NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

NETWORK TD-SOM, USING SELF-ORGANIZING MAPS AND NETWORK ANALYSIS TO MAKE SENSE OF LARGE COLLECTIONS OF DOCUMENTS: THE CASE OF NOVA IMS MASTER'S THESES

Venâncio Tobias António Munhangane

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NETWORK TD-SOM, USING SELF-ORGANIZING MAPS AND NETWORK ANALYSIS TO MAKE SENSE OF LARGE COLLECTIONS OF DOCUMENTS: THE CASE OF NOVA IMS MASTER'S THESES

by

Venâncio Tobias António Munhangane

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialisation in Data Science

Supervisor: Fernando José Ferreira Lucas Bação

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I fully acknowledge the Rules of Conduct and Code of Honour from the NOVA Information Management School.

Lisboa, 17/11/2022

DEDICATION

FAMÍLIA, AMIGOS E TODOS OUTROS QUE SEMPRE ESTIVERAM DO MEU LADO.

ACKNOWLEDGEMENTS

Agradeço à DEUS pela força de vontade, pela saúde, coragem para poder realizar mais um sonho.

Agradeço a toda minha família que sempre me apoiou nos momentos mais difíceis e acreditou que eu fosse completar esta jornada com sucesso. Em especial para Regina, David, Camal e Pedro pelo especial apoio durante este trabalho.

Agradeço ao dedicado Professor Doutor Fernando José Ferreira Lucas Bação que orientou este trabalho, acreditou na minha capacidade e pela oportunidade de tê-lo como supervisor. O meu MUITO OBRIGADO. Grato pela paciência, pelo seu ensinamento e pelos seminarios promovidos com todos os seus alunos.

Agradeço à todos os meus professores pelo aprendizado durante o curso, em especial Joao Pedro Martins Ribeiro da Fonseca, Carina Isabel Andrade Albuquerque, David Fontes Henriques Silvestre da Silva, Davide Montali.

Agradeço à todos funcionários da NOVA IMS, em especial para Elvira Costa e Isabel Pinto;

Agradeço aos meus dois grandes companheiros e amigos da MozHouse, Geraldo Timbe (Paizao) e Allan Nhapulo (Danone) pelo apoio, amizade, ensinamentos e brincadeiras durante o curso e a maravilhosa temporada na MozHouse.

Agradeço à Manuel Carreiras (o leão) grande amigo, colega e companheiro; pela amizade, paciência, incentivos e pelos ensinamentos durante o curso e os grandes momentos que passamos juntos.

Agradeço a todos os meus colegas do curso em especial a Davide Farinati, Hiromi Nakashima, Nadine Aldesouky, Mariana Domingues, Lynder Swale, Catarina Candeias, Rita Ferreira, Catarina Urbano.

Agradeço ao Professor doutor Onofre Simões o meu MUITO OBRIGADO pelos concelhos dados quando a minha chegada.

Agradeço à todos meus amigos por ter a vossa amizade;

Agradeço à Marie Beyer pelo apoio, incentivo nos difíceis momentos durante este trabalho;

Agradeço ao Tony Mangue pelo seu apoio neste trabalho;

Agradeço à todos que directamente ao indirectamente contribuíram para esta conquista.

Kanimanbo, Obrigado a todos, Thanks to all!

ABSTRACT

Digital libraries are a central technology for the dissemination and sharing of knowledge, endless quantities of documents are stored and accessed through them. However, the efficiency of the associated search systems and their ability to identify relevant documents continues to be a bottleneck, and are not keeping pace with the ever-increasing volume of stored data. In this thesis, we present Network TD-SOM, a systematic process that offers a practical method for organizing, searching, visualising, discovering, and extracting knowledge from a vast corpus. Network TD-SOM combines topic modelling with Self-Organizing Maps and Network Analysis algorithms, to provide a visually rich environment where the user can explore and interact with a corpus, and find relevant documents. We test two different topic modelling algorithms separately and use their topic vectors to produce a Self-Organizing Map, which in turn is simplified through the use of a hierarchical clustering algorithm. We apply Network Analysis to the documents using the 3 best topics of each document and visualise the relations between the different documents. Finally, the Network TD-SOM methodology is evaluated on the master's thesis dataset from NOVA IMS. LDA and BERTopic successfully uncovered the thematic structure and extracted helpful knowledge from the dataset. In this context, BERTopic achieves better results and provides a more meaningful clustering solution. On the contrary, when it comes to the network analysis, and although the arrangement of the two network theses had similarities, the one modelled by using features/topics from LDA presents better results.

KEYWORDS

Corpus; Visualisation; Topic modelling; Clustering; Network analysis.

Sustainable Development Goals (SGD):

Index

List of Figures

List of figures (Appendix)

List of tables

List of tables (appendix)

LIST OF ABBREVIATIONS AND ACRONYMS

BMUs	Best match units;
BOW	Bag of words;
c-TF-IDF	Class-based Term Frequency-Inverse Document Frequency;
id2word	Dictionary of words;
GA	Genetic Algorithms;
GIS	Geographic information system;
GP	Genetic Programming;
GSGP	Geometric Semantic Genetic Programming;
GT	Geospatial Technologies;
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise;
K-NN	K-nearest neighbours algorithm;
LDA	Latent Dirichlet Allocation;
LSA	Latent Semantic Analysis;
LSI	Latent Semantic Indexing;
MapIntel	Intelligence Map;
ML	Machine Learning;
MI	Marketing Intelligence;
NLP	Natural Language Processing;
NMF	Non-Negative Matrix Factorization;
NOVA IMS	NOVA Information Management School;
QE	Quantization errors;
RENATES	Registo Nacional de Teses e Dissertações;
SOM	Self-organized maps;
SBERT	Sentence Bidirectional Encoder Representations from Transformers;
t-SNE	t-distributed stochastic neighbour embedding;
TF-IDF	Term Frequency-Inverse Document Frequency;
TE	Topographic error;
TDA	Topological data analysis;
UMAP	Uniform Manifold Approximation and Projection;
UTAUT	Unified theory of acceptance and use of technology**.**

1. INTRODUCTION

The digital library concept was initially discussed in the 1960s when Vannevar Bush and J.C.R Licklider wrote/spoke about the importance of having innovative technologies and approaches toward knowledge sharing as fundamental instruments for progress (Candela et al., 2015). However, the development activity on digital libraries only arose in the early 1990s, and as a result, the manner of accessing document collections started to change (Caillou et al., 2020; Lafia et al., 2019). The access systems have been gradually moving from physical to digital spaces. Today, library shelves with thematically collocated material are largely accessed through digital repositories organised in faceted categories (Lafia et al., 2019). This movement has increased the accessibility, quantity, and intensity of the exchange of information worldwide. However, the efficiency of document search systems and their ability to identify the desired and related information are not keeping pace with the ever-increasing volume of stored information.

The above-mentioned efficiency is limited compared to the libraries, which are naturally built to accommodate related documents on the same bookshelves (Lafia et al., 2019). Fabrikant (2000), offers a good illustration of this: "Typically, a library user looking for plays by Shakespeare will find the items of interest in the English Literature section, where other books by English authors and/or English plays will be placed on nearby shelves". In contrast, receiving an endless set of unwanted results is expected while using digital repositories to query for information or a specific document.

Based on this inefficiency, it has become critical to bring out new ideas for more practical methods of searching, visualising, organising, discovering related documents, and extracting knowledge from a vast collection of documents. Fabrikant (2000), says that the user's interaction with a vast data archive may be more efficient if some sort of graphical display is provided while searching for information. Therefore, a combination of spatialisation and machine learning (ML) can offer practical ways to improve the above-mentioned issues. Spatialisation combines powerful visualisation techniques to create a lower-dimension representation of higher-dimensional data sets (Fabrikant, 2000). At the same time, ML uses dimensional reduction techniques to compress the representation of document collections, preserving the meaningful arrangements of the material (Caillou et al., 2020).

We take on the challenge of combining spatialisation and ML to improve the points stated previously. We propose a Network TD-SOM, which means Network and Self-organised maps of the topic documents. The approach sets a systematic process to extract helpful knowledge and represent the relatedness of documents in a corpus. Network TD-SOM combines topic modelling algorithms, cluster algorithms and network analysis. The topic modelling algorithms have a significant role in this combined methodology. They are utilized to extract helpful knowledge from a corpus and to work as feature extractors for cluster algorithms and network analysis. Two different algorithms are selected and implemented separately for that. The first is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which uses a bag-of-words representation. The order of words in a document is neglected, and the context of the words in a sentence is not considered. The second is BERTopic (Grootendorst, 2022), which generates contextual document embeddings with pre-trained transformer-base language

models, and clusters these embeddings. Ultimately, it generates topic representations with the class-based Term Frequency-Inverse Document Frequency (c-TF-IDF)[1] procedure.

For document clustering, we selected the two-level hybrid clustering approach. First, we use a large SOM (Kohonen, 1990) to organize the documents; secondly, we use a hierarchical clustering algorithm, using the Ward method (Ward, 1963), to simplify the SOM structure, and get a summary of the overall clustering structure. Network analysis (Barabási, 2016), can reveal the direct connections between the documents and their centrality measures. Network TD-SOM has a final output of a two-dimensional web-based interactive network visualisation created using Gephi[2]. In this visualisation, the main results of the techniques stated previously are embedded. Besides these results, the helpful information of the metadata and useful text statistic of the documents can also be embedded. It shows the interlinkages between documents, the document clusters, and the subjects covered in each document. Additionally, it supports exploration, browsing, navigation as well as zoom.

We use the master's thesis dataset from NOVA Information Management School (NOVA IMS) as an opportunity to evaluate the performance of Network TD-SOM. We exploit the results to design various kinds of spatialisation dedicated to different purposes. For example, the topical distribution of each course/specialisation quickly uncovers the thematic structure of the master's thesis dataset, reveals the relationships between the various master's programs and allows for the analysis of the evolution of topical distribution over time. The topical distribution within each cluster and the numerical distribution of the masters's theses from each course/specialisation within each cluster are two spatialisations that can facilitate profiling the clusters. Together they reveal the relationships between the various master's programs. The network of the master's theses shows the direct connection between the theses. The remainder of this thesis is organised as follows. Section 2 presents an outline of previous related works. Section 3 provides a full explanation of the Network TD-SOM. Section 4 presents the theoretical background of the techniques used in Network TD-SOM. Section 5 presents the research methodology applied. Section 6 presents the results of the Network TD-SOM implementation on the master's thesis dataset from NOVA IMS. Lastly, section 7 summarises the conclusions and section 8 provides the reader with the limitation of the research and recommendations for future works.

[1] c-TF-IDF is a modification of TF-IDF, instead of calculating a significance of a word for a set of documents as it does in the usual TF-IDF process the c-TF-IDF calculates the significance of a word for each cluster of documents (Grootendorst, 2022).

[2] Gephi is the leading visualisation and exploration software for all kinds of graphs and networks. Gephi is open-source and free. (Available at https://gephi.org/ , accessed at 12/03/2022).

2. RELATED WORK

This section addresses the related works on visualising and exploring a vast collection of documents.

According to Zavrel (1996), the idea of having two-dimensional maps of the structure of document collections constructed without human intervention was seen as quite unrealistic before Teuvo Kohonen came up with SOM. The special abilities to perform unsupervised clustering and preserve topological relations of the input space have made several authors study SOM for analysing, clustering and browsing document collections. Honkela et al. (1996), created a WEBSOM method. It is the most notable early method for visualising similar documents on a map which can be utilised for exploring a vast collection of documents. This method uses SOM to organise the documents. Before the organisation, they must be encoded. A document is encoded as a histogram of its words, and the order of the words is neglected. The potential of the WEBSOM method was demonstrated in a case study where 4600 full-text articles written in English were automatically organised. The articles were positioned on a two-dimensional, 15 by 21 nodes large map where related articles appeared close to each other. Additionally, a web-based map was developed to aid in exploring the collection of documents and to allow the user to zoom in by clicking on any map area.

Kohonen et al. (2000), presented an improved version of the WEBSOM, called WEBSOM2. This version was used to demonstrate the scalability of the SOM method in dealing with large text collections. In total, 6 840 568 patent abstracts written in English were successfully organised onto a map composed of 1 002 240 neurons. Five hundred (500) dimensional document vectors were used as the feature. These vectors were obtained by reducing the dimensionality of 43222 weighted word histograms using the random projection method (Kaski, 1998). Additionally, the web-based map was improved, providing a search box and keyword search tools to help locate documents of interest. Furthermore, the zoom level was increased.

Ampazis & Perantonis (2004), proposed the LSISOM, which resembles WEBSOM. Both methods utilise a SOM to cluster the documents. However, one significant difference with WEBSOM is that the LSISOM trains the map with word category histograms obtained by applying latent semantic indexing (LSI) representation of the document terms (Deerwester et al., 1990), which enhances the identification of their semantic aspects. The method was successfully tested by clustering 420 articles but had the drawback of requiring enormous computational resources.

In recent years, much progress has been made in natural language processing (NLP). Researchers have built models for understanding human language's context and ambiguity better. Some of these models have been improving the way of visualising and extracting helpful knowledge in document collections. Some of these improvements can be found in Lafia et al. (2019). They designed two kinds of web-based interactive spatialisation to visualise 1731 master's and doctoral theses from a multidisciplinary university library. Before designing them, they applied LDA on the title and abstract of the theses to reduce dimensionality and extract features. Seventy-one (71) topics were identified in the final solution. After that, each thesis was embedded in an unlabelled vector space of 71 dimensions. These vectors were used as features for SOM and network analysis (Barabási, 2016), to design the two spatialisations. SOM designed the spatialisation of the topic's fields, revealing their similarities and the research topics' distribution and prevalence. Additionally, the theses that were related in terms of

topics were grouped. Network analysis (Barabási, 2016) designed the network of the master's and doctoral theses, revealing their connections.

Caillou et al. (2020), presented Cartolabe which is a web-based multiscale system that combines ML and visualisation techniques. The system's primary goal is to provide an effective and interactive two-dimensional representation that allows visualising, exploring and getting a quick understanding of a large document collection. The two-dimensional representation allows the users to pan, zoom, search, filter, and explore the neighbourhoods of the selected instance or region. Latent Semantic Analysis (LSA) (Deerwester et al., 1990), LDA and K-nearest neighbours algorithm (k-NN) (Keller & Gray, 1985) algorithms are the main techniques used on the ML side and UMAP is the main technique used on the visualisation side. LSA and LDA are used to compress the representation of each document in a large document collection finding the vector of topics that preserves the meaningful content of each document. After, UMAP uses these vectors to project the documents into a two-dimensional map focusing on preserving close neighbourhood relationships. Last, K-NN is used to compute the "k" nearest neighbours of each instance. The performance, versatility and scalability of Cartolabe were evaluated in three use cases separately and each case had good and trustworthy results. The first use case was on a corpus of scientific publications from the hall repository containing about 700K documents. The second was on the Wikipedia dataset which contains about 4.5M documents and the last one was on the *Grand Débat*[3] dataset, which contains about 4M documents. On the hall dataset, Cartolabe mapped the documents and the authors in the same space. Documents with similar topics were grouped in the same regions of the map. Authors writing documents in the same field were close to each other and also close to their topical regions. The documents were coloured in blue while the authors were in red but the most important documents and authors have white labels. The map is divided into thematic regions and the labels of each thematic region are coloured in yellow. Additionally, Caillou et al. (2020), made available a filter combo box to choose visualising articles and authors based on different options. Cartolabe allows also visualising the best ten (10) neighbours of the selected document or author.

Based on the above last two (2) works, other methods than SOM can be used to visualise the relatedness of documents in a corpus. Besides the possibility of using different ways to visualise document collections, it has been usual to manually or automatically assign labels and different colours to each topical cluster. This approach helps to have a good overview and knowledge of the corpus without reading all the documents. One good example can be found in Lafia et al. (2021). They presented a method that produces an interactive map that captures the latent topics and their evolution over time in a vast collection of documents. Additionally, it reveals the similarities between the topics and the documents. The topics are found by choosing the best solution between LDA (Blei et al., 2003), and non-negative matrix factorisation (NMF) (Lee & Seung, 1999). After that, they are labelled to quickly uncover the thematic structure of the vast document collection and are used as features for other techniques. The first three terms occurring in each topic are used to label the topics. To reveal the similarities between the documents, they are projected onto a map through the t-distributed stochastic neighbour embedding (t-SNE) (Van Der Maaten & Hinton, 2008), or Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). The documents with similar

[3] The *Grand Débat* dataset includes the citizens' fulltext contributions in response to the French government questions about societal and political issues (Caillou et al., 2020).

topics form regions of topical clusters, and a colour code is assigned to each topical cluster. This method was successfully evaluated by using 3,770 research documents from bibliometric data. In the end, a web application dashboard with two topic maps was deployed. One was coarse with nine (9) topics, and the other was detailed with 36 topics.

The above last three (3) works are good examples to illustrate some of the progress made in the field of NLP. Comparing their approach to visualising the document collections with WEBSOM, WEBSOM2 and LSISOM, it is possible to conclude that new models allow for extracting lower-dimensional features with better quality. Additionally, these features can be used to summarise and extract knowledge from a large collection of documents. However, in the past few years, one of the most outstanding achievements has been building models capable of understanding a text's semantics or context. These models are considered state-of-the-art and have improved the results of several NLP tasks. Silva & Bacao, (2022), utilised these models. They presented a MapIntel derived from Intelligence Map. It is a system for acquiring intelligence from vast document collections. This system produces an interactive two-dimension map that gives a fast overview of the corpus. The documents are projected onto a map through the UMAP (McInnes et al., 2018), in a manner that semantically related documents form topical clusters. Each topical cluster has one colour code and is automatically and semantically labelled. The topical clusters and the respective labels are obtained using BERTopic (Grootendorst, 2022), which unveils the semantic structure of the data. For that, Sentence Bidirectional Encoder Representations from Transformers (SBERT) (Reimers & Gurevych, 2019), encodes the semantic attributes of each document as a multidimensional vector. MapIntel has a search functionality where a user can type a query string and retrieve the semantically closest documents. The map allows the user to hover, browse documents, and zoom in on particular regions of the semantic space. MapIntel was evaluated on the 20 newsgroups dataset and got astonishing results.

We intend to present a method capable of organising and visualising related documents and knowledge discovery in a vast collection of documents. We based our approach on some studies presented above (Lafia et al., 2019, 2021; Caillou et al., 2020; Silva & Bacao, 2022). The idea is to use LDA to discover the hidden topics and to analyse the topical distribution throughout the years. SOM (Kohonen, 1990) and network analysis are used to organise and discover the related documents. Additionally, we apply BERTopic (Grootendorst, 2022) with the aim to improve the LDA results (Blei et al., 2003). Finally, in our approach, the idea is to gather the main results of each technique and helpful metadata information in one visualisation.

3. PROPOSED METHOD

We propose a Network TD-SOM, which means Network and Self-organised maps of the topic documents. The approach adopts a systematic process (Figure 3.1) to cluster and represent the relatedness of the documents in a corpus. Additionally, it allows for the extraction of helpful knowledge or uncovering the thematic structure in the corpus without needing to read all the documents. Network TD-SOM combines topic modelling algorithms, cluster algorithms and network analysis. Each of the techniques utilised in this project has a specific purpose to complete. The topic modelling algorithms have a significant role in this combined process. They are effective in extracting helpful knowledge from a corpus and they work as feature extractors for cluster algorithms and network analysis. Two different algorithms are selected and implemented separately. The first is LDA (Blei et al., 2003), which is the most popular topic modelling technique. In this algorithm, the order of words in a document is neglected and the context of words in a sentence is not considered. The second is BERTopic (Grootendorst, 2022), the state-of-the-art method for topic modelling. BERTopic generates contextual document embeddings with pre-trained transformer-based language models, clusters these embeddings, and generates topic representations with the c-TF-IDF procedure. The techniques' performance is evaluated by the coherence metrics and by analysing the matching of the topic's distribution with some variables in the metadata through several visualisations. Additionally, each topic modelling algorithm performance working as a feature extractor is evaluated by analysing the quality of the results from the cluster algorithms and network analysis.

The cluster algorithms are implemented to discover some relatedness between documents and how to group them to make it easy to examine other closely related documents. At the same time, the Network analysis (Barabási, 2016), is implemented to reveal the direct connections between the documents and their centrality measures. Network TD-SOM has a final output, a two-dimensional web-based interactive network visualisation of the corpus created using Gephi. It supports exploration, browsing and navigation as well as zoom. The visualisation allows for the incorporation of the main results of topic modelling algorithms, clustering algorithms, and network analysis. It shows the subjects covered in each document, the clusters and the interlinkages between documents. Besides that, the helpful metadata information and text statistics of the documents can be embedded in the visualisation. The Network TD-SOM is composed of 8 main steps: document collections, text descriptive and pre-processing, transformation, topic model training and evaluation, topic model interpretation, clustering documents, Network analysis and visualisation. The objective of each step is explained below.

Document collections

Extract documents from a repository:
- Content (Full text or abstract);
- Important metadata.

Text descriptive

Calculate text statistics

BERTopic pre-processing

Tokenization and lower case; Remove the irrelevant Characters; the stop and short words.

LDA pre-processing

BERTopic pre-processing;
Train the bigram and trigram terms;
Lemmatization.

BERTopic Transformation

Use the pre-processing data to Compute the documents embeddings;

LDA Transformation

Use the pre-processing data to compute the id2word and the BOW (TF-IDF);

Training model and evaluation (BERTopic)

Use the documents embeddings to Train several BERTopic models by performing a grid search on the parameters. Calculate the coherence score for each model.

Evaluation: Select the 5 BERTopic model with highest coherence score; Inspect the top 10 to 25 words of each topic; Revisit the pre-processing if the top words of all models are uninterpretable.

Training model and evaluation (LDA)

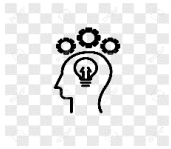
Use the id2word and BOW to: Train several LDA models by performing a grid search on the parameters. Calculate the coherence score for each model.

Evaluation: Select the 5 LDA model with highest coherence score; Inspect the top 10 to 25 words of each topic; Revisit the pre-processing if the top words of all models are uninterpretable.

Topic Interpretation (BERTopic/LDA model)

Use the best model and:
- Visualise the topics through dendrogram, pyLDAvis, etc.
- Label the topics using the top words;
- Analyse the document topic distributions
- Calculate topic trends over time;
- Cross the topics distribution with important metadata variables.
- Produce several visualisations to analyse the distribution and the consistency of the topics;
- If the topics are not consistent back to the pre-processing to re-run the analysis.

Knowledge Discovery in the corpus



Clustering (BERTopic/LDA model)

Use the topic vectors as a feature for the two-level hybrid clustering: At 1st level train several SOM by performing a grid search on the parameters. Calculate the quantization errors (QE) for each SOM.

Evaluation: Select the map with lowest quantization errors (QE);

At 2nd level apply ward method on the BMU of the selected map: Plot the dendrogram to choose the cluster number.

Cluster results (BERTopic/LDA model)

Cluster label of each document and do Cluster profile.

Network Analysis (BERTopic/LDA model)

To centrality measures: calculate the weight by applying the dot product between vectors using all the topics in the vector;

To model the network, use the best 3 topics of each document and calculate the weight by applying the dot product between vectors.

Network visualisation (BERTopic/LDA model)

After model the network use Gephi or other software to create the network visualisation.
Embedded the main results into the network visualisation:
- Embedded the important variables of the metadata;
- Embedded the important text descriptive;
- Embedded the topic distributions of each document;
- Embedded the cluster label of each document.

Figure 3.1 – Network TD-SOM workflow composed of eight main steps.

Step 1: Document Collections
This step aims to get a set of documents and their metadata. They can be extracted from a repository such as libraries, Scopus and others.

Step 2: Text descriptive statistics and pre-processing
The descriptive text statistics of the documents are calculated before the pre-processing. The objective is to calculate various descriptive text statistics and embed them in the visualisation. Since the idea is not to embed the text of the documents on the visualisation, it will incorporate some statistics, which can be helpful while a user is exploring the visualisation. After calculating some text statistics, pre-processing is the next fundamental step for further analysis. The goal is to prepare the vast collection of documents and keep the content that will be helpful for the topic modelling algorithms. Additionally, some visualisations can be provided at this stage. The pre-processing is implemented according to each topic modelling algorithm chosen.

Step 3: Transformation
At this stage, the goal is to prepare the pre-processing data for the topic modelling algorithms. Each algorithm has its way. For BERTopic, the document embeddings are computed. In LDA, the dictionary of words (id2word) and the bag of words (BOW) are computed. Additionally, it is recommended to remove high and low-frequency words.

Step 4: Training model and evaluation
The transformed data are used to train several models to find the parameters' values, which optimizes the models' performance. Grid search is used to perform the hyperparameters tuning of the topic modelling algorithms. During this step, it is crucial to analyse the latent topics from the models because they can have similar topics, and they should be merged. On the other hand, if uninterpretable topics remain, it may be necessary to go back to the pre-processing phase and repeat. At the end of this step, it is crucial to have meaningful and interpretable topics to facilitate a better understanding of the m.

Step 5: Topic Interpretation
The topic interpretation step is closely related to the step above. However, the main goal is to have a more fine-grained understanding of the topics in the context of the domain under study. The possible ways to achieve that are described below:

- Label the topics using the top words on each topic;
- Analysing how the topics are related to other topics;
- Analysing the topical distribution over time.

Additionally, crossing the topics with some variables of the metadata and plotting some visualisations can help to ensure the consistency of the topics and to have a better understanding of the topics. At this step, it is also normal to identify some flaws or aspects to improve. Like in the previous step, in such a case, it would be wise to revisit the pre-processing or try different model parameters and re-run the analyses.

Step 6: Clustering

The Network TD-SOM uses the hybrid clustering algorithms approach with two levels of clustering. According to Cheu et al. (2004), this approach is efficient. It can hit a higher percentage of samples correctly clustered than one algorithm alone, because the clustering at the first level provides data abstraction, reducing the number of samples in the following level. Additionally, the first-level outputs are local data averages and consequently are less sensitive to random variations than the original data. At the first and second levels, the chosen techniques are SOM (Kohonen, 1990), and hierarchical clustering algorithms using Ward's method (Ward, 1963) respectively. The topic vectors of each topic modelling are used as features. To profile the clusters, the average topic weight within each cluster is calculated. Additionally, the clusters and some metadata variables are crossed, and some visualisations are plotted.

Step 7: Network Analysis

After clustering the documents, the topic vectors are used to model the document interlinkages through network analysis (Barabási, 2016). In Network TD-SOM we measure the significance of each cluster. We do that by calculating the average of some centrality measures within each cluster and plotting some visualisations. This approach also gives additional information to better profile the clusters.

Step 8: Visualisation

One of the primary goals of the Network TD-SOM is to produce an interactive visualisation where it is possible to visualise the corpus and incorporate the main results. To achieve this goal, we propose a two-dimensional web-based interactive network visualisation created by using Gephi. In this visualisation, the main results of the above steps are embedded. Furthermore, besides the results, helpful metadata information can also be embedded. Additionally, the corpus network can organise the documents approximating the original topology. It shows the interlinkages between documents, the document clusters, and the topic distribution in each document. The information about the topic distribution can be helpful for the users to understand the subjects covered in each document without a need to read it. The document clusters are colour-coded with the same colour described in step 6. This visualisation allows the user to browse the data and zoom in on a particular region of the network. Additionally, when a document/node is selected, it visualises all the connections of the selected document/node and all its embedded information in a list format.

4. THEORETICAL BACKGROUND

This section presents concepts that are fundamental to this research. These concepts are associated with topic modelling algorithms, clustering algorithms and Network analysis.

4.1. TOPIC EXTRACTION

According to Vayansky & Kumar (2020), during the data analysis, it is essential to determine the characteristics that data points share. In text analysis, this often means choosing a collection of subject matters discussed in the documents. The collection of subject matters may be evident to a human reading the documents, but it is not for a program as it will process only the text and not the collection of subject matters outlined in the documents. To identify the collection of subject matters in the documents, researchers utilise a method called topic modelling.

Topic modelling (TM) is an emerging statistical machine-learning algorithm that can help identify hidden topics and detect common topics in a collection of documents (Choirul Rahmadan et al., 2020). This algorithm has been gaining many exciting new directions and can be used to organise, search, automatically summarise and extract features in a large unstructured collection of documents (Choirul Rahmadan et al., 2020). According to Culmer & Uhlmann (2021), TM has two broad approaches. The first is unsupervised which means that the topics are not known beforehand. Additionally, the models are trained to discover the topics. The second approach is supervised with predetermined topics. This method uses labelled data to train a model that will allow an unseen document to be classified as belonging to a particular topic or multiple topics. In this project, the unsupervised approach was implemented. That is to say, the dataset used to evaluate the performance of Network TD-SOM does not have predetermined topics.

According to Vayansky & Kumar (2020), numerous methods of topic modelling considering many kinds of relationships and restrictions within datasets have been developed. As stated in the previous section, LDA and BERTopic are the topic modelling methods implemented in Network TD-SOM. Table 4.1 presents the reasons behind the selection of these methods. Next is presented the main concepts of each method.

Table 4.1 – Reasons to select LDA and BERTopic on Network TD-SOM (Adapted from Blei et al., 2003; Grootendorst, 2022; Vayansky & Kumar, 2020).

LDA	**BERTopic**
LDA is the most frequently utilised and well-suited method for general topic modelling tasks.	BERTopic is the state-of-the-art method for topic modelling and generates coherent topics considering the context of the words in a sentence.
It has been used for many purposes to understand unstructured text data.	BERTopic can learn coherent patterns of language.
Additionally, it is flexible, and its results are used as a benchmark for all other methods in publications.	BERTopic can generate competitive results to be compared with LDA results.

4.1.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet allocation (LDA) is a well-known topic modelling algorithm introduced by Blei et al. (2003). It is an unsupervised statistical model of document collections that tries to capture topics across the documents (Blei, 2012). The basic idea is that documents are represented as random mixtures over latent topics, where a distribution over words characterises each topic (Blei et al., 2003). The term latent means hidden while Dirichlet refers to the distribution that allows the assignment of topics within documents and words of the documents to topics (Blei et al., 2003; Eletter et al., 2022). Chehal et al. (2021), say that LDA is a good topic modelling and can be used for user reviews classification in recommender systems as well as dimensionality reduction algorithms for document collections. Chapter 2 presents some works where LDA was successfully used for this proposal. LDA identifies important keywords and reduces any document to a fixed set of real-valued features.

Blei (2012), presents a good example to understand the intuition behind LDA which is to find words with a high probability to occur in each topic (Figure 4.1). The article used by Blei (2012), entitled "Seeking Life's Bare (Genetic) Necessities," is about using data analysis to determine the number of genes that an organism needs to survive from an evolutionary perspective. The article blends topics about genetics, data analysis, and evolutionary biology. LDA was used to find words that are related to each of these topics. Words highlighted with the same colour belong to the same topic. It is possible to see that words about data analysis, such as "computer" and "prediction," are highlighted in blue; words about evolutionary biologies, such as "life" and "organism," are highlighted in pink; words about genetics, such as "gene", and "dna" are highlighted in yellow. Knowing the topics within the article helps to uncover the thematic structure of the document.

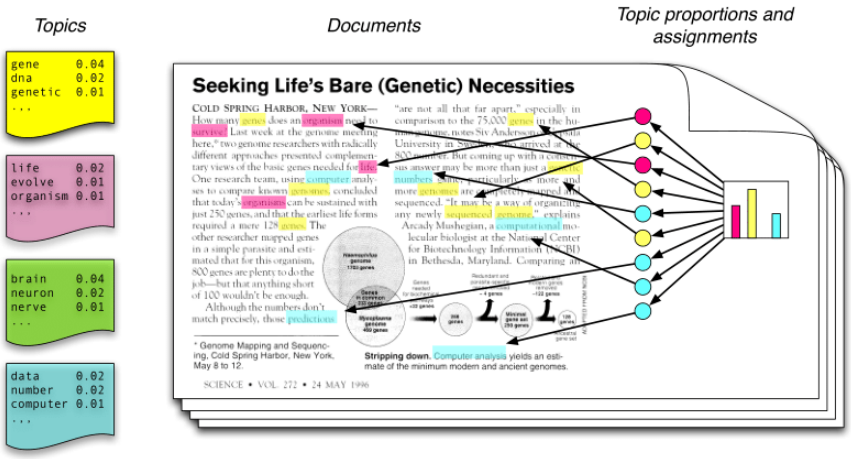


Figure 4.1: Graphical representation of the intuitions behind latent Dirichlet allocation (Taken from Blei, 2012). We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First, choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the coloured coins) and choose the word from the corresponding topic (Blei, 2012).

The LDA process assigns topics to the documents and words to the topics and calculates the respective probabilities based on the following steps. Before applying LDA, the researcher has to exclude words with little topical content such as “and”, “but”, or “the”. He/she also needs to pre-decide the number of topics within the document collection. By assigning random topics to each word of each document, LDA is ultimately assigning random topics to each document. Then, LDA will optimize the random choices of the previous step. LDA applies equation 4.1 and iterates over all documents and all words to find the most important words in each topic. At the end of the process, the researcher can assign a name to each topic based on its most important words.

$$p(w, z\ /\alpha, \beta) = p(w\ /z, \beta) * p(z\ /\alpha) \quad (4.1)$$

Where:

- α: document-topic density;
- z: the topic assigned to each word w;
- β: topic-word density;
- w: word in a topic;
- $p(w\ /z, \beta)$: the probability that word w occurs in topic z;
- $p(z\ /\alpha)$: the probability that topic z is found in document d.

According to Culmer & Uhlmann (2021), α and β are hyperparameters and must be set based on the data being processed. Correct values of α and β are necessary to produce a good model. The α term represents document-topic density. Large values correspond to more topics per document and conversely, smaller values correspond to fewer topics per document. The β term represents topic-word density. Large values correspond to more words per topic and conversely, smaller values correspond to fewer words per topic.

LDA requires two main inputs to be implemented: the dictionary of words (id2word) and the bag of words (BOW). The dictionary consists of giving an index for each token in the corpus while the bag of words consists of a corpus with an index of each word and the respective frequency. The use of the bag of words neglects the order of the words in the document and their context in the sentence (Blei et al., 2003; Chehal et al., 2021; Culmer & Uhlmann, 2021). It is recommended to test several topic numbers and select the number that can best describe the entire document collection.

4.1.2. BERTopic

According to Grootendorst (2022), BERTopic is a topic modelling method that uses semantic document embeddings, cluster techniques and c-TF-IDF to generate coherent document topic representations of the corpus. BERTopic generates topic representations through three main steps: document embeddings, document clustering, and topic representation. Figure 4.2 shows the visual overview of these three main steps. Each step is described in detail in the following paragraphs.

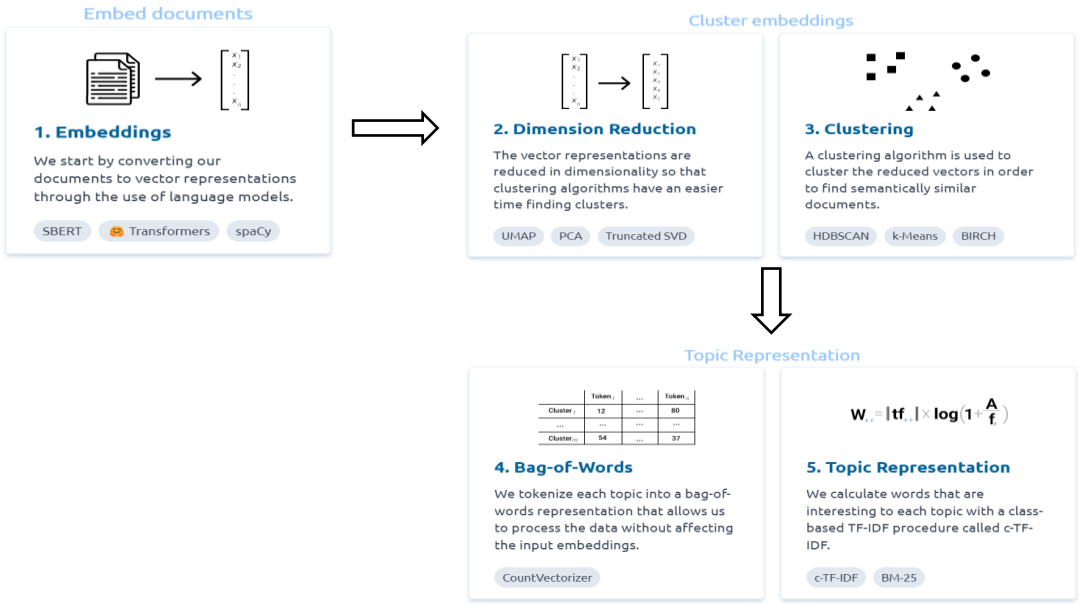


Figure 4.2: Visual overview of the BERTopic three main steps, namely the embedding of documents, the clustering of documents and the topic representation. Taken from https://maartengr.github.io/BERTopic/algorithm/algorithm.html.

According to Grootendorst (2022), the first step is to produce the document embeddings, where the main goal is to convert the documents to numerical representations. The documents are embedded in a vector space that can be compared semantically. BERTopic uses the SBERT framework to perform the embedding, but other frameworks, such as spaCy,[4] can be used. These frameworks use a pre-trained language model to achieve a good performance on the embedding tasks which will increase the quality of clustering. Many pre-trained languages are available, and some work for more than 50 languages. Due to that, it is crucial to select one that fits your use case.

Clustering the documents is the second step in BERTopic (Grootendorst 2022), the main goal is clustering the embeddings to allow extracting the topic representations. The embeddings produced in the previous step are reduced and not directly used to cluster semantically similar documents because clustering algorithms handle poorly high dimensionality. BERTopic, by default, reduces dimensionality through UMAP using the cosine distance. UMAP has the advantage of preserving more of the local and global features, but other techniques that may fit your use case can be used. After reducing the embeddings, they are clustered through Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2015) which uses a soft clustering approach and does not require knowing the number of clusters a priori. However, any cluster model that may fit your use case can be used.

The objective of the final step is to know what makes each cluster different to the others (Grootendorst 2022). Firstly, for each cluster, a topic is assigned. Secondly, in each topic, the most important words are found. BERTopic does that by converting each cluster to a single document and applying c-TF-IDF (equation 4.2) in each document to have the importance score per word. The most important words

[4] Spacy is a free, open-source library for advanced NLP in python used to build information extraction or natural language understanding systems or to pre-process large volumes of text for deep learning (Available at https://spacy.io/usage/spacy-101, accessed at 12/03/2022).

in each topic will be used for a topic representation. Additionally, BERTopic can automatically assign names to the topics or the researcher can manually assign them.

$$w_{x,c} = \left\| tf_{x,c} \right\| \times \log\left(1 + \frac{A}{f_x}\right) \quad (4.2)$$

Where:

$tf_{x,c}$: Frequency of word x in class/topic c;

f_x: Frequency of word x across all the classes/topics;

A: Average number of words per class.

4.2. CLUSTER ALGORITHMS

Clustering algorithms are a well-known class of unsupervised techniques (Xu & Tian 2015), used to group instances based on their similarities. These algorithms use similarity or dissimilarity measures to create groups/clusters of instances. The instances recorded in the same group/cluster must be similar to each other as much as possible, while instances of different groups/clusters must be different to each other as much as possible. The primary goal of clustering is to discover a set of patterns, points, instances or objects from the natural grouping (Oyelade et al., 2019). The standard process of clustering can be divided into the following steps (Xu & Tian, 2015):

1) **Feature extraction and selection**: Extract and select the most representative features from the original data set. In this thesis, this step was done through the topic modelling algorithm. The details are given in section 6.2.
2) **Clustering algorithm design**: Design the clustering algorithm according to the characteristics of the problem. In this thesis, this step was done through the use of SOM combine with hierarchical clustering, in a two-step clustering procedure. The details are given in step 6 in section 3.
3) **Result evaluation**: Evaluate the clustering result. In this thesis, this step was done with quantization error (QE) and several visualisations. The details are given in section 5.3.2.
4) **Result explanation**: Give a practical explanation for the clustering result. The details are given in section 6.3.

Many different clustering algorithms are available today (Cheu et al., 2004). Each has its strengths, and weaknesses and is suitable for different types of data and data distributions (Xu & Tian, 2015). As stated in the previous section, Network TD-SOM uses a hybrid clustering algorithms approach with two levels of clustering. SOM is the technique applied at the first level and it was selected for the reasons described in Table 4.2.

Table 4.2 – Reasons to select SOM on Network TD-SOM (Adapted from Asan & Ercan, 2012; Zhang & Fang, 2012).

It does not make assumptions regarding the distributions of variables and nor does it require independence among variables;
The SOM spatialisations can display complex interdependencies of the variables, and facilitate the visualization of similar relationships of the input data and their clustering tendency;
It is easy to implement and can solve nonlinear problems of very high complexity. Additionally, it is easy to combine SOM with other methods. This approach can facilitate the downstream clustering analysis and improve the cluster performance;
It more effectively copes with noisy and missing data, very small dimensionality, and samples of unlimited size.

At the same time, the hierarchical clustering algorithm using Ward's method is applied at the second level to simplify the SOM structure and to get a summary of the overall clustering structure. It was selected for the reasons described in Table 4.3.

Table 4.3 – Reasons to select hierarchical clustering algorithm (Ward's method) on Network TD-SOM. (Adapted from Ward, 1963).

It is a clustering method based on a classical sum-of-squares criterion, producing groups that minimise within-group dispersion.
It allows for the construction of the dendrogram, which helps to select the best number of clusters to be chosen. Additionally, the dendrogram illustrates the mergers or divisions made at successive levels.

4.2.1. SOM

Between 1981-82, Teuvo Kohonen came up with an automatic data-analysis method/technique denominated self-organizing maps (SOM). SOM is an unsupervised machine learning algorithm that belongs to a group of techniques known as artificial neural networks (ANNs). It is used as an exploratory data analysis tool for visualising and clustering high-dimensional data (Asan & Ercan, 2012). According to Bação et al. (2004), the SOM's basic idea is to map high-dimensional input data onto a usually two-dimensional rectangular or hexagonal (Figure 4.2) grid of units/neurons trying to preserve topological relations. This means that closer patterns in the input space will be mapped to closer units/neurons in the output space and vice-versa. Each instance of the data is an input layer or a vector with weights or coefficients.

According to Asan & Ercan (2012), SOM applies a competitive learning rule where the output nodes compete among themselves for the opportunity to represent distinct patterns within the input space. The SOM training process can be summarized in three steps. The first one is the competition where the goal is to find the Best matching unit (BMU) for each input. The second is cooperation where the BMU and its neighbours activate each other to learn something from the same input. The last step is adaptation where the BMU and its neighbours become similar to a specific input pattern and will have more chance of responding to a similar input pattern. Bação et al. (2004), present a basic algorithm for the SOM training process:

i. Calculate the distance between the pattern and all the units of the SOM ($d_{ij} = ||x_k - w_{ij}||$);
ii. Select the nearest unit as the winner (w_{ij}:$mim(d_{ij})$);
iii. Update each unit of the SOM according to the update function $w_{ij} = w_{ij} + \alpha h$;
iv. Repeat steps (i) to (iii), and update the learning parameters, until a stopping criterion is met.

Where:
w_{ij}: The weight vector associated with the unit positioned at column i and row j;
x_k: The vector associated with pattern k;
d_{ij}: The distance between the weight vector (w_{ij}) and a given pattern;
h: The neighbourhood function assumes values in [0,1] and is a function of the position of the two units (winner unit and another unit), and radius. It is large for units that are close to the output space and small for units far away;
α: The learning rate varies in the [0,1] interval and must converge to 0.

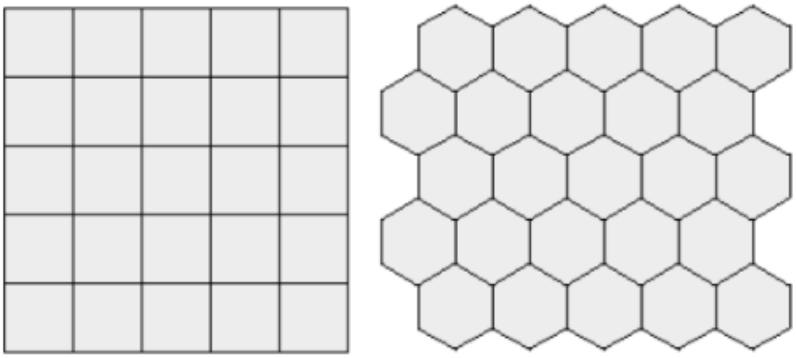


Figure 4.3: Rectangular (left image) and Hexagonal (right image) grid topology (Taken from Schmidt et al., 2011). Each internal node on the rectangular and hexagonal grid has four (4) and six (6) neighbours respectively. A hexagonal grid is preferred since it displays greater variance in the neighbourhood Asan & Ercan (2012).

4.2.2. Hierarchical clustering (Ward's method)

According to Ward (1963), Ward's method is an agglomerative hierarchical clustering algorithm in which the distance between two clusters is the sum of squares between the two clusters summed over all variables. The grouping process starts with n members, termed groups or subsets, although they may contain only one member. The first step in the grouping is to select two of these n subsets, which, when united, will reduce by one the number of subsets while producing the least impairment of the optimal value of the objective function. The n-1 resulting subsets are examined to determine if a third member should be united with the first pair or if another pairing may be made to secure the optimal value of the objective function for n-2 groups. This procedure continues until all n members of the original array are united in one group.

4.3. NETWORK ANALYSIS

Network analysis is an analytical technique evolving from graph theory used in various fields (Gobov & Yanchuk, 2022). Graph theory is a collection of nodes (or vertices) and weighted or unweighted edges (or links) that connect the nodes (Bali Swain & Ranganathan, 2021). The edges of a network can be directed or undirected which will mark the network to be classified as directed or undirected - or as a mix of both if the network simultaneously has directed or undirected links. Network analysis was implemented in this project because it allows us to analyse and view the direct connections between documents through a clean visualisation. Additionally, the use of network statistics, called centrality measures, enables us to weigh the importance of any node (i.e. document) in the network.

Network TD-SOM uses a few common centrality measures namely the degree centrality, the eigen centrality, the betweenness and the closeness centrality. The first two are considered neighbour-based measures and the last two are considered path-based ones. The main concepts of each of the centrality measures used in this thesis are described below (Bali Swain & Ranganathan, 2021; Hansen et al., 2020):

- **Degree centrality** is a simple count of the total number of connections linked to a node. It can be thought of as a kind of popularity measure but does not recognize a difference between quantity and quality.

- **Eigen centrality** is more sophisticated than degree centrality. It recognizes the difference between quantity and quality. A node with few connections can have a very high eigenvector centrality when is very well connected or connected to important nodes.
- **Closeness centrality** takes a different perspective from the above-mentioned network metrics. It captures the average distance between a node and every other node in the network. A node with high closeness centrality is, on average, closer to all the nodes of the network than a node with low closeness centrality.

5. RESEARCH METHODOLOGY

This section describes the evaluation process and experimental procedure of Network TD-SOM. The evaluation process and experimental procedure of the topic modelling algorithms, the two-level hybrid clustering and the network analysis are provided. Additionally, a description of the experimental data, some text statistics to be calculated, the pre-processing pipeline and the software implementation are also provided.

5.1. EXPERIMENTAL DATA

The dataset used in the analysis consists of master's theses of all courses/specialisations that are/were lectured in the master programs of NOVA IMS. The dataset was provided by the NOVA IMS library services. However, it is also available on the web page of RENATES[5] (*Registo Nacional de Teses e Dissertações*). The dataset contains 1352 research theses from 2011 to 2021. It includes 24 variables, but only 14 were retained for this project. Table 5.1 describes them.

Table 5.1 – Description of the variables.

Variables	Description
Thesis Id	The ID of the thesis
Author Name	Name of the student who wrote the thesis
Author email	Email of the student who wrote the thesis
Advisor	Name (s) of the teacher (s) who supervised the thesis
Issued date	The date that the student defends the thesis
Rights info	Information if the thesis is open, restricted or embargoed access
Embargo info	Reasons for the thesis being restricted or embargoed
Embargo date	Date when the thesis is not going to be restricted or embargoed
Thesis title	Title of the thesis
Thesis type	If the thesis is a dissertation, work project and internship report
Thesis language	Language in which the thesis was written
Thesis URL	Link to access the thesis
Thesis course/Specialisation	Course/Specialisation taken by the student
Thesis abstract	The abstract of the thesis

The abstract is one of the essential variables. It was used to achieve this project's main goal instead of the full text. The reasons that justify this decision are:

- Firstly, according to Glasman-Deal (2020), the abstract includes all relevant information while being concise, clear, and coherent. It enhances the study's visibility by highlighting its core and application without further interest in technical details.
- Secondly, the abstract is less computationally expensive than the full text.
- Lastly, it should be noted that 8.7% (118) of the theses are not accessible because of confidentiality clauses related to their content. However, respective abstracts are available and can be used to identify the topics depicted. The topics can be used to identify similar theses that are accessible.

[5] The RENATES platform collects official information on doctoral theses and master's dissertations carried out in Portugal. It also includes records of doctoral theses in progress of theses carried out abroad and recognized in Portugal. https://renates2.dgeec.mec.pt/.

5.2. TEXT DESCRIPTIVE STATISTICS AND PRE-PROCESSING

Pre-processing the abstracts is a fundamental step and was done using *spaCy*. Before that, the abstracts' length and flesh reading score[6] were calculated using TextDescriptives[7], and their histograms were analysed. For the pre-processing, the pipeline below with five steps was built. Additionally, a word cloud of the corpus was analysed.

i. **Train the bigram and trigram terms in the corpus**: The goal was to learn the token combinations. Words that co-occur together frequently were merged into a single new token/word. For bigram, all words that co-occur together at least forty (40) times in the corpus were merged. For trigram, all words and bigrams that co-occur together at least twenty (20) in times the corpus were merged. The gensim[8] was used in this process. This process was undertaken only for the LDA algorithm and not for BERTopic.

ii. **Tokenisation and lowercase:** The text of each abstract was broken into tokens/words after training the bigram and trigram. Each token/word was converted to lowercase to avoid having different representations for the same word (Ex: Data= data) after breaking the text.

iii. **Removing stopwords and short words:** The words such as articles ("a" and "the") without predictive capability (stopwords), were removed. This process was done through *spaCy* using "en_core_web_sm[9]". Additionally, words with less than three characters were also removed.

iv. **Lemmatisation:** The remaining tokens/words were converted into their root form. To avoid having different word representations for ones with similar meanings. This process was undertaken only for the LDA algorithm and not for BERTopic.

v. **Irrelevant Characters (Numbers, punctuation, extra spaces, etc.):** At this stage, all characters that are not considered a word, such as delimiters, white space, numbers, URLs, HTML tags, etc., were also removed.

5.3. EVALUATION MEASURES

The metrics to evaluate the results depend on the project goal and the chosen techniques. In this project, different group techniques were used. The metrics to assess the performance in each group are presented below.

5.3.1. Topic modelling evaluation

The metric CV was the coherence measure used to access the quality of the topics computed by the models trained with LDA and BERTopic. According to Röder et al. (2015), this metric strongly correlates to human topic coherence ratings when the topic is calculated with more than ten (10) words. It ranges from -1 to 1, and higher values indicate an excellent coherent topic description while lower values indicate incoherent topic descriptions. For each algorithm the model which optimizes the topic coherence was selected.

---

[6] Flesch reading score is one of the most widely use and valeted systems for scoring readability. Lower and negative scores indicate that the content of the text is more difficult to read than higher scores (Ceross & Bergmann, 2021). Flesch Reading Ease Score Interpretation: 100-70: Very easy to fairly easy; 69 – 60: Standard; 59 – 30: Fairly difficult to difficult; Bellow 30: Very confusing

[7] TextDescriptives is python library for calculating a large variety of statistics from text(s) using spacy pipeline components and extensions (Available at https://hlasse.github.io/TextDescriptives/, accessed at 12/03/2022).

[8] Gensim is a free open-source python library for representing documents as semantic vectors and is designed to process raw, unstructured digital texts using unsupervised machine learning algorithms. (Available at https://radimrehurek.com/gensim/intro.html#what-is-gensim, accessed at 12/03/2022).

[9] En_core_web_sm: Is a small English pipeline trained on written web text (blogs, news, comments), that includes vocabulary, syntax and entities. (Available at https://spacy.io/models, accessed at 12/03/2022).

5.3.2. Clustering evaluation

The evaluation of the two-level hybrid clustering was done by separately evaluating the results of each stage. SOM was used in the first stage of clustering. The quantisation errors (QE), U-matrix representation, component planes and the Multi-class hexbins charts were analysed to evaluate the performance of mapping the theses. According to Asan & Ercan (2012), QE shows how well the map fits the data and is computed by determining the average distance of the sample vectors to its best matching unit. Small values indicate a good approximation of the input distribution, and higher values indicate the opposite. Topographic error (TE) is another measure to examine the quality of SOM output. However, Asan & Ercan (2012), state that it may be unreliable sometimes, and because of that, it was not taken into consideration in the evaluation. However, the values were calculated and plotted with QE values to help analyse the distribution of QE. In the second stage of clustering, to choose the correct number of clusters and to evaluate the quality of the cluster solution the dendrogram and the SOM map were used respectively. For cluster profiling, the average topic weight within each cluster and the numerical distribution of the master thesis from each course/specialisation within each cluster were calculated and analysed through radar charts.

5.3.3. Network evaluation

The network was not modelled with the entire vector of topics. The network evaluation was based on the network density and the average percentage of all topical coherence explained by selecting a certain amount of the best topics in the thesis. The idea is to use only the important topics in each thesis and to aim at having a network with a good balance between the density and the ability to explain the interlinkages between the master's theses adequately.

5.4. EXPERIMENTAL PROCEDURE

A variety of hyperparameter values were tested to find the ones that optimize the use of each technique. The experimental procedure applied to each technique is described below:

5.4.1. Latent Dirichlet Allocation (LDA)

First, the dictionary of words (id2word) and the bag of words (BOW) were built. The id2word was built using the trigram-trained sentence of the abstracts after the pre-processing. At this stage, tokens/words that were very rare (occurred in less than five (5) theses) or too common (occurred in more than 70% of the theses) were filtered to help find different topics. For BOW, the Term Frequency-Inverse Document Frequency (TF-IDF) was computed to provide better insight. Second, the hyperparameter tuning of the model was performed to find the parameter values that optimize the model's performance. Grid search was used for this goal. In the last step, the twenty-five (25) most important words in each topic were retained for the topic representation. Table 5.2 provides the grid of parameters that were evaluated.

Table 5.2 – Grid of parameters used to train LDA models.

Parameters	Values	Observation
Number of Topics	[4-22]	Step topics =2
Document-topic density (α)	[0.01-1]	Step α=0.02
Word-Topic Density (β)	[0.01-1]	Step β=0.02

5.4.2. BERTopic

The embeddings of the abstracts master's theses were generated through SBERT. The msmarco-distilbert-base-v4[10] was the pre-trained language model to transform each abstract onto a fixed-length embedding vector. After the document embeddings several values for UMAP and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2015), were tested. The hyperparameter tuning of the model was performed to find the parameter values that optimize the model's performance. Grid search was used for this goal. In the last step, the twenty-five (25) most important words in each topic were retained for the topic representation. Table 5.3 provides the grid of parameters that were evaluated.

Table 5.3 – Grid of parameters used to train BERTopic models.

Parameters	Values	Observation
Number of Topics	[4, 22]	Step topics =2
n_neighbors	[5, 30]	Step n_neighbors =5
n_components	[5, 30]	Step n_components =5
min_cluster_size	[5, 30]	Step min_cluster =5
min_samples	[5, 30]	Step min_samples =5

5.4.3. SOM

The topic vectors/features from LDA and BERTopic were separately used as features to map the theses into a hexagonal grid of neurons/units. Several maps with random sizes and different initialisations were trained. Hyperparameter tuning was performed to find the map size and initialisation that optimize the performance of mapping the theses. Grid Search was used for hyperparameter tuning. Table 5.4 provides the grid of parameters that were evaluated.

Table 5.4 – Grid of parameters used to train SOM models.

Parameters	Values
Number of random maps	150 for each vector of topics
Range of possible random map size values	Lowest value map = 15; highest value map = 55
initialisation	PCA, random
lattice	Hexagonal
train_rough_len	400
train_finetune_len	400

5.4.4. Hierarchical clustering (Ward's method)

Ward was applied twice. First, on top of the BMUs found after training SOM with the topic vectors/features from LDA. Second, on top of the BMUs found after training SOM with the topic vectors/features from BERTopic. After that, each dendrogram was cut at a point that allow getting a solution with high inter-cluster heterogeneity and high intra-cluster homogeneity.

[10] msmarco-distilbert-base-v4 is a sentence-transformers model: It maps sentences & paragraphs to a 768-dimensional dense vector space and can be used for tasks like clustering or semantic search. (Available at https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4 , accessed at 12/03/2022).

5.4.5. Network Analysis

As stated previously, some centrality measures were calculated using network analysis. Additionally, a network that allows analysing and viewing a direct connection was modelled. For the centrality measures, the topic vectors of each topic modelling algorithm were separately used. First, the dot product between topic vectors was taken to calculate the weight. Secondly, two neighbour-based and two path-based measures described in section 4.3.1 were calculated. Finally, the average of these measures within each cluster was calculated and analysed through radar charts.

Two networks of the theses were separately modelled by using the topic vectors of each topic modelling algorithm and retaining the three best topics in each thesis. The node, the edges, the weight and the type of the network were defined as described above:

- The theses were treated as nodes;
- The edges were constructed based on the relationship between theses; That is to say, if two theses have at least one of the best three (3) topics in common, they share an edge;
- To calculate the weight of the edges a vector with the three best topics of each thesis was used and the dot product between them was taken. The weight ranges from 0 to 1, and higher values indicate substantial similarity between theses while lower values indicate weak similarity;
- The edges were considered bidirectional with symmetrical weight;
- Yifan hu proportional (Hu, 2011), was the layout used for the network arrangement.

5.4.6. Software implementation

The implementation of the experimental procedure is based on Python programming language and on R programming language. Genism, BERTopic, sompy and networkx were the main libraries used through python and hextri was the main library used through R. In addition, Gephi was used to design the corpus' two-dimensional web-based interactive network visualisation. The reported Network TD-SOM experiments and the results are reproducible using the scripts provided in the GitHub repository of the project.

6. RESULTS AND DISCUSSION

This section presents the exploratory data analysis and discusses the results of topic modelling algorithms. It also analysed which topic modelling algorithm works better as a feature extractor tool in the collection of unstructured documents analysing the cluster and network results.

6.1. EXPLORATORY DATA ANALYSIS

6.1.1. Theses descriptive statistics

The theses from NOVA IMS master programs can be considered as: a dissertation, an internship report or a work project. Figure 6.1 summarises the numbers of each type of master's theses across the last two decades. Overall, the three types of theses show clear upward trends in numbers. However, they marginally decreased in 2021 for dissertations and work projects. The dissertation has been the most preferred choice by students about to complete their master's degree. Although the Internship report is the least preferred choice, its number rose slightly to about forty (40) theses in 2021.

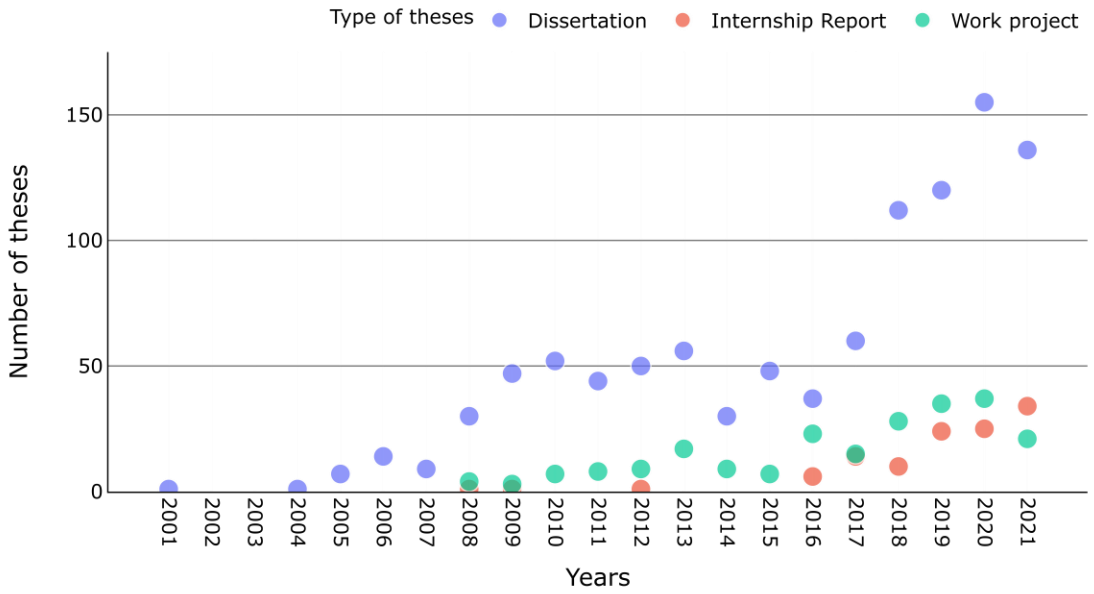


Figure 6.1 – Distribution of master's theses type per year.

The theses are distributed in twelve courses/specialisations. Geospatial technologies has the highest percentage of the theses (20.2%), and Knowledge Management and Business Intelligence has the second-largest percentage (15.4%). Clinical research management has the least (1.3%) (Table 1, Appendix). Figure 6.2 summarises the number of theses in four different courses/specialisations throughout the years. Overall, excluding Geographic information systems and science, the number of theses showed upward trends. Data Science and advanced analytics is the one which increased steadily throughout the years, including in 2021. Geographic information systems and science was the first to graduate students in 2004, while Data science and advanced analytics was the last in 2016. Additionally, knowledge management and business intelligence had the highest number of theses defended in a year (60) in 2020. Figure 1 and Figure 2 in the Appendix provide information about other courses/specialisations.

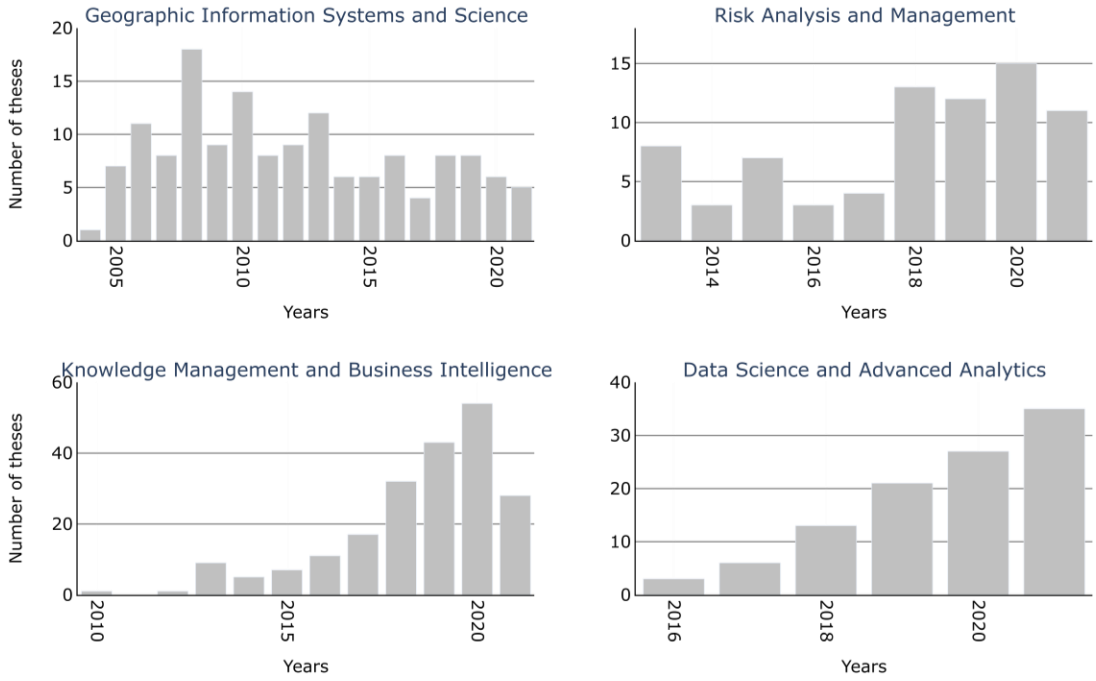


Figure 6.2 – Distribution of thesis by four courses/specialisations per year.

6.1.2. Abstract descriptive statistics

Figure 6.3 shows the distribution of the word numbers in the abstracts related to each thesis and the annual average of the word numbers. Overall, the number of words follows a normal distribution. Most abstracts have a total number of words within the expected range (around 80-250). Consequently, the average in most of the years is below 250.

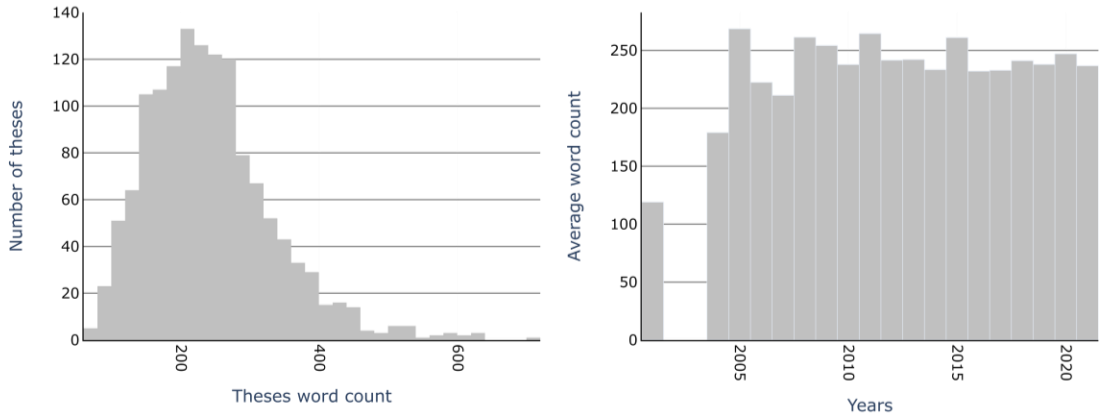


Figure 6.3 – Distribution of word count (on the left) and the annual average length of abstracts (on the right) from the theses.

Figure 6.4 shows the distribution of the Flesch reading ease score in the abstracts related to each thesis and the annual average of the score. Overall, the score follows a normal distribution. In most abstracts, it stands below 60, while the standard average in most years is around 40. Put together these results indicate that the abstract contents have a fairly difficult to complex level of readability.

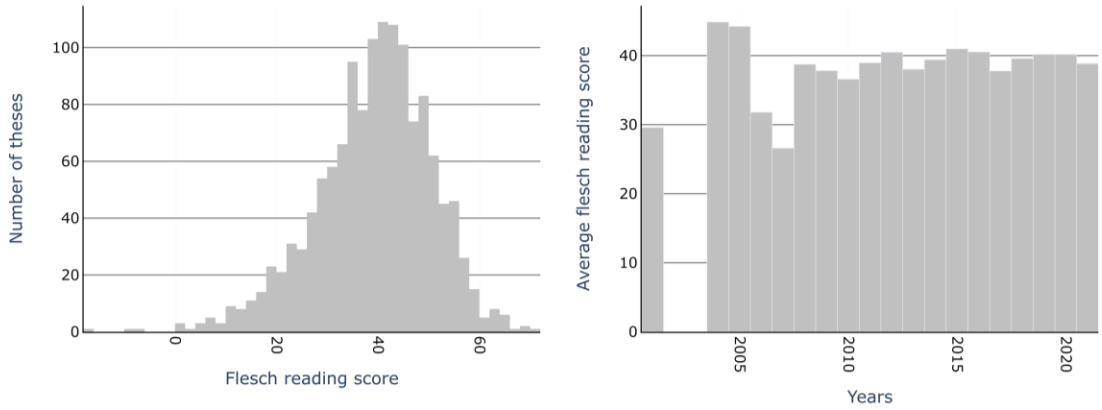


Figure 6.4 – Distribution of Flesch reading score (on the left) and annual score average of abstracts (on the right) from the theses.

Figure 6.5 illustrates the most popular 200 words in the abstracts after training the bigram and the trigram terms. Overall, “Model”, “data”, “result”, “area”, “analysis”, “system”, “based”, “project”, “information”, “study”, “land_cover”, “algorithm”, “management” and “application knowledge” are some examples of words that are more frequently mentioned in the abstracts of the master’s theses from NOVA IMS.

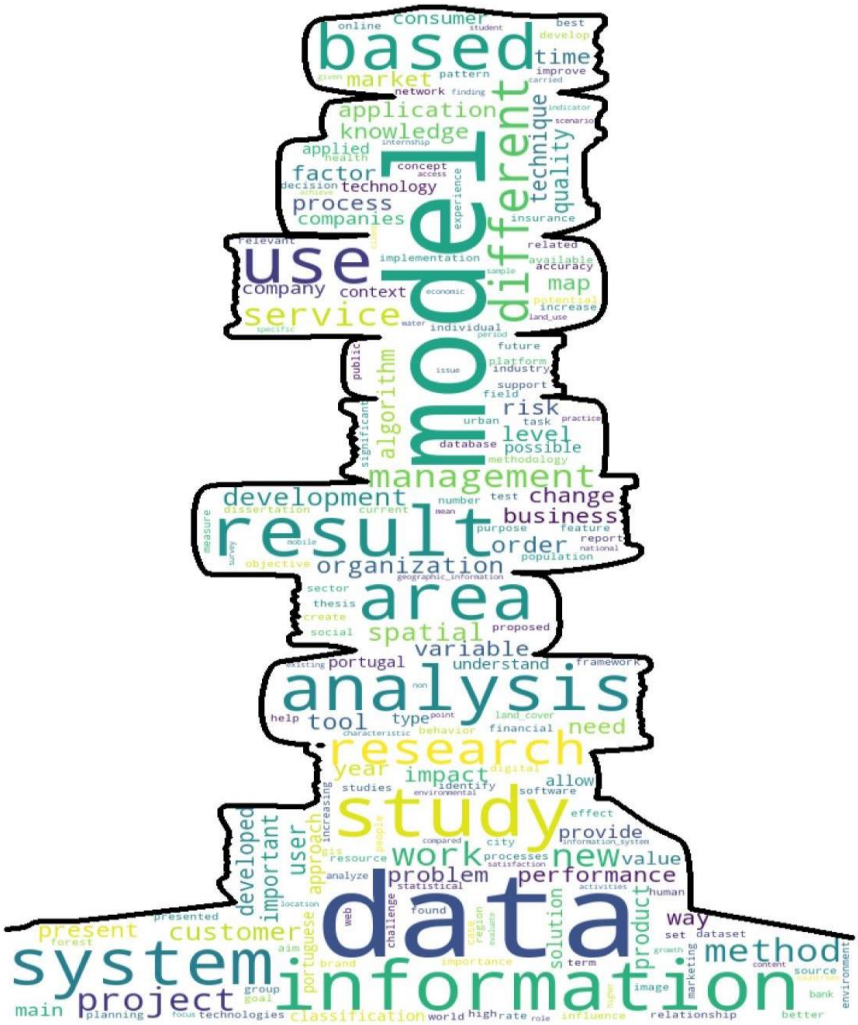


Figure 6.5 – Abstracts word cloud.

6.2. TOPIC MODELLING RESULTS

To find an optimal number of topics in each algorithm the experimental procedure described in section 4.4 was followed. Tables 6.1 and 6.2 show the best five values of the parameters that optimize the performance of the topic modelling algorithms. In total, 378 and 5625 models were trained using LDA and BERTopic, respectively. The best solution turned out to be the one with eight topics in both algorithms. BERTopic had higher values of topic coherence (Cv). The following two chapters analyse the topics found and their distribution by each course/specialisation for each algorithm.

Table 6.1 – Five best parameter values obtained from the hyperparameter tuning process on LDA.

Number of topics	Alpha	Beta	Topic Coherence (Cv)
8	**symmetric**	**0.81**	**0.490**
16	0.21	0.61	0.480
20	symmetric	0.81	0.477
8	0.41	symmetric	0.474
16	0.01	0.81	0.473

Table 6.2 – Five best parameter values obtained from the hyperparameter tuning process on BERTopic.

Number of topics	n_neighbors	n_components	min_cluster_size	min_samples	cluster_selection_method	Topic Coherence (Cv)
8	**20**	**10**	**10**	**15**	**eom**	**0.511**
8	10	5	5	5	eom	0.506
10	25	15	10	10	eom	0.502
18	25	5	10	15	eom	0.499
14	25	5	15	15	eom	0.499

6.2.1. LDA results

In topic modelling, the topics or subjects in a collection of documents are represented by a distribution of commonly occurring words. Figure 6.6 provides the top 25 terms for each topic. Overall, it is possible to see that some commonly occurring words are appearing in more than one topic which indicates some correlation between the topics. Words such as "datum", "study", "model", "research", and "system" are some examples of words repeated across the topics. Furthermore, the top 25 terms for each topic were used to assign a name to the topics as shown below:

- Topic1 was named "Risk/Bank/Insurance/Investments/Markets/Law" because of the presence of words such as "risk", "insurance", "bank", "legal", and "law".
- Topic3 was named "Health//Management/Education" because of the presence of words such as "health", "clinical", "master", "student", "management", and "coordination".
- Topic4 was named "Land_cover/Maps/Urbanism/Population/Environmental", which can be explained by words such as "land", "maps", "urban", "population", and "water".

The other topics were named using the above approach, and the name of each topic is as follows:

- Topic0: Self organising maps;
- Topic2: Business/Customers/Companies/Management/Technology;
- Topic5: Machine_learning_algorithms;
- Topic6: Customer_satisfaction_&_Behaviour/Marketing/Products/Brands;
- Topic7: GIS[11]/Spatial/Smart_cities/Maps/Technology.

[11] Geographic Information System

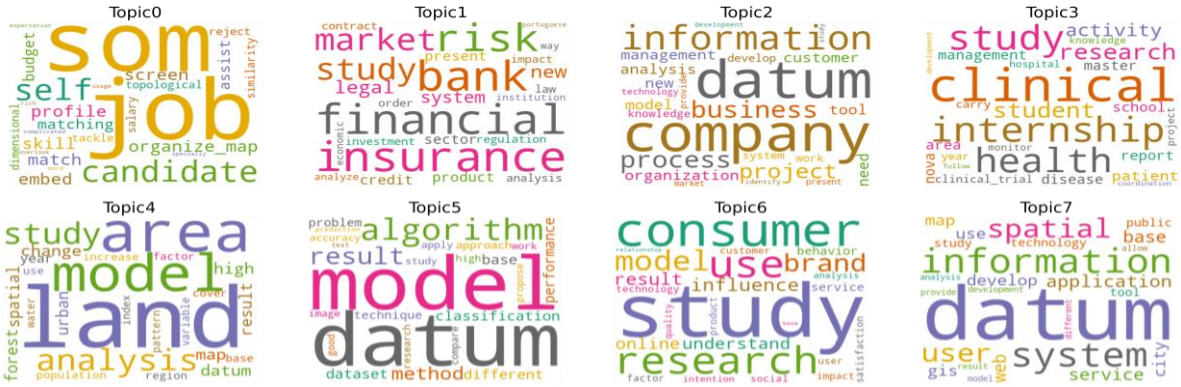


Figure 6.6 – Word cloud of the top 25 keywords in each topic (LDA).

Figure 6.7 allows for uncovering the thematic structure of the master's thesis dataset and shows the average prevalence of the topics for four different courses/specialisations in a given year. Overall, in each course/specialisation, there is a good match with the respective most dominant topic. This indicates that LDA has correctly discovered the latent topics in an unsupervised way. As an example, "GIS/Spatial/Smart_cities/Maps/Technology" and "Land_cover/Maps/Urbanism/Population/Environmental" are the dominant topics in Geographic Information Systems and Science but not in the other courses/specialisations. Another good example is "Machine_learning_algorithms", the dominant topic in "Data Science and Advanced Analytics". It also has a good prevalence in "Knowledge Management and business intelligence" and "Risk Analysis and Management", but it has less prevalence in "Geographic Information Systems and Science". Additionally, it is possible to see the evolution of the topics in each course/specialisation. One good example is "Knowledge management and business intelligence", which had very few topics initially. However, in recent years a considerable increase in topics in its theses can be identified. Figure 4 in the appendix provides information about other courses/specialisations. In each course/specialisation, there is good matching with the respective most dominant topic.

Figure 6.7 – Distribution of weight topics by courses/specialisations per year (LDA).

6.2.2. BERTopic results

BERTopic is state-of-the-art, and it can extract topics capturing the semantic meaning in a text. Although most of the topics found through BERTopic are similar to LDA, figure 6.8 shows that the top 25 words in each topic are more semantically coherent than the results from LDA. One good example is the reduction of repeated words across the topics. Another example is topic5 from LDA and topic6 from BERTopic, which are similar. However, the top words from BERTopic are more semantically coherent, making it easier to assign a name to the topic.

To assign the name of each topic the same approach used in LDA was applied to BERTopic. Each topic was named as follows:

- Topic0: Land_cover/GIS/Geospatial/Maps/Urbanism/Cities;
- Topic1: Business_Intelligence_&_Process/Knowledge_Management/Organizations;
- Topic2: Risk/Insurance/Banks/Markets/Financial/Law;
- Topic3: Customer_satisfaction_&_behaviour/Marketing/Brands/Products;
- Topic4: Predictive_marketing/ICT[12]/UTAUT[13]/Online_usage;
- Topic5: Machine_learning_algorithms/Data_mining/Customers/Insurance/Business/Marketing;
- Topic6: GA[14]/GP[15]/GSGP[16]/Machine_learning_algorithms;
- Topic7: Health/Management/Education.

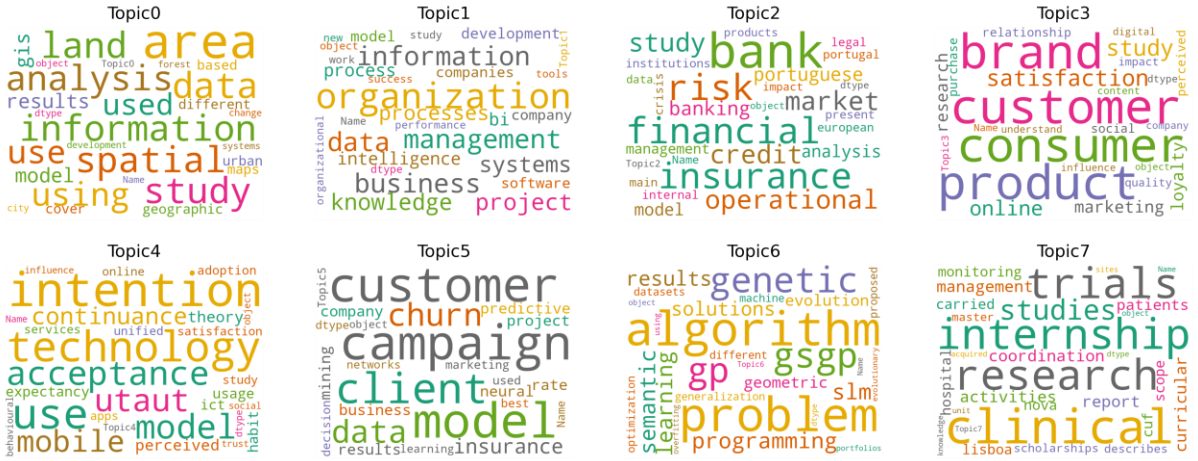


Figure 6.8 – Word cloud of the top 25 keywords in each topic (BERTopic).

Figure 6.9 shows the average prevalence of the topics for four different courses/specialisations in a given year. Overall, it is possible to see that the courses/specialisations and their dominant topics match. Additionally, the top 25 words are semantically linked with the names of the courses/specialisations where they are dominant. This indicates that BERTopic has correctly and semantically discovered the latent topics in an unsupervised way. Some good examples to illustrate this are:

[12] Information and communications technology.
[13] Unified theory of acceptance and use of technology.
[14] Genetic algorithm
[15] Genetic programming
[16] Geometric Semantic Genetic Programming

- Geographic Information System and Science now has one dominant topic ("Land_cover/GIS/Geospatial/Maps/Urbanism/Cities") throughout the period. It is different from the LDA results, which had two;
- All topics discovered using BERTopic have a significant average weight in at least one of the courses/specialisations, different from LDA, where the topic "Self organising maps" does not have a significant average weight in any of the courses/specialisations;
- Another good example is topic6 "GA/GP/GSGP/Machine_learning_algorithms", which is one of the dominant topics in "Data Science and Advanced Analytics". Additionally, the top 25 words are highly semantically linked with the course/specialisation.

Figure 5 in the Appendix provides information about other courses/specialisations. In each course/specialisation, there is good matching with the respective most dominant topic.

Figure 6.9 – Distribution of weight topics by courses/specialisations per year (BERTopic).

Unlike LDA, the BERTopic has some documents where the sum of the topic vector weight is less than 1 (Figure 6, Appendix). The values are even lower for documents considered outliers by the algorithm. In the present work, 341 observations were classified as outliers (Figure 7, Appendix). Based on the two points mentioned above, the stacked bars distribution of the topics in almost all the courses/specialisations does not reach 1 (Figure 6.9 and Figure 5 in the Appendix).

6.3. SOM RESULTS

A hybrid clustering approach was used to find the clusters of the theses. In the first stage, SOM was applied. In total, 300 random maps were trained: 150 maps using topic vectors/features from LDA and 150 maps using topic vectors/features from BERTopic. In general, the maps using the topic vectors/features from BERTopic have lower values of QE than those from LDA (Figure 8, Appendix). The component planes of the topic vectors/features from BERTopic also show a better distribution of the theses than those from LDA. In most of the features from BERTopic, the theses are concentrated

on one specific region of the map (Figure 9, Appendix). In contrast, in features from LDA, the theses tend to be dispersed all over the map (Figure 10, Appendix). Table 6.3 presents three (3) best parameter values that optimize the performance of mapping the theses using features from LDA and BERTopic. The solution with the lowest QE was selected for features from LDA and BERTopic.

Table 6.3 – Three best parameter values obtained from the hyperparameter tuning process on SOM.

Topic vector name	Lattice	Initialization	Map size	Quantization error
	hexa	**pca**	**(53, 46)**	**0.0224**
BERTopic vector	hexa	pca	(54, 48)	0.0228
	hexa	random	(53, 44)	0.0230
	hexa	**random**	**(54, 48)**	**0.0456**
LDA vector	hexa	random	(53, 46)	0.0468
	hexa	pca	(54, 48)	0.0499

6.3.1. SOM Results (LDA topic vectors)

Figure 6.10 represents the distribution of the BMU in the U-matrix (Figure 14, Appendix) and the average prevalence of the topics within each BMU of the best-trained SOM using the topic vectors from LDA. The figure provides useful information, and combining it with the LDA results allows to make sense of the SOM results quickly. It allows for building clusters and visualising distances between neurons and regions with substantial similarity. Additionally, it allows visualising the distribution of the topic weights and the theses quantity across the neurons. In this context, each colour area in the neurons corresponds to the prevalence of the respective topic colour code, and the size of neurons helps to indicate the theses quantity in the neurons.

The arrangement of the neurons in figure 6.10 corresponds to the component planes results (Figure 10, Appendix) and to the LDA results. Some examples of the valuable information that can be extracted from these figures are provided below:

- Neurons with a significant topic prevalence of "Land_cover/Maps/Urbanism/Population/Environmental" have many theses within them. The same happens for neurons with a significant topic prevalence of "Business/Customers/Companies/Management/Technology".
- Neurons located in the middle have overall relatedness. That is to say, the neurons have theses that are similar to the ones in the neurons surrounding them. These neurons have a good topic prevalence of "Machine_learning_algorithms" and "Business_Intelligence_&_Process/Knowledge_Management/Organizations". Additionally, these topics also pertain to most of the neurons, which indicates they exist in many theses.
- Contrarily, the neurons located in the down-right corner and up-left corner have less relatedness with the other ones. Each group of these neurons has a topic prevalence of "Health/Management/Education" and "Self organising maps" respectively. The topic "Health/Management/Education" does not exist in many theses and the topic "Self organising maps" does not have a significant average weight in any of the courses/specialisations (LDA results).
- Most of the neurons with a significant topic prevalence of "Land_cover/Maps/Urbanism/Population/Environmental" and

"GIS/Spatial/Smart_cities/Maps/Technology" are located in the same region. This indicates some relatedness between these neurons, the theses within them and, ultimately, the topics themselves. The same happens for neurons with a significant topic prevalence of "Business/Customers/Companies/Management/Technology" and "Customer_satisfaction_&_behaviour/Marketing/Products/Brands".

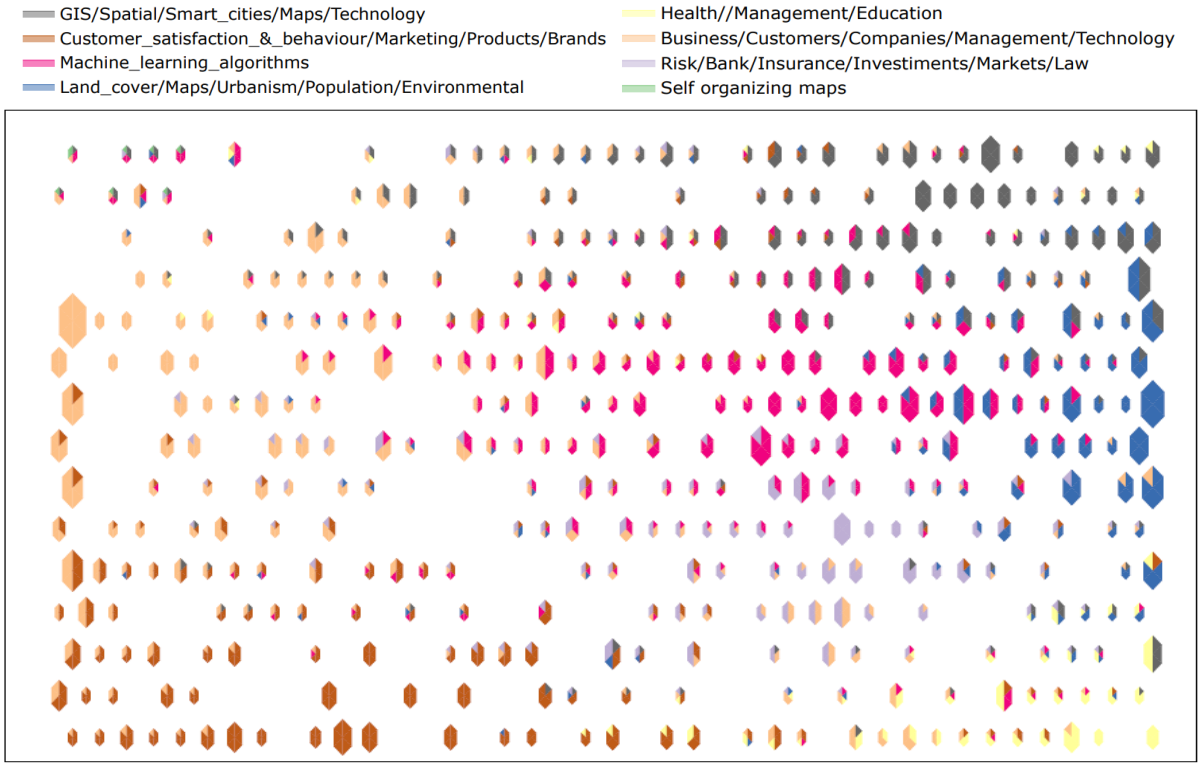


Figure 6.10 – Distribution of the BMUs displaying the prevalent topics weights and a visual indication of the quantity of theses they contain (LDA).

6.3.2. SOM Results (BERTopic topic vectors)

Figure 6.11 represents the distribution of the BMU in the U-matrix (Figure 15, Appendix) and the average prevalence of the topics within each BMU of the best-trained SOM using topic vectors from BERTopic. Figure 6.11 was analysed similarly to the one previously presented. They have some similarities, but the topic vectors from BERTopic leverage the SOM results. It is possible to see that neurons with unrelated topics are well separated, and the distance is more highlighted (Figure 6.11, Figure 9, Appendix and Figure 15, Appendix). Some examples of the valuable information that can be extracted from this figure are provided below:

- As previously presented in the results, neurons with a significant topic prevalence of "Health/Management/Education" are also isolated, but the distance to other neurons is highlighted in figure 6.11. Additionally, neurons with a significant topic prevalence of "GA/GP/GSGP/Machine_learning_algorithms", a specific data science topic, are also isolated.
- Most of the neurons in the middle have a good diversity of topics, indicating overall relatedness. These neurons have a significant topic prevalence of "Machine_learning_algorithms/Data_mining/Customers/Insurance/Business/Marketing" and "Customer_satisfaction_&_behaviour/Marketing/Brands/Products". These topics pertain to most of the neurons which indicates that they also pertain to many theses.

- Neurons with a topic prevalence of "Land_cover/GIS/Geospatial/Maps/Urbanism/Cities" have many theses. The same happens for neurons with a topic prevalence of "Customer_satisfaction_&_behaviour/Marketing/Brands/Products".

Figure 6.11 – Distribution of the BMUs displaying the prevalent topics weights and a visual indication of the quantity of theses they contain (BERTopic).

6.4. HIERARCHICAL CLUSTERING (WARD'S METHOD) RESULTS

The best match units (BMUs) found in the first stage, were used in the second stage. Ward method was applied to the BMUs. After analysing the dendrogram (Figure 11, Appendix), the solution of 6 and 5 clusters using the topic vectors/features from LDA and BERTopic were selected, respectively. Figure 6.12 shows how the clusters are organised using features from each topic modelling algorithm. The final cluster solution using features from LDA shows that most of the theses are well clustered. However, some theses from clusters 2 and 4 are not contiguous, indicating some type of twist or bend in the manifold. In the case of BERTopic, all clusters are contiguously leading to the conclusion that there are no major bends in the manifold and thus we can assume that it provides a better clustering solution. The lack of bends or twists in the arrangement of clusters enables the detection of the relatedness between them, meaning neurons located in the same region/nearby are more related. That is to say, the neurons have similar theses or theses with some similar topics, and neurons which are located far away are unrelated. Clusters 3 and 0 (Figure 6.12: SOM from LDA topic vectors) are an example of related clusters, and clusters 1 and 3 (Figure 6.12: SOM from LDA topic vectors) are unrelated.

To sum it up, the cluster located in the middle/centre of the other clusters has overall relatedness. That is to say, the cluster has theses that are similar to the theses from the clusters surrounding it, and, in other words, it has theses with a good diversity of topics. At the same time, the clusters located on the edge of the map do not present overall relatedness to the other clusters. They have theses that are different to theses from the other clusters. In other words, the clusters have theses with specific topics. Clusters 0 and 4 (Figure 6.12: SOM from BERTopic topic vectors) are good examples of clusters with and without overall relatedness, respectively.

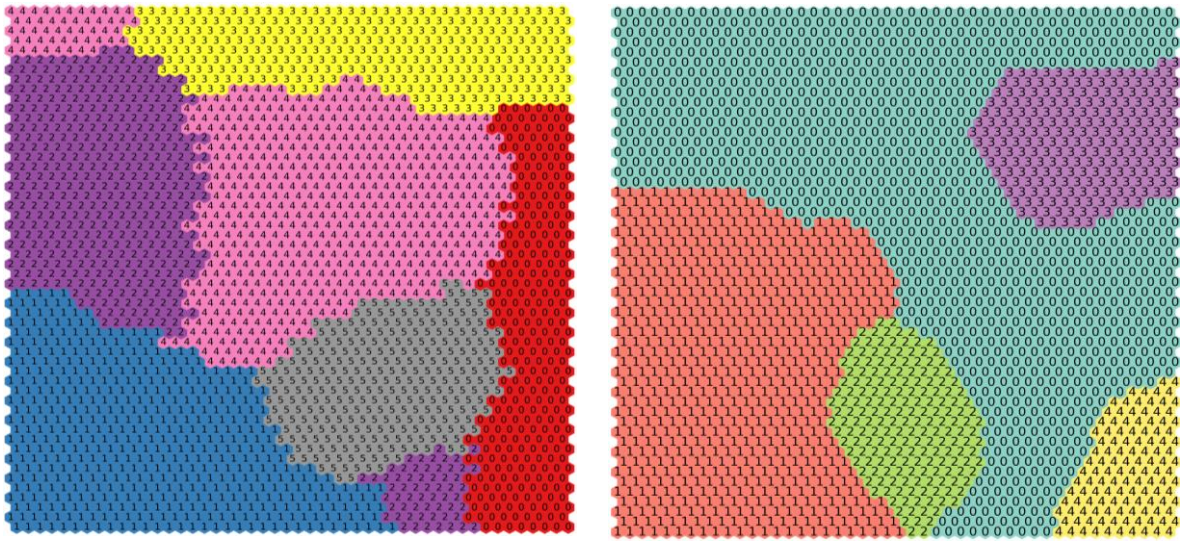


Figure 6.12 – SOM from LDA (left) and BERTopic (right) topic vectors.

6.4.1. Hierarchical cluster results (LDA topic vectors)

Each cluster has more than 100 theses (Figure 12, Appendix). Figures 6.13 and 6.14 represent the average weight of topics and the number of theses from each course/specialisation in each cluster. Combining these figures allows us to discover related topics, clusters, courses/specialisations with related theses and to support the results from SOM from LDA features (Figure 6.12, left image). Additionally, it provides a mechanism to organise the theses so that the researchers who are interested in a particular area of research can easily see related theses by examining those in the cluster.

Topics with a higher weight average in the same cluster can be considered related. Furthermore, suppose one cluster has a higher average weight in one or more topics, it indicates that these respective topics characterise this cluster. Additionally, suppose the cluster has a higher number of theses from one or more courses/specialisations. In that case, the cluster is characterised by the theses from these respective courses/specialisations. Theses in the same cluster indicate some relatedness between them. They can be grouped to make it easy to examine other related theseseven if they are from different courses/specialisations. Some examples are provided below.

- **Cluster 0:** is characterised by high average topic weights on “Health/Management/Education” and “Land_cover/Maps/Urbanism/Populations/Environmental”, and as expected, most of the theses in this cluster are from “Clinical Research Management”, “Geographic Information Systems and Science” as well as “Geospatial Technologies”. These characteristics reveal some relatedness between the topics, the theses, and the courses/specialisations. Additionally,

grouping them could help examine similar theses even though they are from different courses/specialisations.

- **Cluster 1:** is characterized by a higher average topic weight on "Customer_satisfaction_&_behaviour/Marketing/Products/Brands" and "Business/Customers/Companies/Management/Technology". The cluster has most of the theses from "Marketing Intelligence" and "Marketing Research and CRM". Additionally, it has many theses from the "Information Systems and Technologies Management" and "Knowledge Management and Business Intelligence". These characteristics reveal relatedness between the topics, theses, and courses/specialisations.
- **Clusters 1 and 3:** As stated before they are not related clusters. Analysing the dominant topic and the course/specialisation with the majority of the theses within them, it is possible to conclude the same. The same analysis for **Clusters 3 and 0** leads to the conclusion that they are related clusters.

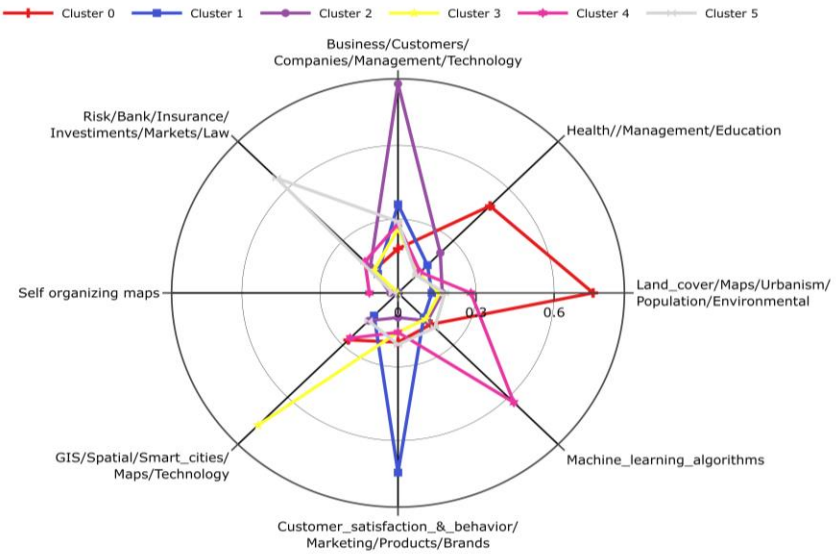


Figure 6.13 – Average topic weights distribution of each cluster (LDA).

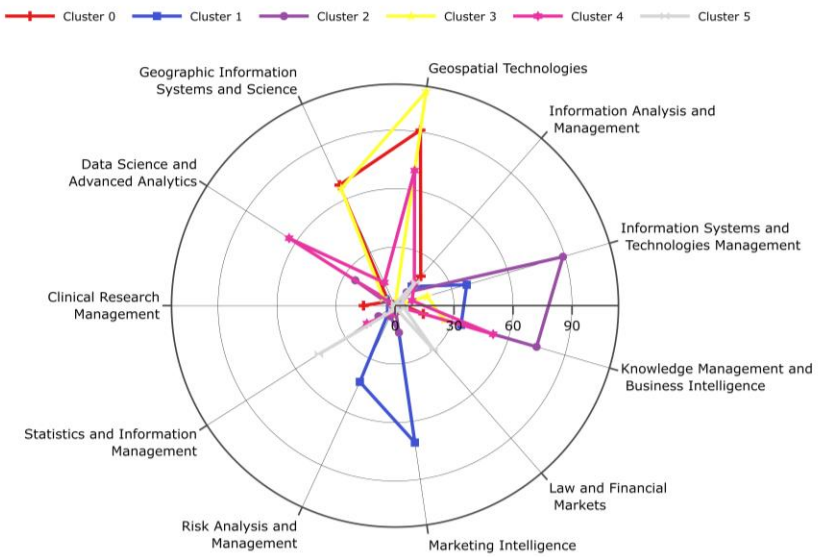


Figure 6.14 – The thesis numbers in each cluster by course/specialisation (LDA).

6.4.2. Hierarchical cluster results (BERTopic topic vectors)

As stated previously, the BERTopic topic vectors provide a better cluster solution than the LDA. The final solution has two small clusters with twenty (20) theses and three big clusters with more than a hundred (100) theses (Figure 12, Appendix). Furthermore, Figures 6.15 and 6.16 were analysed similarly to Figures 6.13 and 6.14. The analysis leads to the conclusion that the clusters are better discriminated. Figure 6.15 shows that most clusters are characterised by only one different topic, indicating that the topics differentiate them well.

Regarding figure 6.16, it is possible to see that the majority of the theses from each course/specialisation are grouped in only one cluster. Additionally, the main topic of the cluster is also strongly linked to the dominant course/specialisation. A good example is theses from "Geographic Information Systems and Science" and "Geospatial Technologies". They are grouped only in cluster 1, which has "Land_cover/Maps/Urbanism/Populations/Environmental" as the main topic.

Figures 6.15 and 6.16 also support the cluster arrangement of SOM from BERTopic features (Figure 6.12, right image). Cluster 4 is at the edge of the SOM from BERTopic features (Figure 6.12, right image), indicating that it contains theses with a specific topic. The dominant topic is linked to genetic algorithms, genetic programming, and geometric semantic genetic programming which are specific subjects in "Data Science and Advanced Analytics". Additionally, only 20 out of more than 100 theses from "Data Science and Advanced Analytics" are in this cluster. Cluster 0, located in the middle/centre, has no dominant topic but rather many theses from many courses/specialisations.

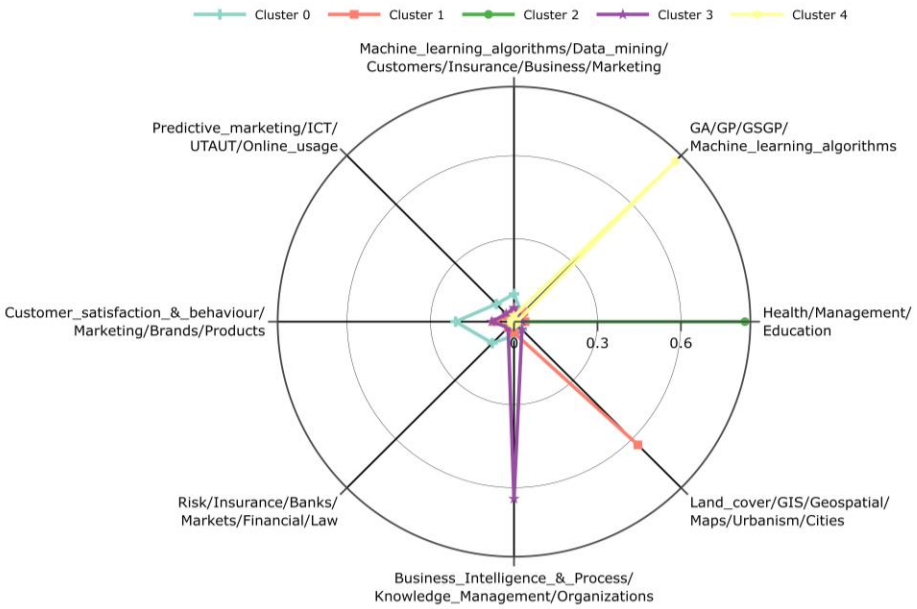


Figure 6.15 – Average topic weights distribution of each cluster (BERTopic).

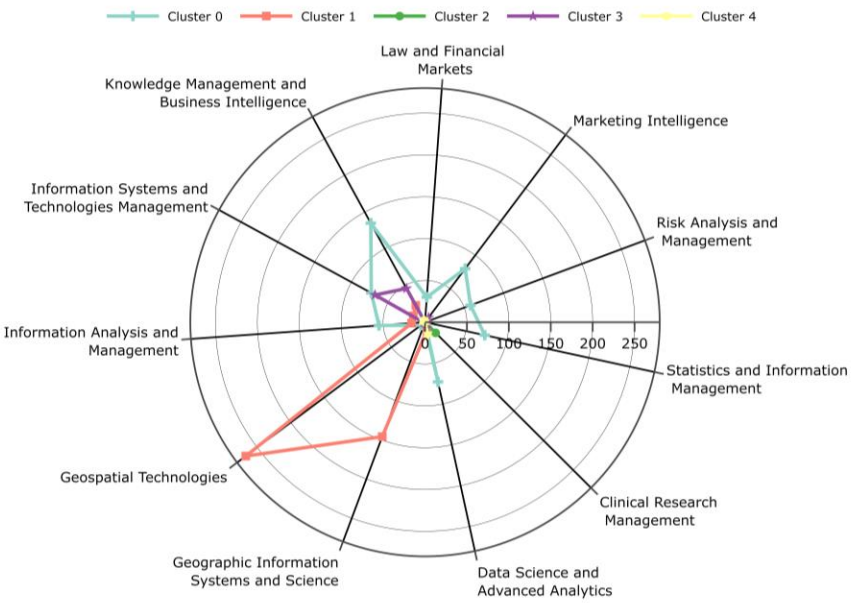


Figure 6.16 –The thesis numbers in each cluster by course/specialisation (BERTopic).

6.5. NETWORK CENTRALITY MEASURES RESULTS

Questions about topic distributions, topics similarities, thesis organisations, and similarities were handled in the above chapters. This chapter will address questions about the direct relationship between the theses and the results of the centrality measures pointed out in section 4.31.

Figure 6.17 illustrates the average centrality measures of each cluster calculated using topic vectors/features from LDA. Higher values indicate higher importance based on the definition of each centrality measure. The comparison between the results of cluster 2 and the other clusters' results leads to the conclusion that it contains the theses with the most connections in the network, it has more links with essential theses in the network, and, finally, it has more theses closer to others, meaning they have a good diversity of topics.

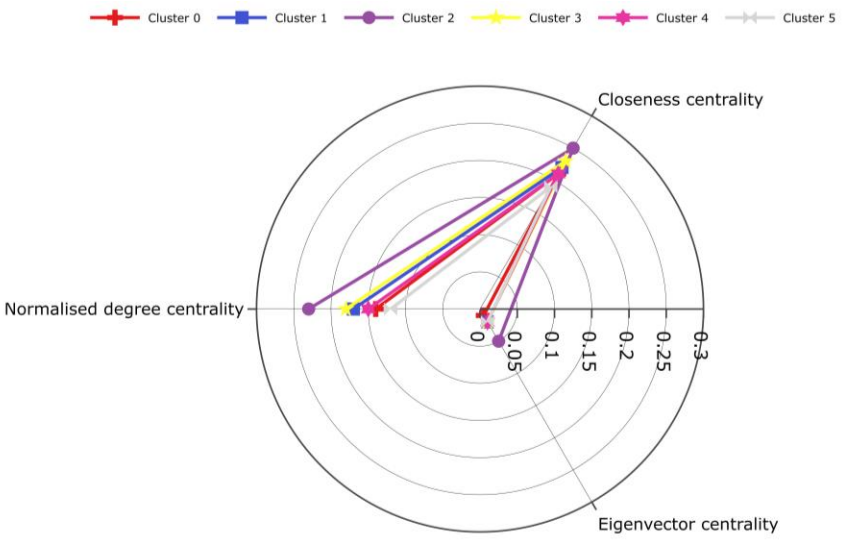


Figure 6.17 – Average centrality measures in each cluster (LDA).

Figure 6.18 illustrates the average centrality measures of each cluster calculated using topic vectors/features from BERTopic. The comparison between the results of cluster 1 and the other clusters' results shows it contains the theses with the most connections in the network, it has more links with essential theses in the network, it has more theses closer to others, meaning they have a good diversity of topics.

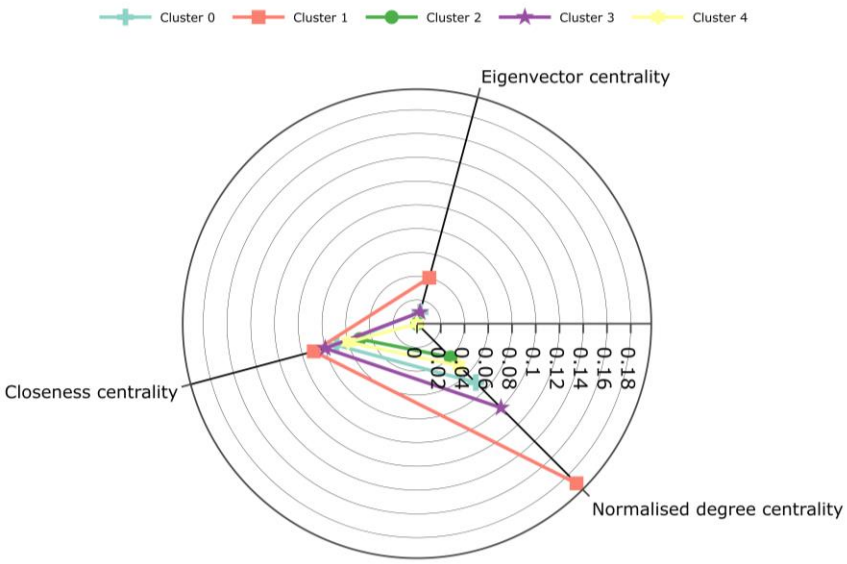


Figure 6.18 – Average centrality measures in each cluster (BERTopic).

Table 6.4 provides information about the density and the average percentage of all topical coherence explained by the two networks of the theses. As stated previously, those networks were separately modelled by using the topic vectors of each topic modelling algorithm and retaining the three best topics in each thesis. According to the results, the theses network using the topic vectors/features from BERTopic had 84.95% of density, and it can explain, on average, 98% of all the topical coherence. The result of LDA indicates a good balance between the explanation ability of the theses interlinkages and the network's density.

Table 6.4– Density and the average percentage of all topical coherence explained by the network from each topic vector algorithm (%).

Topic vector algorithm	The average percentage of all topical coherence picking the three best topics in each thesis	Network Density
LDA	63.35	57.35
BERTopic	98.00	84.95

6.6. VISUALISATION

An important goal of this project is to produce an interactive visualisation of all the theses and incorporate the analysis of the main results found from each technique. A theses network was built using Gephi to achieve this goal. The created network visualisation allows us to visualise all the theses, the clusters, the direct connections between the theses and to check the information like the ID, the author, the supervisors, the title, the type, the URL, the Course/Specialisation, the token numbers,the score readability, the cluster label and the topic distributions on each thesis. Each node on the visualisation represents one thesis. The size of nodes helps to indicate the number of connections that the thesis has in the network. A larger node size means the respective thesis is connected to many other theses. The edge colour intensity indicates the connection level between theses. More colour intensity of the edge means that the two theses have a strong connection.

As stated above, each node colour represents the cluster label of the thesis and theses within the same cluster share some similarities. Nodes/theses closer to each other in the network indicate that these nodes/theses share some similarities. The above statements can point to some similarities between network and cluster results. Additionally, nodes placed in the middle of the network represent theses that are closer/similar to many other nodes/theses. They have a good diversity of topics. Many of these theses belong to clusters with a higher closeness centrality average. Nodes Isolated at the extremities of the network represent theses with few connections, i.e. with topics not commonly found in other theses. Analysing the above information can also point to some similarities between the arrangement of the network and SOM from cluster results.

As stated above, the topic vectors of each thesis were embedded in the visualisation instead of the abstract. The information about the topic distribution can be helpful for the users to understand the subjects covered in each thesis without needing to read the abstract. Due to that, no keywords search engine was embedded. However, suppose the user wants more detail about the thesis. In that case, the network allows access to the abstract and the entire document through its URL. If the document is restricted, it is possible to search in the network for other documents connected to this one.

With no search engine incorporated into the visualisation, a combo box, where the user can select the cluster that he/she wants to examine, was made available (Figure 6.19, left image). Furthermore, a description of each cluster was embedded in the visualisation to guide the user. To access this information, the user only needs to click on the link “more information” at the middle left of the page (Figure 6.19, left image). After choosing the cluster of interest, he/she can select one of the theses to visualise the embedded information and all the thesis connections.

The network offers the user much versatility when searching for the theses. Focussing on the node size allows us to know if the thesis has many connections. The proximity between nodes/theses and the colour intensity of the edge point to the similarities and connection levels between theses. This information can help the user identify essential theses and strategic areas. Furthermore, they can find similar theses without reading the entire abstract by examining the node size or position.

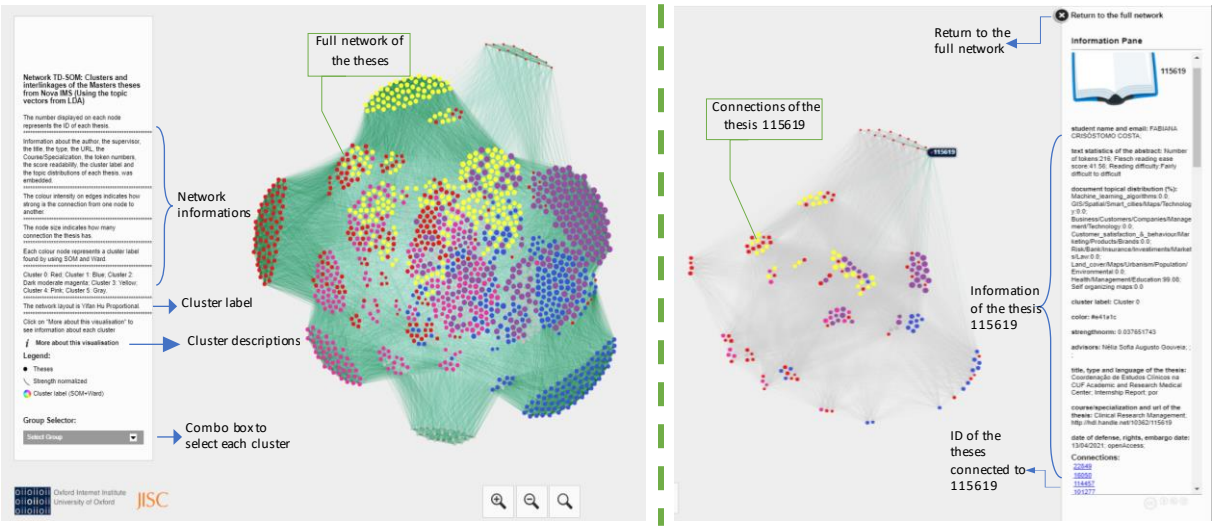


Figure 6.19 – Web-based interactive network visualisation interfaces. It is composed of two interfaces (The first interface (left) and the second interface (right)) and four components.

Figure 6.19 presents the two user interfaces of the visualisation. The first interface (the left image) has two components. The first component presents the information of the theses network, namely the corresponding colour of each cluster label, a link to access the description of each cluster and a combo box to select each cluster to visualise its theses and their connection. The second component is aside from the first. It represents all the nodes/theses of the dataset and their connections.

The second interface is visualised if one thesis is selected and It has two components (figure 6.19, the right image). The first component represents the connections of the selected thesis. The second component presents all embedded information of the selected thesis. In case the user wants to get back to the initial network, he/she needs to click the button “x return to the full network” at the top right-hand side of the page.

This visualisation allows the user to rapidly browse the theses and understand the degree of interconnection between the clusters, topics and theses. The user can also drag the network and interact with the network, zooming to visualise a particular region of the network.

6.6.1. The theses network results (LDA topic vector)

Figure 6.20 shows the interlinkages of the master’s theses from NOVA IMS modelled by using the LDA topic vectors and retaining the three best topics in each thesis. The arrangement of the network has a good similarity with cluster results. Several closer nodes/theses are part of the same cluster. The network of the theses and SOM from LDA topic vectors (Figure 6.12, left image) arrangements also have similarities. Clusters that share the same region on the SOM have their nodes/theses located in the same region/nearby in the network. At the same time, clusters located far away in the SOM have their nodes/theses located far away in the network. Additionally, several nodes/theses in the middle of the network also belong to the clusters located in the middle of the other clusters in the SOM. Finally, isolated nodes at the extremities of the network belong to the clusters located at the edge of the SOM. For instance:

- Cluster 0 (red nodes) and cluster 3 (yellow nodes) are located in the same region/nearby on the SOM from LDA topic vectors (Figure 6.12, left image), and most of their nodes/theses are located in the same region/nearby in the network;
- Cluster 1 (blue nodes) and cluster 3 (yellow nodes) are the opposite of the above example;
- Cluster 2 (Dark moderate magenta nodes) and Cluster 3 (yellow nodes) have a higher average for closeness centrality (Figure 6.17). Some nodes/theses from these clusters are located in the middle of the network. The SOM from LDA topic vectors (Figure 6.12, left image) illustrates that these clusters are also in the middle of other clusters;
- Theses from "Clinical Research Management" belong to cluster 0 (red nodes). They have topics not commonly found in other theses and are located in the extremities of the network. Moreover, analysing the SOM (Figure 6.12, left image) is possible to see that this cluster is also located at the edge of the map.

Besides these similarities in the results, a comparison between SOM and network visualisation shows that the network can reveal additional similarities between the theses even if they are from different clusters.

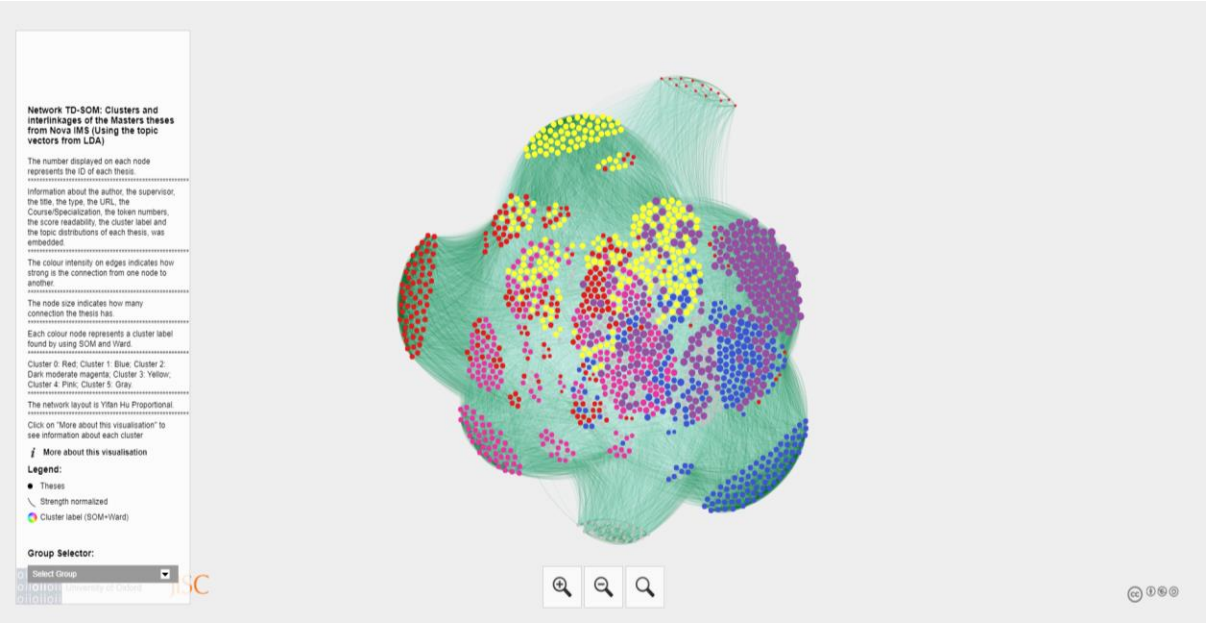


Figure 6.20 – The master's theses network by using LDA topic vectors.
Link to access the interactive visualisation: https://vmunhangane.github.io/Thesis_vis/network/

6.6.2. The theses network results (BERTopic topic vector)

Figure 6.21 shows the interlinkages of the master's theses from NOVA IMS modelled by using the BERTopic topic vectors and retaining the best three topics in each thesis. Overall, the network is dense, and some nodes are overlapping. Besides that, several closer nodes/theses are part of the same cluster. The network and SOM from BERTopic topic vectors (Figure 6.12, right image) arrangements also have similarities. As an example, clusters 0 (Slightly desaturated cyan) and 1 (Soft red) are located in the middle of the other clusters in the SOM from BERTopic (Figure 6.12, right image). Therefore, several of their nodes/theses are also located in the middle of the network.

Cluster 0 (Slightly desaturated cyan) and cluster 3 (Dark moderate magenta) are located in the same region/nearby on the SOM from BERTopic topic vectors (Figure 6.12, right image). Therefore, most of their nodes/theses are located in the same region/nearby in the network. Cluster 1 (Soft red) and cluster 3 (Dark moderate magenta) are located far away in the SOM from BERTopic (Figure 6.12, right image). Therefore, most of their nodes/theses are located far away in the network. Additionally, cluster 4 (Very light Yellow), characterised by the topic "Health Management Education" has several isolated nodes/theses at the extremities of the network. This cluster is also located at the edge of the SOM from BERTopic (Figure 6.12, right image).

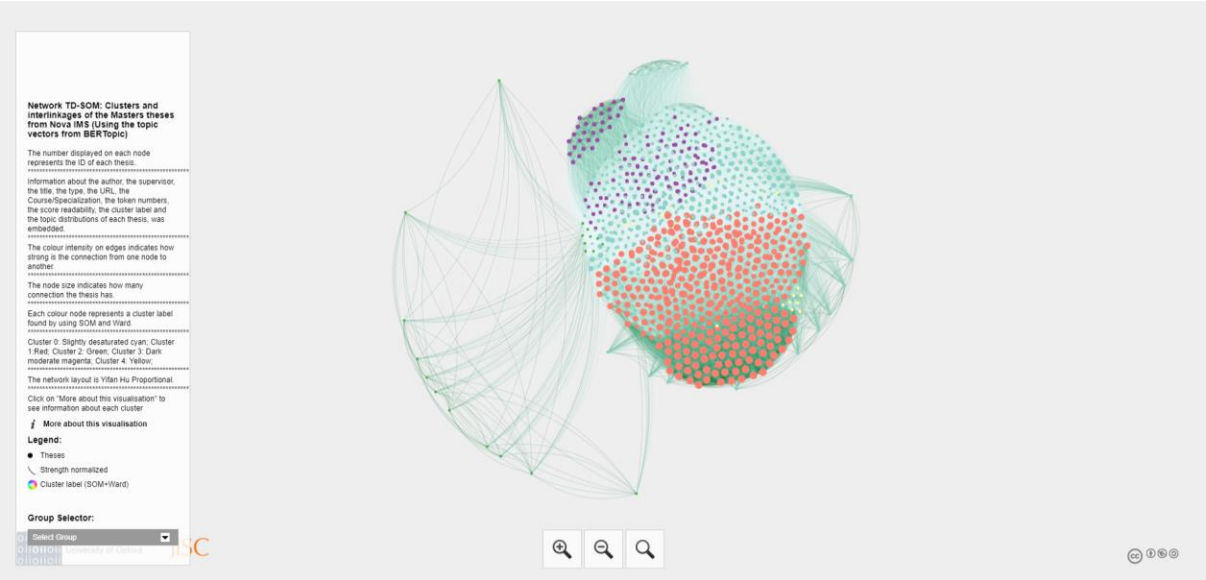


Figure 6.21 – The master's theses network by using BERTopic topic vectors.
Link to access the interactive visualisation:
https://vmunhangane.github.io/Theses_bertopic/network/

7. CONCLUSION

In this work, we present Network TD-SOM, a systematic process combining topic modelling algorithms, cluster algorithms and network analysis, to explore a large set of documents. We test the use of Network TD-SOM in a set of master's theses from NOVA IMS and explore the results. We use the topic modelling algorithms to capture the latent topics in a corpus and then use these topics as features for cluster algorithms and network analysis. We implemented LDA and BERTopic separately, which are two different techniques of topic modelling. For clustering, we selected the two-level hybrid approach, combining SOM and a hierarchical clustering algorithm, using the Ward method for the first and second levels of clustering respectively. The hierarchical clustering algorithm was used to simplify the SOM structure and get a summary of the overall clustering structure. Network analysis was used to reveal the direct connections between the documents and their centrality measures. The final goal is to extract helpful knowledge and represent the relatedness of documents in a corpus. To achieve this goal, we designed various kinds of 2-dimensional visualization/spatialisation. Some of them allow for the uncovering of the thematic structure distribution over time and for profiling/understanding the clusters in a Corpus. The main visualisation is an interactive corpus network that supports exploration, browsing, navigation, and zoom. Additionally, the main results of each technique and the useful metadata information are incorporated into the visualisation. These spatialisations are easily exploited and do not require users to be familiar with visualisation and ML.

Network TD-SOM had a great performance on the master's thesis dataset from NOVA IMS. LDA and BERTopic successfully uncovered the thematic structure and extracted helpful knowledge of the dataset. Although the two techniques presented similarities in the results, their comparison demonstrates the superiority of BERTopic as a solution. It highlights that, in each NOVA IMS master program, the dominant topic and the top 25 words are highly semantically linked. Regarding the topic modelling algorithms' performance as a feature extraction tool, we concluded that the features/topics extracted from BERTopic leverage the cluster results over features from LDA. For the network visualisations of the theses modelled through the topic vectors of each topic modelling algorithm and retaining the three best topics in each thesis, the arrangement of the network modelled through LDA topic vectors was better than the one modelled through BERTopic topic vectors. It has a good balance between the explanation ability of the interlinkages and the density of the network. While the network visualisation modelled through BERTopic features/topics is dense, some nodes are overlapping. However, the two networks of the theses had some similarities with their cluster results.

8. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

While implementing Network TD-SOM on the master's theses from NOVA IMS, we had some limitations. They are described below as well as some recommendations for future work.

BERTopic can semantically discover latent topics in the corpus but has a weakness in assuming that each document only contains a single topic and not multiple topics. LDA does not consider the context of the text to discover the latent topics but assumes that documents contain multiple topics. Furthermore, in BERTopic is normal to have several documents in which the sum of the topic vectors' weight is less than one (1), even if they are not outliers while in LDA it is the opposite. Based on that, we recommend a combination of the two techniques to overcome the issues mentioned previously, which may lead to better results. Additionally, other topic modelling algorithms can be implemented and their performance in discovering latent topics, knowledge discovery and working as feature extractors can be evaluated.

As stated above, BERTopic does not assume that documents contain multiple topics. Most of the time, the algorithm assigns the most weight to the first topic, and the rest will have little weight. We selected the best three topics in each thesis to model the network. Then we took the dot product between vectors to calculate the weight between the theses. This approach, combined with the mentioned weakness, may lead to a dense network with some overlapping nodes. Therefore, we recommend using relation extraction (Ali et al., 2021) which defines if two documents have a relationship capturing the context of the text.

The visualisation produced through Gephi, in which we used the Yifan hu proportional layout, seems to preserve the topological relations of the corpus. Still, it does not allow embedding a slicer in the visualisation to control the number of edges interactively yet. We did not perform any technique to keep only the most essential edges and leverage the network visualisation. We recommend the use of simple methods like thresholds, testing several percentage values of the edges to keep and choosing one which leverages the visualisation. Additionally, it is possible to go for more sophisticated or complex methods like topological data analysis (TDA)[17].

The Gephi allows embedding some metadata information or results in the visualisation. However, the software does not allow changing the order presentation of the embedded information. Furthermore, we noticed that the software produces a JSON[18] file that contains all information. Therefore, we recommend reordering the information using other software.

Network TD-SOM uses Gephi to create web-based interactive network visualisations. The recommended network size of nodes and edges for Gephi is below one million. Based on that, the scalability of Network TD-SOM is not up to million documents.

[17] Topological data analysis (TDA) (Centeno et al., 2021), is usually used to study the data shape by analysing the vertice interactions and it does not only concentrate on the simple-wise connections.

[18] A JSON file is a file that stores simple data structures and objects in JavaScript Object Notation (JSON) format, which is a standard data interchange format. It is primarily used for transmitting data between a web application and a server. JSON files are lightweight, text-based, human-readable, and can be edited using a text editor. (Available at https://fileinfo.com/extension/json , accessed at 23/06/2022)

9. REFERENCES

Ali, M., Saleem, M., & Ngomo, A. C. N. (2021). Unsupervised Relation Extraction Using Sentence Encoding. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12739 LNCS*, 136–140. https://doi.org/10.1007/978-3-030-80418-3_25

Ampazis, N., & Perantonis, S. J. (2004). LSISOM - A latent semantic indexing approach to self-organizing maps of document collections. *Neural Processing Letters*, *19*(2), 157–173. https://doi.org/10.1023/B:NEPL.0000023449.95030.8F

Asan, U., & Ercan, S. (2012). *An Introduction to Self-Organizing Maps*. 295–315. https://doi.org/10.2991/978-94-91216-77-0_14

Bação, F., Lobo, V., & Painho, M. (2004). *The Self-Organizing Map and it's variants as tools for geodemographical data analysis: the case of Lisbon's Metropolitan Area*.

Bali Swain, R., & Ranganathan, S. (2021). Modeling interlinkages between sustainable development goals using network analysis. *World Development*, *138*. https://doi.org/10.1016/J.WORLDDEV.2020.105136

Barabási, A.-L. (2016). Network science introduction. *Network Science*, 1–27.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Caillou, P., Renault, J., Fekete, J. D., Letournel, A. C., & Sebag, M. (2020). Cartolabe: A Web-Based Scalable Visualization of Large Document Collections. *IEEE Computer Graphics and Applications*, *41*(2), 76–88. https://doi.org/10.48550/arxiv.2003.00975

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *10*(1). https://doi.org/10.1145/2733381

Candela, L., Castelli, D., & Pagano, P. (2015). *History, Evolution and Impact of Digital Libraries*. http://www.researchgate.net/publication/229422428

Centeno, E. G. Z., Moreni, G., Vriend, C., Douw, L., & Santos, F. A. N. (2021). A hands-on tutorial on network and topological neuroscience. *BioRxiv*, 2021.02.15.431255. https://doi.org/10.1101/2021.02.15.431255

Ceross, A., & Bergmann, J. (2021). A machine learning approach for medical device classification. *ACM International Conference Proceeding Series*, 285–291. https://doi.org/10.1145/3494193.3494232

Chehal, D., Gupta, P., & Gulati, P. (2021). Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. *Journal of Ambient Intelligence and Humanized Computing*, *12*(5), 5055–5070. https://doi.org/10.1007/S12652-020-01956-6

Cheu, E., Keongg, C., & Zhou, Z. (2004). On the two-level hybrid clustering algorithm. *… Conference on Artificial Intelligence in …*. http://www3.ntu.edu.sg/home/asckkwoh/My Conferences/c2004 AISAT - Zhouzl - full paper.pdf

Choirul Rahmadan, M., Nizar Hidayanto, A., Swadani Ekasari, D., Purwandari, B., & Theresiawati. (2020). Sentiment Analysis and Topic Modelling Using the LDA Method related to the Flood Disaster in Jakarta on Twitter. *Proceedings - 2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020*, 126–130. https://doi.org/10.1109/ICIMCIS51567.2020.9354320

Culmer, K., & Uhlmann, J. (2021). Examining LDA2Vec and Tweet Pooling for Topic Modeling on Twitter Data. *WSEAS TRANSACTIONS ON INFORMATION SCIENCE AND APPLICATIONS*, *18*, 102–115. https://doi.org/10.37394/23209.2021.18.13

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Undefined*. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6

Eletter, S. F., AlQeisi, K. I., & Elrefae, G. A. (2022). *The Use of Topic Modeling in Mining Customers' Reviews*. 1–4. https://doi.org/10.1109/ACIT53391.2021.9677049

Fabrikant, S. I. (2000). Spatialized browsing in large data archives. *Transactions in GIS*, *4*(1), 65–78. https://doi.org/10.1111/1467-9671.00038

Glasman-Deal, H. (2020). *Science research writing : for native and non-native speakers of English*. *Second Edition*, 356.

Gobov, D., & Yanchuk, V. (2022). *Network Analysis Application to Analyze the Activities and Artifacts in the Core Business Analysis Cycle*. 1–6. https://doi.org/10.1109/IISEC54230.2021.9672373

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://doi.org/10.48550/arxiv.2203.05794

Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. (2020). Social network analysis: Measuring, mapping, and modeling collections of connections. *Undefined*, 31–50. https://doi.org/10.1016/B978-0-12-382229-1.00003-5

Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1996). *Newsgroup Exploration with WEBSOM Method and Browsing Interface. Research rept.*

Hu, Y. (2011). *Algorithms for Visualizing Large Networks*.

Kaski, S. (1998). *Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering*. https://doi.org/10.1109/IJCNN.1998.682302

Keller, J. M., & Gray, M. R. (1985). A Fuzzy K-Nearest Neighbor Algorithm. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-15*(4), 580–585. https://doi.org/10.1109/TSMC.1985.6313426

Kohonen, T. (1990). *The Self-Organizing Map*. https://sci2s.ugr.es/keel/pdf/algorithm/articulo/1990-Kohonen-PIEEE.pdf

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, *11*(3), 574–585. https://doi.org/10.1109/72.846729

Lafia, S., Kuhn, W., Caylor, K., & Hemphill, L. (2021). Mapping research topics at multiple levels of detail. *Patterns*, *2*(3). https://doi.org/10.1016/J.PATTER.2021.100210

Lafia, S., Last, C., & Kuhn, W. (2019). Enabling the discovery of thematically related research objects with systematic spatializations. *Leibniz International Proceedings in Informatics, LIPIcs*, *142*. https://doi.org/10.4230/LIPICS.COSIT.2019.18

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 1999 401:6755*, *401*(6755), 788–791. https://doi.org/10.1038/44565

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/JOSS.00861

Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., Aromolaran, O., Uwoghiren, E., Olaniyan, D., & Olawole, O. (2019). Data Clustering: Algorithms and Its Applications. *Proceedings - 2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019*, 71–81. https://doi.org/10.1109/ICCSA.2019.000-1

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.48550/arxiv.1908.10084

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Schmidt, C. R., Rey, S. J., & Skupin, A. (2011). Effects of irregular topology in spherical self-organizing maps. *International Regional Science Review*, *34*(2), 215–229. https://doi.org/10.1177/0160017610387297

Silva, D., & Bacao, F. (2022). MapIntel: Enhancing Competitive Intelligence Acquisition Through Embeddings and Visual Analytics. *Lecture Notes in Computer Science (Including Subseries Lecture*

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *13566 LNAI*, 599–610. https://doi.org/10.1007/978-3-031-16474-3_49/FIGURES/3

Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*. https://doi.org/10.1016/J.IS.2020.101582

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/S40745-015-0040-1

Zavrel, J. (1996). Neural navigation interfaces for Information Retrieval: Are they more than an appealing idea? *Artificial Intelligence Review 1996 10:5*, *10*(5), 477–504. https://doi.org/10.1007/BF00130695

APPENDIX

Table 1 – Distribution of the theses by each course/specialization.

Courses/Specializations	Number | Percentage (%)
Geospatial Technologies	273|20.2
Knowledge Management and Business Intelligence	208|15.4
Information Systems and Technologies Management	159|11.8
Geographic Information Systems and Science	148|10.9
Statistics and Information Management	110|8.1
Data Science and Advanced Analytics	105|7.8
Marketing Intelligence	91|6.7
Risk Analysis and Management	76|5.6
Information Analysis and Management	73|5.4
Marketing Research and CRM	61|4.5
Law and Financial Markets	30|2.2
Clinical Research Management	18|1.3

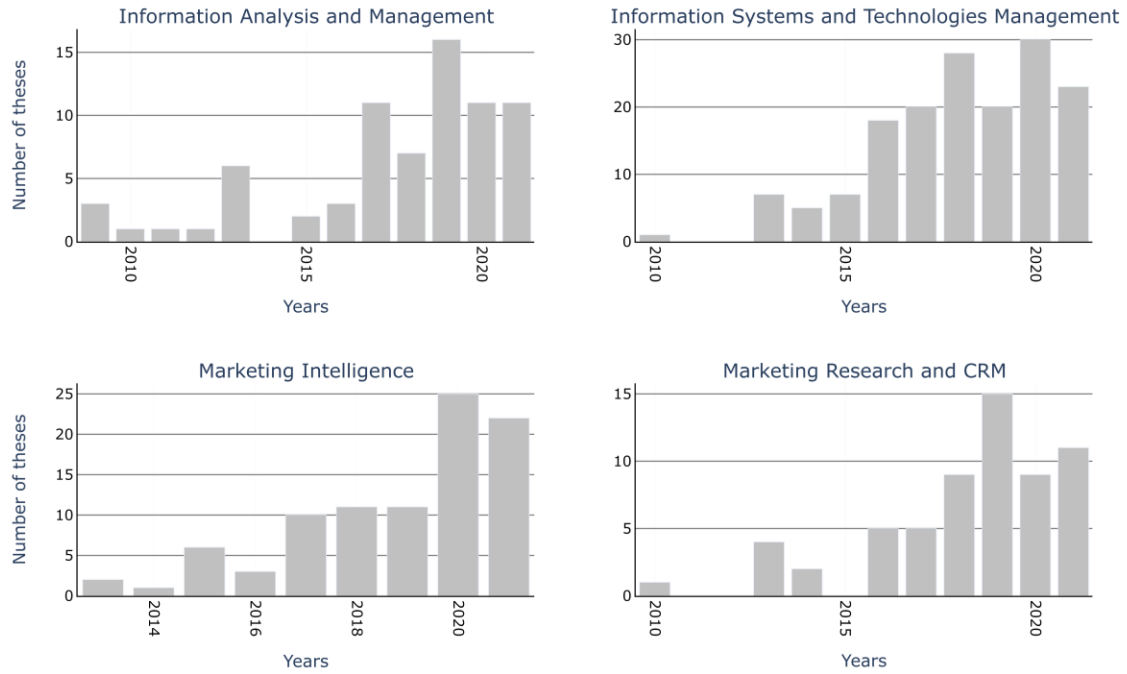


Figure 1 – Distribution of thesis by four courses/specialisation per year (1).

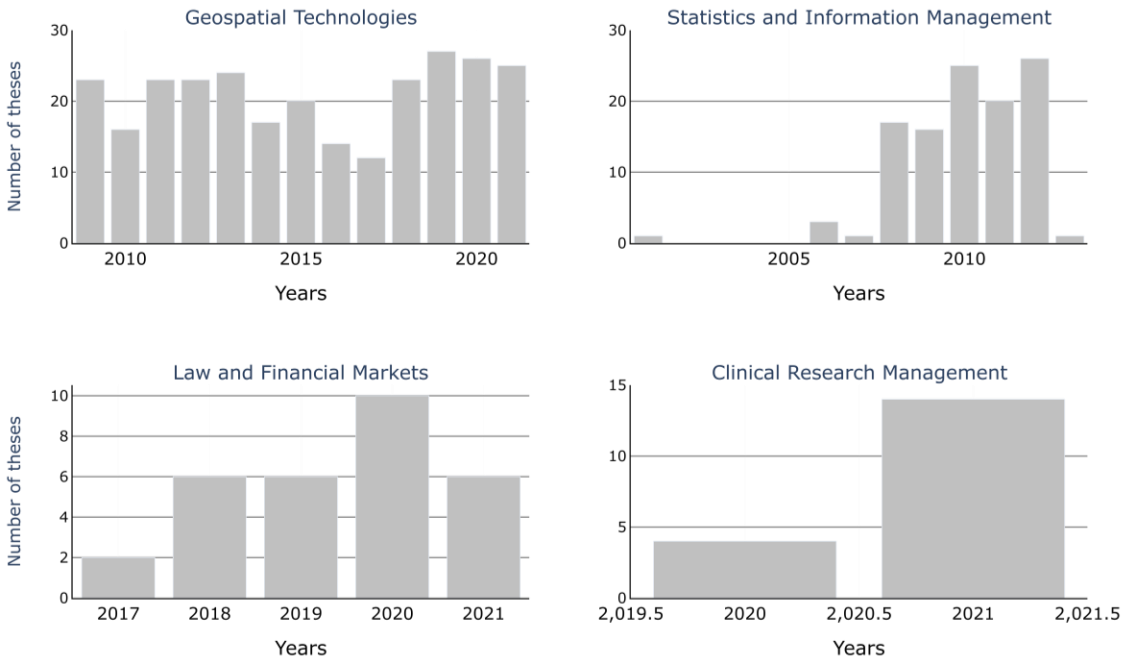


Figure 2 – Distribution of thesis by four courses/specialisation per year (2).

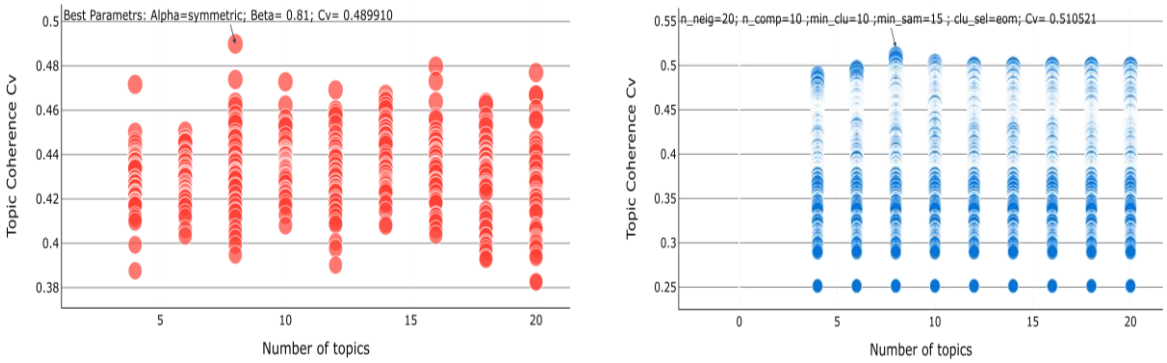


Figure 3 – Distribution of topic coherence (Cv) from LDA (left) and BERTopic (right).

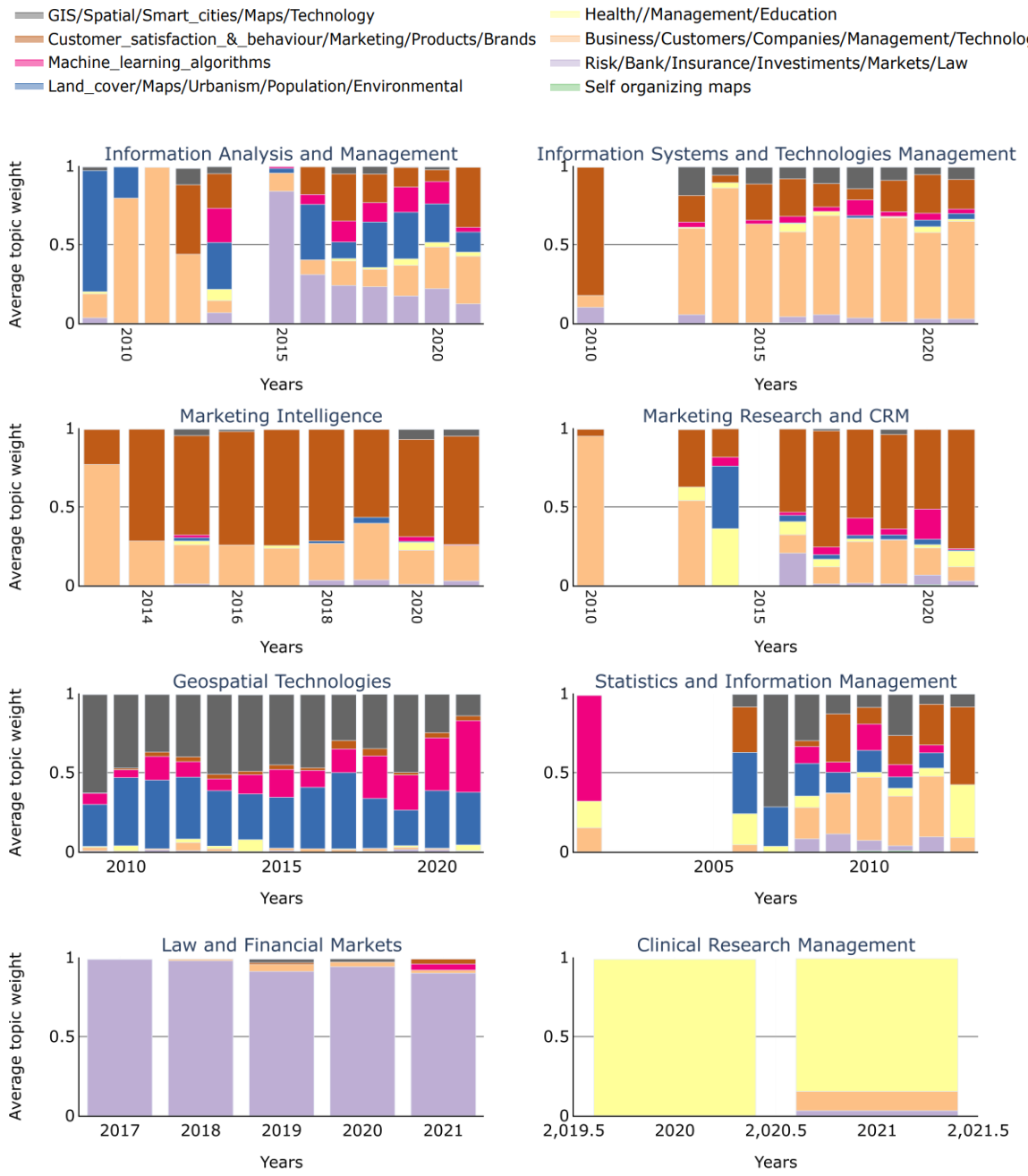


Figure 4 – Distribution of Weight topics by courses/specialisations per year (LDA).

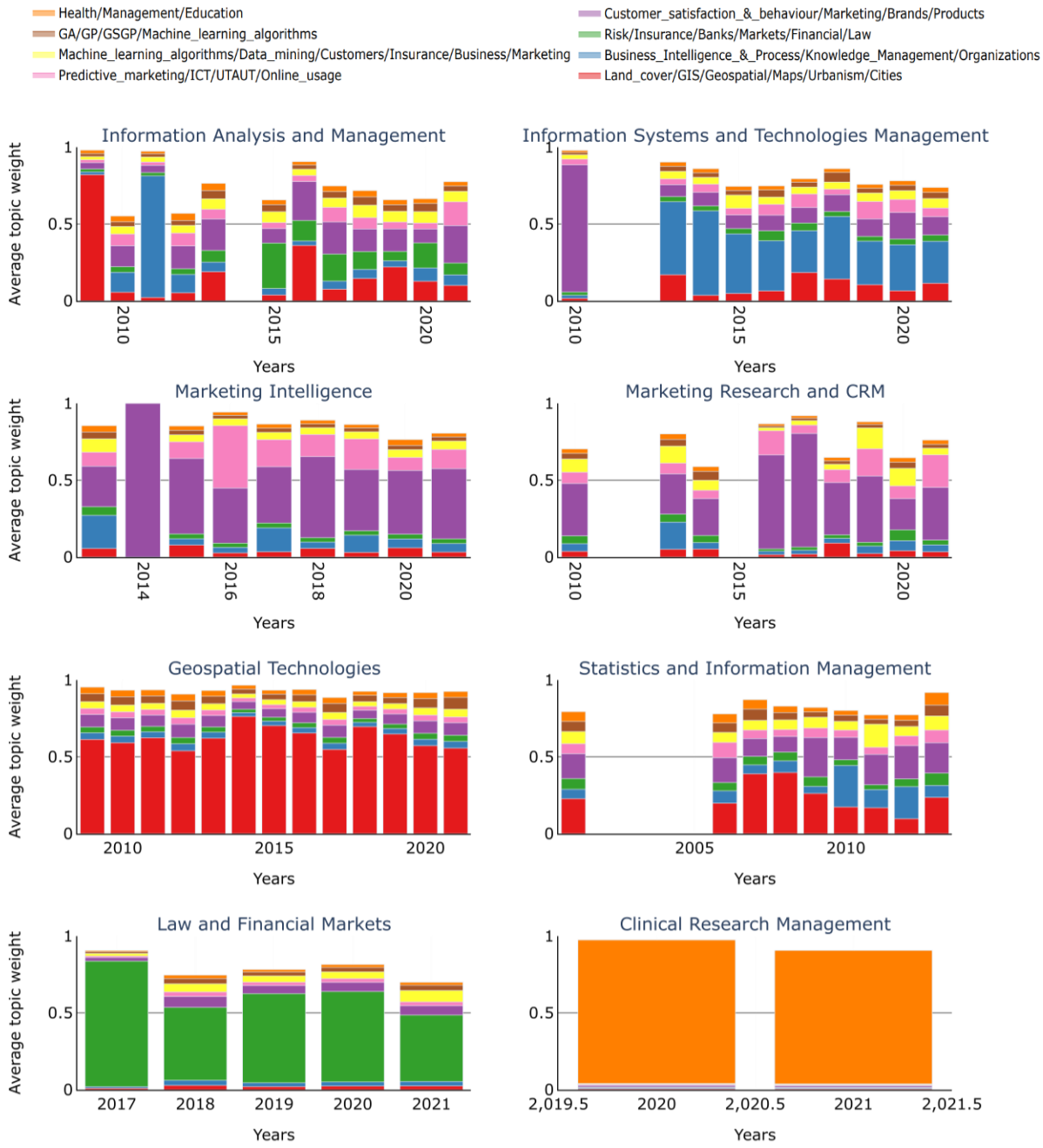


Figure 5 – Distribution of average topic weight by courses/specialisations per year (BERTopic).

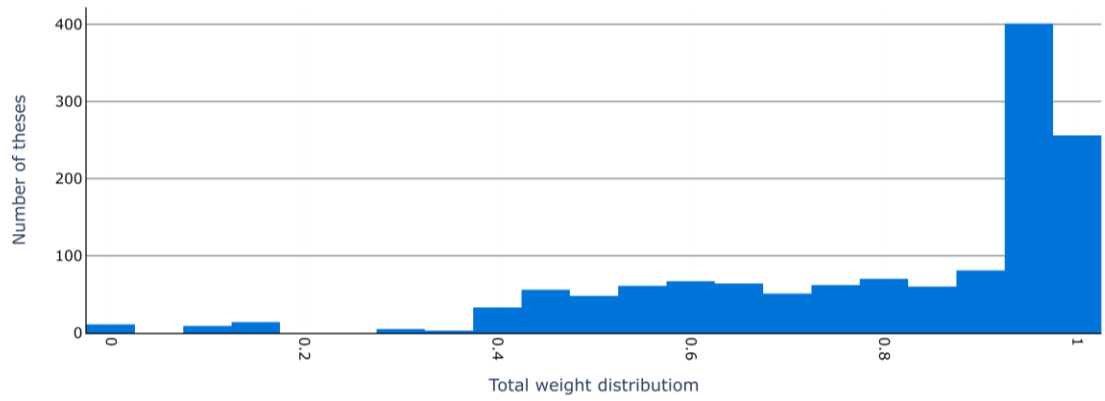


Figure 6 – Distribution of total topic weight in each thesis (BERTopic).

```
In [72]: 1 final_model.get_topic_info()
```

Out[72]:

	Topic	Count	Name
0	0	540	0_data_information_spatial_land
1	-1	345	-1_data_results_study_model
2	1	170	1_business_information_management_data
3	2	110	2_financial_risk_insurance_banks
4	3	88	3_satisfaction_consumers_customer_study
5	4	30	4_intention_technology_use_acceptance
6	5	26	5_customers_models_model_campaign
7	6	22	6_gp_genetic_gsgp_algorithm
8	7	21	7_clinical_internship_research_trials

Figure 7 – Distribution of thesis numbers in each topic (BERTopic).

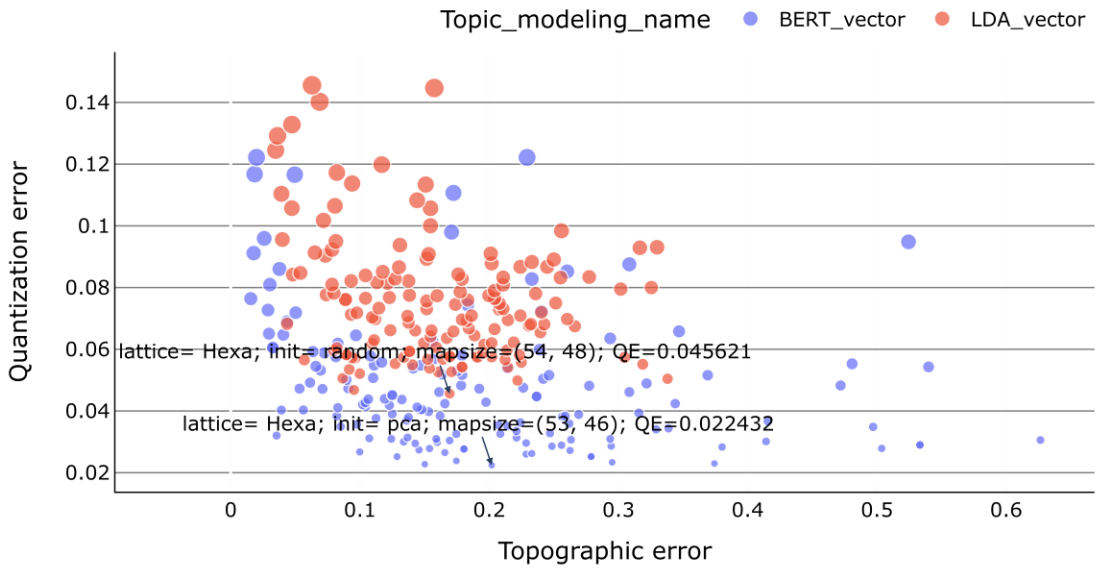


Figure 8 – Distribution of QE and TE from LDA and BERTopic topic vectors.

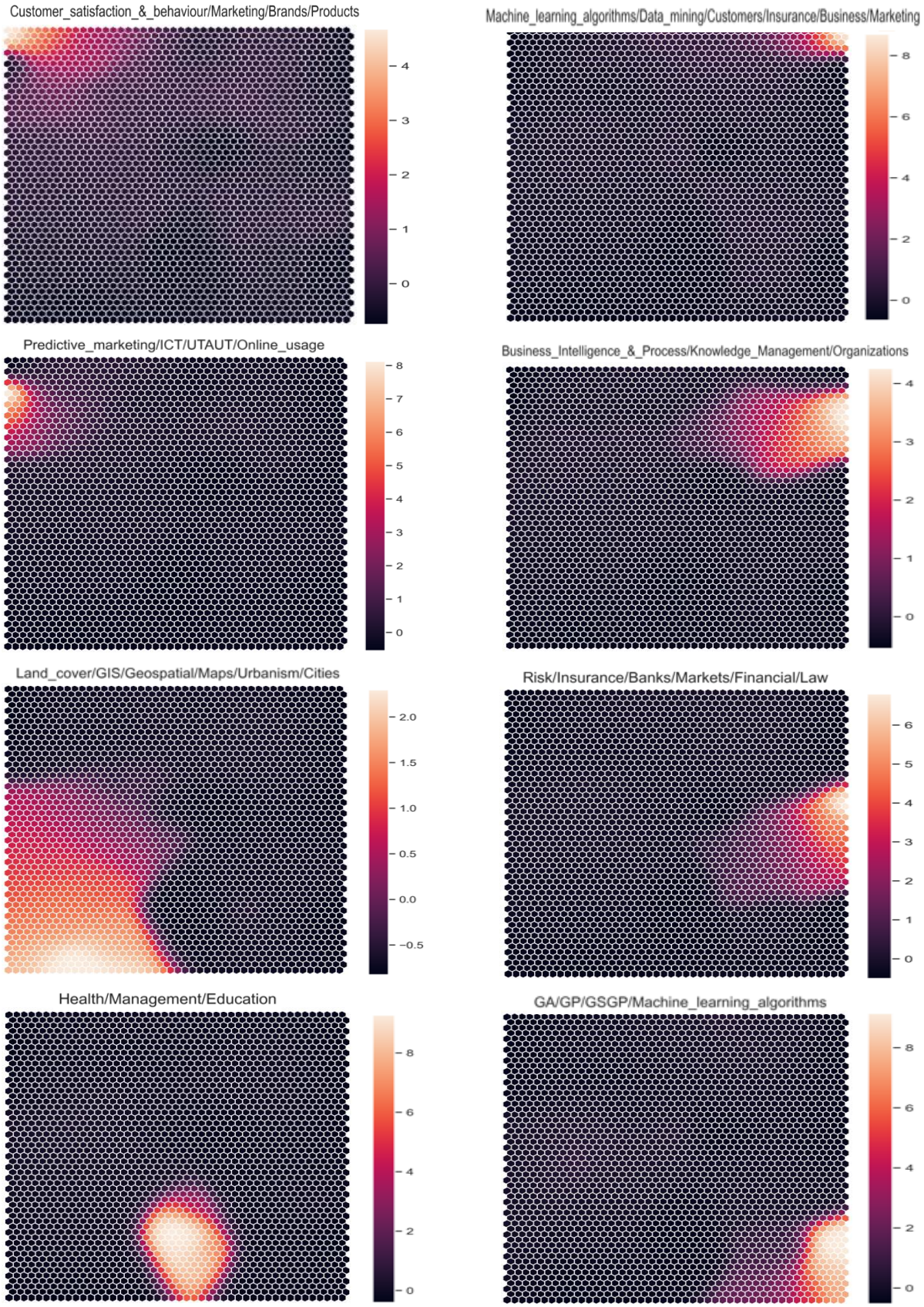


Figure 9 – Component Planes of the topics/features from BERTopic.

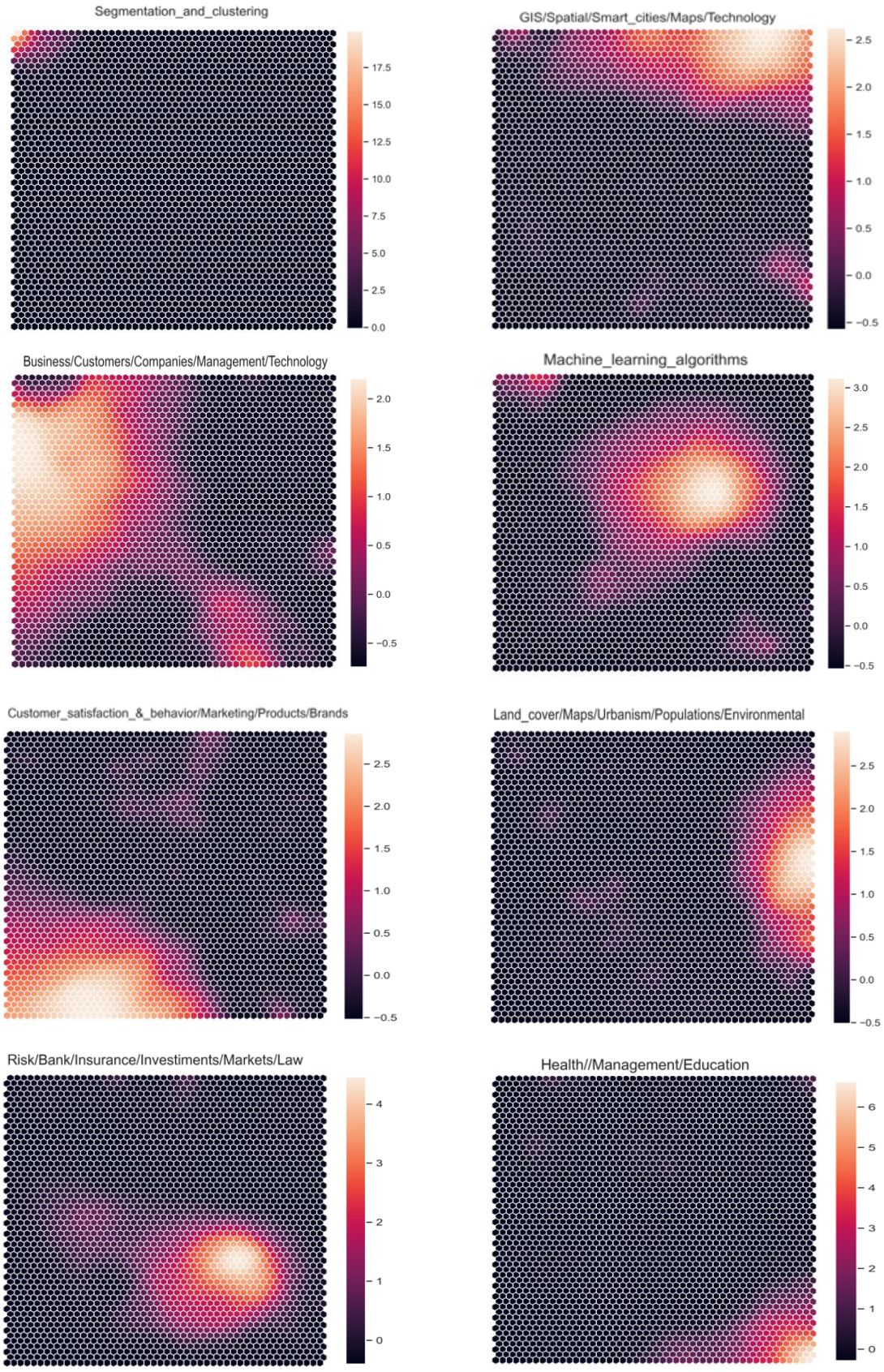


Figure 10 – Component Planes of the topics/features from LDA.

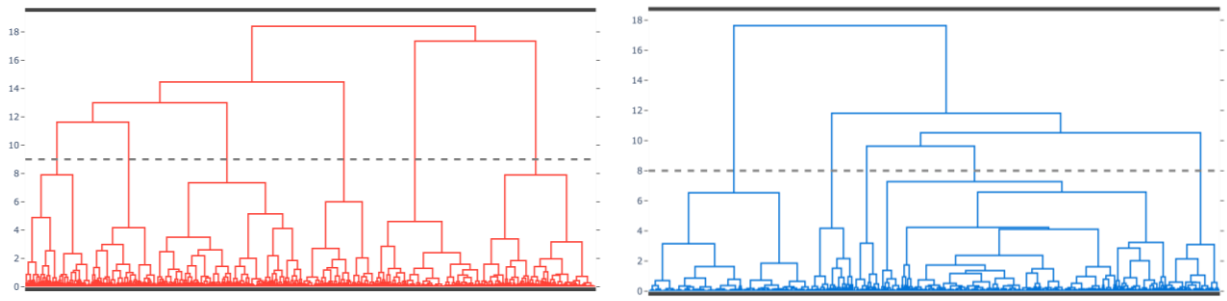


Figure 11 – Dendrogram of the BMUs from LDA (red) and BERTopic (blue) respectively.

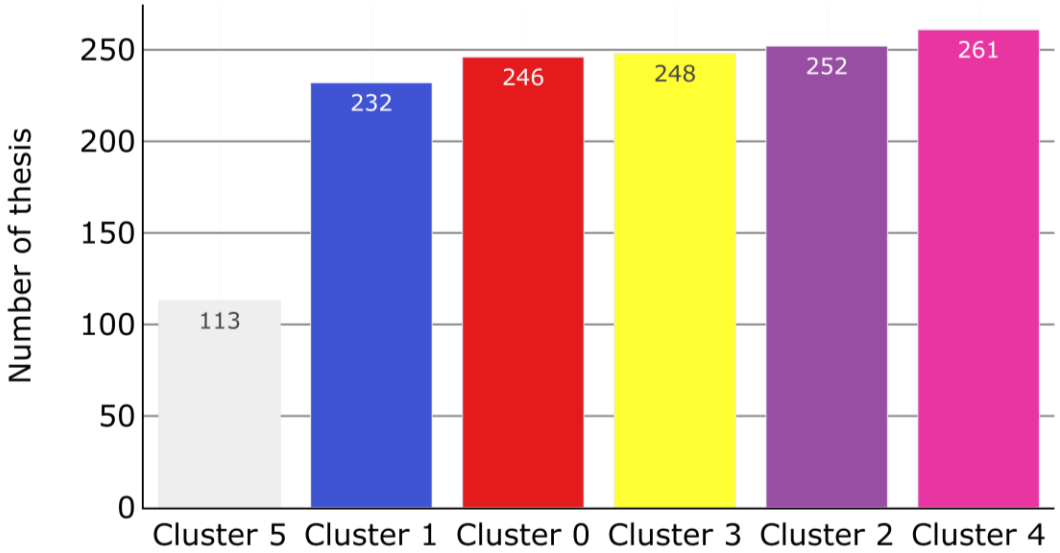


Figure 12 – Number of theses in each cluster (LDA vector).

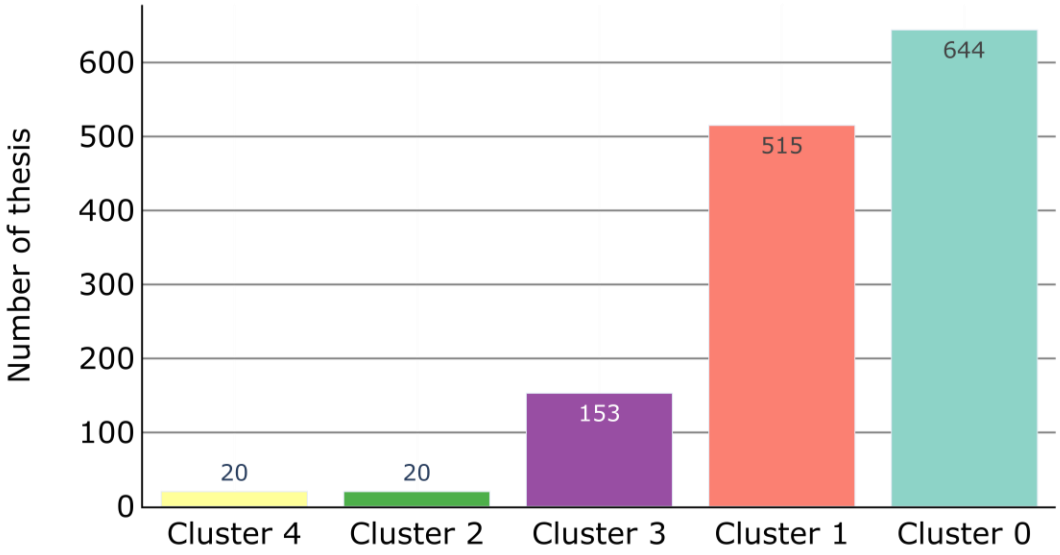


Figure 13 – Number of theses in each cluster (BERTopic vector).

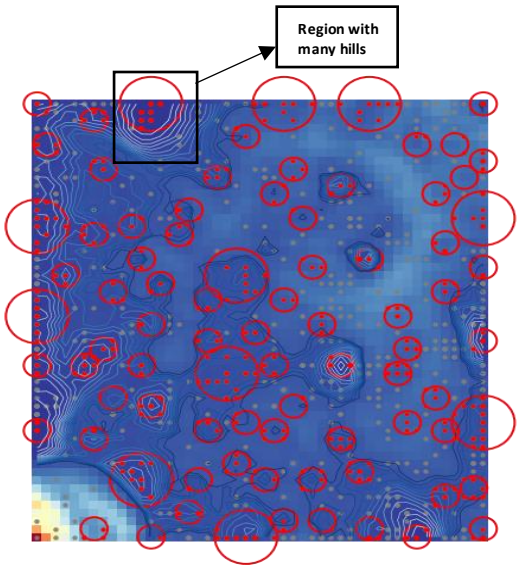


Figure 14 – U-matrix of the best-trained SOM using vector topics from LDA.

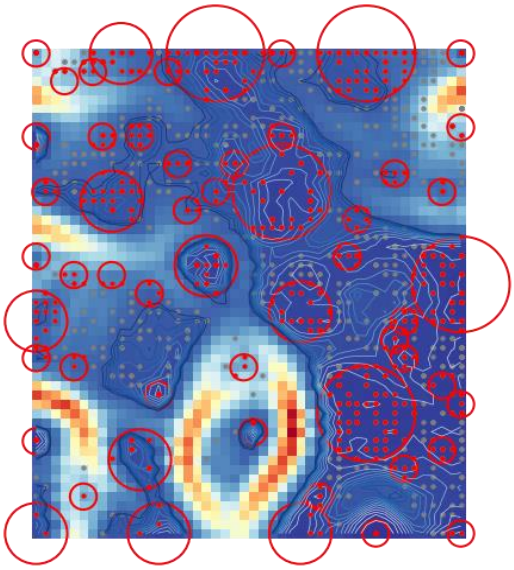


Figure 15 – U-matrix of the best-trained SOM using vector topics from BERTopic.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa