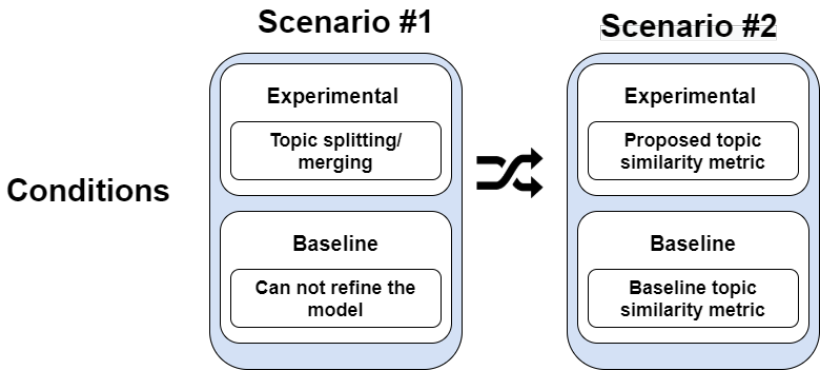


Figure 4.1: Conditions in user study

4.1. First scenario: Topic modeling refinement

The first scenario of the user study (see Figure 4.1) is designed to test H_{0a}: *There are no differences in the observed coherence of topics between people who apply TopicVisExplorer's mechanisms for topic splitting and topic merging and people who do not*. Here, I ask all participants to interpret LDA-generated topics using the first layout of the tool (see Figure 3.1).

The task is completed when users assign a new name to all of the LDA-generated topics as a result of their interpretation. To achieve this task, users visualize the most relevant keywords and the most relevant documents for each topic.

I ask users from the experimental group to execute at least one topic splitting operation and one topic merging operation. Users can undo the operations if the results are not suitable. These topic modeling refinement operations are not available for users from the baseline group.

I collect metrics of the users' experience such as task completion rate (the number of labeled topics) and the require time to complete the task. Also, to capture the quality of topics, I ask participants to rate each topic on a 5-point scale indicating how coherent the topic is. A higher value indicates a higher coherence.

I also evaluate the refined topic models from the experimental group in terms of four automatic coherence measures [58]: C_{pmi} (which is also know as C_{uci}), C_{npmi}, C_v, and U_{mass}. All

these metrics are based on the co-occurrence of terms. C_{pmi} identifies the coherence of the model based on a sliding window and the pointwise mutual information (PMI) of all word pairs of the given topics' top words. C_{npmi} is an enhanced version of C_{pmi} using the normalized pointwise mutual information (NPMI). C_v is based on previous approaches. It retrieves the co-occurrence counts for the top words using a sliding window and generates a set of vectors after calculating NPMI over these terms. Then, it measures the similarity between these vectors using cosine similarity. Finally, U_{mass} identifies the coherence of the model based on the number of documents in which topics' top terms appear together. Thus, if two terms are related, it is expected to see them together in a set of documents.

Each automatic coherence metric gives a score for an entire topic model. The automatic coherence is calculated considering a window size of 20 terms. Notice that while *perplexity* measure has been widely used for topic models evaluation, I do not consider it because recent studies have shown that this metric is not correlated with human judgments [70].

Finally, I use the NASA Task Load Index (NASA-TLX)[1] [24] 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. The average score for these dimensions is called the *unweighted NASA-TLX* score. It is the most common method to evaluate and report the overall workload level perceived during the task [8].

Before the experiment, I ask participants to complete an interactive tutorial on the user interface (see Figure 4.2). This tutorial explains the use of the different components of the TopicVisExplorer first layout. I do not evaluate the users' performance during the tutorial. This tutorial is available at `http://topicvisexplorer.tk/singlecorpus`.

4.2. Second scenario: Topic models comparison

The second scenario is designed to test: H_{0b}:, *There are no differences in the performance and error rate when comparing topics between people who use a topic similarity metric*

[1] https://human-factors.arc.nasa.gov/groups/TLX/downloads/HFES_2006_Paper.pdf

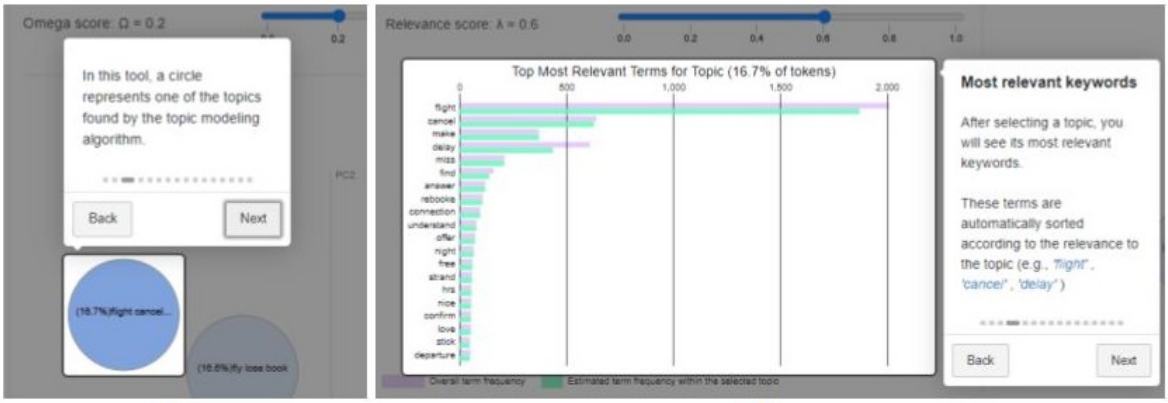


Figure 4.2: Snapshots of the interactive tutorial to explain (a) the representation of topics and the (b) chart with relevant terms

based on keywords and documents and those who use a keyword-based similarity metric. Here I ask users to interpret and compare topics using the second TopicVisExplorer layout (see Figure 3.5). Users must report topics that are similar. Users visualize the same topics of the first user study scenario, and also other LDA-generated topics from a different dataset.

I show participants from the experimental group a layout that uses the proposed topic similarity metric. On the other hand, users from the baseline group see a layout based on the keyword-based topic similarity metric proposed by [68].

I collect metrics for the users' experience, such as time to task completion and the number of labeled topics. Moreover, users self-report the level of workload perceived during this task completing the NASA Task Load Index (NASA-TLX) questionnaire.

Before conducting this scenario, users watch another interactive tutorial (see Figure 4.3) to learn the components of this second layout. Here, I request users to identify similar topics between these two datasets. Again, I do not evaluate the users' performance during this tutorial. This tutorial is available at `http://topicvisexplorer.tk/multicorpora`

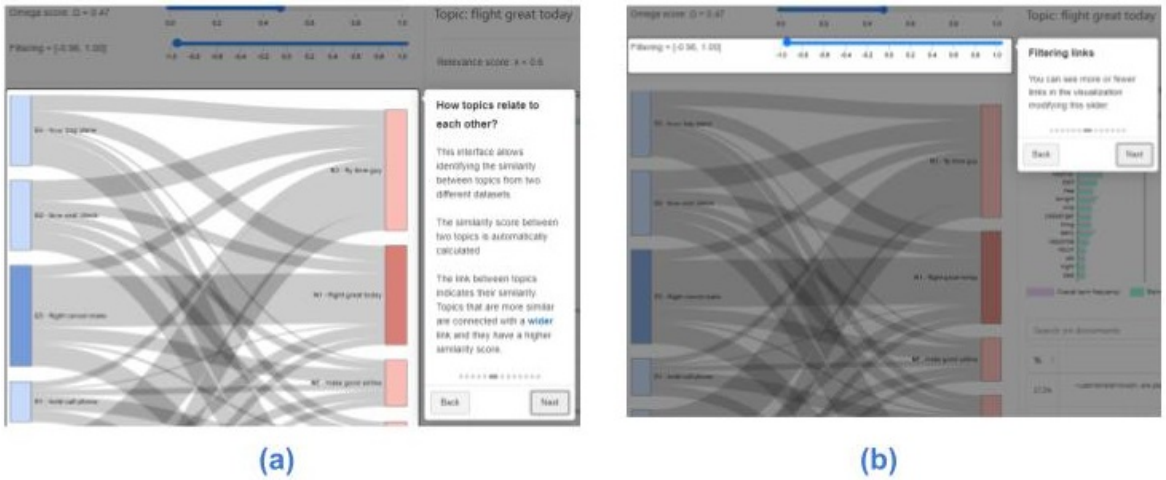


Figure 4.3: Snapshots of the interactive tutorial to explain (a) how topics relate to each other and (b) how filtering links

4.3. User study implementation

This section describes the implementation details of the between-subjects user study.

4.3.1. Users' data collection

TopicVisExplorer has been made accessible to user study participants through a web-browser-based, interactive, visual interface. The user study was done entirely online due to COVID-19 restrictions. The user study scenarios are available at `http://topicvisexplorer.tk/singlecorpus?&scenario=singlecorpus`, and `http://topicvisexplorer.tk/multicorpora?&scenario=multicorpora`.

I used two methods to collect users participants' data (see Figure 4.4). First, in the first scenario, I used a Surveymonkey[2] survey to collect topics' names, Raw NASA-TLX scores, and topics' coherence. Second, a Javascript script inserted in TopicVisExplorer collect metrics associated with users' interactions, such as task completion time, the number of merges and splits per topic, and the refined topic model. After finishing the scenario, users had to click

[2]https://www.surveymonkey.com/

on the *send results* button (see Figure 3.1) (d)) to register their data to a DigitalOcean[3] server.

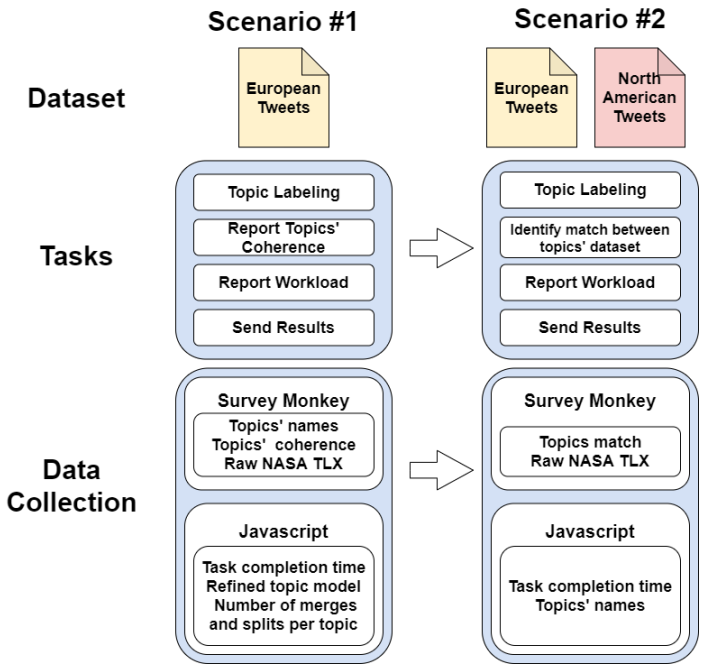


Figure 4.4: User study implementation details

For the second scenario, I used a Surveymonkey survey to collect topics match and Raw NASA-TLX scores. The Javascript script inserted in TopicVisExplorer collects user interactions, such as task completion time and topics' names. As in the previous scenario, participants had to click the *send results* button (see Figure 3.1 (d)) to register their data to a server.

4.3.2. Dataset

In both scenarios, participants analyzed LDA-generated topics from a real-world, large-scale dataset (see Figure 4.4). I used a Facebook-Cambridge Analytica dataset that contains English tweets related to a major data breach scandal collected between April 1st and July 10th from 2018.

The dataset's tweets were collected using Tweepy,[4] a Python library for accessing to the

[3]https://www.digitalocean.com/
[4]http://www.tweepy.org/

standard streaming Twitter API. Using this library, I was able to capture tweets that include hashtags or keywords related to the Facebook-Cambridge Analytica scandal or data privacy, such as: "#CambridgeAnalytica", "#DeleteFacebook", "#Zuckerberg", "#bigdata", "Facebook", "Facebook Cambridge". The complete list of terms used to retrieve this data is available online[5].

This Twitter dataset contains different subsets regarding Twitter user's location. On Twitter, users can self-report the city or country of residence. In order to deal with ambiguous locations (e.g., over 60 different places around the world are called "Paris")[29], I employed the GeoNames[6] API to identify users' geographical regions [22]

In both scenarios, I used the subset that contains 111,745 tweets from 46,927 European Twitter users. During the second scenario, I employed the subset of 342,400 tweets generated by 142,719 North American Twitter users.

One of the most critical parameters for building a topic model is the number of topics [2]. Models with very few topics would result in broad topic definitions that could be a mixture of two or more distributions [3]. On the other hand, models with too many topics are expected to have very specific descriptions that are uninterpretable [3]. Thus, I created different topic models over the European subset, considering a different number of topics. I evaluated the quality of each topic model using the C_v coherence metric[7], which relies on the co-occurrence of terms and has a high correlation with human judgments [74, 58]. The metric returns a score between -1 and 1. A higher score indicates a much coherent model.

The results show that the coherence score is over 0.4 when the number of topics is equal to or higher than five (see Figure 4.5). Taking this into account, a different number of topics were considered during a pilot user study. I selected topic models with six topics, given that a higher number of topics would require participants to spend more than one hour completing the experimental tasks.

Table 4.1 shows the id, name, and top eight terms for the topics used in the user study. Topics from the European subset are identified as E1, E2, ..., E6, while topics from the North

[5]https://github.com/gonzalezf/Regional-Differences-on-Information-Privacy-Concerns
[6]http://www.geonames.org/
[7]https://radimrehurek.com/gensim/models/coherencemodel.html

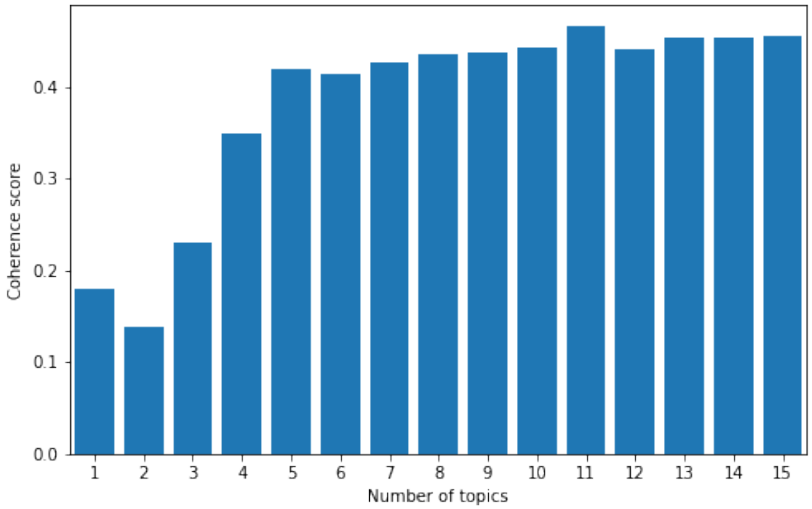


Figure 4.5: Coherence score for LDA topic models from the European subset

America subset are identified as N1, N2, ..., N6. After carefully examining their top 20 terms and their top 20 documents, I assigned a name to all the topics. The ranking of the top terms was determined considering a relevance score equal to 0.6, as suggested by [59]. The complete list of topics' terms and documents are available online[8].

Table 4.1: Topics from the European and North America subset

ID	Name	Top eight terms
E1	Facebook selling user's private data	datum, user, give, change, data, personal, information, ad
E2	Privacy concerns with Facebook features	facebook, privacy, scandal, delete, page, app, twitter, post
E3	Cambridge Analytica's data to manipulate political decisions	make, good, time, company, share, work, story, leave
E4	Popular tech: big data and AI	bigdata, ai, big, late, technology, find, analytic, datascience
E5	Mark Zuckerberg's testimony and the effect of social media over the US presidential political campaign	question, trump, follow, today, day, account, live, news
E6	Business and government's responsibility regarding users' privacy on the Internet	people, business, read, tech, thing, year, great, watch
N1	Stop using Facebook, and delete Facebook campaing	people, delete, today, account, find, day, year, page
N2	Users' privacy on Facebook, privacy costs, personal data protection, and GDPR compliance and Facebook	datum, privacy, user, company, read, sell, pay, tech
N3	Facebook's data collecting practices and data sharing practices	Facebook, make, give, share, ad, friend, live, post
N4	Big data and AI applied to business	bigdata, big, ai, great, business, data, change, late
N5	Facebook scandal & Politics: Mark Zuckerberg being questioned in congress & influence of the scandal on politics	work, question, thing, watch, election, start, talk, campaign
N6	Trump's controversial statements and decisions as an US president	trump, good, call, money, stop, report, follow, love

The topic models contain several topics related to Facebook: (E1) *Facebook selling users' private data*; (E2) *Privacy concerns with Facebook features*; (N1) *Stop using Facebook, and*

[8]https://github.com/gonzalezf/TopicVisExplorer

delete Facebook campaign; (N2) *Users' privacy on Facebook, privacy costs, personal data protection, and GDPR compliance and Facebook*; (N3) *Facebook's data collecting practices and data sharing practices.*

There are also topics related to the influence of the Facebook-Cambridge Analytica scandal on politics: (N5) *Facebook scandal & politics: Mark zuckerberg being questioned in congress & influence of the scandal on politics*; (E5) *Mark Zuckerberg's testimony and the effect of social media over the US presidential political campaign*; (E3) *Cambridge Analytica's data to manipulate political decisions*.

There is a topic related to the responsibility of business and government regarding users' privacy: (E6) *Business and Government's responsibility regarding users' privacy on the Internet*. There are two topics related to big data and artificial intelligence: (E4) *Popular tech: big data and AI*; and (N4) *Big data and AI applied to business*. There is also a topic related to Trump administration: (N6) *Trump's controversial statements and decisions as a US president*.

I used a different dataset in the tutorials. I choose the Twitter US Airline Sentiment dataset[9], which contains 14,640 tweets related to six major US Airlines and labeled according to their sentiment [56]. For the first tutorial, I created a six topics LDA topic model considering all these documents. I divided the dataset into two subsets for the second tutorial: one with only negative tweets and another with neutral and positive tweets. I applied LDA over these subsets, considering six topics in each of them.

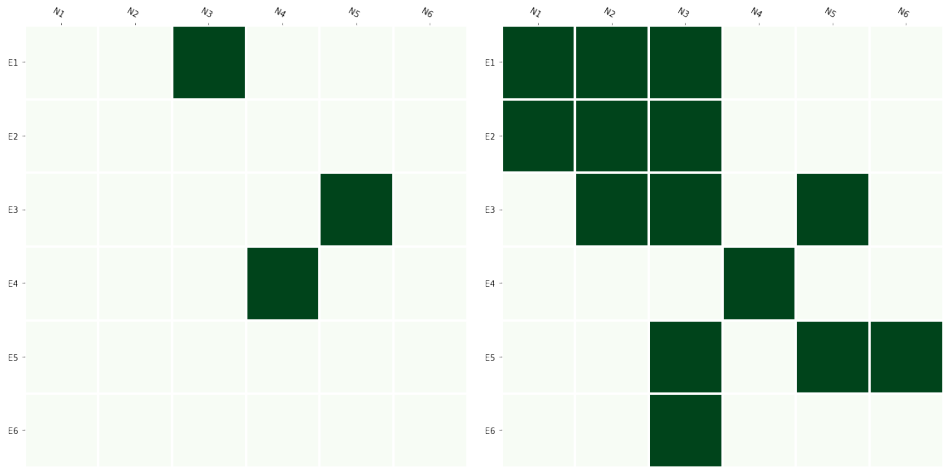
4.3.3. Ground truth

This research proposes a topic similarity metric. Validating and evaluating the results of these metrics is a challenging task because the threshold regarding whether a topic is or is not related to another one is highly subjective [20]. After all, it depends on end users' criteria. Thus, I used three criteria to evaluate the performance of the proposed topic similarity metric.

First, I created a ground truth to identify similar topics. For this purpose, I asked three

[9]https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Computer science students, who did not participate in the final user study, to interpret each topic and report similar ones. These individuals visualized the 20 most relevant documents for each topic using the second TopicVisExplorer layout (see Figure 3.5). They did not visualize any output of an automatic similarity metric, therefore the central panel of the layout was empty (see Figure 3.5 (a)). An inter-coder reliability measure (Cohen's kappa) of 0.339 was obtained, which indicates a *fair agreement* between these individuals [44]. With their answers, I created a strict and moderate ground truth (see Figure 4.6). The strict ground truth corresponds to the intersection of the matches of the three annotators (see Figure 4.6 (a)). The moderate ground truth corresponds to the union of their answers (see Figure 4.6 (b)).



(a) Strict ground truth (b) Moderate ground truth

Figure 4.6: Cells with green color indicate similarity between a pair of topics in the (a) strict ground truth, and (b) moderate ground truth

Moreover, I created a criterion to measure the error rate of the topic similarity metric. I call *errors* to matches between topics that do not share a semantic relationship. For instance, it is challenging to justify a match between a topic about *"humans in space"* and another related to "*TV shows during the last ten years*".

To identify the topics that do not match at all, I applied the following steps. First, I read the 20 most relevant documents and the 20 most relevant keywords for each topic, and I assigned a name to each of them (see Table 4.1). Then, being very flexible, I indicate a match between

all the topics where a relationship between their meaning can be justified. Thus, *errors* are all the matches not identified by this approach (see Figure 4.7). It is important to note that matches not identified by this approach were also not identified by the three annotators (see Figure 4.6).

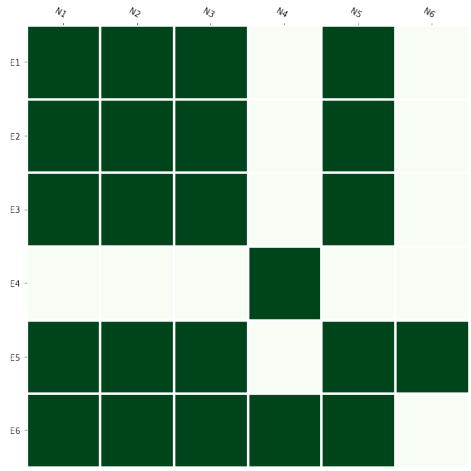


Figure 4.7: Error ground truth. Cells without a green color indicate topics that do not match at all

In particular, it can not be justified a relationship between *(E4): Popular tech: big data and AI* and the following topics from the North America subset: *(N1) Stop using Facebook and Delete Facebook campaign*; *(N2) Users' privacy on Facebook, privacy costs, personal data protection, and GDPR compliance and Facebook*; *(N3) Facebook's data collecting practices and data sharing practices*; *(N5) Facebook scandal & Politics: Mark Zuckerberg being questioned in congress & influence of the scandal on politics*; and *(N6): Trump's controversial statements and decision as a US president*.

It also can not be justified a relationship between the topic: *(N4): Big data and AI applied to business* , and the following topics from the European subset: *(E1) Facebook selling users' private data*; *(E2) Privacy concerns with Facebook features*; *(E3) Cambridge Analytica's data to manipulate political decision*; and *(E5) Mark Zuckerberg's testimony and the effect of social media over the US presidential political campaign*.

Finally, there is not a relationship between *(N6) Trump's controversial statements and decisions as an US president* and the following topics from the European subset: *(E1) Facebook*

selling users' private data; *(E2) Privacy concerns with Facebook features*; *(E3) Cambridge Analytica's data to manipulate political decision*; and *(E4) Popular tech: Big data and AI*.

4.3.4. Word embedding

The proposed topic similarity metric uses a word embedding [47] to identify the semantic context in which terms are framed. I build a word embedding trained solely on the Facebook-Cambridge Analytica dataset, considering the European and North American subsets. Before creating it, I removed stopwords and transformed the text to lowercase. I customized the stopwords to ensure that digits and symbols like "#" were removed but not the words containing them. Links and usernames were removed. Also, I identified bigrams and trigrams, which were incorporated into the word embedding. Moreover, I applied lemmatization. As a result, the corpus comprised 67,706 unique words.

The word embedding was created considering a Word2Vec CBOW architecture with 300 dimensions, and it was trained during 50 epochs. I considered negative sampling and windows size parameters equal to 5. I chose this word embedding architecture because it achieved the best performance for this dataset, according to a prior study that I led [22].

4.3.5. Recruitment

Before starting the user study, all study aspects, instructions, and set up went through several evaluations and pilot testing iterations with three users who did not participate in the actual study. As a result, the user study's instructions were clarified, and the number of topics for the tasks was set up to allow participants to complete both tasks within one hour.

For the user study, I recruited computer science students with no prior knowledge of topic modeling. I required that participants understand the English language, thus be able to read the top keywords and top documents for each topic from the selected dataset.

Informed consent was provided for all participants. Participation was voluntary. In order to safeguard the user study participants' well-being, this study design was evaluated and

approved by the Research Ethics Board from Dalhousie University.

4.3.6. Statistical analysis

In the first scenario, I made several comparisons between the experimental and baseline groups regarding scenario completion time, number of labeled topics, topics' coherence, and Raw NASA TLX scores. I also made evaluations considering only subjects from the experimental group regarding the number of splits and merge operations, the percentage of users who applied a refinement operation per topic, and the automatic coherence scores of their refined topic models.

In the second scenario, I also made comparisons between the experimental and baseline group regarding the number of labeled topics, scenario completion time, number of wrong topics' matches, precision and recall scores regarding the topics' matches, and the Raw NASA TLX scores.

Table 4.2 shows the independent variable (IV), dependent variable, and statistical analysis method used for the data collected in the user study. For all the dependent variables, the normal distribution was confirmed by Shapiro-Wilk's test. I also verified the homogeneity of the variances for each dependent variable with Barlett's test (when the data is normally distributed) and Levene's test (when the data is skewed). To compare data from two groups, I used a t-test when the data follows a normal distribution and a Mann-Whitney U test when it does not. When data follow a normal distribution, but the variances are not equal between groups, I used Welch's two-sample t-test.

Moreover, I used Chi-square goodness of fit tests to determine whether a variable is likely to come from a specified distribution or not. That was used when I evaluated the percentage of users from the experimental group who applied a refinement operation in each topic. It also was the case when I evaluated the NPMI coherence scores per topic.

Furthermore, I used a generalized linear mixed-effects modeling approach, controlling by participant and topic, to evaluate differences between the distribution of coherence scores reported by users from the baseline (*Bas.*) and experimental (*Exp.*) groups.

Table 4.2: Independent variable (IV), dependent variable, and statistical analysis method for the data collected during the user study.

Scenario	IV	Dependent variable	Statistical analysis method
#1	Split and Merge	Number of operations	Mann–Whitney U test
	Bas. and Exp.	Number of seconds in completing the tasks	Welch's two sample t-test
	Bas. and Exp.	Number of labeled topics	Mann–Whitney U test
	Topic ID	Percentage of users who applied a refinement operation	Chi-square goodness of fit
	Topic ID	NPMI Coherence score	Chi-square goodness of fit
	Bas. and Exp.	Coherence reported by users	GLMM
	Bas. and Exp.	Raw NASA TLX scores for each dimension	Mann–Whitney U test
#2	Bas. and Exp.	Number of labeled topics	Mann–Whitney U test
	Bas. and Exp.	Number of seconds in completing the tasks	Mann–Whitney U test
	Bas. and Exp.	Number of wrong topics' matches	Mann–Whitney U test
	Bas. and Exp.	Precision scores regarding the topics' matches	Mann–Whitney U test
	Bas. and Exp.	Recall scores regarding the topics'matches	Mann–Whitney U test
	Bas. and Exp.	Raw NASA TLX scores for each dimension	Mann-Whitney U test

All statistical procedures were performed with a cut-off for significance at 0.05 using Python and R. In all statistical hypothesis tests; I accounted for multiple comparisons by applying alpha adjustment according to Šidák [77, 25]. This method allows controlling the probability of making false discoveries when performing multiple hypotheses tests.

.

Chapter 5

Findings

This section describes how the user study participants refined, and compared topics.

5.1. First scenario: Topic modeling refinement

I invited 120 computer science students to the user study, and 95 (79.16%) agreed to participate. I filtered their answers in two ways. First, I removed all the answers from individuals who did not complete the Surveymonkey survey resulting in 79 valid answers (see Table 5.1). I used this survey to collect topics' names, Raw NASA-TLX scores, and topics' coherence (see Figure 4.4). Then, I removed users who did not send their final topic model. This step is necessary because I could only get secondary data such as the task completion time for users who sent their final topic model. As a result, 71 participants were considered to analyze the first hypothesis (see Table 5.1).

The user study participants saw a model with six LDA-generated topics from the European subset of the Facebook-Cambridge Analytica dataset. Participants interacted with the topic model using the first layout of TopicVisExplorer (see Figure 3.1). They were asked to assign a new name to each topic, and assess topics' coherence. I also requested the experimental group to apply topic modeling refinement operations.

Table 5.1: Number of answers in the topic modeling refinement scenario

	Baseline	Experimental
Invited	60	60
Participated	46	49
Completed SurveyMonkey survey	39	40
Sent topic model	36	35
Performed topic refinement	36	31
Within tolerable time frame	34	28

5.1.1. Answer quality check

I applied two means to evaluate the quality of the answers and remove poor-quality data.

First, I required users of the experimental group to at least perform one topic merging and one topic splitting. I deleted answers from users who did not follow that instruction. This process removed four subjects from the experimental group (see Table 5.1).

Second, I identified the amount of time (in seconds) that users required to complete the first scenario. I expected participants to complete the scenario at once without interruptions. I assume that individuals who required excessive time to complete the tasks lost their focus due to external factors; thus, their answers are not comparable with the rest of them. Therefore, I deleted these outliers from each group. Participants were considered outliers if their completion time was beyond three standard deviations over the mean. As a result, I kept 34 answers from individuals from the baseline group and 28 from the experimental group (see Table 5.1).

5.1.2. Ratio completion task

I asked users to assign a name to all six LDA-generated topics. The distribution of the number of labeled topics do not follow a normal distribution in the baseline group ($W(33) = 0.55$, $p < .001$), and experimental group ($W(27) = 0.76$, $p < .001$). There was no statistically

significant differences in the number of labeled topics between baseline and experimental groups (U = 393.5, N_{bas}=34,N_{exp} = 28, p = .08).

The Bartlett's test for equal variances [4] suggests that these distributions do not have equal variances ($\chi^2(1, 62) = 7.78, p = 005$). Under these conditions, a Welch's two sample t-test [69] found significant differences between these distributions (t(42.13) = -2.28 p =.04). The result suggests that users from the baseline group (M= 1563, SD= 858) required less time to complete the scenario compared with users from the experimental group (M= 2236, SD=1441). Cohen's effect size value (w = .41) suggests a small to medium practical significance [14].

These results suggest that while there are no differences in ratio task completion between groups, participants from the experimental group required more time.

5.1.3. Topic refinement

Participants from the experimental group could apply topic merging and topic splitting operations to refine the initial topic model. The distribution of number of topic merging operations (W(35) = 0.57, p <.001) and topic splitting operations (W(30) = 0.55, p <.001) done by users do not follow a normal distribution. There was no statistically significant differences between the number of topic merging and topic splitting operations performed by users (U($N_{merging}$ = 36, $N_{splitting}$ = 31) = 476.5, p=.47).

Table 5.2 shows the percentage of users who applied any of these topic model refinement operations in each topic. I observed significant differences in those proportions ($\chi^2(5, 50) = 108.24, p < .001$). Cohen's effect size value (w = .59) suggests a high practical significance [14]. Here, I calculated the standard residuals to determine which topics make the greater contribution to this result. I found that compared with other topics, the number of refinement operations is smaller in *(E4) "Popular tech: Big data and AI"* and *(E6) "Business and government's responsibility regarding users' privacy on the Internet"*. The chi-square standard residuals for those topics are -7.00, and -3.64, respectively. The opposite is found

regarding the number of refinement operations applied over the topis *(E2) "Privacy concerns with Facebook features"*, *(E3) "Cambridge Analytica's data to manipulate political decisions"*, and *(E5) "Mark Zuckerberg's testimony and the effect of social media over the US presidential political campaign"*. The chi-square standard residuals are 5.60, 3.08, and 2.24, respectively.

To further examine these results, I evaluated the quality of each topic using a coherence NPMI-based metric[1], as is suggested by [40, 37, 36]. Table 5.2 shows the results. Each value is the average of the NPMI coherence scores [37] considering 5, 10, 15, and 20 terms. There are significant differences in the NPMI coherence scores ($\chi^2(5, 24) = 13.47, p = .02$). Cohen's effect size value (w = .60) suggests a high practical significance [14]. I found that compared with other topics, the NPMI coherence score in *(E4) "Popular tech: Big data and AI"* is significantly higher (chi-square standard residual = 2.90) .

Table 5.2: Percentage of users who applied a topic refinement operation by topic, and topics' automatic coherence. Darker color indicates a higher value

	E1	E2	E3	E4	E5	E6
% of refinement operations	47.06	88.24	70.59	0.0	64.71	23.53
NPMI Coherence score	0.09	0.07	0.03	0.13	0.03	0.03

These results hint that TopicVisExplorer functionalities may help users to identify topics that need further refinement. The percentage of users who decided to apply a topic modeling refinement operation was significantly different across topics. For instance, while most users applied operations over topics related to the Facebook-Cambridge Analytica scandal such as *(E2) "Privacy concerns with Facebook features", (E3) "Cambridge Analytica's data to manipulate political decisions", and (E5) "Mark Zuckerberg's testimony and the effect of social media over the US presidiantial political campaign", none applied an operation over the topic (E4) "Popular tech: Big data and AI"*

The topic *(E4) "Popular tech: Big data and AI"* has the highest automatic coherence score. Additionally, I observed that this topic always appears distant from other topics at different

[1]topic coherence computed using the implementation at: https://github.com/jhlau/topic-coherence-sensitivity

omega score values (see Figure 5.1). The proposed topic similarity metric compares topics considering their most relevant documents and keywords. The parameter *omega* score gives the relationship between these two entities. When calculating the similarity between two topics, a higher *omega* score implies higher importance of their most relevant documents but a lower significance of their most relevant keywords. Thus, the results show that topic E4 differs from the other topics either by its top keywords or its topics' top documents.

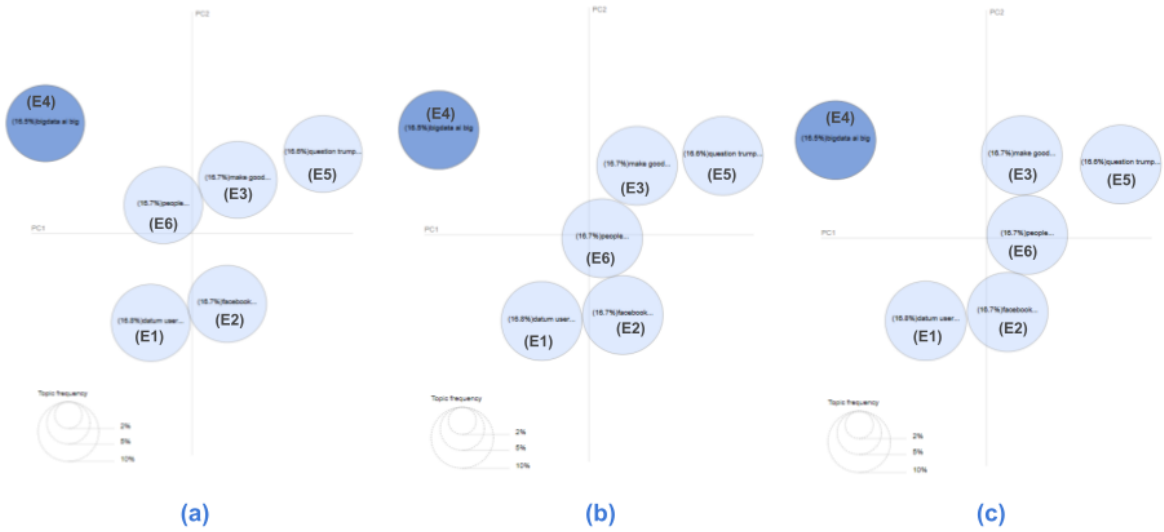


Figure 5.1: Inter-topic similarity in the first scenario for the omega scores: (a) 0.0 , (b) 0.5, and (c) 1.0. The topic E4 is highlighted

The results show that the number of topic refinement operations is also significantly lower for the topic *(E6) " Business and government's responsibility regarding users' privacy on the Internet"*. This topic is data privacy-related. However, it is not directly related to the data breach scandal or with the data breach scandal companies: *Facebook* and *Cambridge Analytica*. That it sets apart from the topics: *(E1) "Facebook selling user's private data"*, *(E2) "Privacy concerns with Facebook features"*, *(E3) "Cambridge Analytica's data to manipulate political decisions"*, *(E5) " Mark Zuckerberg's testimony and the effect of social media over the US presidential political campaign"*. Participants preferred to apply topic refinement operations in topics more related to the Facebook-Cambridge Analytica scandal (see Table 5.2).

5.1.4. Reported topics' coherence

Figures 5.2 and 5.3 show the distribution and relative frequency of the topics' coherence reported by users from the baseline and experimental groups. The results show that for the baseline and experimental groups, the median of the self-reported coherence is above 4.0 out of 5.0 for the topics: *(E1) "Facebook selling user's private data"*, *(E2) "Privacy concerns with Facebook features"*, and *(E4) "Popular tech: big data and AI"* (see Figure 5.2). These topics also have a high NPMI coherence score compared with the other topics (see Table 5.2).

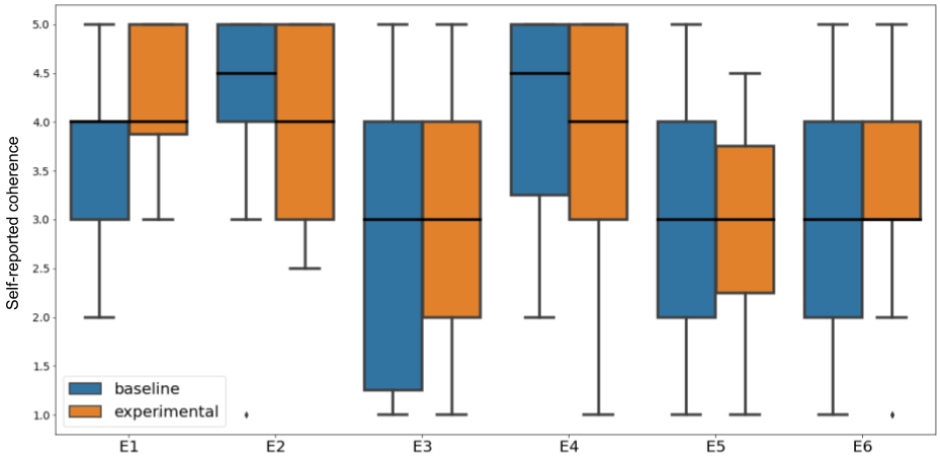


Figure 5.2: Distribution of coherence scores by topic as reported by users. A higher score indicates a higher coherence

In order to evaluate differences between topics' self-reported coherence between the baseline and experimental groups, I used a generalized linear mixed-effects modeling approach (GLMM), controlling by participant and topic. Due to lack of normality, I modeled the dependent variable as a binomial distribution, where 1 indicates a reported coherence score over the global median across conditions (Mdn = 4.0), and 0 otherwise. I did not find statistically significant differences between groups (Experimental group, Estimate=-0.19, Std error=0.29, z value= -0.640, p=.52).

Using the same approach, I also compared the coherence scores between refined and non-refined topics. I did not find significant differences between these conditions (non-refined topics, Estimate = 0.10, Std error = 0.37, z value = 0.27, p=.787)). Therefore, I did not

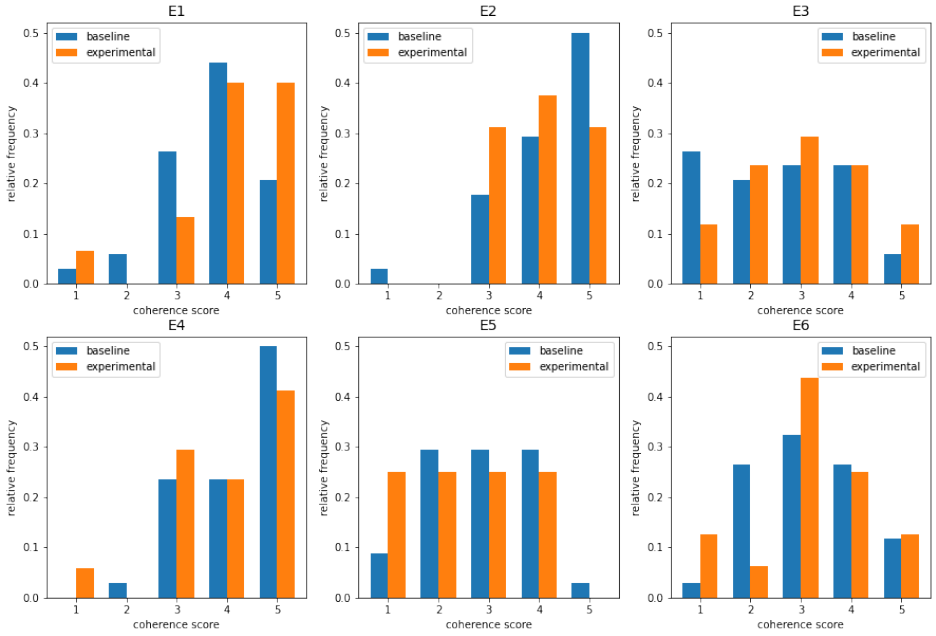


Figure 5.3: Relative frequency of coherence scores per topic

find evidence to reject the first null hypothesis, H_{0a}: *There are no differences in the observed coherence of topics between people who apply TopicVisExplorer's mechanisms for topic splitting and topic merging and people who do not.*

5.1.5. Automatic model coherence in experimental group

Figure 5.4 shows the initial and final automatic coherence scores for the topic models from the experimental group before and after applying topic merging and splitting operations. Table 5.3 reports the percentage of users from the experimental group who achieved a higher automatic coherence score than the initial one. For all the coherence metrics, between 40% and 60% of participants improved the score. These results provide evidence that non-expert users can improve the quality of the topic model using topic modeling refinement operations.

Table 5.3 shows that 60.71% of participants improved the coherence of the model according to the C_v coherence metric. This metric relies on the co-occurrence of topics' top terms and is the automatic coherence metric that mostly correlates with human judgments [74, 58].

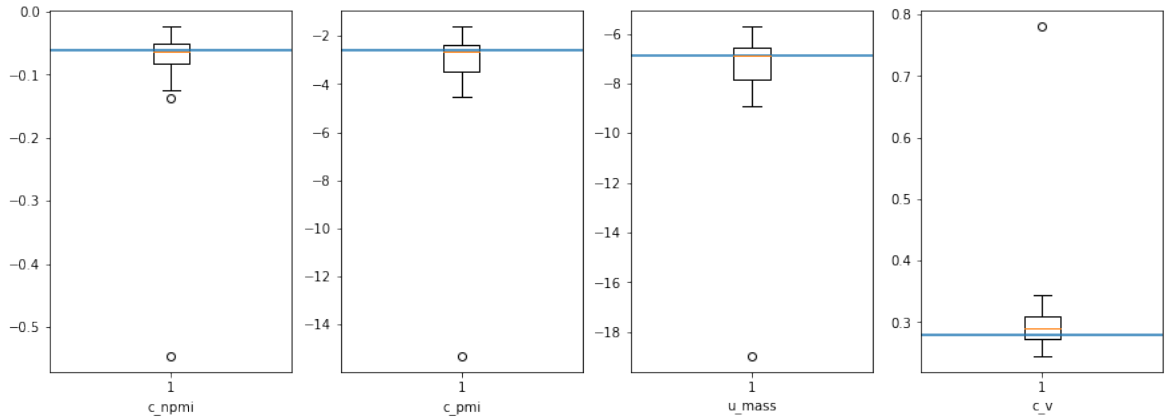


Figure 5.4: Coherence scores of the refined topic models in the experimental group. Horizontal blue line indicates the initial coherence score. A higher score indicates a higher coherence

Table 5.3: Percentage of users whom, after applied topic modeling refinement operations improved the automatic coherence score of the refined topic model

c_npmi	c_pmi	u_mass	c_v
42.86	42.86	50.0	60.71

To further examine this result, I identified the topics that modified users who improved the coherence of the model. I did the same for the users who did not achieve a higher C_v coherence score than the initial. (see Table 5.4). I observed significant differences in the proportion of users that modified the topic *(E1) "Facebook selling user's private data"* ($\chi^2(1, 26) = 8.22, p = .004$). Cohen's effect size value (w=.32) suggests a moderate to high practical significance. I also observed differences in the proportion of users who chose to modified the topic: *(E6) "Business and government's responsibility regarding users' privacy on the Internet"* ($\chi^2(1, 26) = 6.39, p = .01$). Cohen's effect size value (w=.44) suggests a moderate to high practical significance.

I did not find statistically significant differences regarding the proportion of users who applied topic refinement operations on the topics (see Table 5.4): *(E2) "Privacy concerns with Facebook features"* ($\chi^2(1, 26) = 0.09, p = .76$), *(E3) "Cambridge Analytica's data to manipulate political decisions"* ($\chi^2(1, 26) = 0.27, p = .60$), *(E4) "Popular tech: big data and AI"* ($\chi^2(1, 26) = 0.0, p = 1.0$), and *(E5) "Mark Zuckerberg's testimony and the effect*

Table 5.4: Percentage of users who applied a topic refinement operation by topic according to if they achieved or not a higher C_v coherence score than the initial. Darker color indicates a higher value

	E1	E2	E3	E4	E5	E6
Users with high C_v score	52.94	76.47	58.82	0.0	64.71	23.53
Users with low C_v score	27.27	72.73	64.64	0.0	45.45	9.09

of social media over the US presidential political campaign" ($\chi^2(1, 26) = 3.37, p = .07$).

These results suggest that depending on which topic non-expert users apply a refinement operation, the overall coherence of the topic model varies.

5.1.6. Workload reported in the topic modeling refinement scenario

Besides the coherence reported by users, I was interested in investigating possible differences in the perception of task workload. Figure 5.5 presents the scores obtained from the Raw NASA Task Load Index (TLX) questionnaire. I did not find statistically significant differences in the unweighted TLX score (t(60)=0.81, p=0.42), neither in the distribution of the scores of its components: Mental demand (Mann– Whitney U = 440.0, N_{bas}=34, N_{exp} = 28, p = .30); Physical demand (Mann– Whitney U = 378.5, N_{bas}=34, N_{exp} = 28, p = .08); Temporal demand (t(60)=0.10, p=0.92); Performance (t(60)=1.47, p=0.15); Effort (Mann–Whitney U = 428.5, N_{bas}=34, N_{exp} = 28, p = .25), and Frustration (Mann– Whitney U = 455.5, N_{bas}=34,N_{exp} = 28, p = .39). Therefore, users who applied topic merging and splitting operations did not report a higher workload demand than those who completed the task using a static model. This is a positive result considering that the proposed document-based topic splitting operation requires users to carefully evaluate the documents associated with the new subtopics they wish to create.

In this work, I present a visual analytics tool that allows non-expert users to adjust the visualized topic model according to their needs without understanding the inner-working of the topic model algorithm. Overall, the results show that user study participants who applied

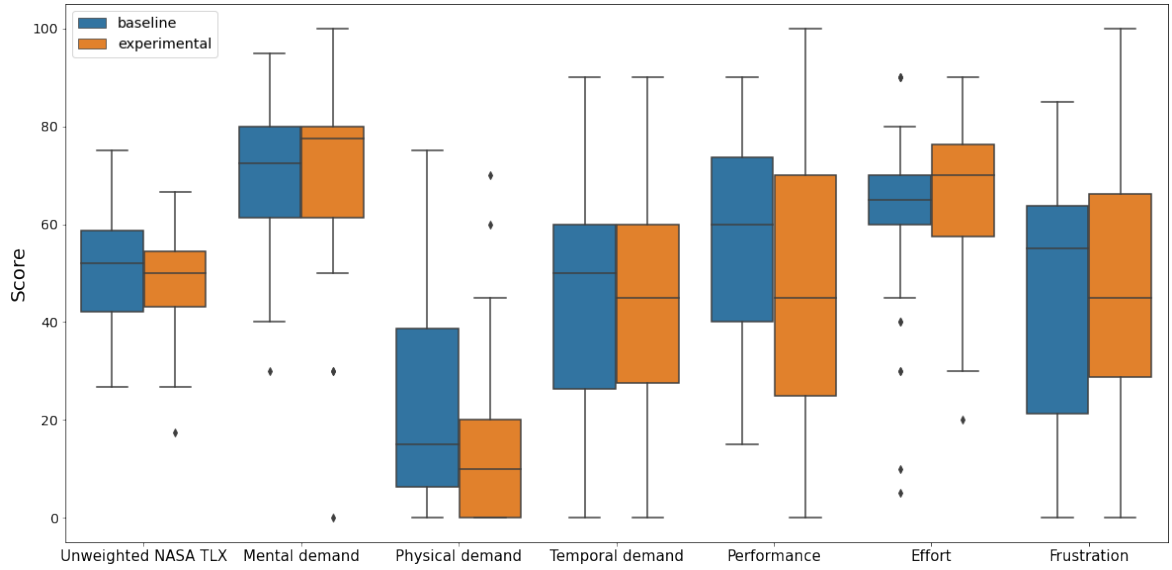


Figure 5.5: Distribution of participant responses to the NASA TLX questionnaire regarding the topic modeling refinement scenario. A lower score indicates a better result.

topic modeling refinement operations did not label more topics than participants who did not apply these functionalities. They also did not self-report higher coherence scores per topic than their counterparts. However, automatic coherence metrics on their refined topic models show that a noticeable percentage of them could improve their quality without reporting a higher workload than users who could not apply these functionalities.

5.2. Second scenario: Topic models comparison

In the second scenario, participants evaluated the similarity of two topic models about the Facebook-Cambridge Analytica scandal: one from European Twitter users and another from North American Twitter users. Using the second layout of TopicVisExplorer (see Figure 3.5), users had to assign a name to each topic and report similar topics between these two topic models. The experimental group used the proposed topic similarity metric, while the baseline group used a keyword-based topic similarity metric proposed by [68].

After completing the previous scenario, participants were again randomly assigned to an

experimental or baseline group. I filtered their answers by excluding those who did not complete the SurveyMonkey survey and those who did not send their final topic model (see Figure 4.4). I used the survey to collect topics match and Raw NASA-TLX scores. Moreover, I required participants to send their topic model to access secondary data such as the task completion time and topics' names. As a result, I considered 37 answers in the baseline group and 42 in the experimental group (see Table 5.5).

Table 5.5: Number of answers in the topic models comparison scenario

	Baseline	Experimental
Invited	60	60
Participated	47	48
Completed Surveymonkey surveys	37	42
Sent topic model	32	36
Labeled topics	26	30
Within tolerable time-frame	23	27

5.2.1. Answer quality check

I applied two means to evaluate the answers' quality. I excluded answers from users who labeled less than eight topics and those who spent excessive time completing the tasks.

First, I assume that users assign a name to each topic after interpreting it. This is an essential step before starting to compare it with another one. In order to safeguard the quality of the results, I excluded answers from users who labeled less than eight out of twelve topics. Participants that labeled less than eight topics are considered outliers because the number of labeled topics was beyond three standard deviations over the mean. As a result, I kept 26 answers from users of the baseline group and 30 answers from users of the experimental group (see Table 5.5).

Second, a limitation of performing the user study online is that I can not control the participants' environment. I expected participants to complete the task at once without any

interruption. I assume that individuals who required excessive time to complete the task lost their focus during the activity due to external factors. I believe that their answers can not be compared with the answers from users who complete the task in a reasonable amount of time. Thus, I excluded outliers from each group. Participants were considered outliers if their completion time was beyond three standard deviations over the mean. As a result, after this process, I kept 23 answers from individuals from the baseline group and 27 answers from individuals of the experimental group (see Table 5.5).

5.2.2. Topics labeling

I required users to assign a name to all twelve LDA-generated topics. The distribution of the number of labeled topics do not follow a normal distribution in the baseline group (W(25) = 0.69, p <.001), and experimental group W(29) = 0.62, p <.001). I did not find differences in the number of labeled topics between experimental and baseline groups (U(N_{bas} = 26, N_{exp} = 30) = 329.5, p=.13)

I also compared the amount of time (in seconds) that users required to complete the user study scenario. The distribution of time required for users from the baseline group (W(22) = 0.83, p =.001), and experimental group (W(26) = 0.77, p <.001) do not follow a normal distribution. I did not find differences in the amount of time participants from each group took to complete the scenario (U(N_{bas} = 23, N_{exp} = 27) = 244.0, p=.09).

5.2.3. Matching topics

Users reported similar topics between the European and North America subset from the Facebook-Cambridge Analytica dataset. Figure 5.6 (a) and Figure 5.6 (b) summarize the answers from the baseline and experimental group, respectively. Each cell indicates the percentage of users who identified those topics as similar. Moreover, Figure 5.6 (c) shows the percentage difference between the answers from these two groups. While there are some similarities between the answers (for instance, 100% of the users in the baseline and the experimental group reported similarity between E4 and N4), there are also some notorious

differences. In fact, for three matches, the absolute percentage differences between the answers from the groups are higher than 30.0%: E6-N1 (-69.08%), E3-N3 (-51.21%), E5-N5 (35.10%). Nevertheless, no differences between the distribution of number of matches by users were found between experimental and baseline groups (U(N_{bas} = 23, N_{exp} = 27) = 278.0, p=.27).

(a) Match between topics in baseline group

(b) Match between topics in experimental group

(c) Difference between groups

Figure 5.6: (a) and (b) indicate the percentage of users who reported those topics as similar in the baseline and experimental group, respectively. (c) shows difference between groups.

While 77.78% of users from the experimental group reported the topics *(E6) "Business and Government's responsibility regarding users' privacy on the Internet"* and *(N1) "Stop using Facebook & Delete Facebook campaign"* as similar, only 8.7% of the users from the baseline group did the same. In this case, the similarity between these topics is not evident at first glance; however, N1 corresponds to a more fine-grained topic that E6 can absorb. While this connection does not appear in the strict ground truth, it appears in the moderate ground truth (see Figure 4.6).

A similar pattern appears for the topics *(E3) "Cambridge Analytica's data to manipulate political decisions"* and *(N3) "Facebook's data collecting practices and data sharing practices"*, where 51% of the users from the experimental group reported as similar while only 4.35% of users from the baseline group did the same. Again, while the similarity between these topics is not evident, it can be justified considering that the topics' names describe the reasons that sparked the Facebook-Cambridge Analytica scandal [21]. As in the previous case, while this connection does not appear in the strict ground truth, it appears in the

moderate ground truth (see Figure 4.6).

On the other hand, 86.96% of users from the baseline group reported the topics *(E5) "Mark Zuckerberg's testimony & the effect of social media over the U.S. presidential political campaign"* and *(N5) "Facebook scandal & Politics: Mark Zuckerberg being questioned in congress & influence of the scandal on politics"* as similar. In contrast, 51.75% of users from the experimental group did the same. In this case, both topic's names coincide with the influence of the data privacy scandal on politics and Mark Zuckerberg's testimony in front of the U.S. Congress. This connection does not appear in the strict ground truth but it does in the moderate ground truth (see Figure 4.6). These results provide evidence that while the proposed topic similarity metric can identify some not evident matches, there is still room for improvement.

In the comparison of topic models different phenomena can appear [2]: (1) some topics will be (close to) direct matches of one other, sharing distributions of both words and documents, (2) a topic from one model will occasionally split into multiple topics in another model (or, multiple topics may merge, depending on the direction of comparison), and (3) some topics from one model will have no correlated counterpart in the other model. In this user study, all the topics should be in some way related to the Facebook-Cambridge Analytica scandal. As a side effect, while there are evident similar topics, finding other matches requires further inspection and can vary according to users' needs and criteria. These conditions make it challenging to identify a valid threshold to determine a similarity between two topics. In this direction, I implemented several approaches to evaluate the results.

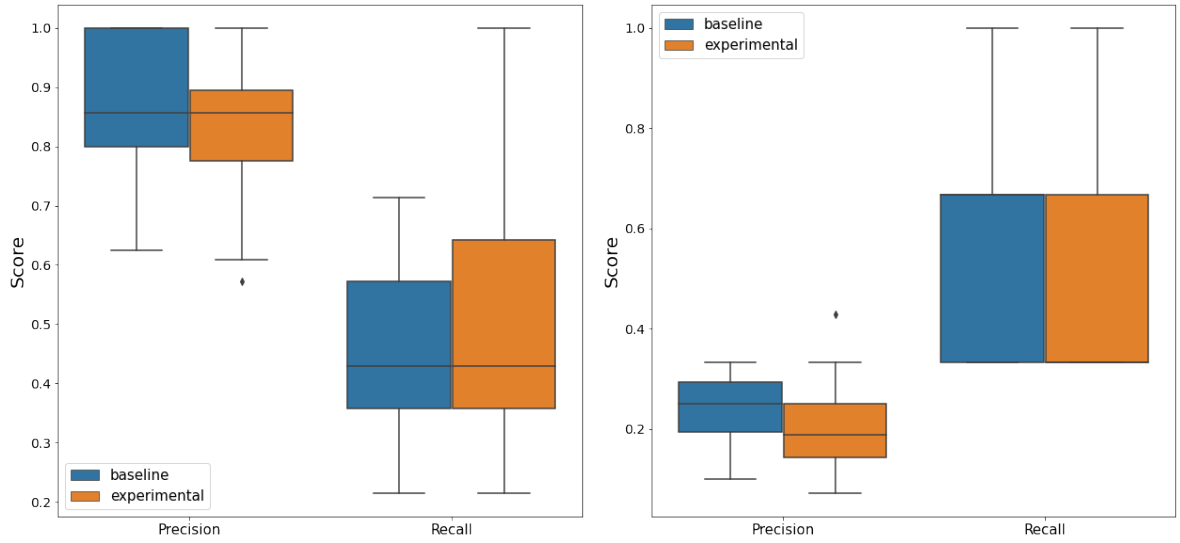
5.2.4. Match error rate

I computed the number of match errors considering the error ground truth (see Figure 4.6 (c)). The number of errors from users from the baseline group (W(22) =0.62, p <.001), and from users from the experimental group (W(26)=0.30, p <.001) do not follow a normal distribution. The failure rate between these groups differ significantly (U(N_{bas} = 23, N_{exp} = 27) = 244.0, p=.027). While 30.44% of the users from the baseline group made at least one mistake, only 7.41% of the users from the experimental group did the same.

Thus, these results provide support to reject the second null hypothesis in terms of error rate, H_{0b}: *There are no differences in the performance and* ***error rate*** *when comparing topics between people who use a topic similarity metric based on keywords and documents and those who use a keyword-based similarity metric.*

5.2.5. Topic similarity metric precision and recall

I also computed widely used information retrieval evaluation metrics, precision and recall, where precision is the fraction of correct answers over the total number of answers given, and recall is the fraction of retrieved correct answers out of all correct ones. Figure 5.7 shows the precision and recall for the baseline and experimental group when comparing with the moderate ground truth (see Figure 5.7 (a)) and strict ground truth (see Figure (5.7 (b)).



(a) Moderate ground truth

(b) Strict ground truth

Figure 5.7: Precision and recall for the baseline and experimental group after comparing their answers with (a) moderate ground truth, and (b) strict ground truth

When comparing participants' answers with the moderate ground truth, the precision scores from the baseline group, (W(22) =0.88, p=.01) and experimental group (W(26) =0.92, p=.04) do not follow a normal distribution. The distributions of the precision scores in the

two groups did not differ significantly (Mann– Whitney U = 273.5, N_{bas}=23, N_{exp} = 27, p = .24).

Regarding the recall scores, I also did not find statistically significant differences between experimental and baseline groups (t(48) = -0.93, p=.36). When the approach to determining a match between topics is somewhat flexible (for instance, using the moderate ground truth), the results provide evidence that a topic similarity metric that considers keywords and documents does not lead to a different performance than a keyword-based topic similarity metric.

I also evaluated the topic similarity metric performance regarding the strict ground truth. The precision scores from the experimental group (W(26) =0.92, p=.03) do not follow a normal distribution. I found that the precision scores obtained for the experimental group were lower than those from the baseline group (U = 215.5, N_{bas}=23, N_{exp} = 27, p = .03).

In this ground truth, the recall scores from the experimental group (W(26) =0.72, p <.001) and baseline group (W(22) =0.80, p <.001) do not follow a normal distribution. These distributions did not differ significantly (Mann– Whitney U = 271.5, N_{bas}=23, N_{exp} = 27, p = .21). When the approach to determining a match between topics is strict; the results show that the ability to identify all correct matches is not different between the proposed topic similarity metric and the baseline. However, the ability to return only correct matches is lower in the proposed metric.

Compared with what was expected, when a strict approach is used to match topics, the comparison of precision scores between groups provides support to reject the second null hypothesis in terms of the performance, H_{0b}: *There are no differences in the* ***performance*** *and error rate when comparing topics between people who use a topic similarity metric based on keywords and documents and those who use a keyword-based similarity metric.*

5.2.6. Workload reported in the topic models comparison scenario

I was also interested in investigating possible differences in the perception of workload during the comparison of topic models. Figure 5.8 presents the scores obtained from the Raw

NASA (TLX) questionnaire filled after completing the second scenario. I did not find statistically significant differences in terms of the unweighted TLX score (t(48)=-0.76, p=.45), and in the distribution of the scores of its components–Mental demand (t(48)=-0.33, p=.74); Physical demand (Mann– Whitney U = 290.0, N_{bas}=23, N_{exp} = 27, p = .35); Temporal demand(t(48)=-1.7, p=.10); Performance (t(48)=-0.56, p=.58); Effort (t(48)=-0.93, p=.35); and Frustration (Mann– Whitney U = 302.5, N_{bas}=23, N_{exp} = 27, p = .44). Therefore, users who identified topics using the proposed topic similarity metric did not report a different workload demand than those who completed the tasks using the baseline metric.

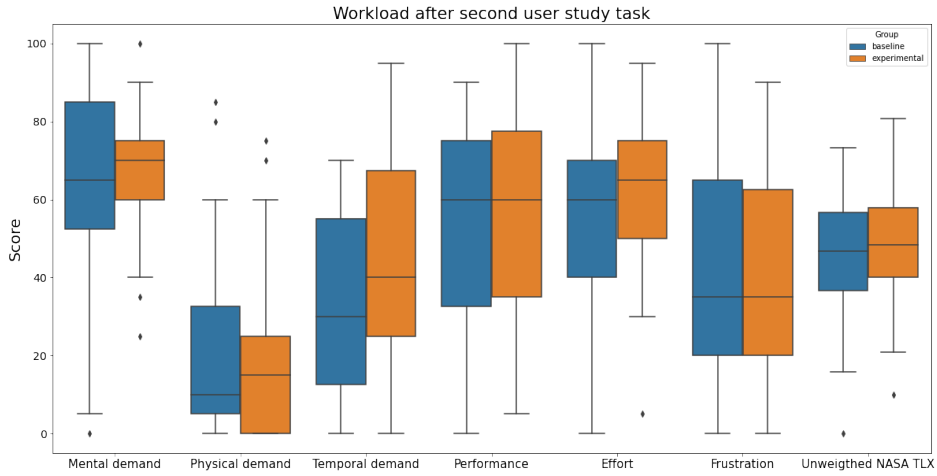


Figure 5.8: Distribution of participant responses to the NASA TLX questionnaire regarding the topic models comparison scenario. A lower score indicates a better result

In this work, I introduce the second layout of TopicVisExplorer that allows users to compare and identify similar topics from two datasets. I also proposed a topic similarity metric that evaluates the similarity between topics regarding their most relevant keywords and their most relevant documents. Overall, the results show that the failure rate when comparing topics is lower in participants who used the proposed metric. The results also show that the metric's performance depends on the approach used to determine the match between topics. While the proposed metric does not have a different performance in identifying loose matches than the baseline metric, the ability to return only correct matches is lower when a strict approach is used to match topics. The results also evidence that users' workload did not differ between the experimental and baseline groups.

Chapter 6

Discussion and conclusions

I seek to understand how non-expert users refine and compare topics. This section discusses and provides conclusions regarding the implications of the thesis results, their limitations, and future work.

6.1. Topic modeling refinement

TopicVisExplorer allows users to incorporate the semantics of a domain knowledge for topic model refinement without understanding the inner workings of the topic model. This work expected that the coherence reported by users per topic would improve significantly after applying topic merging and topic splitting. I did not find enough evidence to support this hypothesis. Several factors might explain this situation.

First, non-expert users do not have preconceptions of how refinements operations might impact the topic model [40]. This is a disadvantage, because the refined topic might not always result in a more coherent topic. This finding goes in line with Hu et al. [28], where it was found that users sometimes create inscrutable correlations, such as connecting unrelated words, and that even sensible feedback did not always lead to successful topic changes. In our case, this interpretation is supported by the automatic coherence scores obtained from the final topic models of the experimental group. While a noticeable percentage of users

improved the initial coherence, for instance, 60.71% of users from the experimental group achieved a better C_v coherence score, there is still an important percentage of users who did not achieve the same. While topic modeling refinement operations are helpful and powerful, there are still entry barriers for non-expert users. I hypothesize that teaching the impact of these operations in the refined topic model and conducting exercises with different datasets might help users get more coherent topics after applying topic modeling refinement operations.

Topic model quality and the way users refine topic models are context-dependent [40]. While the vast majority of the topics are related to a data privacy scandal, some of them are more semantically different. That is the case of *(E4) "Popular tech: Big data and AI"* which in the TopicVisExplorer layout appears more distant from the other topics. None applied a refinement operation over this topic. Instead, non-expert users focused on topics more related to the scandal which also appear closer to each other: *(E3) "Cambridge Analytica's data to manipulate political decisions"*, and *(E5) "Mark Zuckerberg's testimony and the effect of social media over the US presidential political campaign"*. Users' trust and support in the visualization might explain the reasons for these interactions [74].

Moreover, non-expert users tended to apply refinement operations on topics with lower coherence scores, suggesting that topic modeling visualizations tools such as TopicVisExplorer can help users identify topics that need further refinement to improve their quality. Future work should confirm the reasons that make users apply a refinement operation over a topic.

6.2. Topic models comparison

This research offers a visual representation to reflect the similarity between topic models. I also propose a new topic similarity metric. Validating and evaluating the results of topic similarity metrics is a challenging task, given that the threshold regarding whether a topic is or not related to another one is highly subjective [20] because it depends on end users' criteria. In order to reduce the subjectivity, in future studies it is necessary to explicit the criteria that makes two topic match. For instance, match topics that largely share the distribution

of both words and documents, or consider the match between topics that share a common semantic meaning at a high level but with minor differences in their details. Furthermore, one of the main problems is the absence of benchmark datasets for comparative evaluation between models, making the application of automated, unsupervised evaluation methods challenging or unfeasible [20]. Given this situation, in this project, I designed several evaluation methods in order to identify the performance of the proposed topic similarity metric.

First, I compared the proposed topic similarity metric with a baseline [68] regarding the number of errors made by non-expert users while comparing two LDA topic models. The results suggest that the proposed topic similarity metric can significantly reduce non-expert users' errors. The answers from the NASA TLX questionnaire indicate that non-expert who performed the comparison of topics with the proposed metric did not report a higher mental, physical, temporal demand than those who performed the same activity using the topic similarity metric baseline. The same pattern is found for the perceived performance, effort, and frustration level. These results show that non-expert users can benefit from visualization tools to compare topics models and an automatic metric that evaluates the similarity between topics considering their keywords and documents.

Second, I evaluated the precision and recall for the baseline and experimental groups considering a moderate and strict ground truth generated by three annotators. When comparing the answers between the three annotators, I confirmed that the similarity between topics is a subjective task. This can explain why the inter-coder reliability score of these individuals was just *fair*.

While not statistically significant differences were found between the topic similarity metrics when the moderate ground truth was considered, I found that the proposed topic similarity metric achieved a lower precision score than the baseline when the strict ground truth is considered. Several factors can explain this finding. First, the strict ground truth considers only three out of thirty-six possible theoretical matches. Therefore failing to retrieve one of these matches highly impacts the precision score. Second, while the proposed metric reduces the number of erroneous matches, it also suggests that two topics are similar in cases where there was no total consensus among annotators.

Comparative analysis of topic models is an open research challenge that has gained little attention so far [5, 20]. This research introduces an algorithm for evaluating the similarity between two topic models, whether from the same or different corpora. Our results confirm our technique's usefulness in comparing topics models generated from a real-world, large-scale dataset.

6.3. Limitations and future work

As in any study, this research has limitations that need to be taken into consideration. First, the user study only considered data privacy-related LDA topic models. While my approach is helpful in the analysis of large-scale Twitter discussions about a data privacy leak, I plan to investigate its use in other domains. Second, further studies can consider other topic modeling algorithms, such as LDA2VEC [49] or Non-Negative Matrix Factorization [38, 53].

There are several important avenues for extending TopicVisExplorer. First, future work can seek a solution to compactly visualize a larger number of topics (e.g., 100 topics). Second, future TopicVisExplorer versions can incorporate mechanisms to allow users to visualize the relationship between documents and topics. In LDA, every document is a mixture of topics. Thus, each document contains words from several topics in different proportions. This relationship might be identified by users, for instance, by highlighting with different colors document's words.

While TopicVisExplorer allows users to perform the most relevant topic refinement operations, future versions can also support other topic refinement methods such as adding and removing words, removing documents, and adding stop words.

Currently, the default topic's name is determined by its three most relevant keywords. Future work can incorporate automatic topic labeling algorithms to provide users with potentially more representative topic names.

Last, because the time and space complexity of TopicVisExplorer can considerably increase

in topic models with a high number of topics, future action should be taken to improve the tool's performance.

6.4. Conclusions

In this manuscript, I presented an interactive visualization system to address some limitations of previous topic modeling visualizations tools related to the refinement and comparison of topics, which were based only on their most relevant keywords. TopicVisExplorer supports users in refining topics of one topic model and evaluating the similarity between topics from one or two topic models.

I conducted a user study with 95 non-expert users to evaluate TopicVisVisExplorer functionalities for refining and comparing topics from a large-scale real-world Twitter dataset. The results show that participants recognized topics that need further refinement in order to improve their coherence. Non-experts applied topic merging and topic splitting operations to adjust the visualized topic model to their needs. The analyses show that subjects tended to improve the automatic coherence of a topic model after applying topic merging and topic splitting.

When comparing two data-privacy-related topic models, I confirmed that finding the similarity between them is highly subjective. While the similarity between two topics is evident in some cases, other matches require further inspection, and there is no easy consensus among people. I identified that the functionalities of TopicVisExplorer support users during the comparison of topics. In fact, the results suggest that the proposed topic similarity metric can significantly reduce the number of erroneous matches during the comparison of topic models.

Bibliography

[1] Nikolaos Aletras and Mark Stevenson. Measuring the similarity between automatically generated topics. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 22–27, 2014.

[2] Eric Alexander and Michael Gleicher. Task-driven comparison of topic models. *IEEE transactions on visualization and computer graphics*, 22(1):320–329, 2015.

[3] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic significance ranking of lda generative models. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 67–82. Springer, 2009.

[4] Maurice Stevenson Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282, 1937.

[5] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

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

[8] Alex Cao, Keshav K Chintamani, Abhilash K Pandya, and R Darin Ellis. Nasa tlx: Software for assessing subjective mental workload. *Behavior research methods*, 41(1):113–117, 2009.

[9] A. J. Chaney and David M. Blei. Visualizing topic models. In *ICWSM*, 2012.

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

[11] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. *Advances in neural information processing systems*, 19:241–248, 2006.

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

[13] Jason Chuang, Christopher D Manning, and Jeffrey 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.

[14] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[15] Wenwen Dou, Xiaoyu Wang, Remco Chang, and William 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.

[16] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013.

[17] Miles Efron, Peter Organisciak, and Katrina 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.

[18] Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. Semantic concept spaces: Guided topic model refinement using word-embedding projections. *IEEE transactions on visualization and computer graphics*, 26(1):1001–1011, 2019.

[19] Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen, Daniel Keim, and Christopher Collins. Visual analytics for topic model optimization based on user-steerable speculative execution. *IEEE transactions on visualization and computer graphics*, 25(1):374–384, 2018.

[20] Mennatallah El-Assady, Fabian Sperrle, Rita Sevastjanova, Michael Sedlmair, and Daniel 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.

[21] Felipe González, Yihan Yu, Andrea Figueroa, Claudia López, and Cecilia Aragon. Global reactions to the cambridge analytica scandal: An inter-language social media study. 2019.

[22] Felipe González, Andrea Figueroa, Claudia López, and Cecilia Aragón. Regional differences in information privacy concerns after the facebook-cambridge analytica data scandal. *(under review) Computer Supported Cooperative Work: the journal of collaborative computing*, 2021.

[23] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[24] Sandra G Hart and Lowell 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.

[25] Winston Haynes. Bonferroni Correction. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*, pages 154–154. Springer New York, New York, NY, 2013.

[26] Enamul Hoque and Giuseppe Carenini. Convis: A visual text analytic system for exploring blog conversations. In *Computer Graphics Forum*, volume 33, pages 221–230. Wiley Online Library, 2014.

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

[28] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine learning*, 95(3):423–469, 2014.

[29] Alan Jackoway, Hanan Samet, and Jagan Sankaranarayanan. Identification of live news events using twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 25–32, 2011.

[30] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

[31] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[32] Pooja Kherwa and Poonam Bansal. Topic modeling: A comprehensive review. 2019.

[33] Dongwoo Kim and Alice Oh. Topic chains for understanding a news corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 163–176. Springer, 2011.

[34] Hannah Kim, Barry Drake, Alex Endert, and Haesun Park. Architext: Interactive hierarchical topic modeling. *IEEE transactions on visualization and computer graphics*, 2020.

[35] Wojtek Krzanowski. *Principles of multivariate analysis*, volume 23. OUP Oxford, 2000.

[36] Jey Han Lau and Timothy Baldwin. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 483–487, 2016.

[37] Jey Han Lau, David Newman, and Timothy 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*, pages 530–539, 2014.

[38] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[39] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. ivisclustering: An interactive visual document clustering via topic modeling. In *Computer graphics forum*, volume 31, pages 1155–1164. Wiley Online Library, 2012.

[40] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42, 2017.

[41] Shixia Liu, Xiting Wang, Jianfei Chen, Jim Zhu, and Baining Guo. Topicpanorama: A full picture of relevant topics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 183–192. IEEE, 2014.

[42] Arun S Maiya and Robert M Rolfe. Topic similarity networks: visual analytics for large document sets. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 364–372. IEEE, 2014.

[43] Sana Malik, Alison Smith, Timothy Hawes, Panagis Papadatos, Jianyu Li, Cody Dunne, and Ben 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.

[44] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

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

[46] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[47] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

[49] Christopher E Moody. Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*, 2016.

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

[51] David Newman, Jey Han Lau, Karl Grieser, and Timothy 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*, pages 100–108. Association for Computational Linguistics, 2010.

[52] William Liam O'Brien. Preliminary investigation of the use of sankey diagrams to enhance building performance simulation-supported design. In *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design*, page 15. Society for Computer Simulation International, 2012.

[53] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[54] Jessica Peter, Steve Szigeti, Ana Jofre, and Sara 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.

[55] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[56] Ankita Rane and Anand Kumar. Sentiment classification system of twitter data for us airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 769–773. IEEE, 2018.

[57] Eugenia Ha Rim Rho, Gloria Mark, and Melissa 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.

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

[59] Carson Sievert and Kenneth 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.

[60] Alison Smith, Timothy Hawes, and Meredith Myers. Hiearchie: Visualization for hierarchical topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 71–78, 2014.

[61] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah 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.

[62] Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, and Leah 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.

[63] Alison Smith, Sana Malik, and Ben Shneiderman. Visual analysis of topical evolution in unstructured text: Design and evaluation of topicflow. In *Applications of Social Media and Social Network Analysis*, pages 159–175. Springer, 2015.

[64] Matt Taddy. On estimation and selection for topic models. In *Artificial Intelligence and Statistics*, pages 1184–1193, 2012.

[65] Silvia Terragni, Elisabetta Fersini, and Enza Messina. Word embedding-based topic similarity measures. In *International Conference on Applications of Natural Language to Information Systems*, pages 33–45. Springer, 2021.

[66] Silvia Terragni, Debora Nozza, Elisabetta Fersini, and Messina Enza. Which matters most? comparing the impact of concept and document relationships in topic models. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 32–40, 2020.

[67] Jun Wang, Changsheng Zhao, Junfu Xiang, and Kanji Uchino. Interactive topic model with enhanced interpretability. In *IUI Workshops*, 2019.

[68] Xi Wang, Anjie Fang, Iadh Ounis, and Craig Macdonald. Evaluating similarity metrics for latent twitter topics. In *European Conference on Information Retrieval*, pages 787–794. Springer, 2019.

[69] Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[70] Linzi Xing, Michael J Paul, and Giuseppe Carenini. Evaluating topic quality with posterior variability. *arXiv preprint arXiv:1909.03524*, 2019.

[71] Shuo Xu, Lijun Zhu, Xiaodong Qiao, Qingwei Shi, and Jie Gui. Topic linkages between papers and patents. In *Proceedings of the 4th International Conference on Advanced Science and Technology*, pages 176–183, 2012.

[72] Yi Yang, Doug Downey, and Jordan Boyd-Graber. Efficient methods for incorporating knowledge into topic models. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 308–317, 2015.

[73] Yi Yang, Quanming Yao, and Huamin Qu. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics*, 1(1):40–47, 2017.

[74] Guopeng Yin and Jian Chen. Improving causal inference with text as data in empirical is research: A machine learning approach. In *The 48th International Conference on Information Systems*, 2020.

[75] Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L Alkhouja. Mr. lda: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st international conference on World Wide Web*, pages 879–888. ACM, 2012.

[76] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. Constrained lda for grouping product features in opinion mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 448–459. Springer, 2011.

[77] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.