Comparative Analysis of Generative and Neural Topic Modeling Techniques Applied to Modern Greek Political Corpora

Nikolaos Giannakopoulos and **Dionysios Rigatos**
Norwegian University of Science and Technology
Athens University of Economics and Business
{nikolagi, dionysir}@stud.ntnu.no

Abstract

In this project, we explore a variety of topic modeling techniques with a focus on their application to a corpus of modern Greek political texts. We analyze and compare several classic generative models with state-of-the-art neural models, which provides an extensively modular framework for deep learning based topic modeling. By leveraging these sophisticated models, we aim to address the challenges of unsupervised topic modeling, such as their interpretability and incompatibility with quantitative metrics. Our approach involves a sophisticated experimental setup where models are fine-tuned and compared using a variety of topic coherence, diversity and similarity metrics. This study not only tests the effectiveness of different topic modeling techniques but also examines their application in a linguistically diverse dataset and the challenges that come with it - such as text pre-processing - providing insights that could benefit further research in the area of topic modeling for Greek corpora.

1 Introduction

In such a rapidly expanding field like text analytics and natural language processing and with the power of deep neural models, tackling multilingual tasks is more accessible than ever. The integration of such technologies is rapidly transforming the landscape, opening up a great amount of possibilities for both advancing multilingual research as well as creating useful models with specific applications in these languages. In our case, we leverage an array of topic models using the Optimizing and Comparing Topic Models Is Simple (OCTIS) (Terragni et al., 2021) library and BERTopic (Grootendorst, 2022) so as to perform an analysis on a corpus of modern Greek political speeches and statements issued by the Prime Minister's office from 2012 to 2024. Our frozen-in-time model approach will make full use of the unlabeled data and allowing these models to achieve optimal performance in extracting topics for offline, unsupervised tasks aimed primarily at dataset evaluation.

This project not only presents an analytical study but also sets the foundational framework by outlining the dataset selection process, addressing the unique challenges of pre-processing texts in the Greek language, and understanding the rationale behind our methodological choices. With our approach, we aim to contribute to both the academic and practical stakeholders in the field of modern Greek text analysis.

2 Theoretical Background

2.1 Topic Modeling

Topic modeling is a statistical modeling technique utilized in the fields of natural language processing and text mining among others so as to identify and extract the topics or themes present within a large collection of documents. It operates on the premise of the distributional hypothesis as explained by Harris (1954) - documents encompass a mixture of various topics that consist of frequently co-occurring words.

Implementations of such models are usually trained using unsupervised learning methods, thus allowing them to uncover more than what is known about a document by providing insights into large and unstructured text datasets by discovering the latent structures within them.

2.2 Topic Diversity

The Topic Diversity score measures the lexical diversity across topics in a corpora by examining the ratio of unique top-K words across all of the topics to the total number of top-K words.

$$TD = \frac{|\text{Unique\_TopK\_Words}|}{K \times |\text{opics}|}$$

2.3 Topic Coherence Metrics

Topic Coherence is a quantitative assessment of how well topics can be explained by the corpora they were generated from.

2.3.1 UMass Coherence

The UMass coherence score (Mimno et al. (2011)) assesses topic quality by measuring how words that belong to a coherent topic are more likely to appear together in the same documents more frequently than words that do not belong to a coherent topic. As such, coherent topics will have word pairs that show a higher degree of co-occurrence and higher $C(t)$ values. It is formulated as such:

$$C(t; V(t)) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \left(\frac{D(v_{t,m}, v_{t,l}) + 1}{D(v_{t,l})} \right)$$

2.3.2 CV Coherence

The CV coherence score (Röder et al. (2015)) measures the semantic similarity between high-scoring words in topics, considering a wide window of words instead of just adjacent words. This score uses Normalized Pointwise Mutual Information (NPMI) (Aletras and Stevenson (2013)) and the cosine similarity between word vectors of the top-ranked words within the topics.

2.4 Topic Similarity Metrics

Topic similarity is a quantitative assessment of how similar topics are to each other.

2.4.1 Pairwise Jaccard Similarity

The Pairwise Jaccard Similarity score, in the context of topic modeling, measures the similarity between different topics by comparing the sets of documents associated with each topic. The Jaccard Similarity Index (Jaccard (1940)), calculated for two sets, is the ratio of the number of elements in the intersection of the sets to the number of elements in their union. Applied to topic modeling, if each topic is represented as a set of documents that prominently feature the topic, the Jaccard Index can quantify how similar two topics are based on their document overlap. A higher Jaccard score indicates a greater overlap and thus higher similarity between the topics.

2.5 Word Embeddings

Word embeddings (Mikolov et al., 2013) are a class of techniques where words are represented by vectors of real numbers in a low-dimensional space. These vectors try to capture the semantic meaning of words and are constructed in such a way that semantically similar words are closer together in that vector space. Techniques like Continuous Bag-of-Words (CBOW) and skip-gram are used to learn these embeddings from large text corpora.

2.6 Clustering

Clustering refers to unsupervised machine learning techniques that group data points into clusters based on distance (e.g. Euclidean distance), or similarity (e.g. cosine similarity) measures. The goal is to organize data in such a way that those who share similar attributes are put in the same cluster.

2.7 Dimensionality Reduction

As dimensions increase we face something often referred to as the curse of dimensionality (Bellman, 1957). Traditional distance measures don't work and finding meaningful clusters becomes challenging. Dimensionality reduction refers to a set of techniques aimed at transforming data from a high-dimensional space to a lower one, while preserving the important characteristics of the data, so we can apply traditional distance measuring and clustering techniques.

2.8 Word Weighting / TF-IDF

Word weighting refers to techniques used for calculating the importance of words within a set of texts (corpus).

The most popular amongst these techniques is Term Frequency - Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988). Term Frequency $tf_{t,d}$ is a measure for how often a word appears in single document, while document frequency counts how rare a word is within a corpus by measuring the amount of documents that contain that word.

$$W_{t,d} = tf_{t,d} \cdot \log(\frac{N}{df_t}) \tag{1}$$

2.9 Transformer Architecture

Transformers (Vaswani et al., 2023) are a deep learning architecture relying on the concept of Attention (Bahdanau et al., 2016). Self-attention allows the model to focus on different parts of the input sequence, by assigning different weights to different input tokens. These weights, in contrast to other types of Neural Networks, are not learned during training, but are computed based on the input sequence and can change for each input token. This allows the model to learn dependencies between tokens that are far apart in the input sequence, which is a key advantage of transformers compared to other architectures.

3 Related Work

As it was previously mentioned, machine learning has become more versatile than ever when it comes to applications in under-represented languages. Topic Modeling heavily benefits from this versatility as it has applications in both supervised and unsupervised tasks - providing a relevant solution to the extraction of information from large corpora.

As shown by Panagiotis (2022), state-of-the-art (SOTA) solutions such as BERTopic can lead to impressive results in language-specific tasks - something confirmed thanks to the availability of labels in their experiments. Their approach yielded acceptable quantitative metrics in both topic quality and classification of labeled corpora, while the qualitative aspect was also satisfactory.

Our approach will rely heavily on manually gathered data from a specific domain (political publications) and in a strictly unsupervised context. On top of that, we aim to show that a thorough exploration of preprocessing techniques can significantly boost the performance of classic topic modeling algorithms. Additionally, we look for the optimal language-specific tools to help us achieve that which - in a way - evaluates them in a practical scenario. Finally, we will experiment generously with different modular components and hyperparameters in BERTopic, as well as with topic representation models for improved clarity in the outputs. Tackling this task with the big image in mind is very likely to yield interesting results at every single step of the analysis.

4 Architecture & Approach

4.1 Exploratory Data Analysis

In this section we will go over the dataset chosen for this task, the preprocessing pipeline implemented as well as the a-priori insights we can extract regarding its contents.

4.1.1 Dataset

The dataset used in this analysis consists of speeches and statements gathered from the official website of the Office of the Greek Prime Minister in the time period between 2012 and 2024. It was fetched and assembled by a custom-built web page scraper, yielding approximately 2030 distinct publications.

The dataset is primarily in Greek, with a minority of English or French phrases/words. These publications are official transcripts and/or summaries of speeches and statements given by the active prime minister at the period. No other filtering was applied in the initial construction of the dataset other than the aforementioned time frame. While the speeches and the statements were loaded independently, a merged version of the data will be used in all of the tasks presented as they have an identical format, as shown in Table 1.

Field	Type	Description
date	str	Publication Date
id	int	Publication ID
url	str	Publication URL
title	str	Publication Title
text	str	Publication Content

Table 1: Dataset Attributes

4.1.2 Data Preprocessing

Our dataset consists of long texts in Greek, with a lot of repetitive language and - due to its context - legalese. OCTIS provides classes for easily preprocessing and storing the dataset, which may then be used as input in any of its algorithms for training and evaluation. One drawback of OCTIS' provided preprocessing is that it is black-box - it uses spaCy (Honnibal and Montani, 2017) to modify the data and construct the expected dataset class. While spaCy provides support for processing Greek corpora, its lemmatization pipeline was unable to reduce most of the words to their root forms sufficiently. This was especially apparent in less mainstream words, which are quite common in official speeches and statements, as well as verbs.

In order to overcome this issue we built a custom and OCTIS-compatible preprocessor based on Stanford's Stanza (Qi et al. (2020)). Stanza supports the usage of external, custom-made models for the processing steps. While Greek is natively supported, we opted for a more specialized approach on lemmatization for the Greek language. As presented by Prokopidis and Piperidis (2020) in A Neural NLP toolkit for Greek - their hybrid lexicon and part-of-speech based lemmatizer will be used so as to achieve more accurate conversions to root forms and therefore avoid the repetition of lexical cognates when generating topics. The difference between the two pipelines was not trivial and almost eliminated the issue of monolectic topics. With our pipeline prepared, there is a series of additional steps to be performed so as to build our training-ready dataset.

Stopword removal will be applied using the union of spaCy's Greek el_core_news_sm and English en_core_news_sm. In fact, stopwords will be lemmatized and will be removed from the post-lemmatized text for higher coverage as Greek is quite a versatile language and these lists were not exhaustive. Removal of numbers and punctuation marks was also applied so as to have a more relevant vocabulary and ensure that words are not entangled between punctuation marks.

Word weighting with TF-IDF was used in order to maximize the relevance of the words present in our vocabulary in order to assist with the generation of topics. Specifically, all words present in more than 20% or less than 1% of the texts were omitted from the vocabulary. Additionally, a text had to consist of at least 20 words in order to be counted in as a data point. Finally, the minimum word length was set to 4, immediately excluding a plethora of context-lacking phrases.

While the OCTIS models require preprocessing in order to produce acceptable result, unlike BERTopic which according to its documentation performs better when the input data is unprocessed and performs data processing later in its pipeline.

4.1.3 Understanding The Data

Instead of tackling the task blind, it is generally a good idea to extract some statistics prior to the analysis. This helps in spotting possible biases, imbalances or errors in the dataset.

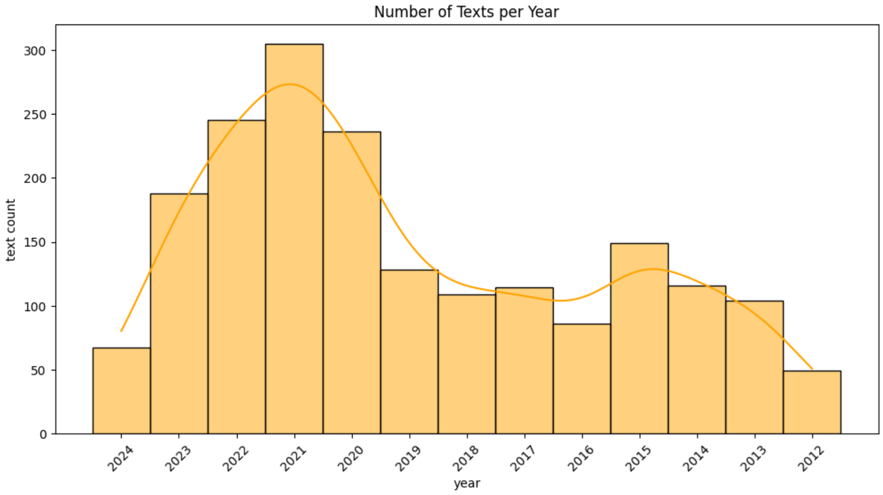


Figure 1: Count of data samples per year

As seen in Figure 1, our data is imbalanced in the sense that there has been a lot of publication activity from the Office of the Greek Prime Minister in the post-COVID era, meaning we might see an increased amount of recent topics compared to the mid-2010's.

Table 2: Selected Common Words by Year

Year	Common Words	Topic Description
2022	ενεργειακός, ψηφιακός, τουρισμός, πηγή, πόλεμος	Energy, Digital, Tourism, War
2021	εμβολιασμός, ψηφιακός, εμβολιάζω, υφυπουργός, γραμματέας	Vaccines, Digital Campaigns
2018	μνημόνιο, περιφέρεια, παραγωγικός, νησί	Memorandum, County, Island
2013	πλεόνασμα, κόμμα, ανταγωνιστικότητα, ανεργία	Surplus, Competitiveness, Unemployment

Complete table available at Appendix A Table 10

As we see in Table 2, we can also take a quick look at the most common words that appear in our texts, per year, as a taste of what to expect when generating topics.

4.2 OCTIS Models

Optimizing and Comparing Topic Models Is Simple (OCTIS) (Terragni et al. (2021)) is a library that encapsulates the preprocessing, optimization, training, evaluation and comparison of topic models. It contains a plethora of topic model algorithm implementations and metrics and thus will be used throughout so as to provide a platform for our chosen methods in this analysis. In this sub-section we will briefly go over the OCTIS-provided algorithms we will use in our experiments.

4.2.1 Latent Semantic Indexing

Latent Semantic Indexing (also known as Latent Semantic Analysis, Landauer et al. (1998)) is one of the first approaches in topic modeling that attempts to overcome the issue of variability in language when trying to retrieve relevant topics. It is able to extract context and capture semantic similarities between different words.

The algorithm initially constructs a weighted document-term matrix which stores the importance of each word based both on term frequency and inverse document frequency (tf-idf). This high-dimensional sparse matrix is then decomposed into three intermediate matrices using Singular Value Decomposition (SVD). These matrices are the term-concept vector matrix W, the singular value matrix S and the concept-document vector matrix P. By picking the most significant singular values (columns) from S, the original matrix is essentially truncated and thus its dimensionality is reduced while keeping the maximum possible context.

4.2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei et al. (2003)) is a generative probabilistic model designed for the analysis and decomposition of collections of discrete data sets into a predetermined number of topics. It is predicated on the construction of a Bayesian model where documents are conceptualized as random mixtures over latent topics and each topic is characterized by a specific distribution over words. Documents are generated by sampling words from these topic-specific distributions.

It fundamentally operates using a fixed K-dimension Dirichlet distribution. As seen in Figure 2, the parameters include α and β, which are hyperparameters that set the prior distributions over document-topic mixtures and topic-word mixtures respectively. Additionally, θ represents the distribution of topics in a specific document. Finally, z assigns topics to individual words and w denotes the distribution of words for each topic, capturing the semantic essence. This assignment over the N words is repeated for all the documents D.

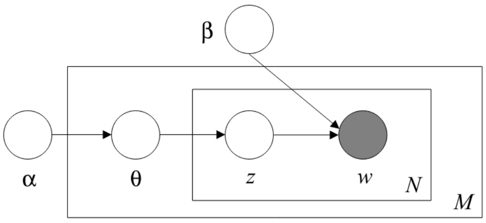


Figure 2: Plate representation of LDA (adapted from Blei et al. (2003))

4.2.3 Hierarchical Dirichlet Process

Hierarchical Dirichlet Process (HDP) (Teh et al., 2006) is a topic modeling technique that extends the capabilities of LDA by allowing for an infinite number of topics across documents. While LDA requires pre-specifying dimensionality of the Dirichlet distribution (number of topics), HDP overcomes this limitation by leveraging the Dirichlet Process (DP) to model a potentially infinite number of topics, thereby providing a more flexible modeling approach. DP is an extra level added to the model.

4.2.4 Product of Experts Latent Dirichlet Allocation

Product of Experts Latent Dirichlet Allocation (ProdLDA) (Srivastava and Sutton, 2017), is a topic model that enhances the conventional LDA by modifying its underlying generative process. Unlike LDA, which represents documents as mixtures of topics where each word's generation is attributed to a single topic, ProdLDA redefines this structure by modeling the generation of words in a document as a Product of Experts (Hinton, 1999).

This simple yet important distinction allows ProdLDA to overcome a limitation of LDA where its generative process tends to yield less specific, and at times, less interpretable topics because the likelihood of word generation is averaged over the topics.

4.2.5 Honorable Mentions

Non-negative Matrix Factorization (NMF) is a machine learning algorithm aimed at reducing the dimensionality of non-negative data by breaking it down into two matrices revealing underlying patterns, similar to LSI. While not natively a topic model, it is popular thanks to its ability to reduce dimensionality and produce coherent topics.

4.3 BERTopic

The previously mentioned algorithms rely on bag-of-word representations, which ignore the semantic relationships between words. Word embeddings and the emergence of Transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) allow texts to be represented in such a way that similar texts are put together in the vector space. Then, as proposed by Sia et al. (2020), those embeddings can be clustered with each cluster representing a different topic.

BERTopic (Grootendorst, 2022) builds on this and combines clustering with a novelty class-based TF-IDF (c-TF-IDF) to create coherent topics from the created clusters.

The steps of a BERTopic process, as seen in Figure 4.3, include creating embeddings of the documents, reducing the dimensions of those embeddings and then performing clustering. After that c-TF-IDF weighs the words present in each cluster to create topic representations.

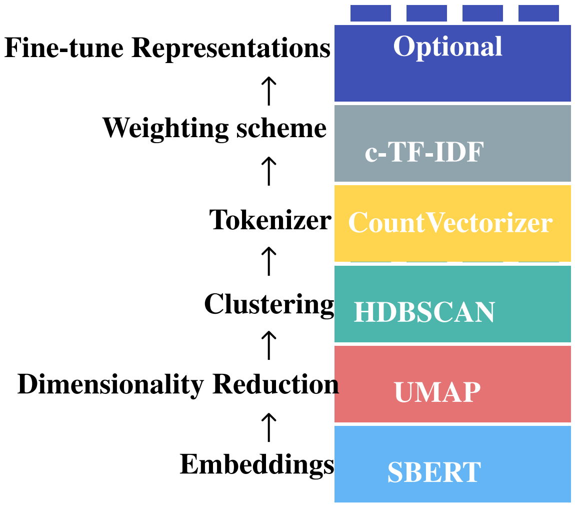


Figure 3: BERTopic Architecture (figure from Grootendorst Maarten)[1]

BERTopic's other advantage lies in its modularity where each step algorithm of each step can be replaced with a different one.

4.3.1 Sentence Transformers

Similar to how traditional word embedding techniques create numerical vectors to make word representations, sentence transformers, as the name suggests, are able to generate embeddings for a large collections of tokens, creating vector representations for whole documents, sentences or paragraphs.

BERTopic for this step uses the Sentence-BERT (SBERT) framework (Reimers and Gurevych, 2019b) which consists of pre-trained Transformer-based language models. The SBERT framework provides many pre-trained models but any model can be used for creating text representations, such as models provided by Spacy, Gensim and Scikit-Learn Embeddings,

4.3.2 Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)

Sentence embeddings created by SBERT exist in a high dimensional space and, as found by Steinbach et al. and Beyer et al., traditional clustering techniques don't work well in high dimensions. The solution to that is reducing the dimensions.

Techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are widely used dimensionality reduction but BERTopic uses UMAP (McInnes et al., 2020), which has shown to better preserve local features when reducing to as low as 2 dimensions.

4.3.3 HDBSCAN

After the document embeddings have reduced in dimensions, the clustering is performed with a hierarchical variation of Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

[1] https://maartengr.github.io/BERTopic/algorithm/algorithm.html

DBSCAN (Ester et al., 1996) is a clustering algorithm that identifies dense regions of points that exist within a certain distance from each other and adding them to the same cluster. More specifically, it works with two parameters: ε, which specifies the radius of the circle around each data point and minimum points, which specifies the least amount of points needed to form a cluster. Points with enough neighbors are core points of a cluster, whereas points far from any neighborhood are classified as noise.

Hierarchical DBSCAN (HDBSCAN) (McInnes et al., 2017) performs DBSCAN with varying epsilon values, trying to find the one that gives the best stability. In this way, clusters of varying densities are created.

4.3.4 c-TF-IDF

Each cluster created by HDBSCAN represents a single topic. Depending on our corpus, each cluster may include hundreds or more words and we need to find a meaningful way to represent each cluster.

$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{tf_t}) \tag{2}$$

As seen in Equation (2) the frequency of a term t is calculated for an entire cluster c. Then, instead of the inverse document frequency, we calculate the inverse *class* frequency, measuring how much information a term provides in an entire cluster.

4.3.5 Topic Representation Fine-Tuning

BERTopic's final component focuses on how the created topics can be represented. Although the default topic representations created with c-TF-IDF are representative of the general theme of each topic they can sometimes be misleading as, due to c-TF-IDF's nature, rely on word frequency. As a result the most semantically relevant words may be ignored and many topics end up with a similar representation. While all the previous components played a crucial role in how the topic clusters are created, representation models play a more qualitative role. These models can vary from Part of Speech matchers, Maximal Marginal Relevance Calculation, keyword extractors such as KeyBERTInspired and even Large Language Models such as Generative Pretrained Transformers. Any combination of these different models can be combined in a chain model, where multiple representation models are used in succession with the output of one being the input of the next one.

KeyBERTInspired is an adaptation of KeyBERT (Grootendorst, 2020). KeyBERT creates embeddings of both each document and each word in the document. Afterwards it calculates the cosine similarities between the word and document embeddings to find the words that are most similar to the document. Those words are considered to be the most representative.

Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) on the other hand works by trying to calculate both the relevance of a word within a document, and its diversity among other words in the document. This helps avoid having very similar words as the topic representation.

5 Experimental Setup

In this section we will go over how the models were set up with their parameters as well as justify and reason about the choices made for their initialization.

5.1 Evaluation Metrics in Topic Modeling

As seen in Sections 2.3, 2.2 and 2.4 there is a plethora of quantitative metrics for evaluation of topic models. However, these metrics are oftentimes incoherent and have no universal ground-truth value - as explained by Hoyle et al. (2021). It is the case that theoretically perfect values sometimes produce questionable topics and theoretically not-so-perfect values produce topics that make much more sense to a human evaluator.

This curse of topic model evaluation was quite apparent in our experiments and a compromise was required so as to maintain scientific integrity. We observed that these metrics were quite successful in highlighting low-quality topic models, unlike when it came to finding the best model, something we used as a filtering measure. Throughout our experiments, and especially in hyper-parameter optimization, quantitative metrics were a necessary factor in finding the best model configuration. As such, our evaluation approach makes use of these metrics so as to highlight theoretically well-performing models and then using empirical evaluation so as to pick the best among the top-few best. Despite the array of metrics available to us, we decided to focus on two metrics for topic coherence, specificially the CV and UMass metrics, one metric for similarity, specifically the Pairwise Jaccard Similarity and the self-explanatory Topic Diversity metric. To better understand the relationships between these metrics as well as why others were omitted to avoid redundancy, see Appendix B Figure 5.

While this introduces an undesired subjectivity to our task, our research indicated that human evaluation was a necessary evil - one we are willing to accept for the sake of interpretable and cohesive results.

5.2 General Experimental Setup

Since we will be comparing the results of a few different algorithms, it is important to have a common baseline so as not to render the comparisons useless. As seen in Table 3, when the algorithm requires an amount of topics, or is able to limit them to that number, we opt for a maximum of 30, with BERTopic being the only one that follows the maximum rule and the rest of the algorithms generating exactly 30.

Parameter	Value
Number of Topics	<=30
Top K Words	7

Table 3: General Hyperparameters

The reason behind that choice - or compromise for some - is that while we are certain that there has been a plethora of significant events in the last decade, we would like to explore the bigger image of it all. At the same time, setting the amount of topics to a lower number would only result in vague and general topics such as "economy", "issues" or "crisis", which is not analytically interesting or unique.

5.3 OCTIS Experimental Setup

The OCTIS topic models at our availability provide a lot of customization and tuning capabilities through their hyperparameters. While this task seems infeasible for a large amount of algorithms and parameters, OCTIS offers a hyperparameter optimization foundation which can be used with the built-in algorithms so as to find the best possible setup. We built a wrapper which uses OCTIS' optimizer so that it automatically extracts the best parameters for multiple models and stores them in a file - ensuring that the general hyperparameters, as shown in Table 3, are common across all the experiments. The optimization's objective was the CV Coherence (2.3.2) score, and specifically its maximization.

Table 4: Hyperparameters for selected OCTIS Models

(a) LSI

Parameter	Value
power_iters	6
extra_samples	100

(b) LDA

Parameter	Value
passes	10
alpha	0.1179
eta	None

(c) HDP

Parameter	Value
alpha	0.1341
eta	0.5
gamma	0.5
tau	32
kappa	0.5

(d) Prod LDA

Parameter	Value
batch_size	64
lr	0.0037
dropout	0.0438
num_epochs	100
momentum	0.6117
num_layers	1
num_neurons	236
activation	softplus
solver	adam

In Table 4, we see the hyperparameters chosen for final experiments. The amount of experiments ran for OCTIS models was enormous - consuming approximately 5 hours of state-of-the-art hardware compute. The final values were selected not only based on quantitative measures but also empirical evaluation of the topics.

5.4 BERTopic Experimental Setup

As mentioned in Section 4.3, the BERTopic pipeline allows for users to replace any of its submodels and components, as well as their parameters. This will be the basis for our model and hyperparameter tuning. In similar fashion to OCTIS in Section 5.3, we have created a custom optimization class that trains and evaluates multiple models in a hyperparameter space - extracting their results for further analysis. As BERTopic has a multi-step pipeline, we focused on the tuning of the dimensionality reduction and clustering models as they have the largest impact on the output topics. A total of 260 combination of hyper-parameters, with different dimensionality reduction models (PCA, t-SVD, None), were tested in the process of finding the best possible set - occupying SOTA graphics processing unit (NVIDIA RTX4090) for a total of 10 hours. Other models and parameters were initialized independently or prior to the tuning so as to set the baseline.

5.4.1 Embeddings and Sentence Transformers

Picking the right sentence transformer is key to getting the best out of BERTopic. The choices are infinite, however we narrowed it down to three options; BERTopic's default multilingual "paraphrase-multilingual-MiniLM-L12-v2" (Reimers and Gurevych, 2019a), a Greek-Media-BERT-based sentence transformer "dimitriz/st-greek-media-bert-base-uncased" (Zaikis et al., 2023) and a Greek/English XLM-Roberta-based sentence transformer "lighteternal/stsb-xlm-r-greek-transfer" (Papadopoulos et al., 2021). Additionally, BERTopic documentation recommends finer granularity when using large texts. We will also experiment with sentence-level granularity, where each sentence becomes a document as far as the model is concerned. Sentence transformers are evaluated on a default BERTopic model (and sub-models) for uniformity.

Through empirical evaluation, we concluded that "dimitriz/st-greek-media-bert-base-uncased" for document-level granularity performed the best - something not backed up by the metrics as it quantitatively performed the worst. Metrics and further insights on sentence-level granularity are present in Appendix B Table 12.

5.4.2 Tokenizer and Weighting Scheme

For our tokenizer we will use the SKLearn's (Pedregosa et al., 2011) CountVectorizer with the configuration from Table 5a, with the same stopwords as in the OCTIS configuration. As we see, we'll use soft max/min document frequency limits so as to filter extremes. Additionally, we'll have unigrams and bigrams so as to capture phrases in the corpus.

BERTopic's default ClassTfidfTransformer will be the main weighting scheme, with the configuration shown in Table 5b - essentially reducing common words like a black-box.

(a) CountVectorizer

Parameter	Value
stopwords	$GR \cup EN$
ngram_range	(1,2)
max_df	.95
min_df	.005

(b) ClassTfidfTransformer

Parameter	Value
reduce_frequent_words	True

5.4.3 UMAP and HDBSCAN

The hour-long, compute-intensive optimization proved BERTopic's default choice solid when it comes to algorithms - UMAP and HDBSCAN performed the best. A set of optimal hyper-parameters was also extracted from the large search space.

Specifically, for UMAP, we optimized on the number of components, the number of neighbours and the minimum distance between the neighbours. More components results in a richer representation, but large dimensionalities affect the clustering process negatively. In Table 6 we can look at the optimal configuration for UMAP.

Parameter	Value
n_components	15
n_neighbors	15
min_dist	0.2

Table 6: UMAP Configuration

HDBSCAN's parameters yield much more significance, as they directly impact the amount and the quality of topics we will generate. The minimum cluster size directly affects the the number of clusters that will be generated, with lower values leading to more microclusters. Metric regards the distance metric that will be used for the construction of clusters, and the boolean on whether we want to use this model for predictions, which is out of our scope. The values can be seen in Table 7.

Parameter	Value
min_cluster_size	7
metric	euclidean
prediction_data	False

Table 7: HDBSCAN Configuration

5.4.4 Topic Representation Model

For the topic representation model we used a chain model consisting of KeyBERTInspired and MMR. More specifically we first extracted the top 50 words from each topic with KeyBERTInspired and we filtered those through MMR. We chose a diversity of 0.5 striking a balance between how relevant they should be for their corresponding topic and how diverse they need to be to each other. Additionally, for the labels of each topic we chose the top 3 words found through our chain model, after filtering to only keep words of a maximum length of 10. Since the representation model plays a purely qualitative role we found that topic labels with smaller words help with both the visualizations and the reading comprehension. More about our reasoning for topic representation model selection on Appendix B.4.

6 Experiments

In this section, all the results derived from our analysis and experimentation will be presented.

6.1 Quantitative Results

In this section, result metrics from all the models tested will be presented.

Table 8: Quantitative Topic Evaluation Metrics for All Models

Model	Coherence_{CV}	Coherence_{UMASS}	Diversity_{TOPIC}	Similarity_{PJS}
LSI	0.5585	-1.3672	0.5714	0.0334
LDA	0.6418	-1.3635	0.7762	0.0132
HDP	0.4650	-2.2493	0.4905	0.0162
NMF	0.5703	-1.6358	0.5143	0.0350
ProdLDA	**0.6782**	-2.8220	0.8857	0.0069
BERTopic	0.5996	**-0.5838**	0.9523	0.0020
BERTopic+	0.2903	-0.6087	**0.9714**	**0.0012**

BERTopic+ refers to the BERTopic model with additional Topic Representation Tuning as mentioned in 5.4.4.

6.2 Qualitative Results

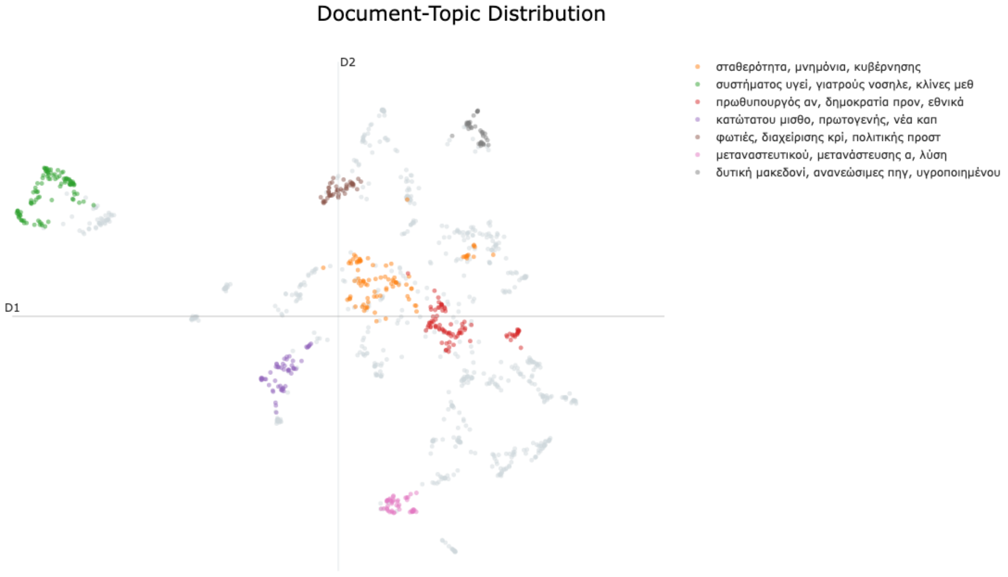
In this section, qualitative results that hold a weight in the empirical evaluation will be presented.

Table 9: Top 8 most common topics for the best performing models

Topic	ProdLDA	BERTopic+	BERTopic
1	υποχρεούμαι εκκαθάριση προθύρτω λεπτό υπερχρεώνω στόμα απαριθμώ αναλυτής αντικειμενικά	κυβέρνησης λύση κράτος μεταρρυθμίσεις χρέους ξέρετε εθνική	democracy reforms ευρώ περιοχή ανάπτυξης ευρωπαϊκό συμφωνία
2	ενεργειακός προεδρία οδηγία κανονισμός απώτερος έτοιμος πολυετής υιοθέτηση μετανάστευση ανεργία	σταθερότητα μνημόνια κυβέρνησης βουλή μεταρρυθμίσεις χρέους ελληνικός λαός	τσίπρα εκλογές μητσοτάκη συριζα ευρώ κύριε μητσοτάκη κυρίες κύριοι
3	εταίρος διαπραγμάτευση ρύθμιση ριζοσπαστικός διοίκηση ασφυκτικός θεσμικός ευρωζώνη διεκδίκηση φορολογικός	συστήματος υγείας γιατρούς νοσηλευτές κλίνες μεθ νοσοκομείων υγείας βασίλης εμβολιασμός εντατικής	υγείας νοσοκομείο σύστημα υγείας νοσοκομεία εμβολιασμού εθνικό σύστημα σύστημα
4	ελευθερία δημοκρατικός ανοχή ενάντια ρατσισμός τάξη στερεότυπα εξτρεμισμός εχθρός ήθος	πρωθυπουργός αντώνης δημοκρατία προνομιακός εθνικά συμφέροντα απλώς εαβ προεδρία	que de en la el σαμαράς δημοκρατίες
5	εκκαθάριση υποχρεούμαι καταστροφικά στόμα προθύρτω γραφήχρηματιστήριο όμοια επιμερίζω	κατώτατου μισθού πρωτογενής νέα καπ κατώτατου κτηνοτρόφων κλάδο εστίασης επιχειρήσεις	αγρότες αγροτικής προϊόντα αγροτική πρωτογενή τομέα προϊόντων αναπηρία
6	νοσοκομείο προσωπικό γιατρός περίθαλψη πρωτοβάθμιος ασθενής καρκίνος διακυβέρνησης νοσηλευτής διοικητής	φωτιές διαχείρισης κρίσεων πολιτικής προστασίας εκαβ πυροσβεστικής προστασίας πολίτη πλημμύρες	πολιτικής προστασίας προστασίας εκαβ πολιτική προστασία πυροσβεστικής πυρκαγιές αστυνομίας
7	πόλη μετρό μουσείο δρομολογώ χιλιόμετρο ολυμπιακός οδικός σταθμός αθλητισμός ανάπλαση	μεταναστευτικού μετανάστευσης ασύλου λύση σύμφωνο μετανάστευση εε τουρκίας ταυτοποίησης ροές	frontex ροές πρόσφυγες τουρκία σύνορα μετανάστευση συνόρων
8	χρειάζονται προσθέστε πίνακας των εμφανίσεων	δυτική μακεδονία ανανεώσιμες πηγές υγροποιημένου ελπε αγωγός	αερίου ενέργειας φυσικού αερίου φυσικού αέριο φυσικό αέριο φυσικό

BERTopic+ refers to the BERTopic model with additional Topic Representation Tuning as mentioned in 5.4.4.
Table is available in English at Appendix C.1 Table 15.

Figure 4: Document-Topic Distribution Plot for BERTopic+



7 Evaluation and Discussion

Going into this task, we were aware of the challenges when it comes to evaluating topic models - the results, however, are even more concerning than what we initially expected. As we see in Table 8, the quantitative evaluation of in our models seems to be especially confusing. We made the choice of tuning our models based on CV Coherence, which in retrospect appears to be far from the ideal coherence measure - raising concerns about whether our optimization achieved the best possible, and interpretable, results. However, this also seems to be the case with UMass. According to our empirical evaluation based on Table 9, BERTopic+ and ProdLDA produced the best topics. This is where the incoherence of coherence starts - with ProdLDA having the worst UMass and the best CV coherence scores, and BERTopic+ having a CV score that we would, unless we had seen the topics, discard immediately. On the other hand, Topic Diversity and Pairwise-Jaccard Similarity seem to be directly correlated with our empirical evaluation.

Qualitative results are presented in Table 9, where the strictness of the ProdLDA preprocessing is apparent due to the absence of irrelevant words and the diversity of topics. On the other hand, BERTopic seems to suffer from poor topic representations. This is visible on Topic 4, with the words "que, de, en, la, el" being the prime representation words for that topic. BERTopic+ has eliminated such noise and was able to produce richer topics. On the analytical side, as expected, all of the algorithms produced topics related to economy, external policies, healthcare, infrastructure, democracy, COVID-19, immigration and other political cliches. For brevity reasons, not all topics are shown, with more niche topics being shadowed. Finally, Figure 4 presents how the top topic clusters of BERTopic+ are distributed. They are clearly distinct with no overlap, something attributed to the optimized UMAP dimensionality reduction as well as the HDBSCAN clustering. It is safe to say that our top-performing models, in combination with our preliminary work, are able to achieve notable results.

8 Conclusion and Future Work

This project started off as an analytical task, but quickly transitioned to a optimization and research project, as the fine-tuning and optimization of topic modeling algorithms proved to be far more exciting with a lot of depth to explore. We broke away from the defaultness of black-box pre-processing algorithms by constructing tailored pipelines and optimizing legacy algorithms to their best capabilities, allowing them to generate topics up to par with the state-of-the-art. Algorithms such as LSI, LDA, HDP and ProdLDA were put on the spot against BERTopic, which offers a variety of conveniences and is justifiably sitting at the SOTA throne for topic models. The ambiguity of quantitatively evaluating unsupervised topic models was a constant concern that led to a lot of second-guessing and the human factor was crucial to finding the best model.

Our work sets the baseline for unsupervised topic modeling in Hellenistic corpora, with a key takeaway being that English-based approaches such as spaCy and multilingual sentence transformers are outperformed by niche, yet well tailored, toolkits and models. The analysis was not groundbreaking - it was no surprise to anyone that Greece went over an economic crisis or that COVID-19 was a very hot topic - but the frozen-in-time model approach offered a retrospective view that while not as interesting, is certainly accurate and can prove useful in real-world scenarios.

Due to limitations in both time and project length, only the iceberg of our findings was presented. As an additional contribution to the field, all of our analysis is available in the project's open-source repository (Rigatos and Giannakopoulos, 2024). Further experimentation with an alternative input data granularity (sentences, paragraphs) and a more generous topic allowance could result in deeper insights with regards to understated events. Finally, additional approaches such as zero-shot, multi-modal and dynamic (over time) topic modeling make for interesting tasks to be tackled, especially in Greek.

References

Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In *International Conference on Computational Semantics*, 2013. URL https://api.semanticscholar.org/CorpusID:15651747.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

Richard Bellman. *Dynamic Programming*. Dover Publications, 1957. ISBN 9780486428093.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In Catriel Beeri and Peter Buneman, editors, *Database Theory — ICDT'99*, pages 217–235, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. ISSN 1532-4435.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155. doi:10.1145/290941.291025. URL https://doi.org/10.1145/290941.291025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD'96, page 226–231. AAAI Press, 1996.

Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. URL https://doi.org/10.5281/zenodo.4461265.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/pdf/2203.05794.pdf.

Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. doi:10.1080/00437956.1954.11659520. URL https://doi.org/10.1080/00437956.1954.11659520.

Geoffrey E. Hinton. Products of experts. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 1–6. IET, 1999.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Alexander Miserlis Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is automated topic evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2107.02173.

Paul Jaccard. Coefficient générique réel et coefficient générique probable. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 61:117–36, 01 1940. doi:10.5169/seals-272981.

Thomas Landauer, Peter Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 01 1998. doi:10.1080/01638539809545028.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi:10.21105/joss.00205. URL https://doi.org/10.21105/joss.00205.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In Regina Barzilay and Mark Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1024.

Kosmas Panagiotis. Topic modeling techniques in corpora and news speeches. MSc Thesis, Aristotle University of Thessaloniki, Thessaloniki, Greece, November 2022. URL https://ikee.lib.auth.gr/record/342983/files/Kosmas%20Panagiotis.pdf.

Dimitris Papadopoulos, Technical University of Crete, and Hellenic Army Academy. Semantic textual similarity for the greek language using transformers and transfer learning. 2021. URL https://huggingface.co/lighteternal/stsb-xlm-r-greek-transfer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Prokopis Prokopidis and Stelios Piperidis. A neural nlp toolkit for greek. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 125–128, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450388788. doi:10.1145/3411408.3411430. URL https://doi.org/10.1145/3411408.3411430.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. URL https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019a. URL http://arxiv.org/abs/1908.10084.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019b.

Dionysios Rigatos and Nikolaos Giannakopoulos. A depth-first approach in topic modeling with hellenistic corpora for prime ministerial speeches - github repository. 2024. URL https://github.com/DionGR/greek-pm-topic-modeling.

Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450333177. doi:10.1145/2684822.2685324. URL https://doi.org/10.1145/2684822.2685324.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi:https://doi.org/10.1016/0306-4573(88)90021-0. URL https://www.sciencedirect.com/science/article/pii/0306457388900210.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736. Association for Computational Linguistics, November 2020. doi:10.18653/v1/2020.emnlp-main.135.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models, 2017.

Michael Steinbach, Levent Ertöz, and Vipin Kumar. The Challenges of Clustering High Dimensional Data. In Luc T. Wille, editor, *New Directions in Statistical Physics*, pages 273–309. Springer Berlin Heidelberg. doi:10.1007/978-3-662-08968-2_16. URL http://link.springer.com/10.1007/978-3-662-08968-2_16.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Hierarchical dirichlet processes. *Machine Learning*, pages 1–30, 12 2006. doi:10.1198/016214506000000302.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics, April 2021. URL https://www.aclweb.org/anthology/2021.eacl-demos.31.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Dimitrios Zaikis, Stylianos Kokkas, and Ioannis Vlahavas. Dacl: A domain-adapted contrastive learning approach to low resource language representations for document clustering tasks. In Lazaros Iliadis, Ilias Maglogiannis, Serafin Alonso, Chrisina Jayne, and Elias Pimenidis, editors, *Engineering Applications of Neural Networks*, pages 585–598, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-34204-2.

A Architecture

A.1 Common Words Per Year

In Table 10 we present a more exhaustive list of the most common words found per year:

Table 10: Most Common Words by Year

Year	**Common Words**	**Topic Description**
2024	πανεπιστήμιο, εταιρεία, εκλογή, σχετικά, αγρότης	University, Elections, Farmers
2023	φωτιά, τετραετία, τουρισμός, περιφέρεια, εκλογή	Fire, Tourism, Elections
2022	ενεργειακός, ψηφιακός, τουρισμός, πηγή, πόλεμος	Energy, Digital, Tourism, War

Year	Common Words	Topic Description
2021	εμβολιασμός, ψηφιακός, εμβολιάζω, υφυπουργός, γραμματέας	Vaccines, Digital Campaigns
2020	τουρισμός, νησί, νοσοκομείο, ψηφιακός, κορονοϊός	Tourism, COVID-19
2019	βουλευτής, εκλογή, ψηφίζω, πλειοψηφία, κόμμα	Elections
2018	αναπτυξιακός, μνημόνιο, περιφέρεια, παραγωγικός, δημοσιονομικός	Development, Memorandum
2017	παραγωγικός, περιφέρεια, αναπτυξιακός, παραγωγή, σχεδιασμός	Production, Regional Development
2016	νησί, διαπραγμάτευση, δικαιοσύνη, κόμμα, αξιολόγηση	Island Politics, Justice
2015	κυπραίικος, βέβαιος, διαπραγμάτευση, βουλευτής, τράπεζα	Cyprus Issue, Banking
2014	ευρώ, οικονομικό, κρίση, σχέδιο, ανάκτηση	Economic Crisis, Recovery Plan
2013	πλεόνασμα, κόμμα, ανταγωνιστικότητα, πρωτογενής, ανεργία	Surplus, Competitiveness, Unemployment
2012	ανταγωνιστικότητα, ανταγωνισμός, ςομπετιτιενεσς, αποκαθιστώ, διατηρώ	Competitiveness, Market Dynamics

B Experimental Setup

B.1 Evaluation Metrics

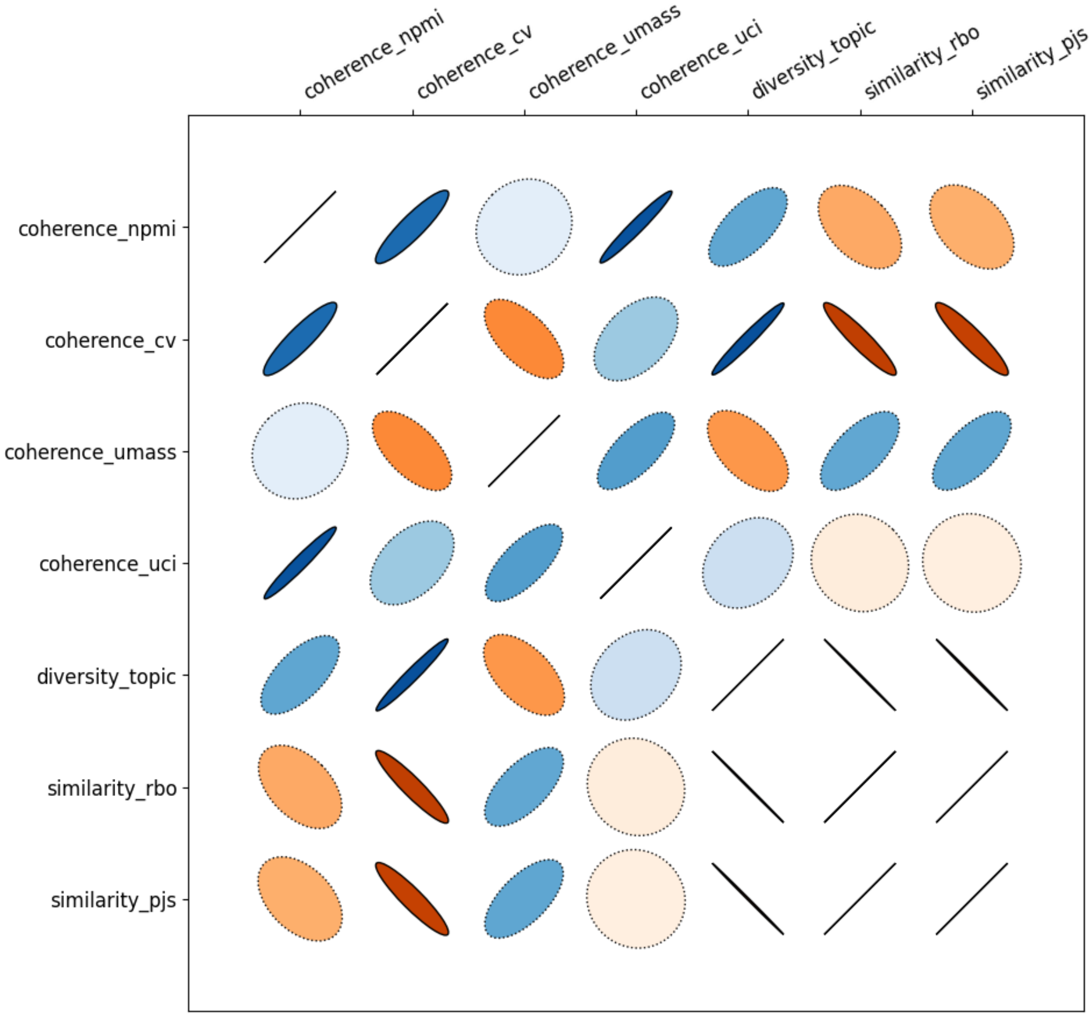


Figure 5: Metric Correlation Matrix

B.2 Hyperparameters

Below we present the lists of hyperparameters for the NMF model:

Parameter	Value
kappa	1.0
minimum_probability	0.0543

Table 11: NMF Hyperparameters

B.3 Sentence Transformers

Document-level granularity models performed significantly better when it comes to the metrics, something we expected because of the constrained amount of topics. Without the topic limitation, sentence-level models produced thousands of topics - with BERTopic's automatic topic reduction not reducing them below four digits. This bottleneck renders their purpose useless and while it would certainly be an interesting topic (pun intended) to explore, it seems out of scope for this analysis.

Model	CV Coherence	Topic Diversity
gr_r_xlm_sentences	0.5019	0.9793
gr_media_sentences	**0.6041**	**1.0000**
multilingual_sentences	0.5826	0.9724
gr_r_xlm_docs	0.6826	**0.7310**
gr_media_docs	0.6865	0.7241
multilingual_docs	**0.7024**	0.6087

Table 12: Model Evaluation on CV Coherence and Topic Diversity

B.4 Topic Representations

As previously mentioned, quantitative metrics aren't reliable measures of the quality of the topics. This becomes more apparent when changing the representation model. The default c-TF-IDF model, as seen in Table 13 yields better metrics, but different representations, as seen in Table 14, make the topics much more human interpretable.

	CV Coherence	UMass Coherence	Topic Diversity
Default Representation Model	0.646	-0.518	0.935
Custom Representation Model	0.335	-0.592	0.957

Table 13: Topic Coherence Metrics

Default representation	Custom representation
que, de, en, la, el	δημοκρατία, προνομιακός, μεταρρυθμίσεις
ευρώ, μητσοτάκη, κύριε μητσοτάκη	σταθερότητα, μνημόνια, κυβέρνηση

Table 14: Comparison of Default and Custom Representations

C Experiments

C.1 Best Model Topic Representations

Table 15: Top 8 most common topics for the best performing models (in English)

Topic	ProdLDA	BERTopic+	BERTopic
1	Obligated, clearance, entrance, minute, overcharge, mouth, enumerate, analyst, objectively	Government, solution, state, reforms, debt, you know, national	*democracy, reforms*, euro, region, development, European, agreement
2	Energy, presidency, directive, regulation, ultimate, ready, multi-year, adoption, migration, unemployment	Stability, memorandums, government, parliament, reforms, debt, Greek people	Tsipras, elections, Mitsotakis, SYRIZA, euro, Mr. Mitsotakis, ladies gentlemen
3	Partner, negotiation, regulation, radical, administration, suffocating, institutional, eurozone, claiming, tax	Health system, doctors, nurses, beds, ICU, hospitals, health, Vasilis, vaccination, intensive care	Health, hospital, health system, hospitals, vaccination, national system, system
4	Freedom, democratic, tolerance, against, racism, order, stereotypes, extremism, enemy, ethos	Prime Minister, Antonis, democracy, privileged, national, interests, simply, EAB, presidency	*que, de, en, la, el*, Samaras, democracies
5	Clearance, obligated, disastrous, mouth, entrance, writing, stock market, similar, apportion	Minimum wage, primary, new, CAP, minimum, livestock farmers, restaurant industry, businesses	Farmers, agricultural, products, agricultural, primary, sector, products, disability
6	Hospital, personnel, doctor, care, primary, patient, cancer, governance, nurse, manager	Fires, crisis management, civil protection, EKAB, firefighting, civil protection, citizen, floods	Civil protection, protection, EKAB, civil protection, policy, firefighting, fires, police
7	City, metro, museum, schedule, kilometer, Olympic, road, station, sports, regeneration	Immigration, asylum, solution, pact, migration, EU, Turkey, identification, flows	Flows, Frontex, refugees, Turkey, border, migration, borders
8	Needs, add, table of appearances	Western Macedonia, renewable sources, liquefied, ELPE, FSRU, pipeline, ADMIE	Gas, energy, natural gas, natural gas, natural gas, natural

Words in italics were found in the same language as presented here