NOVA IMS
Information Management School

MMAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

MAPINTEL: ENHANCING COMPETITIVE INTELLIGENCE ACQUISITION THROUGH EMBEDDINGS AND VISUAL ANALYTICS

David Fontes Henriques Silvestre da Silva

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics

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

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

MAPINTEL: ENHANCING COMPETITIVE INTELLIGENCE ACQUISITION THROUGH EMBEDDINGS AND VISUAL ANALYTICS

by

David Fontes Henriques Silvestre da Silva

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Adviser: Fernando Bação

February, 2022

MapIntel: Enhancing Competitive Intelligence Acquisition Through Embeddings and Visual Analytics

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.5) [1].

To Joana, Carlos, Cristina, and Cecília,
for constantly reminding me life is wider than a computer screen.

Acknowledgements

A big thank you to my adviser, Professor Fernando Bação, for the support and motivation provided. It has been a long journey and without him this work would not have been possible. Thank you for the brainstorming, weekly meetings and countless reviews of this document. I would also like to show my gratitude to the MapIntel team, João Fonseca, Georgios Douzas, and Iñigo de Troya, for the ideas shared and debated.

I also want to thank NOVA IMS for being my home for the last 5 years, where I have grown both personally and professionally, and for allowing me to participate in the higher education of countless students. In addition, I would like to acknowledge *Fundação para a Ciência e Tecnologia* of *Ministério da Ciência e Tecnologia e Ensino Superior* for the support, through a Master grant under the DSAIPA/DS/0116/2019 project. Finally, a big thank you to AICEP - *Agência para o Investimento e Comércio Externo de Portugal* - for promoting this research project.

A word of appreciation to my friends, Pedro, David, Rita, Sofia, Ana, and Marisa, for being by my side during this long quest. You definitely had a part on this work. To my family, who supported me with love and encouragement, my sincerest thank you.

Last but not the least, I should thank my girlfriend Joana, for the love and friendship, and for your distinct ability of cheering me even when everything else isn't going great. This work would not have been possible without your words of motivation and affection.

Abstract

Competitive Intelligence allows an organization to keep up with market trends and foresee business opportunities. This practice is mostly performed by analysts scanning for any piece of valuable information in a myriad of dispersed and unstructured sources. Here we present MapIntel, a system for acquiring intelligence from large collections of text data by representing each document as a multidimensional vector that captures its own semantics. The system is designed to handle complex Natural Language queries and visual exploration of the corpus, potentially aiding overburdened analysts in finding meaningful insights to help decision-making. The *searching* module of the system uses a retriever and re-ranker engine that first finds the closest neighbors to the query embedding, and then sifts the results through a cross-encoder model that identifies the most relevant documents. The *browsing* or visualization module also leverages the embeddings by projecting them onto 2 dimensions while preserving the multidimensional landscape, resulting in a map where semantically related documents form topical clusters which we capture using topic modeling. This map aims at promoting a fast overview of the corpus while allowing a more detailed exploration and interactive information encountering process. We evaluate the system and its components on the 20 newsgroups dataset, making use of the semantic document labels provided, and we demonstrate the superiority of Transformer-based components. Finally, we present a prototype of the system in Python and show how some of its features can be used to acquire intelligence from a news article corpus we collected during a period of 8 months.

Keywords: Natural Language Processing, Transformer Architecture, UMAP, Information Encountering, Information Retrieval, Competitive Intelligence, Sentence Embeddings, Topic Modeling, Unsupervised Learning

Contents

List of Figures

List of Tables

1 Introduction

Competitive Intelligence (CI) is the process and forward-looking practices used in producing knowledge about the competitive environment to improve organizational performance [2]. According to [3], "Companies with competitive intelligence programs have better knowledge of their markets, better cross-functional relationships between their business units and a greater ability to develop proactive competitive strategies." CI has a fundamental role in helping businesses remain competitive, influencing a wide range of decision-making areas, and leading to substantial improvements such as the increase of revenue, new products or services, cost savings, time savings, profit increases, and achievement of financial goals [4].

Competitive Intelligence analysts are responsible for developing the CI task through a combination of gathering data, processing it, and communicating information. The digitalization of the market and the growth of the data economy have pushed the business environment to an online realm where every action and event is public and thus potentially relevant for decision-making. This shift has produced a large volume of data about products, customers, competitors, and any aspect of the business environment that can be used to foresee opportunities and risks. Given the vastness and diversity of this data, it has become a necessity to design tools that can aid analysts in the CI gathering and analysis process. Therefore, the goal is to enhance the analyst's task by providing a tool to explore, organize and visualize the environmental data present in the array of existing sources.

Traditionally, the most important sources of CI have been, respectively, news providers, corporate websites, and trade publications [5]. With the advent of the internet, new sources, such as social networks [6], have emerged, while existing ones have become enriched and easily accessible. Despite the increased availability, CI resources are dispersed through a variety of websites and the underlying data is unstructured and noisy. These characteristics add to the difficulty of the analyst's task and exacerbate the need for tools to support it.

Various studies have attempted to create systems for exploring and gathering intelligence from large collections of textual data [6–10]. These studies have consistently

applied Natural Language Processing (NLP) techniques for helping users comprehend large volumes of text without requiring to sift through every document. [6] designs a system for CI that captures data from multiple sources, cleans it, uses NLP to identify and tag the relevant content, stores it, generates consolidated reports, and produces alerts on pre-defined triggers.

Although the previously mentioned systems have successfully been used for dealing with large amounts of text, insufficient attention has been paid to the exploratory and serendipitous aspects of the analyst's task. Accordingly, we propose an information environment that supports analysts in having stimulating and productive information encounters. Thus, contrarily to previous systems, we center ours around Information Encountering which is defined by [11] to encompass "finding interesting, useful or potentially useful information when looking for some other information, not looking for any information in particular, or not looking for information at all". This is achieved by incorporating two types of information acquisition tasks: *searching*, consisting of an information retrieval module that allows *ad hoc* queries on the entire document collection, giving the user the ability to actively seek information, and *browsing*, consisting of a visualization module that equips the user with tools to actively or passively acquire information through the visual exploration of the document corpus (and its thematic cohorts) in a two-dimensional map.

With the recent emergence of the Transformer architecture [12], significant improvements were made in several NLP subdomains, having reached state-of-the-art results in a wide range of tasks [12]. This new architecture is based solely on the attention mechanism, providing parallelization capabilities and thus avoiding the sequential nature of existing Recurrence models [13, 14]. The attention mechanism allows incorporating information from the input sequence words into the one it is currently being processed, thus providing "context" to the word from the rest of the sequence. Language models like Bidirectional Encoder Representations from Transformers (BERT) [15] leverage this architecture, making up a large part of the modern NLP landscape by providing an off-the-shelf, powerful way to create state-of-the-art models for a wide range of tasks. The Transformer flexibility, allied to its reduced training times and improved ability to learn long-range relationships between terms in a sequence, make it one of the pillars of modern NLP research, and we intend to apply this architecture in our work.

In this thesis, we explore Transformer-based models for representing documents as semantic vectors. These vectorial representations are commonly denominated as embeddings, and we intend to use them in a CI system as a mechanism for extracting information from environmental data. Furthermore, the system facilitates Information Encountering by incorporating *searching* and *browsing* mechanisms that leverage the document embeddings. We name the proposed system as MapIntel from (Competitive) Intelligence Map.

2

Related Work

The process of extracting business-related information for anticipating risks and opportunities is an important task for many companies, yet analysts are overwhelmed with large amounts of unstructured data. To support CI analysts, we propose an NLP system for exploring and gathering intelligence from large collections of textual data. To situate our contribution, we review, in this section, existing work in similar systems applied in CI as well as in other domains.

Early work on visualizing and interpreting large collections of documents can be found in [16], where WEBSOM - a system that organizes a document collection using Self Organizing Maps (SOM) [17], mapping each document into the node of a two-dimensional grid that best represents it, thus providing a reliable and visual representation of the collection - is presented. An improved version of the system is given in [18], where users can perform queries using either a set of keywords or a descriptive sentence, a zooming feature to explore specific regions of the map with finer detail is provided and, when selecting a particular node in the map, the titles of the corresponding documents are displayed. [8] also uses SOM together with Latent Dirichlet Allocation (LDA) [19] to convey the relatedness of research themes in a multidisciplinary university library. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. That said, each document is embedded in a vector space of n dimensions, corresponding to the number of topics selected. SOM produces a landscape for exploring the topic space and provides users with an overview of the document collection and the ability to navigate, discover items of interest, change the level of detail, select individual documents and discover relationships between documents.

Arguably, the closest system to ours in terms of domain application is [6]. They formulate a system for acquiring competitive intelligence from different types of web resources, including social media, using a wide array of text mining techniques. They also show how the system can be integrated with the business data and adopted for future decision-making. Their goal is to help the analyst in the task of reading, extracting information, and organizing the data. The system is composed of four distinct

modules: *content acquisition and assimilation* gathers and extracts relevant content from an array of sources, *data pre-processing* is responsible for cleaning and extracting relevant content from different sources, *content processing* identifies and tags the relevant content from the vast collection, and *alerting* notifies the user on pre-defined triggers such as competitor's product launch. The paper presents an approach for labeling news articles according to CI-related topics by applying LDA clustering and extracting entities and relations within each cluster, which are then used to define the respective labels. The labeling contributes to the organization of the collection and facilitates the information extraction process.

[9] proposes a method for modeling and mapping topics from bibliometric data and builds a web application based on this method. The map produced allows users to read a body of research "at a distance" while providing multiple levels of detail of the documents' topics. They also incorporate a time dimension, allowing users to understand the evolution of the topics over time. They apply Non-negative Matrix Factorization [20] to discover the underlying topics in the data and obtain vectorial representations of the documents, and they employ t-distributed Stochastic Neighbor Embedding (t-SNE) [21] for visualizing the documents, resulting in a two-dimensional representation of the corpus. To allow for different detail levels, the authors produce two maps: a coarse map of 9 topics that gives a general overview of the topics within the data and a detailed map of 36 topics that captures more specific research themes. The web application consists of an interactive dashboard that allows users to explore the map of documents and easily extract information.

We base our *searching* module on the Vector Space Model (VSM) [22, p. 120-126], a common framework in Information Retrieval, consisting of representing a set of documents as vectors in a vector space, while also allowing full-text queries to be represented in the same space. The model then ranks each document in decreasing order of their similarity with the query. The fundamental assumption of the model is that similar documents will be placed close together in the vector space, whereas dissimilar documents will be far away. An application of VSM for medical imaging can be found in [23]. They propose a methodology for Content-based Medical Image Retrieval where Magnetic Resonance Imaging (MRI) images are represented using features such as color, shape, and texture, and the k items with the smallest euclidean distance to a given query image are retrieved. They report that by using this approach they can classify dementia-affected MRI images with an average precision of 97.5% and a maximum recall of 95%, making it an effective way to diagnose new images. [7] presents a different application of VSM for querying COVID-19 literature. They propose Co-Search, an Information Retrieval system that combines semantic search, question answering, and abstractive summarization. The system uses Sentence-BERT (SBERT) [24], a Transformer-based model for representing documents as semantic vectors, combining it with approximate nearest neighbors and cosine similarity to return the relevant results for a query.

A more recent work focusing on the frontier between Computer Graphics and Machine Learning is *Cartolabe* [10]. *Cartolabe* is a web-based, scalable and efficient system for visualizing and exploring large textual corpora, relying on topic modeling algorithms like LDA [19] and LSA [25] to represent documents as vectors of topics and on UMAP [26] to produce a 2-dimensional plane that preserves the original topology and neighborhood of the documents. Additionally, they provide an interactive high-level visualization that allows exploration of the corpus in real-time by offloading most of the computations to the data pre-processing pipeline making the system highly scalable to large collections of documents. We intend to apply the same idea of performing the pre-processing offline to improve the responsiveness of the system and user interaction. Contrarily to *Cartolabe*, we aim to explore Transformer-based embeddings instead of topic vectors due to the novelty aspect of this architecture and the improved results it has shown in multiple benchmarks in other NLP subdomains.

3

MapIntel

We propose MapIntel - Figure 3.1, a system that supports the exploration of a document collection while promoting serendipity and satisfying emerging information needs by allowing full-text queries over the entire collection. The system is scalable to large amounts of data, is dynamic as it regularly integrates new data, and is fast. It is composed of three main pipelines: Indexing, Query, and Visualization which objectives are respectively, to get documents and their metadata from a source to a database, to retrieve the most relevant results to a user query, and to produce an interactive interface for exploring the document collection.

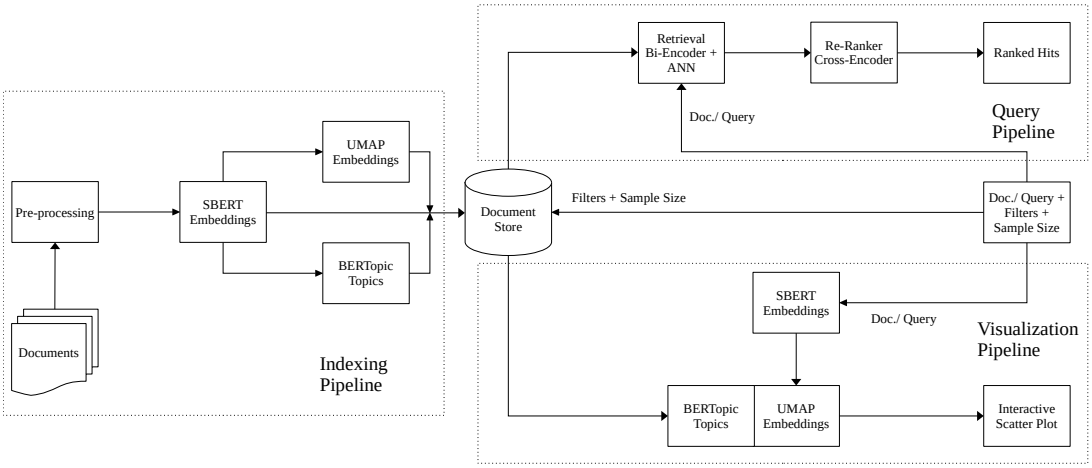


Figure 3.1: **MapIntel architecture**. Composed of three main pipelines: Indexing, Query, and Visualization

3.1 Indexing

In this work we decided to focus on how NLP and particularly sentence embeddings could help in organizing, exploring, and retrieving text documents in the CI domain. Thus, we don't develop as much the precedent tasks of data collection and pre-processing as we believe they are independent of the system and can be easily integrated with it. Nevertheless, it is important to point out that the quality of the

system is extremely reliant on these steps, as if we feed it non-ideal data, we will get non-ideal results. MapIntel is nothing but a set of tools to facilitate the exploration and understanding of a corpus, and it will not give useful insights if the data isn't useful itself.

Once new documents are fed to the system, their respective embeddings are computed. This process is the basis of our work as it allows the encoding of the semantic identity of the document onto a vector of a given dimensionality. This semantic identity describes what is the subject of the document, and it can be used to compare documents between each other *i.e.* documents with the same subject will be close in the semantic space and vice-versa. We use SBERT [24], a derivative of the Transformer-based BERT model, to embed the documents using a pre-trained encoder trained on reducing the distance between queries and relevant results in the MS MARCO dataset [27]. This produces vectors of 768 dimensions, which we then reduce to 2 dimensions using the Uniform Manifold Approximation and Projection (UMAP) [26] algorithm. UMAP constructs a topological representation of the high and low dimensional data and its goal is to minimize their cross-entropy, which measures the difference between the two representations, by adjusting the low-dimensional representation. This is another important component of MapIntel as it allows the organization and localization of the entire document collection in a 2-dimensional map, which can be used to explore and interact with the data. We opted to use UMAP over other dimensionality reduction techniques because of its improved map quality, reduction in time required to produce the output map, support for larger data set sizes, and, most importantly, its ability to update the output map with new data without having to rebuild it [26].

We also apply a topic modeling technique called BERTopic [28], based on the work of [29]. Topic modeling unveils the latent semantic structure of the data and unlike some of the classical techniques such as Latent Dirichlet Allocation [19] and Probabilistic Latent Semantic Indexing [30], BERTopic leverages the SBERT embeddings and their capacity to encode the semantic attributes of a document to find the most representative topics of a corpus. BERTopic clusters the documents using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [31] to find the densest areas of the semantic space while identifying outliers. To overcome the sparsity of the high-dimensional space and the obstacles it creates in finding dense clusters, UMAP is used to reduce the embeddings to a lower dimension (5 dimensions by default) prior to the clustering stage, this UMAP is separate from the one we use to visualize the corpus on 2 dimensions. The main assumption behind BERTopic is that each dense area in the semantic space is generated by a latent topic shared among the documents that comprise it. Finally, a class-based variant of TF-IDF [32] (c-TF-IDF) is used to extract for each cluster an importance value of each word, which can be used to represent each topic as the set of its most important words. Another advantage of BERTopic over the classical approaches is that we can reduce the number of topics obtained by iteratively comparing the c-TF-IDF vectors, merging the least common

topic with its most similar topic, and re-calculate the c-TF-IDF vectors, giving us the option to choose the number of topics.

Finally, we load the documents, their metadata, their SBERT embeddings, their UMAP embeddings, and their topics into a database. We use Open Distro for Elasticsearch[1] — an open-source, RESTful, distributed search and analytics engine based on Apache Lucene[2] — to store the data, organize it in an index and perform full-text search on it. We can think of the approach described as an Indexing Pipeline — Figure 3.2 — that extracts new raw documents from a data source, pre-processes and manipulates them, stores the results in a database, and indexes the documents for future search tasks.

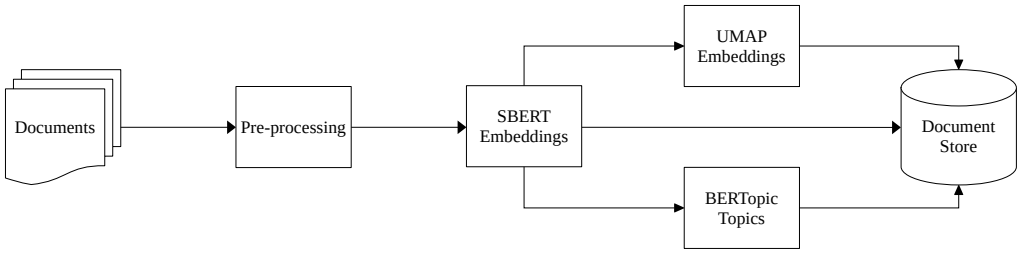


Figure 3.2: **Indexing pipeline**. Documents flow through a pre-processing step, followed by an inference step where their SBERT embeddings, UMAP embeddings and BERTopic topics are computed and indexed into a Database for future search tasks.

3.2 Query

Finding meaningful information within a large amount of data is a big part of the CI task. The ability to retrieve relevant documents from a large collection of news articles through natural language queries empowers the CI analyst with an easy and intuitive interface to scan the environment.

MapIntel provides a search functionality based on Open Distro for Elasticsearch and its *k*-Nearest Neighbor (*k*NN) Search module. By utilizing the *k*NN module, we can leverage the SBERT embeddings by projecting the query string onto the same semantic space as the corpus and computing its *k*-nearest neighbors *i.e.* finding the *k* documents whose embedding vectors are closest to the query embedding vector, according to some pre-defined similarity metric. Since the embedding vectors encode the semantic identity of each document, this method provides semantically relevant results for a given query. Furthermore, the *k*NN module delivers a highly performant and scalable similarity search engine by leveraging Elasticsearch's distributed architecture and by implementing Approximate Nearest Neighbors (ANN) search based on Hierarchical Navigable Small World Graphs [33]. The *k*NN module can also be combined with binary filters that help the user obtain focused results based on characteristics of the

[1] opendistro.github.io/for-elasticsearch
[2] lucene.apache.org

documents such as publication date and topic. These filters are applied directly to the database, reducing the search space as a result and improving the subsequent search time.

Once again, we can think of the search functionality as a pipeline, illustrated in Figure 3.3, where we feed a query string and some binary filters, and we obtain documents ordered by their relevancy to the query. We employ a Retrieve and Re-rank pipeline based on the works of [34, 35] composed by a "Retrieval Bi-Encoder + ANN" node that performs semantic search using Elasticsearch's kNN module as described above, and by a "Re-Ranker Cross-Encoder" node consisting of a BERT model fine-tuned on the MS MARCO dataset that receives a document and query pair as input and predicts the probability of the document being relevant to the query.

The pipeline works by taking advantage of the characteristics of both nodes. The Bi-Encoder together with ANN search can retrieve fairly relevant candidates while dealing efficiently with a large collection of documents. The Cross-Encoder isn't as efficient since it has to be performed independently for each document, given a query. However, since attention is performed across the query and the document, the performance is higher in the second node [36]. Therefore, we combine both nodes by retrieving a large set of candidates from the entire collection using the Bi-Encoder, and by filtering the most relevant candidates with the Cross-Encoder while removing noisy results.

With this pipeline, we can provide relevant documents to the user given a query and binary filters while ranking them according to a relevancy score. The pipeline is efficient and makes use of the SBERT embeddings and the Elasticsearch architecture. As an additional feature, we can input a document instead of a query, allowing us to search for semantically similar documents within the collection.

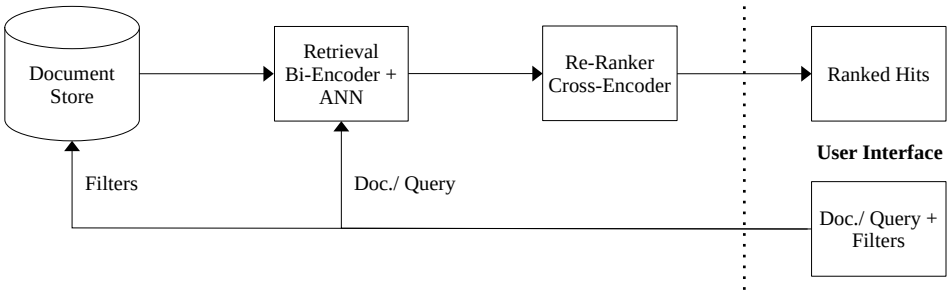


Figure 3.3: **Query pipeline**. Given a user query or selected document, its SBERT embedding is inferred and the closest neighbors are computed, while a set of binary filters is used to reduce the search space. Following, a Cross-Encoder BERT model is used to re-score the neighbors from the previous step, and they are retrieved together with their similarity score.

3.3 Visualization

To facilitate the environment scanning task, we developed a visual interface that organizes and displays the documents, giving the user the ability to browse the data and

zoom on particular regions of the semantic space. The interface uses the UMAP algorithm to reduce the dimensionality of the original semantic space to a 2-dimensional representation that reliably preserves the original topology.

The methodology employed to produce the interface is described in Figure 3.4. It begins by taking the same inputs passed to the Query pipeline: a query, and a set of filters. The common inputs create a connection between the two modules — when the user queries the database, the query text is projected onto the 2-dimensional map and the filters define which documents are displayed in the map. In this way, the map can be seen as a graphical extension of the searching mechanism, where the relevant results reside in the neighborhood of the query, giving the user some insight into how the results are obtained. In addition to the common inputs, we require a relative sample size that defines the percentage of randomly chosen documents (after applying the filters) to be displayed in the map. This is necessary as interaction with the map is hindered by a large number of data points, resulting in a slow and unresponsive experience. Notice that the sample size doesn't affect the query results, as the search is always performed on the entire collection.

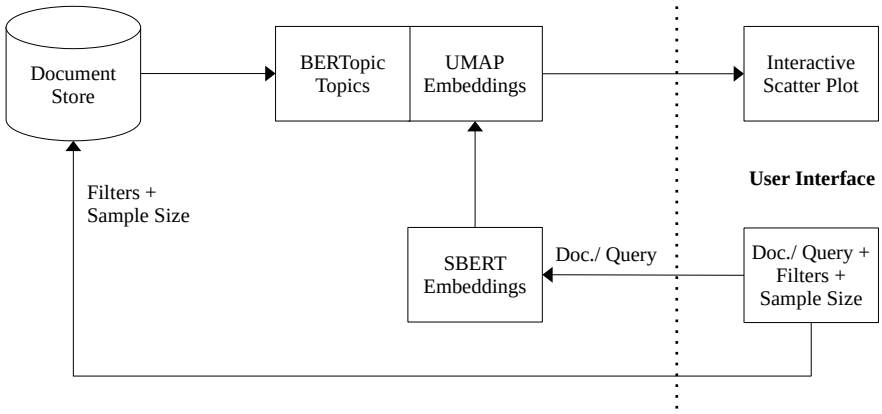


Figure 3.4: **Visualization pipeline**. Given a user query or selected document, its SBERT and UMAP embedding are inferred, while a set of binary filters and a sample size is used to reduce the amount of documents retrieved. Finally, the selected documents' topics and 2-dimensional embeddings, together with the query embedding, are displayed in an interactive scatter plot.

To produce the interactive scatter plot, the filters and sample size are used to select the documents to be displayed from the database. We compute the SBERT embedding of the query, followed by its UMAP embedding, thus being able to locate the query in the same space as the documents. An advantage of these two models is that we can efficiently produce embeddings of new text without having to re-train them, making this process quite fast. Once we have the UMAP embedding of the query, we join it with the pre-computed embeddings of the documents from the Indexing pipeline, and we produce the interactive map.

The map provides a means to explore the documents and the different semantic

cohorts present within the collection. We color-code the points with the documents' topics identified in the Indexing stage, allowing us to visualize the latent semantic structure of the data, and when hovered, the points display their corresponding title and content attributes.

4

Evaluation

Our methodology addresses the issues of information dispersion and overload impacting the CI analysts' task. The proposed system provides searching and browsing capabilities, contributing to an easier understanding of the business environment by supporting analysts in seeking specific information, while promoting undirected information encountering. In this section, we elaborate our choices in the design of the MapIntel system with the results of our experiments and analyze the different components of the system individually.

4.1 Experimental Setup

We evaluate our system quantitatively using the 20 newsgroups [37] dataset and the document labels provided. This dataset consists of around 18,000 newsgroups posts on 20 topics divided into 6 main groups: "Computer", "Recreation", "Science", "Miscellaneous", "Politics" and "Religion". We opted to use this dataset because of the presence of labels that describe the semantic meaning of each document, allowing us to have a reference which we can compare the identified topics with.

Given the inherent difficulty of evaluating the system on its entirety, we decided to deal with each component separately, however since every component of our system depends on the vector representation of the documents, we cannot guarantee an orthogonal evaluation of the components. We focus our experiments in comparing 2 of the main components of the MapIntel system: the Topic Model and the Sentence Embedding. The following algorithms are compared in this thesis:

- **Sentence embeddings**

    1. **Paragraph Vector Model (or Doc2Vec)** [38]: an unsupervised algorithm that learns both word and document vectors by minimizing the error of predicting the next word in a paragraph (a variable-length piece of text) given the paragraph and previous word vectors;

2. **SBERT**: a derivative of the Transformer-based BERT model, to embed the documents using a pre-trained encoder trained on reducing the distance between queries and relevant results in the MS MARCO dataset;

- **Topic modeling**

1. **LDA**: a generative probabilistic model of a corpus in which each document is modeled as a finite mixture over an underlying set of topics, and each topic is characterized by a distribution over words;
2. **BERTopic**: a cluster-based topic model that unveils the latent document-topic distribution responsible for the existing groups of documents in a semantic vector space;
3. **Contextualized Topic Model (CTM)** [39]: combines contextualized representations (like SBERT) with neural topic models resulting in more meaningful and coherent topics;

We use three main metrics to guide our model comparison:

- **kNN classifier accuracy for the UMAP projections**: evaluate the quality of the two-dimensional projections by inferring their labels using a kNN classifier and computing its Accuracy for multiple values of k [26]. We present the average Accuracy over the range of k values we tried: 10, 20, 40, 80, 160.

- **Normalized Mutual Information (NMI) for the topic assignments**: information theoretic based measure that tells us how much knowing about the topics of the documents reduces our uncertainty about the original labels and vice-versa [40]. The NMI ranges between 0 and 1 where the former corresponds to no mutual information, and the latter indicates perfect correlation. The higher the value of this metric, the better we can capture the true topical nature of the documents, reflected by their labels, with the assigned topics.

- **Topic coherence**: measure the quality of the words that describe each topic by applying the C_v metric [41] indicating whether the words that compose a given topic support each other. This metric is shown to be correlated with human ratings on understandability of topic descriptions, given as word sets, and it ranges between -1 and 1, where the former corresponds to incoherent topic descriptions, and the latter indicates a coherent topic description.

We perform hyperparameter tuning using a multi-objective approach to optimize the three metrics specified previously. We use the Tree-structured Parzen Estimator (TPE) algorithm [42, 43] for sampling the hyperparameter space at each trial t of the optimization process. Contrarily to random search, TPE samples values x_t^α for each hyperparameter α by fitting one Gaussian Mixture Model (GMM) $l(x^\alpha)$ to the set of

values associated with the best objective scores in past observed trials, and another GMM $g(x^\alpha)$ to the remaining values. It then chooses the hyperparameter value x_t^α drawn from $l(x^\alpha)$ that maximizes the ratio $\frac{l(x^\alpha)}{g(x^\alpha)}$. For each trial, we evaluate the sampled hyperparameters using a 5-fold cross-validation approach where the folds preserve the percentage of samples of each class. In total, 100 trials were evaluated.

4.2 Results

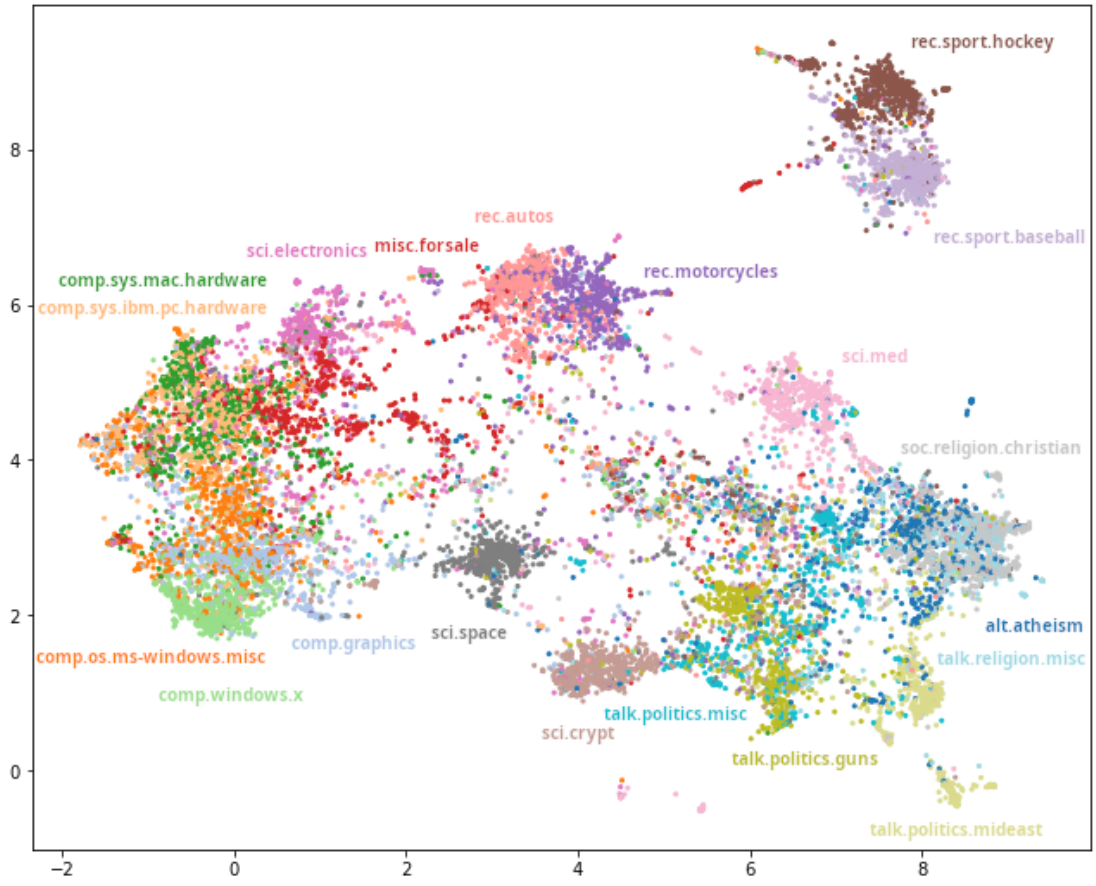
Our results based on the setup described above are shown in Table 4.1. For each trial, we report the average results and standard deviations over the cross-validation folds. The table contains the best trials for each of the Topic/Embedding model combinations according to the average of the three objective metrics, which we applied MinMax scaling to avoid any impact of the metrics scale on the choice of the best model. We can see that the combination that uses BERTopic and SBERT outperform the others with respect to both NMI and C_v while having a within standard deviation kNN Classifier Accuracy to the best value. Another interesting observation is that combinations using SBERT have generally better results. To facilitate results reproduction efforts, we open-sourced the code developed for the experiments at github.com/NOVA-IMS-Innovation-and-Analytics-Lab/mapintel_research.

Topic Model	**Embedding Model**	MinMax Average	NMI	Topic Coherence C_v	kNN Classifier Accuracy
BERTopic	Doc2Vec	0.50	0.11 ± 0.01	0.72 ± 0.02	0.16 ± 0.01
BERTopic	SBERT	**0.94**	**0.36 ± 0.03**	**0.76 ± 0.08**	0.36 ± 0.01
CTM	Doc2Vec	0.56	0.23 ± 0.02	0.55 ± 0.02	0.24 ± 0.03
CTM	SBERT	0.70	0.33 ± 0.02	0.58 ± 0.02	0.28 ± 0.04
LDA	Doc2Vec	0.58	0.25 ± 0.03	0.53 ± 0.03	0.25 ± 0.01
LDA	SBERT	0.71	0.26 ± 0.03	0.53 ± 0.03	**0.37 ± 0.05**

Table 4.1: Hyperparameter tuning best trials per topic and embedding model according to the MinMax average of the multiple objective metrics.

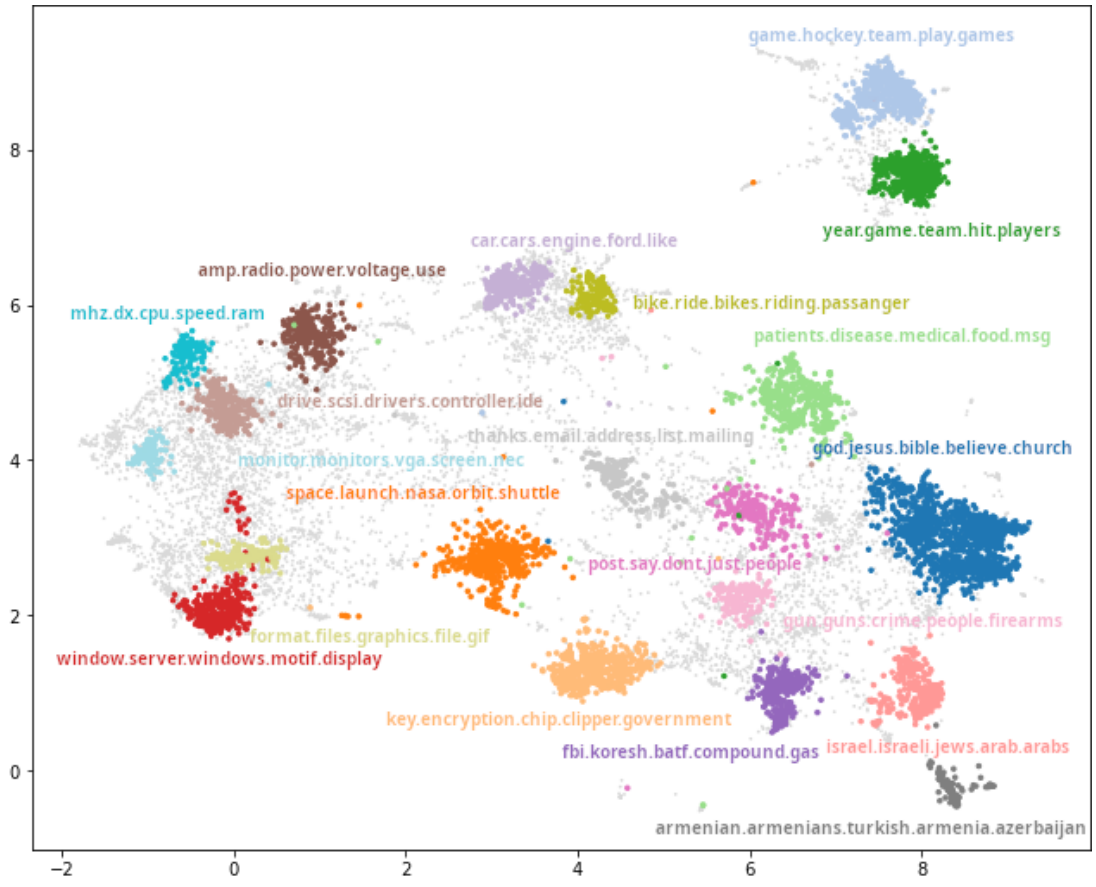
Additionally, we present the UMAP 2-dimensional maps of the documents in the 20 newsgroups dataset. Figure 4.1 shows the comparison between the distribution of the original labels and the topics assigned by the best performing model according to the MinMax average score for the train data (see Table B.1 for exact hyperparameter values). Likewise, Figure 4.2 shows the same comparison for the test data and demonstrates the ability of the model to generalize to unseen samples. We can see that the identified topical cohorts are mostly matching with the original groups, indicating that the embeddings have learned the original labels in a fully unsupervised way. Additionally, it is possible to see that semantic similar topics are located close to each other in the map. This is the case of all the computation related topics such as *window.server.windows.motif.display* and *format.files.graphics.file.gif*. Finally, there is also an agreement between the topic meaning given by the top 5 words describing the topics and the original label description. For example, the same points that have the label *sci.space* also have the topic *space.launch.nasa.orbit.shuttle*. Figure A.1 in the appendix

shows how much correspondence there is both in terms of labels and positioning for technological-related documents - we can see how the distribution of the original label *comp.graphics* matches the distribution of the topic *format.files.graphics.file.gif*, likewise *comp.windows.x* matches *window.server.windows.motif.display*, and *sci.electronics* matches *amp.radio.powed.voltage*
.use, whereas the remaining original labels are lost, giving birth to new, and more semantically coherent, groups of documents (note how the distribution of the original label *comp.sys.ibm.pc.hardware* is very dispersed, even overlapping with other groups and how the distribution of the topics occupying the same region is very clear and compact).



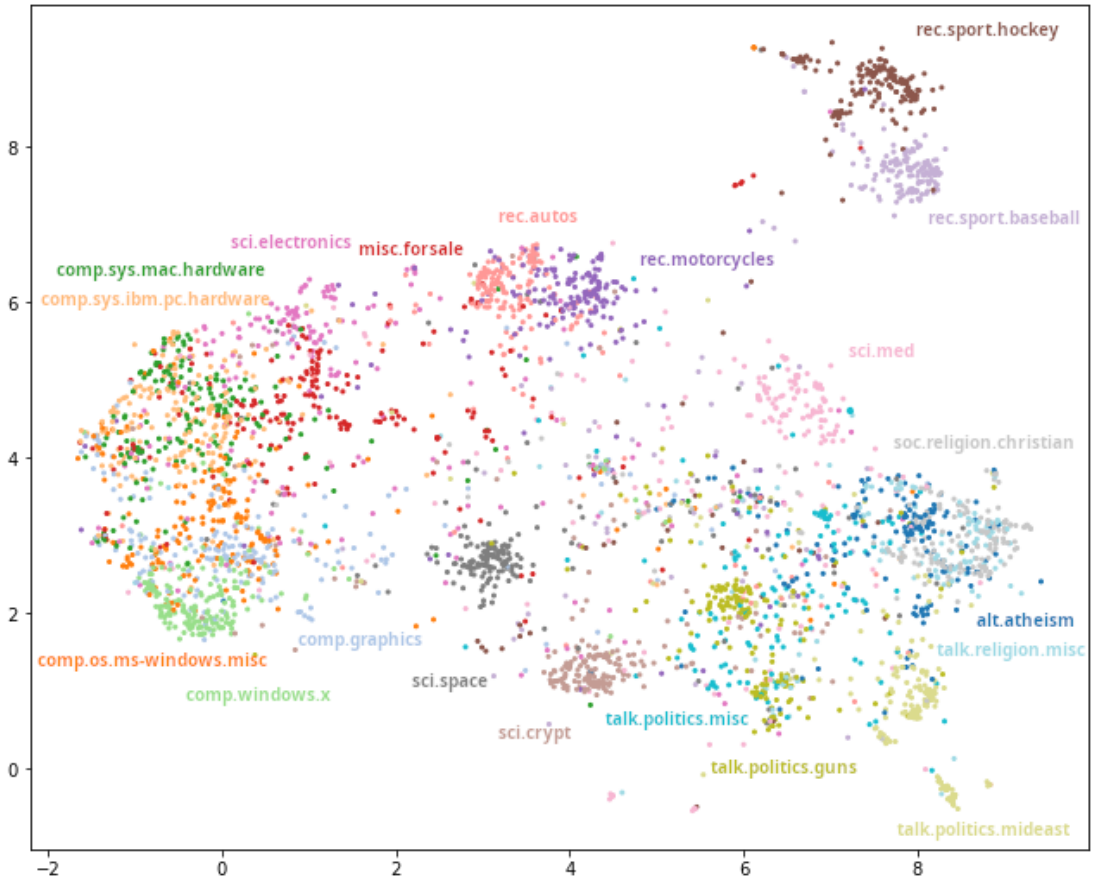
(a) Train Data Original labels

An important characteristic of BERTopic is that it is able to identify noise, leading to a topic assignment where part of the observation are classified as outliers. This produces a cleaner map to explore the documents at the loss of samples that are not given a topic. In Figure 4.1b the percentage of documents classified into the aforementioned category is 51.4% - these are the light grey points scattered across the map.

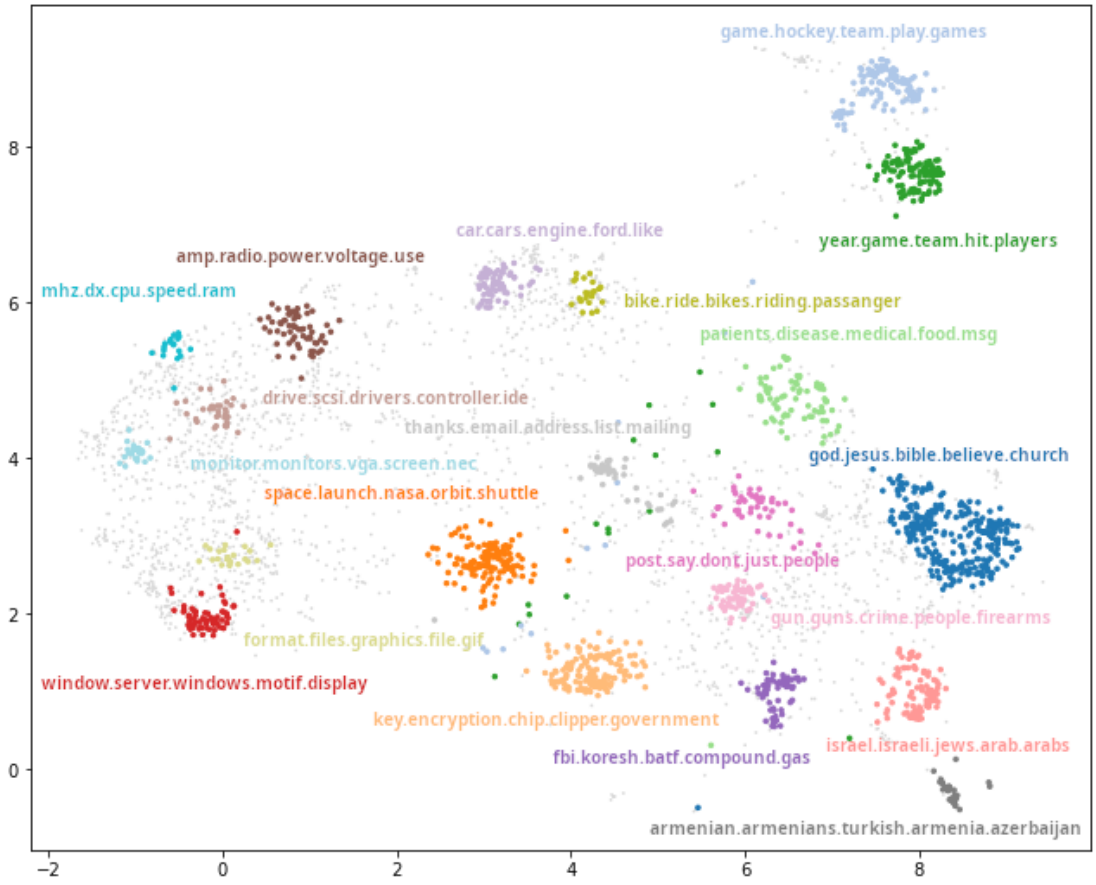


(b) Train Data Topic labels

Figure 4.1: Comparison between UMAP planes of **train data** with original (a) and topic labels (b).



(a) Test Data Original labels



(b) Test Data Topic labels

Figure 4.2: Comparison between UMAP planes of **test data** with original (a) and topic labels (b).

5

Case Study

This use case is based on a real-world example, related with the recurrent tasks that the intelligence analysts at AICEP - Portuguese Trade & Investment Agency have to perform. AICEP "is a government business entity, focused in encouraging the best competitive foreign companies to invest in Portugal and in contributing to the success of Portuguese companies abroad in their internationalization processes or export activities." Given this mandate AICEP needs to monitor the world news and make sure it is updated with world events. Additionally, analysts at AICEP are called upon to produce reports on specific markets and industries, in order to guide the Portuguese state's diplomatic efforts and private company investments. These reports are preformatted, however the use and integration of text data remains a challenge. This is unfortunate as recent news and text documents allows a more detailed understanding of specific problems and nuances that are difficult to capture in an alphanumeric data table. The objective of this use case is to show how MapIntel can be used to mitigate this limitation and allow the inclusion of relevant text data in the report.

In this case study, we will show the most important tools that the MapIntel system offers to explore and select relevant documents to include in a market or industry report. These tools allow the analyst to screen the news and select those that are relevant to include in the report. This process cannot be strictly seen as an information retrieval process, in fact it is much more an exploration process, in which the analyst's criterion plays a crucial role. The process usually starts with a natural language query that the analyst defines based on his/her knowledge of the market or industry. The query is used by the system as the seed in the process of exploring the corpora. The query is projected onto the two-dimensional surface produced by MapIntel, and the neighbor documents are defined in the semantic space, thus generating a number of candidate results relevant for the report. From this starting position the analyst can interact visually with the results and the entire corpus, and guide his/her interest with successive queries, focusing in the most interesting/promising instances to include in the report.

To evaluate the MapIntel system with real data, we gathered news articles from

multiple international sources with the use of an API[1]. As already stated, there are multiple sources of CI, and different information can be obtained from these. [6] shows in Table 5.1 what kind of information can be acquired from these sources, particularly the ones that are easily available through the web. We decided to work mainly with news articles as they provide a general and accessible means of information about the environment, however, our methodology is easily extensible to data from different sources.

Type of Competitive Intelligence	**Web Resources**
People event	News, company web-sites
Competitor strategies, Technology investment, etc.	News, Discussion forum, Blogs, Patent search sites
Consumer sentiments	Review sites, Social networking sites
Promotional events and pricing	Social networking sites
Related real-world events	News, Social networking sites

Table 5.1: Competitive intelligence resources on the web [6].

The API employed retrieves news articles and their metadata, including attributes such as source, author, title, description, content, category, URL, and publication date and time. We used this API to feed the system with updated data on a schedule while focusing on articles written in English from a set of predefined categories (business, entertainment, general, health, science, sports, technology). The system was fed with a total of 334,925 articles during the period between October 2020 and June 2021.

Due to API limitations, the retrieved data has its content truncated to 200 characters. To overcome this, we treat a single document unit as the concatenation of title, description, and content, providing us a semantically loaded piece of text that we can use for testing the system. Despite this limitation, we give the user the possibility of accessing the full article through its URL. We ensure that each document is unique, is written in English, and doesn't have any HTML tags or any strange pattern.

Once the data is cleaned, it follows the Indexing pipeline 3.2, so it can be used downstream by the Query pipeline 3.3 and the Visualization pipeline 3.4.

[1] newsapi.org

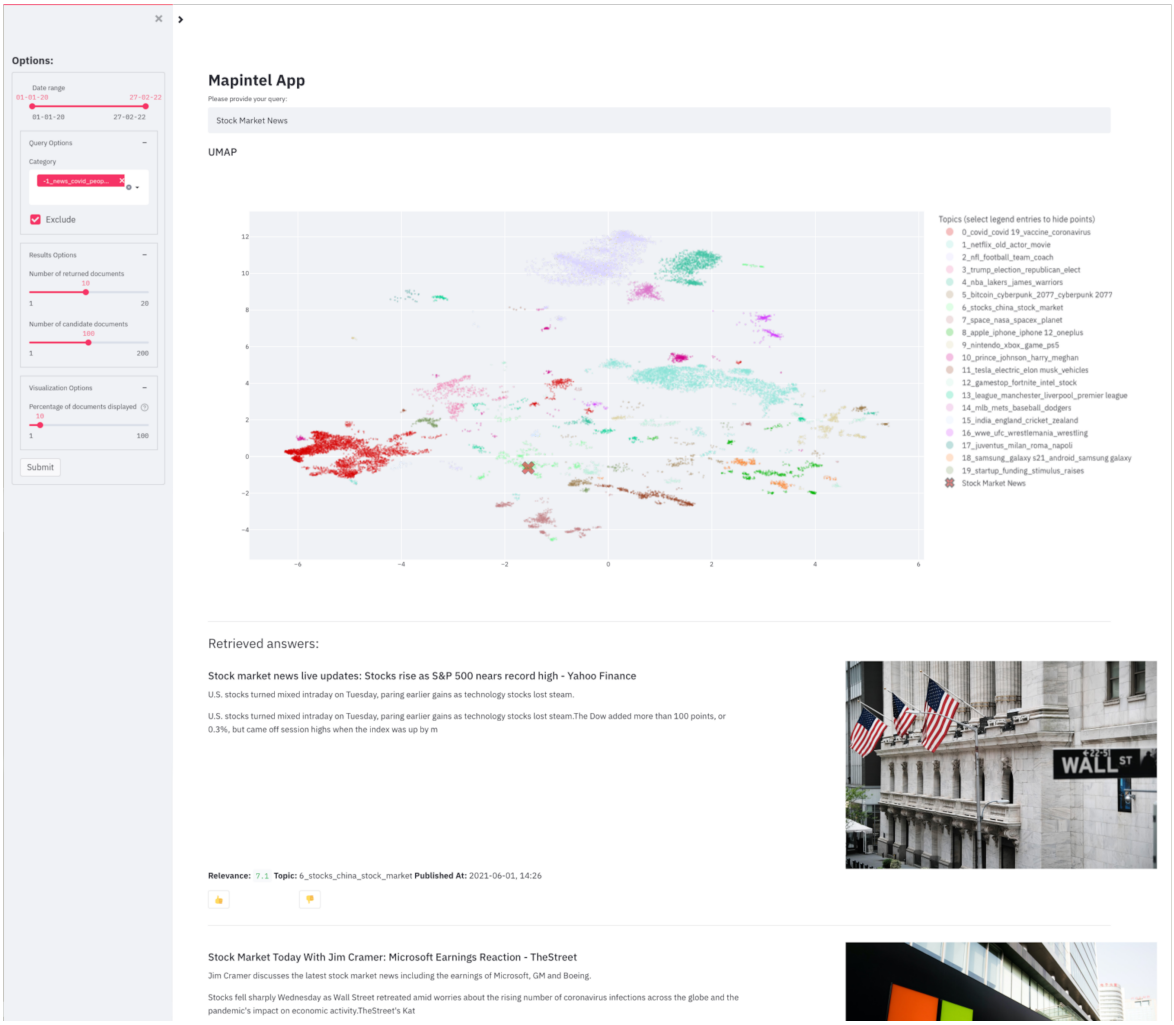


Figure 5.1: **MapIntel webapp interface**. Composed of four main components: search box, interactive scatter plot, parameter sidebar, and list of results.

We also built a simple web application in Python based on the MapIntel system. The web application is containerized and composed of three main modules: an Elasticsearch instance container that stores and indexes the documents, their metadata, their topics and their embeddings, a container responsible for all the computations involved in the three pipelines of the system that communicates with the user interface through a FastAPI[2] server providing the necessary endpoints, and a third container consisting of a Streamlit[3] application that interacts with the data through HTTP requests. For transparency, the code used for the web application is made public at github.com/NOVA-IMS-Innovation-and-Analytics-Lab/mapintel_project.

We highlight the user interface of the app in Figure 5.1. It is composed of 4 main components: a **search box** at the top of the page that the user can use to write any natural language query, an **interactive scatter plot** that shows the UMAP projection of

[2] fastapi.tiangolo.com
[3] streamlit.io

the document embeddings and allows the exploration of the corpus, a **sidebar** at the left where the user can specify any additional parameter or binary filter to the query and map, and a **list of results** at the bottom that contains the documents with the highest scoring for the specified query.

The app enables information searching by receiving a query through the search box together with the filters and parameters selected in the sidebar. This allows the CI analyst to write queries in Natural Language while having a finer control of the results through sidebar parameters like including or excluding specific topics and specifying a date range. We show in Table 5.2 some query examples (without any filters) a CI analyst could make and their respective results. We can see that the results are semantically relevant to the query provided.

Query	Top 3 Results (Title)	Score
"Suez Canal World Trade"	Suez Canal blockage felt across the world as trade comes to a pause.	4.90
	Suez Canal ship partially refloated after huge effort to unblock key global trade route.	4.70
	Egypt 'seizes' container ship over $1bn Suez claim.	4.42
"Indian Elections"	India elections: Modi party defeated in West Bengal battleground - BBC News.	5.36
	At 103, India's first voter casts vote in Himachal panchayat radish polls.	4.77
	Vote count in five Indian states under way as pandemic rages.	4.44
"Oil Prices"	Oil prices near 2-year highs above $70 as investors expect OPEC+ to confirm its supply policies.	7.63
	Oil Prices Rally Towards $70 As Demand Outlook Improves - OilPrice.com.	7.62
	Oil prices to reach $72 by summer: Goldman Sachs - Fox Business.	7.45
"Myanmar Coup"	Myanmar Coup: With Aung San Suu Kyi Detained, Military Takes Control - NPR.	8.03
	Myanmar's military has detained leader Aun San Suu Kyi in a coup. Here's what you need to know - CNN.	8.03
	News24.com | Myanmar military stages coup and declares state of emergency, detains leader Aung San Suu Kyi.	7.94

Table 5.2: Example queries and their top 3 results and respective relevancy scores. The results are computed according to the methodology described in 3.2.

An additional feature of the app is the ability to browse documents through the scatter plot. This visualization shows the 2-dimensional UMAP projection of the 768-dimensional embeddings of the documents. It can be used by the analyst to explore the document collection and the underlying thematic groups. The map can be interacted with by zooming, panning and hovering specific regions and documents, allowing

the analyst to see the content of each data point. A representation of the hovering capability and a more detailed image of the document map can be seen in Figure 5.2. In addition, we integrate the *searching* and *browsing* modules of the system by plotting the query embedding in the UMAP plane - represented with an "X" shaped marker -, allowing a visual exploration of the relevant results *i.e.* the data points close to the query. Finally, we allow searching of similar documents by providing the option to click on any point in the map and obtain their most similar documents in the list of results below. We found that this feature enhances exploration and finding of relevant documents.

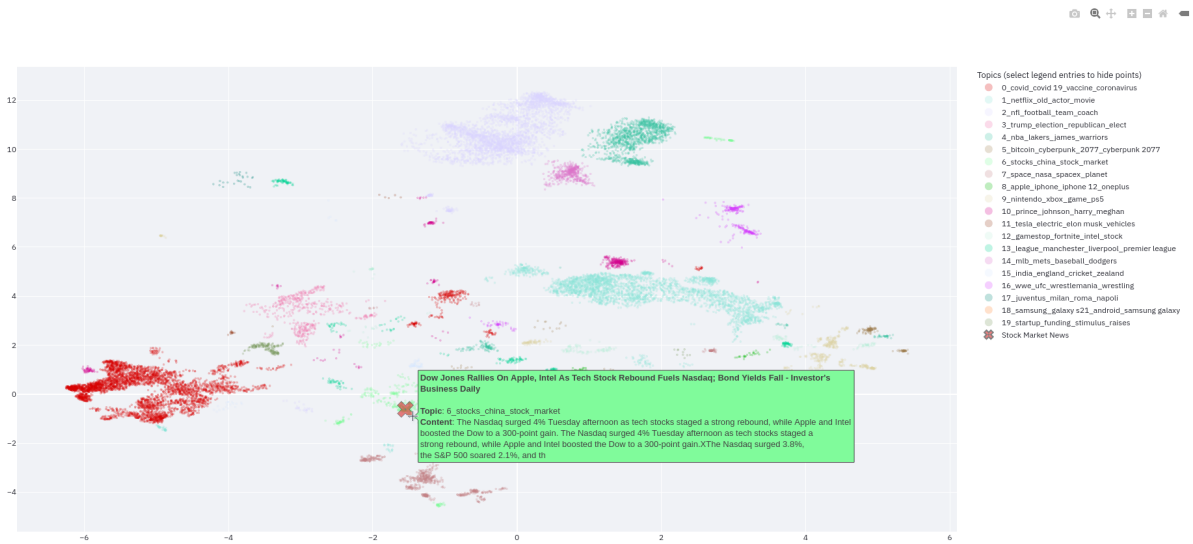


Figure 5.2: **UMAP projection of the corpus**. The documents' SBERT embeddings are projected on two dimensions with UMAP and the data points are colored according to their assigned topic. The map can be used interactively by zooming, panning, hovering, and selecting points. On hovering, the document text and some metadata is displayed and on selection the document is queried against the rest of the corpus. The query is marked with an "X" in the map.

Besides the searching and browsing functionality, the system can also organize the documents into semantically similar cohorts and automatically label them. This is achieved with the help of topic modeling - particularly BERTopic - and the resulting topic labels can be seen in Table 5.3. The topic labels can be particularly useful to the analyst to understand the different subjects covered by the corpus at a glance. In addition, by looking at Figure 5.2 we can notice that these topics represent clusters of documents in the semantic space, which confirms the main assumption that semantically similar documents have similar embeddings.

Topic Number	Topic Words	Topic Coherence
−1	- (outliers)	-
1	covid, covid 19, vaccine, coronavirus	0.644
2	netflix, old, actor, movie	0.223
3	nfl, football, team, coach	0.359
4	trump, election, republican, elect	0.12
5	nba, lakers, james, warriors	1
6	bitcoin, cyberpunk, 2077, cyberpunk 2077	0.336
7	stocks, china, stock, market	0.413
8	space, nasa, spacex, planet	0.314
9	apple, iphone, iphone 12, oneplus	1
10	nintendo, xbox, game, ps5	1
11	prince, johnson, harry, meghan	1
12	tesla, electric, elon musk, vehicles	0.943
13	gamestop, fortnite, intel, stock	0.212
14	league, manchester, liverpool, premier leagu	1
15	mlb, mets, baseball, dodgers	0.356
16	india, england, cricket, zealand	1
17	wwe, ufc, wrestlemania, wrestling	1
18	juventus, milan, roma, napoli	1
19	samsung, galaxy s21, android, samsung galax	1
20	startup, funding, stimulus, raises	0.617

Table 5.3: Topic labels and respective coherence values. We used the 5 words with the highest c-TF-IDF score per topic to label them and extracted the coherence value C_v of these words.

6

Conclusion

In this thesis, we present MapIntel, a new system for extracting knowledge from large corpora of text documents. MapIntel differentiates from previous systems in that it leverages Transformer-based document embeddings to provide efficient, natural language searching of documents. The use of Transformer-based embeddings allows to harness the semantic attributes of the documents, which can then be explored in a 2-dimensional map, produced using UMAP. Additionally, MapIntel also organizes the documents in topical cohorts, providing yet another framework for the interaction of the user with the corpus. The system is centered around the concept of Information Encountering [11], providing *browsing* and *searching* capabilities to acquire information and promote serendipity. MapIntel is aimed at supporting Competitive Intelligence analysts by providing a tool that facilitates the exploration and monitoring of the competitive environment from textual data.

We detailed the methodology proposed, having evaluated it through a well-defined experimental setup. Furthermore, we showed how the MapIntel system can be used in a real-world case and we developed a web application, applying the proposed methodology, while enhancing the interaction with the underlying data through an interactive scatter plot. Finally, we developed and open-sourced the code base of the web application and our experiments, so the work can be easily reproducible and continued.

Our next steps will be to perform a more extensive evaluation of the system with new corpora and develop a case study more closely with CI analysts from AICEP to better understand their needs. Some different directions would be to expand the system to different application domains to test its generality, include multilingual text documents to more easily monitor international events, and provide a way to interact with subsets of documents through manual selection in the interactive map.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on p. iii).

[2] L. Madureira, A. Popovič, and M. Castelli. "Competitive Intelligence: A Unified View and Modular Definition". en. In: *Technological Forecasting and Social Change* 173 (2021-12), p. 121086. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2021.121086 (cit. on p. 1).

[3] S. Brod. "Competitive Intelligence: Harvesting information to compete and market intelligently". In: *Camares Communications, New York, NY* (1999) (cit. on p. 1).

[4] J. Calof, N. Sewdass, and R. Arcos. *Competitive Intelligence: A 10-Year Global Development*. Tech. rep. Competitive Intelligence Foundation, 2017 (cit. on p. 1).

[5] J. Marin and A. Poulter. "Dissemination of Competitive Intelligence". en. In: *Journal of Information Science* 30.2 (2004-04), pp. 165–180. ISSN: 0165-5515. DOI: 10.1177/0165551504042806 (cit. on p. 1).

[6] L. Dey et al. "Acquiring Competitive Intelligence from Social Media". In: *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*. MOCR_AND '11. New York, NY, USA: Association for Computing Machinery, 2011-09, pp. 1–9. ISBN: 978-1-4503-0685-0. DOI: 10.1145/2034617.2034621 (cit. on pp. 1–3, 22).

[7] A. Esteva et al. "CO-Search: COVID-19 Information Retrieval with Semantic Search, Question Answering, and Abstractive Summarization". In: *arXiv:2006.09595 [cs]* (2020-06). arXiv: 2006.09595 [cs] (cit. on pp. 1, 4).

[8] S. Lafia, C. Last, and W. Kuhn. "Enabling the Discovery of Thematically Related Research Objects with Systematic Spatializations". In: *14th International Conference on Spatial Information Theory (COSIT 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019 (cit. on pp. 1, 3).

[9] S. Lafia et al. "Mapping Research Topics at Multiple Levels of Detail". en. In: *Patterns* 2.3 (2021-03), p. 100210. ISSN: 2666-3899. DOI: 10.1016/j.patter.2021.100210 (cit. on pp. 1, 4).

[10] P. Caillou et al. "Cartolabe: A Web-Based Scalable Visualization of Large Document Collections". In: *IEEE Computer Graphics and Applications* 41.2 (2021-03), pp. 76–88. ISSN: 1558-1756. DOI: 10.1109/MCG.2020.3033401 (cit. on pp. 1, 5).

[11] S. Erdelez and S. Makri. "Information Encountering Re-Encountered: A Conceptual Re-Examination of Serendipity in the Context of Information Acquisition". In: *Journal of Documentation* 76.3 (2020-01), pp. 731–751. ISSN: 0022-0418. DOI: 10.1108/JD-08-2019-0151 (cit. on pp. 2, 27).

[12] A. Vaswani et al. "Attention Is All You Need". In: *arXiv:1706.03762 [cs]* (2017-12). arXiv: 1706.03762 [cs] (cit. on p. 2).

[13] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997-11), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735 (cit. on p. 2).

[14] K. Cho et al. "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014-10, pp. 1724–1734. DOI: 10.3115/v1/D14-1179 (cit. on p. 2).

[15] J. Devlin et al. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv:1810.04805 [cs]* (2019-05). arXiv: 1810.04805 [cs] (cit. on p. 2).

[16] S. Kaski et al. "WEBSOM–Self-Organizing Maps of Document Collections". In: *Neurocomputing* 21.1-3 (1998), pp. 101–117 (cit. on p. 3).

[17] T. Kohonen. "Self-Organized Formation of Topologically Correct Feature Maps". In: *Biological cybernetics* 43.1 (1982), pp. 59–69 (cit. on p. 3).

[18] T. Kohonen. "Essentials of the Self-Organizing Map". In: *Neural networks* 37 (2013), pp. 52–65 (cit. on p. 3).

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022 (cit. on pp. 3, 5, 8).

[20] D. D. Lee and H. S. Seung. "Learning the Parts of Objects by Non-Negative Matrix Factorization". en. In: *Nature* 401.6755 (1999-10), pp. 788–791. ISSN: 1476-4687. DOI: 10.1038/44565 (cit. on p. 4).

[21] L. Van der Maaten and G. Hinton. "Visualizing Data Using T-SNE." In: *Journal of machine learning research* 9.11 (2008) (cit. on p. 4).

[22] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to Information Retrieval*. Vol. 39. Cambridge University Press Cambridge, 2008 (cit. on p. 4).

[23] N. Sampathila, Pavithra, and R. J. Martis. "Computational Approach for Content-Based Image Retrieval of K-Similar Images from Brain MR Image Database". en. In: *Expert Systems* n/a.n/a (2020), e12652. ISSN: 1468-0394. DOI: 10.1111/exsy.12652 (cit. on p. 4).

[24] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks". In: *arXiv:1908.10084 [cs]* (2019-08). arXiv: 1908.10084 [cs] (cit. on pp. 4, 8).

[25] S. Deerwester et al. "Indexing by Latent Semantic Analysis". In: *Journal of the American society for information science* 41.6 (1990), pp. 391–407 (cit. on p. 5).

[26] L. McInnes, J. Healy, and J. Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv:1802.03426 [cs, stat]* (2020-09). arXiv: 1802.03426 [cs, stat] (cit. on pp. 5, 8, 14).

[27] P. Bajaj et al. "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset". In: *arXiv:1611.09268 [cs]* (2018-10). arXiv: 1611.09268 [cs] (cit. on p. 8).

[28] M. Grootendorst. *BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.* Version v0.7.0. 2020. DOI: 10.5281/zenodo.4381785. URL: https://doi.org/10.5281/zenodo.4381785 (cit. on p. 8).

[29] D. Angelov. "Top2Vec: Distributed Representations of Topics". In: *arXiv:2008.09470 [cs, stat]* (2020-08). arXiv: 2008.09470 [cs, stat] (cit. on p. 8).

[30] T. Hofmann. "Probabilistic Latent Semantic Indexing". In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1999, pp. 50–57 (cit. on p. 8).

[31] L. McInnes, J. Healy, and S. Astels. "hdbscan: Hierarchical density based clustering". In: *The Journal of Open Source Software* 2.11 (2017), p. 205 (cit. on p. 8).

[32] K. S. Jones. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". In: *Journal of documentation* (1972) (cit. on p. 8).

[33] Y. A. Malkov and D. A. Yashunin. "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs". In: *arXiv:1603.09320 [cs]* (2018-08). arXiv: 1603.09320 [cs] (cit. on p. 9).

[34] R. Nogueira and K. Cho. "Passage Re-Ranking with BERT". In: *arXiv:1901.04085 [cs]* (2020-04). arXiv: 1901.04085 [cs] (cit. on p. 10).

[35] B. Kratzwald, A. Eigenmann, and S. Feuerriegel. "RankQA: Neural Question Answering with Answer Re-Ranking". In: *arXiv:1906.03008 [cs]* (2019-08). arXiv: 1906.03008 [cs] (cit. on p. 10).

[36] S. Humeau et al. "Poly-Encoders: Transformer Architectures and Pre-Training Strategies for Fast and Accurate Multi-Sentence Scoring". en. In: (2019-04) (cit. on p. 10).

[37] F. Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* (2011) (cit. on p. 13).

[38] Q. Le and T. Mikolov. "Distributed Representations of Sentences and Documents". In: *International Conference on Machine Learning*. PMLR, 2014, pp. 1188–1196 (cit. on p. 13).

[39] F. Bianchi, S. Terragni, and D. Hovy. "Pre-Training Is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence". In: *arXiv:2004.03974 [cs]* (2021-06). arXiv: 2004.03974 [cs] (cit. on p. 14).

[40] N. X. Vinh, J. Epps, and J. Bailey. "Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?" In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009-06, pp. 1073–1080. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553511 (cit. on p. 14).

[41] M. Röder, A. Both, and A. Hinneburg. "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. New York, NY, USA: Association for Computing Machinery, 2015-02, pp. 399–408. ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2685324 (cit. on p. 14).

[42] J. Bergstra et al. "Algorithms for Hyper-Parameter Optimization". In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. Red Hook, NY, USA: Curran Associates Inc., 2011-12, pp. 2546–2554. ISBN: 978-1-61839-599-3 (cit. on p. 14).

[43] Y. Ozaki et al. "Multiobjective Tree-Structured Parzen Estimator for Computationally Expensive Optimization Problems". In: *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*. GECCO '20. New York, NY, USA: Association for Computing Machinery, 2020-06, pp. 533–541. ISBN: 978-1-4503-7128-5. DOI: 10.1145/3377930.3389817 (cit. on p. 14).

A

Zoom on UMAP projection

(a) Train Data Original labels

(b) Train Data Topic labels

Figure A.1: **Zoom on technological region of UMAP**. Comparison between UMAP planes of **train data** with original (a) and topic labels (b).

B

Best hyperparameter configuration

Model	Hyperparameter	Value
-	embedding_model	sentence-transformers/msmarco-distilbert-base-v4
-	topic_model	BERTopic
HDBSCAN	cluster_selection_epsilon	0.05516517617078291
HDBSCAN	cluster_selection_method	leaf
HDBSCAN	min_cluster_size	110
BERTopic	min_topic_size	33
UMAP	metric	cosine
UMAP	n_components	10
UMAP	n_neighbors	14

Table B.1: Hyperparameter values with the highest MinMax average of NMI, Topic Coherence C_v, and kNN Classifier Accuracy for training data according to Table 4.1.