VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

SOFTWARE ENGINEERING STUDY PROGRAMME

Master's Thesis

Machine Learning Based Narrative Search in the Information Space

Mašininio mokymosi metodais pagrįsta naratyvų paieška informacinėje erdvėje

Gvidas Pranauskas

Supervisor : assoc. prof. dr. Viktor Medvedev

Reviewer : assoc. prof. dr. Linas Petkevičius

Vilnius
2025

Summary

With the ever increasing amount of information available, the need for new methods to systematize and analyze this data becomes apparent. This research proposes and validates a dynamic topic modeling process for identifying and tracking time-varying narratives in textual datasets. The study analyzes the 2024 Lithuanian parliamentary election in a dataset which consists of 7385 documents collected from both digital and traditional media sources (TV, radio, and press). The analysis is conducted using two algorithmic topic modeling approaches: BERTopic and ANTM, which combine embeddings, dimensionality reduction, clustering and narrative extraction using Large Language Models (LLMs). The research also explores LLMs based evaluation metrics for measuring topic model performance. Key findings reveal that BERTopic excels in better maintaining temporal topic coherence across time steps, while ANTM produces more diverse topics but struggles with thematic consistency. This research show how advanced NLP techniques can systematically identify evolving narratives in complex information spaces, with potential applications for media professionals, political strategists or journalists in investigating the dynamics of public opinion.

Keywords: topic modeling, natural language processing, large language models, dynamic topic modeling, BERTopic, ANTM, embeddings, clustering, algorithmic topic models.

Santrauka

Nuolat didėjant informacijos kiekiui, reikia naujų metodų šiems duomenims sisteminti ir analizuoti. Šiame tyrime analizuojamas dinaminis temų modeliavimo procesas, skirtas nustatyti ir sekti naratyvus tekstiniuose duomenų rinkiniuose. Tyrime analizuojami 2024 m. Lietuvos Seimo rinkimai, remiantis duomenų rinkiniu, kurį sudaro 7385 dokumentai, surinkti tiek iš skaitmeninių, tiek iš tradicinių žiniasklaidos šaltinių (TV, radijo ir spaudos). Analizė atliekama naudojant du algoritminius temų modeliavimo metodus: BERTopic ir ANTM, kurie naudoja įterpinius, dimensijų mažinimą, klasterizavimą ir naratyvų išgavimą naudojant didžiuosius kalbos modelius (LLM). Tyrime taip pat nagrinėjamos LLM pagrindu apskaičiuojamos vertinimo metrikos, skirtos temų modelių rezultatui įvertinti. Rezultatai atskleidžia, kad BERTopic geriau išlaiko temų nuoseklumą laiko atžvilgiu, o ANTM geba išgauti įvairesnes temas, tačiau susiduria su teminiu nuoseklumu laike. Šis tyrimas parodo, kaip natūralios kalbos apdorojimo metodai gali padėti sistemingai identifikuoti besivystančius naratyvus informacinėse erdvėse, su potencialiu pritaikymu žiniasklaidos profesionalams, politikos strategams ar žurnalistams tiriant viešosios nuomonės dinamiką.

Raktiniai žodžiai: temų modeliavimas, natūralios kalbos apdorojimas, dideji kalbos modeliai, dinaminis temų modeliavimas, BERTopic, ANTM, įterpiniai, klasterizavimas, algoritminiai temų modeliai.

List of Figures

List of Tables

Contents

List of Abbreviations

ANTM	Aligned Neural Topic Model - A neural topic model that aligns topics across time steps.
BERTopic	A topic modeling technique that uses BERT embeddings, dimensionality reduction, and clustering.
BERT	Bidirectional Encoder Representation from Transformers - A language representation model.
DTM	Dynamic Topic Modeling - A technique for analyzing topic evolution over time.
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise - A clustering algorithm.
LDA	Latent Dirichlet Allocation - A traditional probabilistic topic modeling method.
LLM	Large Language Model - Advanced AI models trained on vast amounts of text.
NLP	Natural Language Processing - Field focused on interactions between computers and human language.
NPMI	Normalized Pointwise Mutual Information - A metric for evaluating topic coherence.
PCA	Principal Component Analysis - A dimensionality reduction technique.
TF-IDF	Term Frequency-Inverse Document Frequency - A statistical measure for word importance.
c-TF-IDF	Class-based TF-IDF - A modified version of TF-IDF for topic modeling.
t-SNE	t-Distributed Stochastic Neighbor Embedding - A dimensionality reduction technique.
UMAP	Uniform Manifold Approximation and Projection - A dimensionality reduction technique.
CTC	Contextualized Topic Coherence - Metrics that use LLMs to evaluate topic models.

Introduction

The Internet has become a crucial part of modern day society, making the world smaller and facilitating the exchange of information between the most remote points on Earth in a short period and with a vast amount of data. With the ever growing capabilities of Internet, the amount of data being generated and made available to the public is also growing. As humans, we are capable of processing only a fraction of the data available to them on a daily basis. Average human consumes around 34 gigabytes of information per day and this number has been constantly growing by 5.4% from 1980 to 2008 [9].

This clearly shows the growing amount of information that is being constantly processed by our mind. Eventually, we might reach or even have already reached a point where we will not be able to process this extraordinary amount of data. Thus, this requires new methods to systematize and analyze this data to the level which is comprehensible by the human mind, possibly by grouping documents with related topics or narratives, to help filter data which is interesting and required.

A narrative, or a topic, in its essence, encompasses a coherent storyline or theme within a corpus of text. It encapsulates a sequence of events, ideas, or discussions that contribute to a larger overarching concept or storyline. In the context of this and other [22, 23, 26] researches, narratives or topics serve as the fundamental building blocks that help to structure and understand the content buried within extensive textual data, otherwise known as Information Space.

Information Space, as a term, refers to ”a set of concepts, and relations among them, held by an information system; it describes the range of possible values or meanings an entity can have under the given rules and circumstances” [41]. In the context of this study, it can be understood as a dynamic and interconnected digital environment where narratives and topics continuously evolve. It contains diverse digital content, including texts, images, and multimedia, interconnected through hyperlinks and semantic relationships. This space is shaped by user interactions and content creation, forming a complex web of information where narratives emerge, intersect, and influence each other.

One of the emerging methods to systematize this vast Information Space and extract meaningful topics from it is called Topic Modeling. Topic Modeling is a suite of algorithms that aim to discover and annotate large archives of documents with thematic information [6]. One of the main important aspects of topic modeling lies in its ability to identify and categorize patterns in word usage, thus linking together documents that showcase similar patterns. Essentially, topic models operate under the idea that documents are composed of various topics [7]. The term ’topic’ is understood as a probability distribution over words [49].

In order to identify topics over time in text corpus, various strategies of Dynamic Topic Modeling (DTM) are employed, including probabilistic time series models [8], latent Dirichlet allocation [7], ANTM [45], BERTopic [20], and a plethora of other approaches. DTM aims to analyze the time evolution of topics in large document collections. Each topic represents an interpretable semantic concept and is described as a group of related words. They also infer what topics each document contains to understand their underlying semantics [55]. DTM models update their estimates of the underlying topics as new documents are added to the corpus, allowing the identification of evolving trends and

patterns in the archive. Topic evolution analysis has been used in a variety of applications, such as discovering the evolution of research topics and innovations in scientific archives and understanding trends in public opinion on particular issues [45].

This study aims to propose a process based on topic modeling to identify time-varying topics in a textual dataset. To achieve this aim, the **following objectives** must be solved:

1. Conduct an extensive literature review to identify and highlight methods and solutions related to topic modeling;
2. Explore existing datasets which highlights recently emerged topics;
3. Compare different methods, propose an effective solution, and outline a systematic process for extracting topics from textual datasets;
4. Propose process for identifying and tracking emerging topics over time;
5. Evaluate the proposed process based on topic modeling metrics;
6. Summarize the results and findings of extracting topics from textual datasets.

The research object is a dynamic topic-modeling process for extracting, tracking and analyzing evolving themes in large-scale, time-stamped textual corpora by integrating data preparation, multiple modeling methods, temporal alignment strategies, evaluation metrics and visualization techniques. This involves a comprehensive review of existing topic modeling methods, (e.g., Natural Language Processing, Text Embeddings), exploration of relevant datasets, identifying suited evaluation of results metrics, and creation of a process for topic evolution over time analysis. The goal is to enhance understanding and handling of dynamic information withing domains such as journalism, market research, and national security [19]. This improvement will be achieved through the utilization of state-of-the-art machine learning algorithms in the field of natural language processing.

The **defended statement** of this research states that the proposed DTM process demonstrates better performance in certain scenarios for identifying and analyzing time-varying topics in text datasets compared to traditional topic modeling methods. This improvement is attributed to the integration of advanced Natural Language Processing and machine learning algorithms.

The study predominantly utilizes machine learning algorithms for *Natural Language Processing* (NLP) for language processing and data clustering to identify and categorize topics within the textual information. Contextual embeddings [16, 37] enables researches to analyze text as a numerical representations in a way that captures their contextual meanings. Embeddings map words to high-dimensional vector space where similar words or phrases with similar contexts are closer together. This results into words with similar meaning or usage being more likely to have similar numerical representations. Using these algorithmic language processing techniques, research has been conducted on Algorithmic Topic Models [20, 51], which combine several algorithms and take advantage of numerical optimization techniques to represent topics as a set of weighted words extracted from a cluster of semantically similar documents [45]. This study also aims to study topic evolution in time through the algorithmic topic modeling point of view.

To evaluate the output and overall quality of the proposed process, it is important to identify evaluation metrics. Since topic modeling is classified as unsupervised models, the comparison of outputs is considered to be difficult. However, there are proposals to use *contextualized topic coherence* metrics which are context-aware metrics, used to evaluate topic modeling techniques that are based on Large Language Models [44]. Topic coherence metrics aim to measure the interpretability and meaningfulness of the extracted topics. A high value of such metrics suggests that the words in the topic are similar, understandable and are likely to occur when observed under similar circumstances. In the context of this research, which considers the use of NLP with Large Language Models, an important goal is the identification of such metrics.

There are numerous studies done exploring the applicability of topic modeling. One of those researches [23] addresses the issue of disinformation surrounding the Russian invasion of Ukraine, highlighting the inadequacies in tracking and combating this misinformation across diverse online platforms. Existing methods fall short in mapping and responding to evolving false narratives. To fill this gap, the study employs a sentence-level analysis approach, utilizing advanced language models to trace the spread of Russian state media narratives across news sites and social platforms. By uncovering the prevalence of disinformation and its impact on platforms like Reddit, the research demonstrates the efficacy of this method in systematically identifying and tracking false narratives, offering a foundation for future efforts in combating online misinformation.

Another study [25] delves into the surge of coordinated disinformation campaigns amid the Ukrainian conflict, focusing on counterfeit websites posing as legitimate European news sources. It zeroes in on two main sites, RRN and WoF, associated with Russian government backing. Investigating their potential to deceive, the research examines their ownership and role within a broader disinformation network. It stands apart by providing a comprehensive dataset encompassing all WoF posts, conducting linguistic and topical analyses. The paper's key contributions include this extensive dataset.

To continue with datasets, there are many researches done to classify social media texts based on ideology [47] or other criteria. There are also researches done to specifically create datasets regarding the recent global events, such as COVID-19 [5], Russia invasion of Ukraine [12], Israel and Palestine Conflict [13]. Although these datasets may be not directly used in researches on the evolution of topics and narratives of these events, it lays a foundation for this study and provides data to conduct such a study. As the texts in those datasets usually contain a position regarding the global events mentioned before, which suggests that they exhibit a certain topic or narrative, it allows to perform topic modeling. Also, since the majority of the datasets are concluded of social media posts, the datasets might be utilized for dynamic topic modeling, which requires a time dimension to be present in the data. In summary, the amount of research conducted and the number of datasets available imply that there are available datasets to support the execution of this study.

Considering **novelty** and **expected results** of this study, it is important to revisit the research's aim. It suggests that the expected results conclude into a process for identifying emerging and evolving topics within the selected dataset by utilizing topic modeling and evaluating its effectiveness and accuracy. Recent state-of-the-art advancements in NLP, dimensionality reduction, and clustering al-

gorithms open up new fields of research related to topic modeling. This is particularly relevant given the ever growing archive of textual information in the Information Space, in the form of social networks, content sharing platforms, blogs, where texts appear in diverse forms, from short tweets to extensive studies. The established probabilistic topic modeling techniques often fall short [2, 23] in this context. Therefore, there is a need for novel approaches to topic modeling, which includes a dimension of time. This study aims to review new techniques in dynamic topic modeling, identify metrics used to evaluate topic modeling outcomes and as result of the study propose a process to model topics in the selected dataset which would be comprehensible, accurate based on the identified metrics and easy to reproduce across various examples or datasets.

To keep the study scope manageable, the research will be concentrated on textual information analysis to extract topics established in written content available on social media platforms and other textual data sources such as news articles, online platforms, blogs. The scope excludes other media sources like audio, video or images due to specific methods required to analyze these information formats.

In the aim to decode the vast expanse of digital information, the emergence of dynamic topic modeling stands as a important methodology to reveal evolving trends and critical narratives. This innovative approach not only deciphers emerging topics but also tracks their evolution, offering a deeper understanding of our information ecosystem. As we navigate this landscape, the imperative to identify misinformation, as seen in studies on events like the Russian invasion of Ukraine, or the ability to systematize large corpus of texts highlights the need for systematic processes which utilize advanced language models.

1 Literature review

The following section will present an overview of literature on the topic modeling. It begins with introduction to topic modeling and practical examples, focusing on traditional techniques of topic modeling approach. Then, exploration of Large Language Models usage within topic modeling is done, presenting how these models can enhance narrative search in textual datasets. Also, the review discusses esential steps in algorithmic topic modeling such as dimensionality reduction, clustering methods and evaluation metrics. This literature review should establish the necessary background to understand topic modeling and used approaches in this study.

1.1 Introduction to Topic Modeling

Topic modeling is a type of statistical methods for discovering abstract topics or narratives that occur in a collection of documents. Such methods are often utilized for various text-mining tasks in natural language processing (NLP) field that helps organize and summarize large datasets of textual information, gathered from Information Space. Following are some examples of common use-cases of topic modeling:

- **Document Classification.** As the collective knowledge is more and more digitized and stored in information systems - it becomes difficult to navigate and organize this vast amount of digitalized textual information. This promotes a need of new tools to search and understand this amount of information, which is continually growing. It is common to use tools such as basic text search to find the needed information. Resources in the Information Space are often interconnected using links - directions from one resource to another, allowing to navigate to related documents and resources. However, as the amount of information grows, gathering knowledge becomes difficult. This statement can be defended by taking the example of a well-known company **Google**. This company develops and maintains a tool, which has become so familiar and even assimilated into everyday language, known as **Google Search**. Basically, **Google Search** is a big archive of data available on Internet, indexed by **Google Search** crawlers and allows users to quickly discover information based on keywords that we expect to find in text. This is a useful tool to discover and navigate with information available on Information Space, but it lacks the ability to truly discover resources based on topics. Topic modeling, citing D. Blei, would allow us to ”imagine searching and exploring documents based on the themes that run through them. We might ”zoom in” and ”zoom out” to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other” [6]. So, topic modeling is used to automatically categorize documents into predefined themes, narratives or categories based on their content, which allows to identify distinct topics within a text corpus, and these topics can be used as features of a search engine, trend analysis or as an input to classification model.

- **Information Retrieval.** Based on the previous assumption of ever growing amount of archival information, efficient information retrieval is becoming an increasingly difficult task. People

spend more time on gathering the relevant documents and texts than actually reading and analyzing the information itself. With the precedent of propaganda, information wars, it is becoming even more difficult to distinguish between relevant information and attempts to spread misleading information. By mapping documents into a topic space, information retrieval can improve query relevance, as they match the topics inferred from search queries. This use case helps to crystallize relevant information from vast amounts of textual archives, highlighting documents which has relevant information.

- **Content Recommendation.** Once again, as the previous two use cases have already mentioned, we deal with a problem with increasing amount of data available. Processing this information is becoming increasingly difficult for a humans, thus it is easier than ever to lost interest or a track on information. News outlets try to combat this situation in several ways, one of which is click-baiting headlines. This method is an attempt to attract the attention of a reader, suggest an interesting topic and invoke curiosity. There are other variations on how to attract reader to the platform, such as content recommendation. They are rigorously crafted algorithms that collect and analyze patterns of user throughout the Information Space and attempts to predict which content might be interesting to a reader. Topic modeling is also used in this context, as a method to recommend articles, news or papers that are relevant to user's interests. With techniques used in topic modeling, business are able to suggest content that are semantically and thematically similar to those that a user has shown interest in, thereby personalizing content delivery.

- **Trend Analysis.** This use case is closely related to the topic of this work, since the naming of it suggests a narrative or a topic being present in the textual data, that can emerge, evolve and die out over time throughout the text corpus. Trend analysis in the context of topic modeling should be considered as a powerful method for understanding how certain topics of discussion evolve over time. This application is often met in fields such as social sciences, market research, media monitoring, intelligence. Usually any textual information can be analyzed and gathered - social media posts, articles, blog posts, research papers, TV or radio shows, podcasts - anything that has value being converted to text, as long as this information has temporal aspect. Meaning that data should have the possibility to be sliced into time frames. Then, various topic modeling algorithms can be applied over the entire dataset to discover a set of narratives present in it. The application of topic modeling algorithms can have many variations. For example, a static topic modeling algorithm like LDA (Latent Dirichlet Allocation) can be applied to entire dataset to discover a fixed number of topics, which then can be split into the selected time frames and documents assigned to each topic can be calculated. This would show this exact topic trend over time, however, this approach loses the value of modeling evolution of topics. To combat this problem, dynamic topic modeling algorithms can be applied. They allow to analyze not only how topics are distributed in specific time slices, but also how these distributions evolve into the next time slice based on the previously observed data. Topics in each time slice are related to the topics in the precious slices, allowing for topics to evolve, split or merge as the dataset progresses through time. For example, in academic research, dynamic

topic modeling could show how discussions about AI have shifted from theoretical aspects to ethical implications and real-world applications over time, especially capturing the change in trend with recent developments on generative language models.

- **Narrative Search.** Another use-case, closely related to the topic of this work. As stated before, topic modeling can be a useful tool in text mining tasks to discover hidden thematic structure in a large archive of documents. A narrative, or a topic, in its essence, encompasses a coherent storyline or theme within a corpus of text. Topic modeling and narrative search in this case can be considered as almost synonyms. It could be said that topic modeling techniques can be a way to perform narrative search. Dynamic topic models are able to capture evolution of topics over time. DTM can identify how topics change, merge or split over time. This temporal aspect is critical when dealing with narratives that evolve, such as news stories or ideological texts. In fact, one of the papers [23] employ algorithmic topic modeling techniques to first analyze the thematic structure of Russian state media outlets, known for spreading misleading information and performing coordinated information campaigns. Employing sentence level analysis, they discover patterns on how certain websites publish topics that are later echoed by other Russian sites. After this analysis, they collect data from *Reddit* social network and analyze the correspondence of earlier discovered narratives and topics with the submissions and comments on *Reddit*, finding that 39.6% comments on `r/Russia` subreddit corresponded to narratives from pro-Russian disinformation websites.

Moving to analyzing topics over time, it is important to establish the definition and deeper understanding of Dynamic Topic Models (DTM). As a concept, DTM was defined by David Blei and John Lafferty in their paper ”Dynamic Topic Models” [8], published in 2006. In DTM, documents are assumed to be generated from a set of topics that evolve over time. Such models capture both the topical content of documents and the evolution of topics over time by including temporal dependencies between topics and words. Using such approach, the method is able to identify how topics change, merge, split or disappear over time, providing insights into the dynamics of underlying data.

The difference from static topic modeling methods, such as Latent Dirichlet Allocation (LDA) [7], stems from the fact that static topic models assume that topics in the corpus of texts remains static. While static topic models infer topics from the entire corpus without considering temporal aspects underlying in data, DTM explicitly model the evolution of topics over time. Data in this case are grouped by time slice, for example, years, and it is assumed that documents of each group come from a set of topics that evolved from the documents of the previous slice. In this case, DTM can be thought of as an extension of static topic models, where order of the documents matter, which offers more nuanced understanding of temporal dynamics within Information Space.

1.2 Example of Topic Modeling

To illustrate the practical application of topic modeling, consider the case of analyzing international media news coverage on Lithuania. Suppose we have collected news articles over the past year from various international news outlets, focusing on keywords like ”Lithuania”, ”Baltic”, and

”Vilnius”. Using topic modeling, we can identify key themes within these articles. Initially, a static topic model might reveal topics such as ”Lithuanian politics”, ”economic development”, and ”cultural events”. However, this static view does not capture how these discussions change over time.

By applying dynamic topic modeling, we can track the evolution of these topics month by month. For instance, we might observe that discussions on ”Lithuanian politics” peak during election periods or that ”economic development” becomes more prominent following major trade agreements. Visualization tools can then be used to create line graphs that show the frequency of each topic over time, providing a clear picture of trends and shifts in media coverage.

For example, in January, topics might include ”Lithuanian presidency in the EU”, ”energy independence”, and ”NATO exercises”. In March, the focus might shift to ”economic sanctions on neighboring countries”, ”renewable energy projects”, and ”international cultural festivals”. These shifts can be visualized to show how international perceptions and discussions about Lithuania change over time.

This use case not only demonstrates the power of topic modeling in organizing and summarizing vast amounts of textual data but also highlights its ability to uncover temporal patterns and evolving narratives, offering valuable insights for policymakers, media analysts, and researchers. By understanding how Lithuania is portrayed in the international media and how these portrayals evolve, stakeholders can better address issues, leverage positive coverage, and counteract negative narratives.

An illustration of one of the possible techniques of topic modeling is displayed in Figure 1. Considering the algorithmic topic modeling, it can be understood as a collection of different algorithms, where each one is responsible for certain tasks. It starts from getting embeddings of documents, then reducing the dimensionality of embeddings, clustering them and last few steps are responsible for extracting representation of the topics. In fact, when looked at the individual steps and how they connect with each other, topic modeling in this way becomes clear and easy to understand.

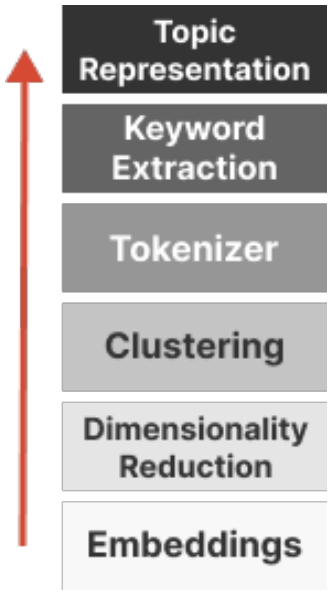

Figure 1. *Example sequence of steps to create topic representations.*

Moving on to Dynamic Topic Modeling, which can also be done relatively simply using the pre-

viously mentioned algorithm. It does not require to recalculate entire model for each timestep. First, we fit the topic model as if there is no temporal aspect in the data. This will create a general topic model. Then, we can use the global representation as map to the main topics at different timesteps. For each topic and timestep, we calculate the only the topic representations. This approach results in a specific topic representation at each timestep without the need to create clusters from embeddings, as they have already been generated.

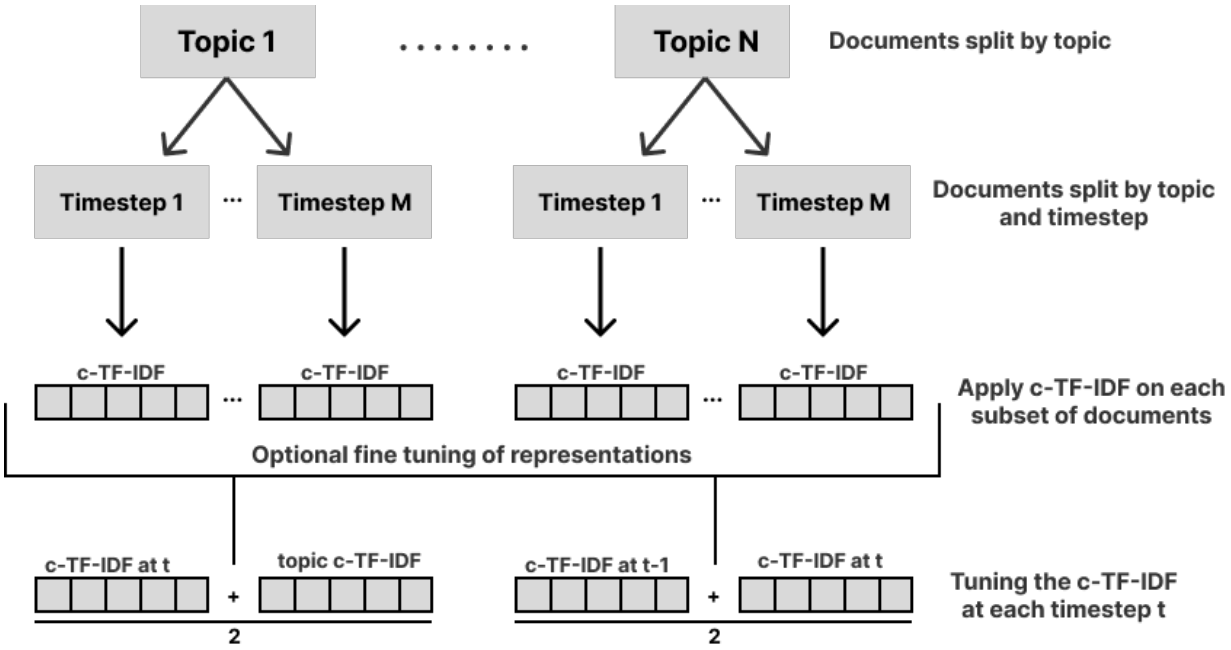

Figure 2. *Example sequence of steps to create dynamic topic representations.*

There are two primary methods to further refine these specific topic representations: **global fine-tuning** and **evolutionary fine-tuning**. For global fine-tuning, a topic representation at timestep t is adjusted by averaging its c-TF-IDF representation with the global representation. This method allows the topic representation to gradually align with the global representation while retaining some of its unique terms. For evolutionary fine-tuning, a topic representation at timestep t is refined by averaging its c-TF-IDF representation with the c-TF-IDF representation from timestep $t-1$. This approach enables each topic representation to evolve over time, reflecting changes incrementally.

Regarding the datasets used in topic modeling research, the arXiv dataset stands as an important resource for topic modeling research, encompassing over 1.5 million articles as of 2019, with this number expected to have increased. This comprehensive repository includes detailed metadata for each article such as the arXiv ID, submitter's publicly visible name, a list of authors, the title, abstract, full text, and publication date [15]. Its extensive coverage across diverse scientific domains — from physics to economics — enables cross-disciplinary studies and comparative analysis, crucial for tracking the narratives across scientific research fields. The inclusion of publication dates makes the arXiv dataset particularly valuable for dynamic topic modeling, as it allows researchers to examine how topics evolve over time, providing insights into shifting scientific paradigms and emerging research trends. Furthermore, the availability of full texts facilitates the application of sophisticated natural language processing techniques, supporting deep semantic analysis, thus providing even more con-

text that can be used with advanced NLP techniques. As an open access resource, the arXiv dataset promotes reproducibility of published research, which makes it even more useful resource for modern scientific research approach. This dataset can be utilized to analyze topic modeling, enabling scientific trend identification, impact assessment, and investigations into scientific collaboration networks.

1.3 Traditional Dynamic Topic Modeling Techniques

Referencing the same paper on Dynamic Topic Models [8], we should define the initial thought process behind DTM. There are newer and more advanced models defined now, but this paper serves as the ground point for direct applications of DTM in machine learning field.

This paper begins with the definition of static topic models, in the paper called "exchangeable topic model". These models assume that the words of each document are independently drawn from a multinomial distribution. The distributions for each document is also drawn randomly. Topics are shared by all documents available in the dataset. For this reason, each document is a mixture of different proportions of all the available topics in the dataset.

Later on, the authors of paper start to describe the idea of DTM. It is stated that for a large textual dataset, treating documents as exchangeable components - in other words, treating the collection of documents as non-ordered collection, could be considered inappropriate. Document collections such as articles, papers, blog posts often contain a distribution over time, so can reflect a change of content over time. Because of this reason, it is of interest to model the dynamics of underlying topics, not just the overall distribution of topics over text corpora.

Let $\beta_{1:K}$ be K topics, each of which is distributed over a fixed vocabulary, available in dataset. Formally, in static topic models, such as LDA, we can assume that each document is drawn from the following generative process [8]:

1. Choose topic proportions θ from a distribution over the $(K-1)$-simplex, such as Dirichlet.
2. For each word:
   (a) Choose a topic assignment $Z \sim Mult(\theta)$.
   (b) Choose a word $W \sim Mult(\beta_z)$.

This process would be correct (and is correct in LDA case) if we assumed that the documents are drawn exchangeably from the same set of topics, available in the initial dataset. However, in case of dynamic topic model, we should assume that the data is divided by time slice, for example by month. Then, the documents of each slice are modeled with a K-component topic model, where topics related to slice t evolve from the topic associated with slice $t-1$.

For a K-component model with V as words in the vocabulary of dataset, let $\beta_{t,k}$ denote the $V-vector$ of natural parameters for specific topic k in time slice t. In dynamic topic modeling context, Dirichlet distribution is not suited for sequential modeling. In this case, the natural parameters of each topic $\beta_{t,k}$ are chained together in a state space model that evolves with Gaussian noise:

$$\beta_{t,k} \mid \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I). \tag{1}$$

Compared to dynamic topic models, in static topic models like LDA, each document is represented as a mixture of topics. The proportions of these topics in a document are drawn from a Dirichlet distribution, specified by a concentration parameter α. In dynamic topic models, the topic proportions also need to evolve over time to reflect how the relevance or popularity of topics changes. For the purpose of topic proportions modeling in dynamic topic models, logistic normal distribution with parameter α is used. This choice stems from the logistic normal's distribution ability to model sequential structure between models, which is captured with a dynamic model:

$$\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I). \tag{2}$$

By combining the models of topics and topic proportion distributions, the authors of DTM paper [8] define a generative sequentially tied collection of topics models for slice t of sequential dataset like this:

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \alpha_{\sqcup -\infty}, \delta^{\in}\mathcal{I}$.
   (b) For each word:
      i. Draw $Z \sim Mult(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.

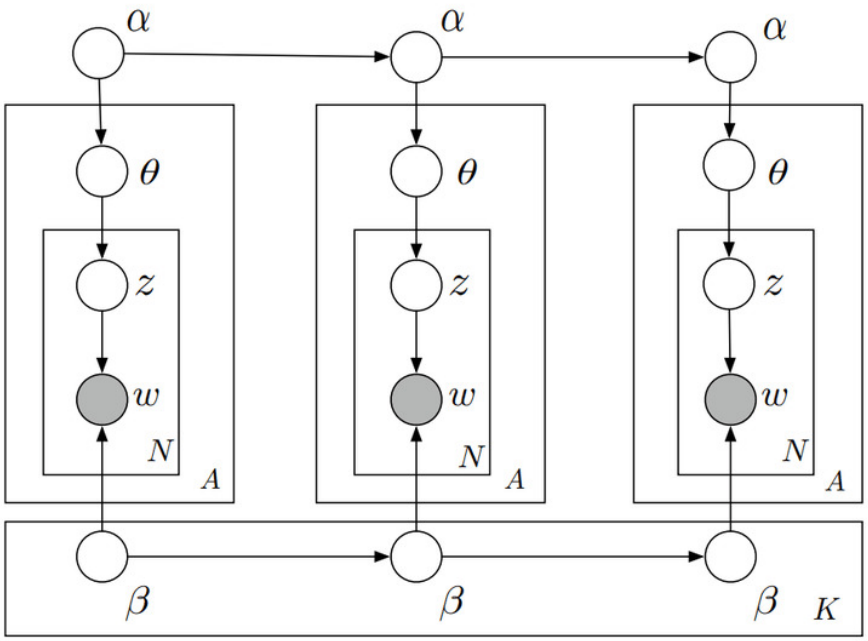

Figure 3. *Plate notation of DTM for three time slices [8].*

The graphical representation of the generative DTM model described earlier is shown in Figure 3. If horizontal arrows would be removed, thus removing the temporal aspect of the model, the

graphical representation would reduce to a set of independent topic models. With time dynamics, it shows how topic kth at time slice t has evolved from topic kth at time slice $t-1$.

1.4 Incorporating Large Language Models in Dynamic Topic Modeling

One of the main aspects related to this thesis is incorporation of Large Language Models (LLM), clustering, and dimensionality reduction algorithms into topic modeling. This presents both opportunities and challenges. One of the challenges or problems related to methods that use LLM to perform topic modeling tasks lies in the deviation from traditional topic modeling understanding. Probabilistic methods, such as the LDA or DTM mentioned above, infer the distribution of topics in one document, not only the latent topics over a collection of documents. This is possible due to the assumption that each document in the collection of documents is drawn from the same sample space, containing all possible topics of the documents, and each document contains a certain distribution of these topics [6].

Algorithmic topic models, which are largely discussed in this work, work on a slightly different assumption. Algorithmic models combine several techniques and algorithms in order to produce topics that may be available in the corpus. Specifically, this approach employs LLM or smaller language models (such as *word2vec* [36]) to convert text into high-dimensional vectors, otherwise known as embeddings [1, 16]. Working with data in high-dimensional spaces might be susceptible to various phenomena that do not occur in low-dimensional settings, and even has its own name, curse of dimensionality. In order to avoid these problems, reducing the dimension of extracted vectors is the next important part in algorithmic topic modeling. Having dimension-reduced vectors allows us to apply clustering algorithms and retrieve patterns of data that were once textual. All of these steps employ an interesting concept of converting text into numerical space. Working on text on its own is tedious process - imagine identifying topics in a very large corpus of texts, for example, science journals just by reading the texts itself. Converting text to numerical representations allows one to perform clustering and identify clusters of documents that are semantically similar. This is the key concept of algorithmic topic modelling, identifying clusters of documents that are semantically similar (embeddings of those documents in the vector space are relatively close together) and extracting topics from these clusters. Compared to probabilistic topic models, algorithms-based models can only produce topics but are not able to infer the distribution of topics in the document itself. When clustering, the document is assigned to one specific cluster (or is considered an outlier), not multiple clusters, which results in this deviation from probabilistic topic models.

1.5 Large Language Models

A large language model (LLM) typically refers to a deep learning model designed to understand, generate, and interpret human language based on the Transformer architecture. These models are characterised by their large number of parameters (often in the billions), sophisticated neural network architectures, and the ability to perform a wide range of natural language processing tasks with high levels of proficiency.

One of these models could be called BERT [16]. Compared to more recent LLM models such as GPT-3 [10] which has 175 billion parameters, nowadays BERT is not always considered LLM, having 110 million parameters as *base* version and 340 million parameters as *large* version. However, based on the context of the publication date of the BERT paper, it can be considered an LLM.

BERT is considered a language representation model, and the acronym stands for Bidirectional Encoder Representation from Transformers. This model is designed to include the context of the words, both in the left and right direction. As a result, the BERT model can be fine-tuned for a variety of tasks, such as entity recognition, sentiment analysis, document classification. Compared to other language models, such as the GPT models, BERT proposed a new idea on how to process language. GPT used a left-to-right architecture, where every token can be related only to previous tokens in the self-attention layers of the Transformer [52] model. BERT was improved compared to other models using a pre-trained objective *of a masked language model* (MLM). The masked language model randomly hides tokens on the input text and has the objective to correctly predict the original word in the place of masked word based on the surrounding context. MLM objective allows the representation of text to be a combination of left and right contexts, which allows training of a bidirectional transformer model such as BERT.

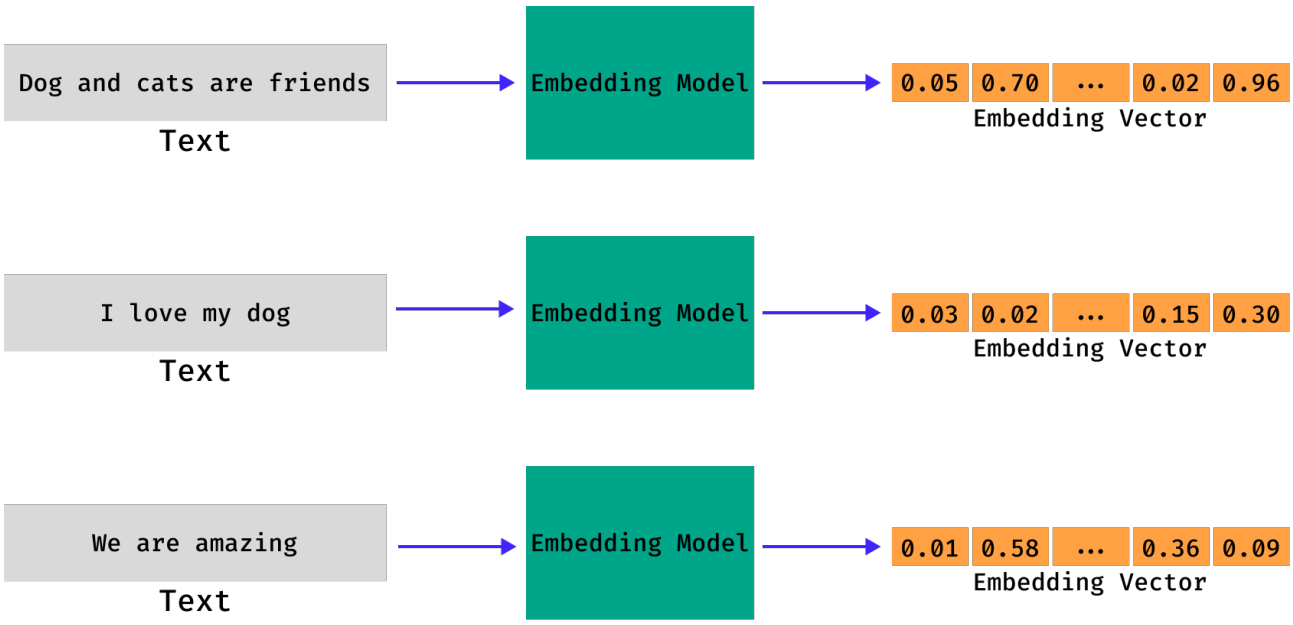

Figure 4. *Graphical representation of embeddings.*

As mentioned before, BERT or basically any other LLM is considered a language representation model, which accepts an input sequence (text) and transforms input to a numerical representation of language, called embeddings. A simple idea of text embeddings is represented in Figure 4. However, since embeddings are a crucial part of algorithmic topic modeling and thus the context of this work, it is important to provide a deeper understanding of this idea.

In the context of machine learning and specifically in natural language processing, an embedding is a representation of discrete variables, like words or phrases, as continuous vectors in a multi-dimensional space. Embeddings has the property of capturing the semantics of the input data by placing semantically similar items closer to each other in the embedding space. This concept facili-

tates machines in effectively performing tasks such as word prediction, sentence classification, and sentence similarity evaluation by converting hard to handle textual information into a form that various algorithms can manipulate more easily.

An embedding is a mapping of a discrete or categorical variable to a vector of continuous numbers. The purpose of this mapping is to approximate the meaning of the objects in a high-dimensional space, as distances and directions preserve relevant properties of the objects. If x is a discrete variable with a unique encoding in a high-dimensional space, a embedding function f maps x to a dense vector $f(x)$ in $\mathbb{R}^d$, where d is significantly smaller than the dimension of x in the original encoding space.

Key contributions to the idea of embeddings include the introduction of Word2Vec [36], a model that utilizes neural networks for word representation in vector spaces. Following this, GloVe [42] aggregated global word-word co-occurrence statistics to produce word embeddings, revealing subtle linear substructures within the word vector space. Embeddings are not limited to word representations; sentences or paragraphs can also be converted to a numerical representation, capturing the context of it. The before mentioned BERT model has a context window of 512 tokens, or roughly speaking 512 words, meaning that it can efficiently capture context and produce embeddings of text no longer than 512 tokens at once.

Referring back to Large Language Models and Algorithmic Topic Modeling, it should become more apparent how these techniques can be combined together to perform Topic Modeling. LLMs are usually fine-tuned for downstream tasks of sentiment analysis, named entity recognition, machine translation. These downstream tasks often take the outputs of LLMs - embeddings - as their input and employ a variety of machine learning techniques to perform different tasks. In the context of algorithmic topic modeling, LLMs are employed just to represent the textual information as embedding vectors, which are then analyzed further, allowing to group documents based on their topical similarity. As a result, the quality of algorithmic topic modeling should constantly improve as new and improved LLMs are developed. This allows algorithmic topic modeling to continuously grow with the current state-of-the-art in embedding techniques.

1.6 Dimensionality Reduction

Dimensionality reduction can be considered as a fundamental tool in the context of machine learning. With the ever-growing amount of data and the high-dimensionality of it, the field of machine learning or data science greatly benefits from the development of algorithms that are both scalable to vast amounts of data and able to handle diverse data in different embedding dimensions.

Considering the embeddings provided by LLMs, they often have the property of having a large number of dimensions. While there are clustering methods (clustering is part of the algorithmic topic modeling approach) available to work on high-dimensional data, a simpler strategy is to reduce the dimensionality of embeddings. As LLMs grow in terms of the number of parameters, the complexity and capacity of these models often also increase. Larger models can potentially capture more contextual information, requiring higher-dimensional embeddings to fully represent the complexity of language. Larger vector spaces can potentially capture more detailed distinctions between features,

which can be considered critical for tasks involving fine semantic distinctions, such as language understanding and generation. However, this is not the only reason to consider dimensionality reduction algorithms that can be versatile and efficient in handling different embedding dimensions. There are multiple available language models, each having a unique architecture, thus producing different variations of embeddings. Because of this, a dimensionality reduction algorithm is needed, which would show the ability to preserve more of the high-dimensional structure after dimensionality reduction despite the form of the initial data.

There are numerous worth mentioning dimensionality reduction algorithms, such as PCA [28] or t-SNE [32]. However, a more recent and fitting the needs of algorithmic topic modeling techniques algorithm is available, called UMAP. **Uniform Manifold Approximation and Projection** (UMAP) [35] is a dimensionality reduction technique that can be used for visualisation, but also for general non-linear dimensionality reduction. Based on the original paper, the UMAP algorithm is founded on three assumptions about the data:

1. The data is uniformly distributed on a Riemannian manifold;
2. The Riemannian metric is locally constant (or can be approximated as such);
3. The manifold is locally connected.

These assumptions can be difficult to understand, but these assumptions allow the model to find a manifold with a fuzzy topological structure. The embedding is found by looking for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure. There are improved variations of the original UMAP already, called **densMAP** [40], which preserves local density information in addition to the topological structure of the data. This new improvement is worth looking at in the algorithmic topic modeling task.

The core idea of this algorithm is to model the high-dimensional data as weighted graph. Each point in the data set is a node in this graph, and edges between these nodes are weighted based on the distance between points. UMAP views the data as a manifold that can be approximated by locally connected regions. This algorithm starts by evaluating the distance between points in the high-dimensional space. Usually, Euclidean distance is used, but the algorithm can be flexible in using other metrics as well. For each point, UMAP then finds a local region and computes how each point in this region is connected to every other point based on distance. Then, UMAP aims to create a low-dimensional graph that as closely as possible reflects the same relationships as discovered before in high-dimensional weighted graph. The initial layout of low-dimensional graph is generated randomly. Then UMAP adjusts this layout to reduce cross-entropy between the connectivity of the high-dimensional graph and the low-dimensional graph, using stochastic gradient descent. UMAP balances the preservation of both local and global data structure, maintaining broader relationships within the data, which is in turn crucial for tasks like clustering.

Overall, UMAP is an algorithm that can work with a variety of distance metrics, making it versatile option for different kinds of data. It is generally faster than other dimensionality reduction algorithms and can work with larger datasets. These features of the UMAP algorithm makes it a suitable choice working with high-dimensional representations of language.

1.7 Clustering

Having the dimension of embeddings reduced, the next step is recognizing patterns in the data via clustering. There are many clustering algorithms available, but in the context of this work, one of the most suited algorithms for clustering embeddings should be considered **Hierarchical Density-Based Spatial Clustering of Applications with Noise** (HDBSCAN) [34]. HDBSCAN is an extension of DBSCAN [18], which added the hierarchical component. DBSCAN on itself is one of the well known clustering algorithms that is able to handle noise in the dataset and evaluate those points as outliers, but it is not able to find the density of the data. DBSCAN forms clusters based on the concept of density reachability. In datasets with variable density clusters, the density difference between clusters can be significant. DBSCAN utilizes parameter ε, which notes the maximum distance between two points for one to be considered as in the neighborhood of the other point. Other parameter is k, which denotes the density threshold or the minimum number of points for region to be considered cluster. Since HDBSCAN could be considered as an extension of DBSCAN, the HDBSCAN can be considered as an algorithm to search over all ε values for DBSCAN to find clusters that are present for various ε values [34]. This modification of DBSCAN provides the benefit of not having to select a predefined ε value, but also helps to mitigate the problem of variable density clustering, a problem which is common to DBSCAN.

To illustrate the advantage of HDBSCAN over regular DBSCAN, an example of dataset with two significant clusters can be used. One of the clusters present in the data has high density (data points are very close together), and other cluster is sparse (low density). If ε is too small, the algorithm of DBSCAN might not identify the low-density cluster. If ε is too large, the algorithm could over-cluster, meaning that both clusters would be recognized as one. In this case, the use of HDBSCAN could solve the problem, since it iterates over the ε values, thus dealing with varying densities much better.

HDBSCAN, as described by the original authors, is particularly beneficial for exploratory data analysis. Exploratory data analysis aims to uncover interesting patterns and generate new hypotheses without making assumptions about the data. Traditional clustering methods often struggle in this area due to their need for parameter fine-tuning, assumptions about data distributions, which can lead to false assumptions about cluster relationships and reveal misleading conclusions. The strength of HDBSCAN lies in its approach to these challenges. This algorithm is part of density-based clustering techniques, which are less dependent on assumptions about the distribution of data points within a space. Because of this reason, HDBSCAN can be a very prominent clustering tool in the context of natural language processing. Natural language is a difficult concept, even for humans, so having a vast amount of textual information and making assumptions about it might be difficult. Especially when this information is transformed into embeddings, the parameter fine-tuning can be even more difficult. In essence, HDBSCAN facilitates the exploration of data without extensive pre-processing or assumptions, making it a suitable tool for clustering embeddings, even in cases where data has a lot of noise and other clustering algorithms are not suited.

Referring back to the section of UMAP, the combination of UMAP and HDBSCAN is particularly powerful for datasets where clusters may vary significantly in density and when the high-dimensionality of data (in case of embeddings) might obscure natural clusters. UMAP's ability to

reduce dimensionality while maintaining local and global structures makes it an ideal preprocessing step for HDBSCAN, which can then effectively identify clusters based on the density of points in this transformed space. This combination of dimensionality reduction and clustering algorithms has been proved effective in number of research papers. One of the researches [4] states that the combination of UMAP and HDBSCAN on short text embeddings is better than baseline methods, evaluated on clustering metrics purity and NMI.

Another approach, which is an algorithm called BERTopic [20], combines the large language models, dimensionality reduction algorithms and clustering algorithms, to prove them effective in the context of topic modeling. They state that the use of HDBSCAN models clusters with a soft-clustering approach, which allows noise in the dataset to be considered outliers. This functionality of HDBSCAN prevents documents that are not in any topic to be excluded, thus improving the performance of topic modeling techniques.

1.8 Topic Extraction

At this step, all the documents in the dataset have been processed and assigned to one particular class or counted as outlier. The process beforehand could be understood as a unsupervised document classification task - trying to categorize documents based on their similarity into different categories. In the case of topic modeling, the task is to figure out what is the common narrative across the documents in a particular category.

For finding the common topic in a textual dataset, the task is basically discovering the most important words in that particular dataset. The simplest way, of course, would be to calculate the frequencies of terms in the dataset and find the most important or common words that way. In reality, languages have words which are usually more common than others, for example, ”of”, ”is”, ”and”. If we utilize this method for topic extraction, we can expect that every single topic would have those common words and wouldn’t be so informative. In best case scenario, such topic model should be penalized based on the evaluation metrics which are discussed later.

1.8.1 TF-IDF

One common numerical statistic to determine how important a word is to a document is called **Term Frequency-Inverse Document Frequency** (TF-IDF) [30]. The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in corpus. The idea of algorithm is to denote each document d as a vector $\overrightarrow{d} = (d^{(1)}, \ldots, d^{(|W|}$) in vector space, where documents with similar content have similar vectors - in this case, similar topics. The values of $\overrightarrow{d}$ elements $d^{(i)}$ are calculated as a combination of *term frequency* (TF) and *inverse document frequency* (IDF). TF measures how frequently the a term occurs in a documents and IDF measures how important a term is in a dataset. IDF is calculated from the *document frequency* (DF) metric, which is a number of documents in which the word w occurred. Then, IDF can be calculated from the DF as such:

$$IDF(w) = \log\left(\frac{|D|}{1 + DF(w)}\right), \tag{3}$$

where $|D|$ represents the total number of documents present in dataset. The IDF has low value if a word occurs in many documents and vice versa if the word occurs in only one document, meaning it is rare. Then, the value of $d^{(i)}$ is calculated as this:

$$d^{(i)} = TF(w_i, d) \cdot IDF(w_i) \tag{4}$$

and can be called as weight of word w_i in document d. This measure of weight is extensively used in topic modeling due to the fact that it helps to measure how critical a word is to a document in corpus, thus helping extract better topic representations.

1.8.2 c-TF-IDF

However, this classical approach of TF-IDF takes into account the whole dataset and measure the importance of each word available in the dataset. Having our data clustering, this approach is not ideal. Instead, we would want to extract topic representations in the cluster level, disregarding other clusters. As described in BERTopic paper [20], a modified TF-IDF algorithm, called class-based TF-IDF procedure, can be used. In this algorithm, the cluster is considered a single document by concatenating all the documents in the cluster. Then, the original TF-IDF is adjusted to this representation:

$$c^{(i)} = TF(w_i, c) \cdot log\left(1 + \frac{A}{CF(w)}\right), \tag{5}$$

where $c^{(i)}$ is a class in a vector of classes $\overrightarrow{c} = (c^{(1)}, \ldots, c^{n})$. Class in this case represents a collection of documents concatenated into a single document for each cluster. Then, CF is *class frequency*, which shows the frequency of word across all classes, and lastly - A is the average number of words per class. By using this algorithm, we are able to model the importance of each word throughout the classes, instead of individual documents, which in turn allows to extract topics for each cluster of documents, instead of the whole dataset.

1.8.3 Generative Language Models

Usually, the topic extraction processes described earlier would be sufficient in certain use cases. When using c-TF-IDF for topic extraction, we are able to extract a set of *keywords*, which represents the latent topic in cluster of documents. A quick example of arbitrary topics extracted using c-TF-IDF is presented in Table 1.

The topic representations in Table 1 would be sufficient enough to understand the general topic of documents cluster, however, the topic of this work is narrative search. In literature, the terms *topic* and *narrative* are often used interchangeably [23, 26], as they are used in the context of this work. But if analyzed deeper, one could argue that topic and narrative have similar meaning, but not the exact same. To defend this viewpoint, topic could be understood as a static representation of documents cluster, as shown in Table 1, but narrative would encompass a coherent storyline, visible over the time within a corpus of text. Utilizing dynamic topic modeling allows us to address the temporal aspect of a narrative, but there is a lack of storyline aspect in the provided topic representations.

Table 1. *Arbitrary c-TF-IDF topic representations.*

Topic ID	Representation
0	uk_politics_brexit_elections
1	global_health_covid_climate_change
2	technology_artificial_intelligence
3	sports_football
4	entertainment_movies_awards_celebrities

As stated in the Artificial Intelligence Index Report 2024 [33], in recent years LLMs have become more performant than humans on traditional English language benchmarks, such as SQuAD [46] (QA benchmark) or SuperGLUE [53] (language understanding). These advancements in LLMs have suggested the need for more extensive benchmarks in order to evaluate the ever-growing capabilities of LLMs. One of such benchmarks is called *Massive Multitask Language Understanding* [24]. Without delving into much details of the benchmark, we can say that its purpose is to assess model performance in zero-shot or few-shot tasks across 57 subjects, including STEM, social, humanities sciences. To obtain high accuracy on this benchmark, models must have extensive world knowledge, problem solving ability, context understanding of the questions asked. Furthermore, in January 2024, Google's Gemini Ultra [50] model has surpassed human performance baseline of 89.8% [33] on Massive Multitask Language Understanding test. Also, based on the generation tasks, where AI models are tested to produce fluent and practical language responses, evaluation done by humans on Chatbot Arena platform [14] is constantly increasing, ranking OpenAI's GPT-4 Turbo model as the most performant.

All these recent advances in LLMs suggest that the possibilities to utilize language models increase and the possible tasks solved by LLMs get broader and more complex. An idea worth exploring might be utilizing LLMs for the single purpose of zero-shot topic modeling tasks. This could be a possible application, however, current LLMs still have quite short context windows of input text. If the dataset has a large amount of texts, entering into the big data range, LLMs are still not possible to analyze the whole input due to limited context and input tokens limit. For this reason, to utilize LLMs in topic modeling, we need to reduce the amount of possible texts with before mentioned techniques - dimensionality reduction, clustering and keyword extraction from documents clusters.

Applying these processes enables us to crystallize the essential information hidden in the dataset, thus reducing the noise and the total amount of data. We can then utilize prompt engineering techniques to achieve the further goal of topic extraction. A new paradigm of *"pre-train, prompt and predict"* [31] is becoming more apparent. In this paradigm, instead of adapting pre-trained language models to downstream tasks such as sentiment analysis, downstream tasks are formulated to resemble the original tasks solved during language model training, using a textual prompt. Prompt is a predefined list of instructions provided to LLM that enables to customize the output and capabilities of it [54]. For example, when predicting the sentiment of headline "Government Audit Reveals Widespread Misuse of Public Funds Amidst Economic Downturn", it can be followed by a prompt "This is ___ news for our economy" and ask language model to replace the blank space with a word

that has emotional weight. In this way, this creates an opportunity to utilize the pre-trained, very large language model to predict the desired output. In the scope of this work, this would mean constructing such a prompt, which would accept *keywords of cluster* and *representative documents of cluster* (documents that are closest to the center of cluster) and in turn output *coherent, comprehensive, diverse representation of narrative*. By utilizing this process, we can expect to fulfill the ultimate goal of this work - narrative search in the Information Space.

1.9 Evaluation Metrics and Benchmarking

Machine learning models often are evaluated based on the accuracy scores of predicted results versus ground truth. However, in the case of topic modeling, it is very hard to define ground truth, since the technique is unsupervised, meaning the ground truth is not always available and other metrics to evaluate the results are necessary. Because of this, we can find two groups of evaluation metrics for topic modeling:

- Metrics which allows to evaluate results of topic model independently of a ground truth. This group consists number of metrics, but *coherence* metrics are the most popular ones [48]. Coherence is a measure which evaluate whether statements or facts support each other. An example of coherent fact set would be {"basketball is a sport played with a ball", "basketball is a team sport", "basketball demands great physical efforts"}.
- Metrics that assess the accuracy of a model's predictions by comparing the generated clusters with actual topic labels assigned to each document.

Since the goal of this work is to utilize LLMs to generate topic labels, the entropy this approach creates makes the comparison of generated labels to ground truth obsolete. Topic coherence metrics will be used to evaluate the results.

Topic Coherence (TC) metrics, as mentioned before, tries to evaluate the interpretability of generated topics. Once again, these metrics can be categorized into two classes, named automated TC metrics and human-annotated TC metrics [27]. Automated TC metrics assess the interpretability of topic models based on factors like co-occurrence or semantic similarity of topic words. In contrast, human-annotated TC metrics involve creating surveys that rate or score the interpretability of topic models, relying on human judgment to validate the semantic coherence and meaningfulness of the topics. While human-annotated TC metrics provide a more accurate and nuanced understanding of a topic model's performance by capturing the underlying themes in a text corpus, they are costly, time-consuming, and require multiple human participants to avoid personal biases. On the other hand, automated metrics are more cost-effective, as they do not necessitate hiring and training human annotators, enabling the evaluation of large datasets and numerous model comparisons.

1.9.1 Automated Topic Coherence Metrics

A high TC value suggests that the words within the topic are semantically similar and tend to co-occur in similar contexts. With this in context, a quick introduction on some of the metrics that are planned to be used in this work will be present.

The first metric in the list is **Normalised PMI (NPMI)** [3]. This method uses context vectors for each topic word *w* to generate the frequency of word co-occurrences within a window of ± 1 words surrounding all instances of *w*. NPMI values range from -1 to 1, with higher values indicating greater coherence. A value of 1 means perfect co-occurrence, 0 indicates independence, and -1 suggests complete anti-correlation. NPMI has been shown to have a stronger correlation with human ratings compared to other metrics. Formula for *NPMI* calculation is provided in Equation 6.

$$NPMI(w_i^r, w_i^s) = \frac{\log_2 \frac{P(w_i^r, w_i^s)+\epsilon}{P(w_i^r)P(w_i^s)}}{-\log_2(P(w_i^r, w_i^s)+\epsilon)} \tag{6}$$

UMass [38] metric proposes an asymmetric confirmation measure to estimate the coherence between words within a given topic by calculating the log ratio frequency of their co-occurrences in a document corpus. UMass counts the number of times a pair of words co-occur in the corpus and compares this to the expected number of co-occurrences if the words were randomly distributed across the entire corpus. More formally, UMass computes the co-document frequency of words w_i^r and w_i^s divided by the document frequency of word w_i^s. UMass values are typically negative, and less negative values indicate higher coherence, meaning the words in the topic co-occur more frequently and are more semantically related. *UMass* is provided in Equation 7.

$$UMass(w_i^r, w_i^s) = \log \frac{D(w_i^r, w_i^s)+\epsilon}{D(w_i^s)}. \tag{7}$$

1.9.2 Contextualized Topic Coherence

Contextualized Topic Coherence (CTC) has been recently brought up in research to evaluate new measures for the quality of topic models. CTC are context-aware family of topic coherence metrics based on the LLMs [44]. The last plan is to propose relatively new topic coherence measures, based on the narratives found in previous steps and utilize LLMs for the evaluation of results.

Intrusion is a measure aimed to identify the words that do not belong to a category of topic. These so called intruder words are usually detected by human evaluation to assess the quality of topic model and gives a higher score if extracted topics do not have intruder words. Replacing humans with LLMs to identify the intruder words is the main idea of this metric. As suggested in the original paper [44], LLM prompt for **Intrusion** evaluation is provided in Listing 1.

Listing 1. Prompt for $CTC_{Intrusion}$ evaluation.

```
I have a topic that is described by the following keywords: [KEYWORDS].
The topic can be described as this: [TOPIC DESCRIPTION]

Identify all intruder words in the list with respect to the topic
   keywords and description provided.
The results should only be in the following format without any
   formatting: intruders: <words in a list>
```

The number of intrusion words ($|I_i|$) returned by LLM for each topic i is used to define $CTC_{Intrusion}$ as follows:

$$CTC_{Intrusion} = \sum_{i=1}^{n} \frac{1 - \frac{|I_i|}{m}}{n}, \tag{8}$$

where n is the number of topics identified by the model and m is the number of words in the specific topic.

Rating measures the usefulness of the topic words for retrieving documents on a given topic. As stated in the original paper [44], human topic ratings are expensive to produce, but they often serve as the basis of evaluation, especially for subjective outputs, such as topics or narratives. Humans (and in this case, LLMs) are asked to evaluate a randomly selected subset of topics for their usefulness and score each topic on a 3 point scale, where 3 equals highly coherent topic and 0 - useless (or less coherent). The prompt used for evaluation of **Rating** is presented in Listing 2 [44].

Listing 2. Prompt for CTC_{Rating} evaluation.

```
I have a topic that is described by the following keywords: [KEYWORDS].
The topic can be described as this: [TOPIC DESCRIPTION]

Evaluate the interpretability of topic words on a 5-point scale where 5
   = 'meaningful and highly coherent' and 0 = 'useless' as topic words
   are usable to search and retrieve documents on a single particular
   subject.
The results should only be in the following format without any
   formatting: score: <score>
\label{prompt:rating}
```

The CTC_{Rating} for a model then can be obtained by the average sum of all ratings over all the topics.

1.9.3 Topic Diversity

Topic Diversity is a rather simple metric. It shows the proportion of unique words across topics. If the diversity is close to 0, it may indicate redundant topics. Otherwise, if diversity is close to 1, it indicates more varied topics [17].

$$\text{Topic Diversity} = \frac{|\text{UniqueWords}|}{k \times |\text{Topics}|}, \tag{9}$$

where k is the number of top words considered from each topic.

1.10 Section Summary

The literature review of this thesis provides a comprehensive exploration of various methodologies and applications of topic modeling, particularly focusing on dynamic topic modeling (DTM)

and the integration of advanced language models. These are the key conclusions that can be drawn from the literature review:

1. **Growing Importance of Topic Modeling**. The vast expansion of digital information requires sophisticated methods for data analysis and retrieval. Topic modeling has emerged as a vital tool for organizing and summarizing large text corpora, facilitating document classification, information retrieval, content recommendation, and trend analysis.

2. **Dynamic Topic Modeling (DTM)**. Traditional topic models, like Latent Dirichlet Allocation (LDA), fail to account for the temporal evolution of topics. DTM addresses this limitation by modeling the evolution of topics over time, capturing how discussions and narratives shift. This capability is particularly useful for analyzing trends in social sciences, market research, and media monitoring.

3. **Advanced Techniques in Topic Modeling.** The integration of advanced techniques, such as embeddings, dimensionality reduction, and clustering algorithms, enhances the effectiveness of topic modeling. These techniques enable more precise and context-aware analysis, improving the identification and tracking of evolving topics.

4. **Application of Large Language Models (LLMs).** LLMs have significantly advanced natural language processing (NLP) capabilities. These models offer sophisticated means for text representation and understanding, facilitating improved topic extraction and narrative search. The use of LLMs in zero-shot or few-shot tasks showcases their potential in enhancing topic modeling.

5. **Evaluation Metrics.** Evaluating the performance of topic models is challenging due to the lack of ground truth in unsupervised learning tasks. Topic coherence metrics, both automated (e.g., NPMI, UMass) and contextualized (e.g., CTC-Intrusion, CTC-Rating), provide valuable insights into the interpretability and meaningfulness of extracted topics. These metrics are crucial for assessing the quality and effectiveness of topic models.

6. **Practical Applications and Case Studies.** The literature review highlights practical applications of topic modeling in various domains, such as analyzing international media coverage, tracking disinformation campaigns, and understanding public opinion trends. These case studies demonstrate the real-world relevance and impact of topic modeling techniques.

After the literature review, we can crystallize the main idea and problem that is solved by this thesis. The aim is to address the challenge of managing and analyzing large amounts of data available in the information space using dynamic topic modeling techniques. Specifically, the goal is to develop a process to identify, track, and analyze time-varying topics in textual datasets. The problem can be summarized as follows: *Information Overload* is a significant issue due to the growing amount of data on the Internet, causing people and systems to struggle with processing and making sense of vast amounts of information efficiently. *Topic Identification and Tracking* aim to implement and improve dynamic topic modeling techniques to not only identify topics within large text corpora, but also track the evolution of these topics over time. Additionally, the *Improvement of Narrative Search*

involves applying advanced natural language processing techniques to refine the process of evolving narrative search within large datasets, allowing the retrieval and understanding of coherent story lines or narratives dynamically and efficiently. The ultimate goal is to extract topics that are easy to understand and accurately reflect the information present in documents within a particular cluster.

In summary, this review of the literature underscores the critical role of topic modeling in managing and analyzing the ever-growing information space. By integrating advanced language models and dynamic modeling techniques, it is possible to better understand and navigate the complex landscape of digital information, uncovering valuable insights and trends.

2 Methodology

The previous section of the literature review provides an overview of topic modeling and how it is usually performed. Since the goal of this work is to define a process based on topic modeling to identify time-varying topics in textual datasets, it is time to define the plan and methods for exploring topic modeling possibilities.

2.1 Dataset

The dataset for this study was collected between mid-September 2024 and late December 2024. It includes data from various Lithuanian sources, such as websites, press, social media, TV, radio, and podcasts. The data was gathered specifically to extract texts related to the 2024 Lithuanian parliamentary election, by filtering relevant media articles via named entity recognition, keyword filtering and manual selection based on the entities provided in Table 2. It was collected using web and social media scraping tools via third-party providers and automatic speech-to-text transcription for audio and video segments. In total, there were **7385** documents collected and constructed into the dataset. The majority of the data were collected in from Lithuanian sources, but to simplify processing, texts are translated into English via cloud translation service, namely Azure AI Translator. This cloud translation service has shown reliability in preserving both the original meaning and the integrity of entities within the translated text over texts used in other projects, not only in this study.

Table 2. *Key entities used for the data collection process.*

Party	Key Entities
Nemuno Aušra	Remigijus Žemaitaitis, Agnė Širinskienė
Taikos Koalicija	Viktoras Uspaskich, Mindaugas Puidokas
Laisvės partija	Aušrinė Armonaitė, Tomas Raskevičius
Liberalų sąjūdis	Eugenijus Gentvilas, Viktorija Čmilytė-Nielsen
Socialdemokratai	Gintautas Paluckas, Vilija Blinkevičiūtė, Juozas Olekas
Demokratų sąjunga „Vardan Lietuvos”	Saulius Skvernelis, Lukas Savickas
Lietuvos valstiečių ir žaliųjų sąjunga	Ramūnas Karbauskis, Aurelijus Veryga
Tėvynės sąjunga-Lietuvos krikščionys demokratai	Ingrida Šimonytė, Gabrielius Landsbergis, Laurynas Kasčiūnas
Tautos ir teisingumo sąjunga	Petras Gražulis, Artūras Orlauskas
Lietuvos liaudies partija	Eduardas Vaitkus

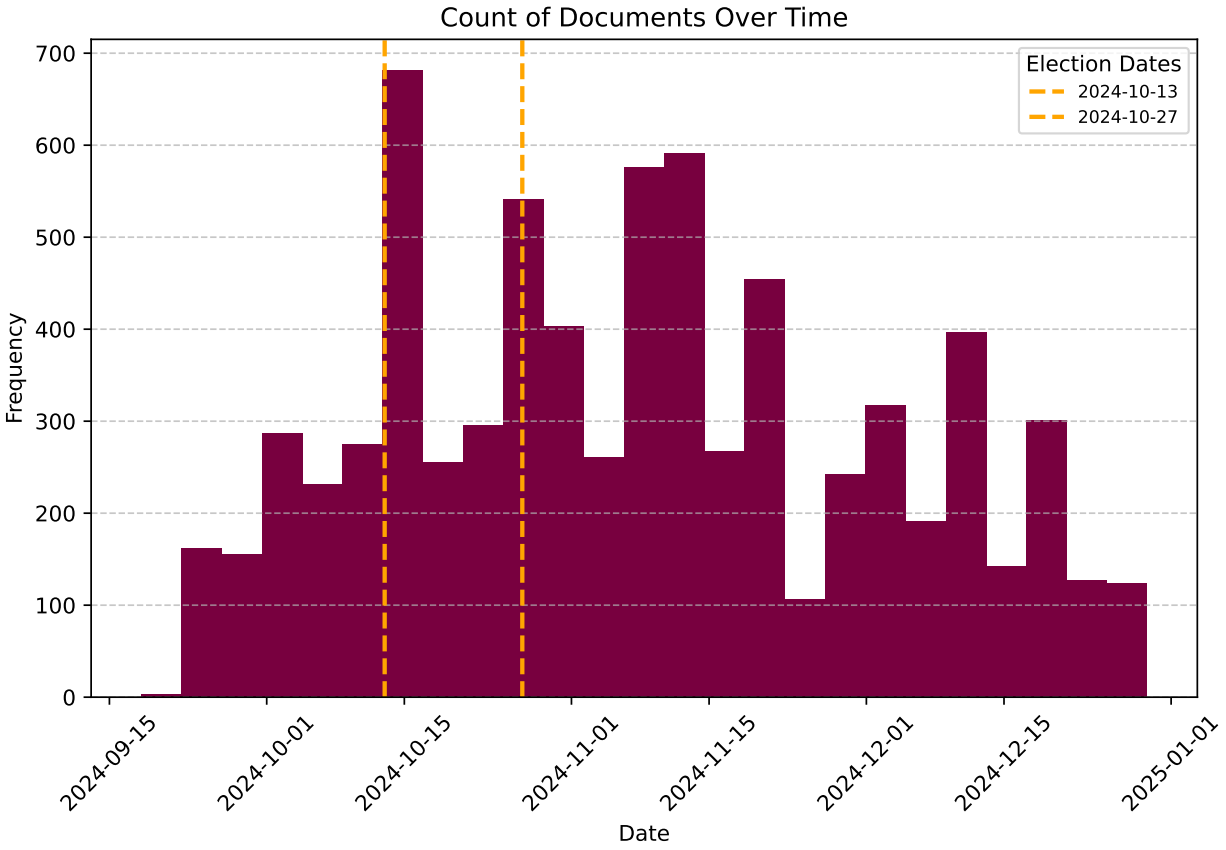

Figure 5. *Count of documents over time in the dataset used.*

The 2024 Lithuanian parliamentary election was held in two rounds on October 13 and 27 to elect 141 members of the Seimas. It marked a significant shift in the political landscape, with the Social Democratic Party emerging as the largest party. Voter turnout increased compared with that in 2020, with 52.2% in the first round and 41.04% in the second round.

In Lithuania, during the first round of elections, people vote for a party, whereas in the second round, they vote for constituency members. An overall picture of the media coverage is shown in Figure 5. As expected, the first round, which took place on October 13, presented the highest volume of published documents. However, during the second round on October 27, media coverage was lower. Considering the context of elections, this trend likely reflects the nature of the public interest and media focus. The first round generally attracts more attention as it determines the broader political landscape, including which parties secure seats in parliament. In contrast, the second round involves a narrower focus on individual constituencies, which may generate less nationwide media attention. Given the data collection methods, it is likely that documents related to the second round focused more on specific individuals, which may have resulted in fewer documents being captured.

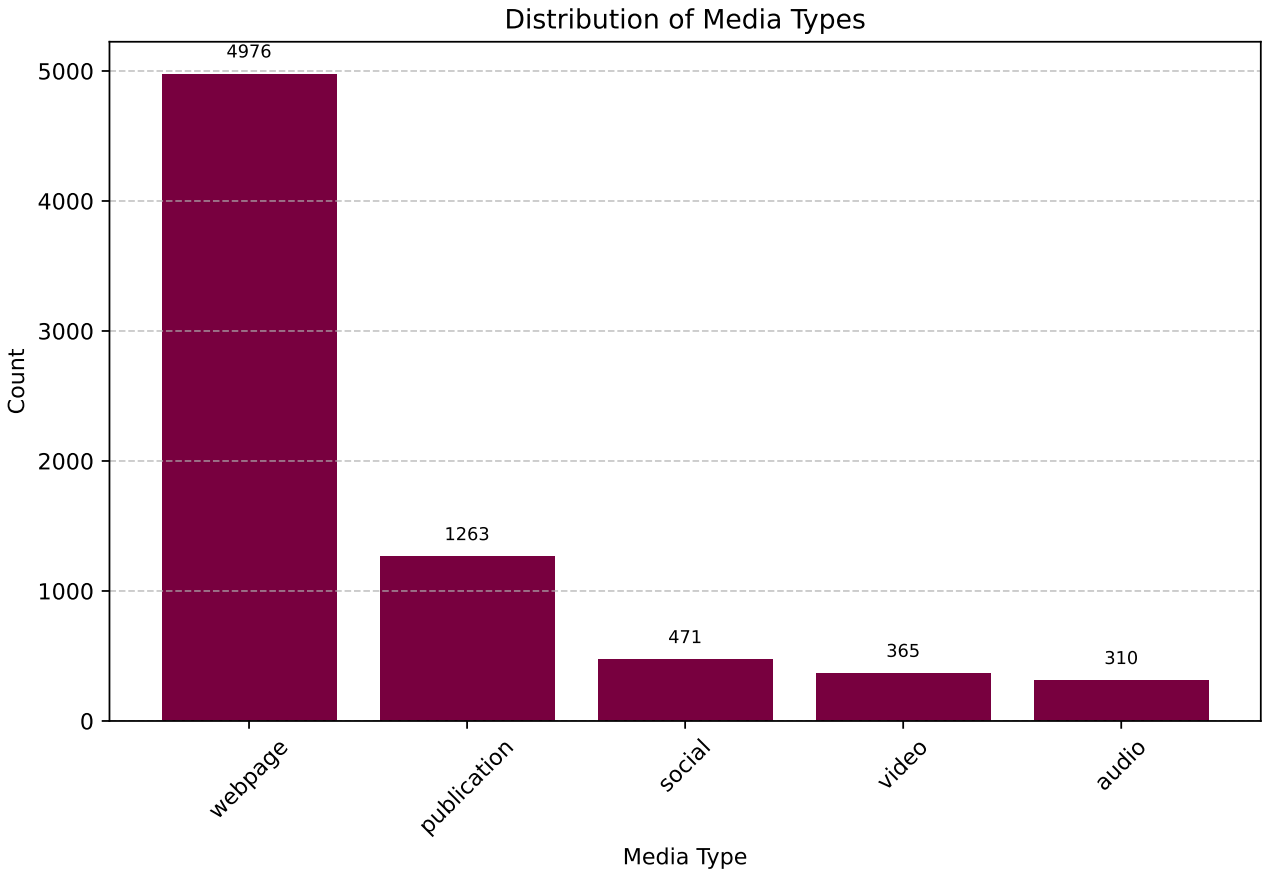

Figure 6. *Distribution of media types in the dataset used.*

The distribution of media types within the dataset is illustrated in Figure 6. It highlights entry counts for each category: webpage, publication, social, video, and audio. The webpage category has the highest count, followed by publication, while audio has the lowest. For clarity, exact counts are displayed above each bar.

To improve processing efficiency and reduce translation costs, the texts were segmented into sentences, including only those relevant to the theme. While this approach reduces computational demands and makes tasks such as calculating embeddings and translation less resource-intensive, it has several drawbacks. By isolating sentences from their original context, the dataset loses the broader thematic and contextual coherence of the full texts. This reduction in topical structure may limit the depth of analysis that could be achieved with complete texts.

2.2 Embedding Model

In this work, Jina Embeddings v2 [21] is used in order to obtain text embeddings, which are later used for downstream tasks. Jina Embeddings are relatively new open-source text embedding models designed to handle extended text sequences up to 8192 tokens, significantly exceeding the 512-token limit common in most embedding models such as BERT. By incorporating innovations such as Attention with Linear Biases (ALiBi) [43] for efficient positional encoding, the model can process long documents while maintaining high performance on standard natural language processing benchmarks.

Jina Embeddings utilizes a modified BERT architecture trained on a large text corpus. The model is fine-tuned using a combination of text pairs and hard negatives, optimizing its performance for tasks such as information retrieval, clustering, and text classification. Compared with chunk-based approaches, the model's ability to embed long texts into single vector representations significantly reduces memory and computational overhead.

Jina Embeddings is a highly effective model for clustering and topic modeling due to its robust semantic encoding capabilities. It demonstrates strong performance in clustering tasks, as evidenced by its high scores on the Massive Text Embedding Benchmark (MTEB) [39], showing its ability to create meaningful semantic groups. The model's fine-tuning approach on diverse datasets improves its adaptability across various text domains, making it versatile for different clustering and topic modeling applications. Additionally, its design prioritizes computational efficiency, enabling this study to use fewer computational resources and faster iterations of different topic models. The final output of the Jina Embeddings v2 model is a 768 element-long embedding vector.

2.3 Topic Models

Topic models are computational methods designed to uncover hidden thematic structures within large text datasets. These models provide a means to organize and summarize textual data, revealing patterns and relationships that are not immediately apparent.

This section explores various approaches to topic modeling, beginning with probabilistic topic models like Dynamic Topic Model (DTM). Following this, algorithmic topic models are analyzed, including BERTopic and the Aligned Neural Topic Model (ANTM), which leverage advancements in neural embeddings and clustering to represent topics dynamically, are subsequently analyzed. These methodologies highlight the evolution of topic modeling, from probabilistic foundations to neural and algorithmic innovations, offering diverse strategies for analyzing complex textual datasets.

2.3.1 BERTopic

The previously introduced Dynamic Topic Model (DTM) is best characterized as a probabilistic dynamic topic model. It assigns probabilities to words and topics over time, enabling researchers to infer evolving themes and patterns in data. Probabilistic topic models like DTM operate under the assumption that each document in a corpus is a mixture of topics, with each topic represented as a probability distribution over the corpus vocabulary.

In contrast, BERTopic and ANTM fall under the category of algorithmic topic models. These models leverage advances in neural document and word representation techniques, such as embeddings, and employ numerical optimization methods to identify topics. Here, topics are represented as weighted word vectors derived from clusters of semantically similar documents.

BERTopic [20] generates topic representations in three steps:

1. **Embedding Documents**: Each document is converted into an embedding using a pre-trained language model. BERTopic's architecture allows flexibility in the choice of language models, as long as they produce the required embeddings.

2. **Dimensionality Reduction and Clustering**: The dimensionality of the embeddings is reduced using the UMAP algorithm [35] to optimize the clustering process. Clustering is performed using the HDBSCAN algorithm [34], as detailed in Sections 1.6 and 1.7.

3. **Topic Extraction**: The topics are derived from document clusters using c-TF-IDF, a class-based variation of the TF-IDF algorithm discussed in Section 1.8.2.

To accommodate dynamic topic modeling, BERTopic integrates temporal information into its analysis. This approach assumes that while the temporal context influences the representation of topics, global topics remain consistent across time. For example, consider a global topic related to climate change. This topic might consistently include general terms such as ”global warming” and ”greenhouse gases” across different periods. However, documents from the 1990s may frequently reference terms like ”Kyoto Protocol” and ”ozone depletion,” while more recent documents might focus on ”Paris Agreement” and ”carbon neutrality.” Although the specific terms change over time, they all relate to the broader topic of climate change.

To generate dynamic topic representations, BERTopic follows a two-step process:

1. The entire corpus is processed without considering temporal aspects, creating a global overview of topics.

2. For each timestamp *i*, local topic representations are created by multiplying the term frequencies of the documents of that timestamp with the global IDF values. This avoids the need for re-embedding or re-clustering documents, enabling efficient computation.

This approach effectively separates global topic identification from temporal dynamics, allowing BERTopic to represent topics consistently while incorporating temporal variability. By doing so, it ensures scalability and computational efficiency in dynamic topic modeling.

Although it is possible to observe how topic representations differ over time, the representation in timestep *t* is independent of the representation in the previous timestep *t-1*. Consequently, this approach to dynamic topic modeling may not naturally produce linearly evolving narratives. By definition, a narrative implies that the representation of the topic at time *t* is influenced by its representation at *t-1*. To address this, the c-TF-IDF vector at each timestep is first normalized using the L1 norm. Then, to compute the topic representation at timestep *t*, it is averaged with the representation from the previous timestep, *t-1*. This process smooths the topic representations across their temporal sequence, ensuring a more coherent temporal evolution.

The experiments in this work will be carried out using the library *BERTopic*, published by the original BERTopic author Maarten Grootendorst, available at GitHub.

2.3.2 Aligned Neural Topic Model

The Aligned Neural Topic Model (ANTM) improves dynamic topic modeling by addressing limitations in BERTopic and introducing a methodologically coherent approach to analyzing evolving themes. BERTopic applies UMAP for dimensionality reduction and HDBSCAN for clustering, producing static clusters that are represented dynamically through term frequency normalization across

time-specific corpora. However, this approach treats each temporal representation independently, resulting in fragmented or redundant topic mappings across time.

ANTM overcomes this limitation through a layered architecture that ensures temporal coherence and interpretability. It consists of three key components:

1. **Contextual Embedding Layer (CEL)**: Pre-trained Transformer-based large language models generate time-aware embeddings, capturing both semantic and temporal contexts of documents.

2. **Aligned Clustering Layer (ACL)**: Documents are grouped into clusters within overlapping time frames using aligned dimensionality reduction (AlignedUMAP [29]) and HDBSCAN. These clusters are then aligned across time to create evolving topics, ensuring continuity and capturing narrative evolution.

3. **Representation Layer (RL)**: Relevant terms are extracted for each cluster using c-TF-IDF.

This architecture enables ANTM to align topic representations across consecutive time steps, ensuring smooth and meaningful transitions while maintaining semantic integrity. Using a neural alignment mechanism, ANTM preserves the narrative progression of the themes, avoiding the disjointed snapshots produced by BERTopic.

Additionally, ANTM incorporates a multi objective optimization framework that balances topic coherence, distinctiveness, and temporal alignment. This approach ensures that high-quality, interpretable topics evolve consistently over time. Experimental evaluations show that ANTM surpasses BERTopic in maintaining topic continuity, reducing redundancy, and improving the interpretability of temporal trends.

The experiments in this study will utilize the *ANTM* library, developed by the original ANTM author, Hamed Rahimi, and accessible on GitHub.

2.4 Prompts for Narrative Extraction

This section outlines the prompts utilized for extracting information about documents and keywords within topics after preprocessing through topic models. Specifically, the prompt used for topic label extraction is detailed in **Listing 3**. This particular prompt was selected because of the characteristics of the dataset, which primarily consists of short sentences. Given the limited context provided by such data, generating accurate descriptions can be difficult. To avoid this issue, the prompt in **Listing 3** was selected as the most suitable prompt. However, for datasets with longer documents and richer context, the prompt illustrated in **Listing 4** would likely produce more coherent and diverse descriptions, potentially yielding more interesting results.

Listing 3. Prompt for topic label extraction.

```
In this topic, the following documents are a small but representative
   subset of all documents in the topic:
[DOCUMENTS]
The topic is described by the following keywords: [KEYWORDS]

The dataset is regarding Parliament Elections in Lithuania. You should
   extract a detailed and descriptive topic label that includes key
   figures, major events, political parties, and significant themes
   related to Parliament Elections in Lithuania.

Based on the information above, provide a detailed and descriptive topic
   label in the following format:
topic: <topic label>
```

Listing 4. Prompt for topic summary extraction.

```
I have a topic that is described by the following keywords: [KEYWORDS]
In this topic, the following documents are a small but representative
   subset of all documents in the topic:
[DOCUMENTS]

The dataset is regarding Parliament Elections in Lithuania. You should
   try to extract topics that includes key figures, major events,
   political parties, and significant themes related to Parliament
   Elections in Lithuania in the topic description.

Based on the information above, please give a concise description of
   this topic in the following format:
topic: <description>
```

2.5 Topic Alignment Metric

Another metric for evaluating the extracted topic keywords and descriptions could be named as **Topic Alignment** or, in the notation from Section 1.9.2, CTC_{TA}. The primary objective of this work is not only to extract keywords from text corpora that represent topics but also to generate human-readable narratives. Ideally, the extracted keywords and the corresponding narrative should align. However, this is not always achieved in topic modeling, particularly when the extracted keywords include stop words or highly frequent terms, making it difficult to comprehend the actual topic of a document cluster. When both keywords and representative documents are provided to a generative

model, the model can infer the thematic structure of the documents and produce a comprehensible topic label, even when the keywords are of poor quality. Examples of such prompts can be found in **Listings 2** and **1**.

After the topic label is obtained, the quality of the extracted keywords can be further evaluated. Using the prompt provided in **Listing 5**, we can assess the alignment between the keywords and the topic description.

Listing 5. Prompt for CTC_{TA} evaluation.

```
The topic is described by the following keywords: [KEYWORDS]
The topic can be described as this: [TOPIC DESCRIPTION]

Evaluate the given keywords and whether they match the provided topic
   description. Evaluate this on a 3-point scale, where 3 = 'high
   keywords and topic description match' and 0 = 'no match'.

The results should be in the following format:
score: <score>
```

The CTC_{TA} metric for a model can then be calculated as the average of all ratings across all topics.

2.6 Proposed Process

The methodology presented earlier can be summarized into a comprehensive process for narrative discovery and analysis in large-scale textual data. Beginning with the collection and preprocessing of diverse data sources, this approach transforms unstructured information into meaningful insights through different stages, as shown in Table 3. By using embedding techniques with dimensionality reduction and clustering, the process is able to identify patterns that might otherwise remain unseen. The application of keyword extraction and LLM-assisted narrative generation transforms these clusters into human-interpretable narratives. This process, supported by evaluation metrics and visualizations, enables to effectively map the landscape of public discourse and track emerging narratives.

Table 3. *Key entities used for the data collection process.*

Stage	Action	Methods
Data collection and preprocessing	Collect data and transform into textual representation	TV, radio recordings; speech-to-text models; internet scraping techniques
Text embeddings	Transform textual dataset into numerical representations known as embeddings	Embedding models

Dimensionality reduction	Remove *curse of dimensionality*, preserve high-dimensional structure in low dimensions	UMAP
Clustering	Discover dense semantic regions, drop outliers	HDBSCAN
Keyword-based topic extraction	Summarize each cluster as a weighted term list	c-TF-IDF
Narrative generation	Turn sparse keywords + sample docs into human-readable labels and summaries	LLM prompts (*pre-train* → *prompt* → *predict* paradigm)
Evaluation	Evaluate quality and coherence of extracted narratives	Automated (**NPMI** and **UMass**), **Topic Diversity**, and LLM-based (**CTC_{TA}**, **$CTC_{Intrusion}$**, **CTC_{Rating}**) metrics
Visual analysis	Topic popularity curves, evolution graphs	BERTopic and custom visualizations

2.7 Section Summary

The methodology section of this study showcases a process to explore time-varying topics in textual datasets, focusing on the 2024 Lithuanian parliamentary election. A dataset of 7,385 documents sourced from Lithuanian media, was translated into English using Azure AI Translator. Text embeddings were generated using Jina Embeddings v2, enabling efficient and scalable processing for clustering and topic modeling tasks.

Dynamic topic modeling approaches BERTopic and ANTM were analyzed for their ability to capture evolving themes. BERTopic relies on temporal c-TF-IDF transformations to represent dynamic topics, while ANTM ensures continuity and coherence through aligned clustering and neural mechanisms.

This study shows the importance of narrative extraction, utilizing LLMs prompts to generate human readable topic labels and descriptions. A new evaluation metric, CTC_{TA}, evaluates the alignment of extracted keywords with generated topic descriptions.

3 Experimental Evaluation

In this section, the application results of the previously mentioned topic models (Sections 2.3.1 and 2.3.2) on dataset mentioned in Section 2.1 will be presented. The aim of each model is to fine-tune the topic model with corresponding parameters, evaluate the base results and later use enhanced narrative extraction methods to extract and present narratives.

3.1 BERTopic Results

This section presents the results of topic modeling using the BERTopic model. BERTopic combines techniques such as dimensionality reduction, clustering, keyword extraction, and narrative extraction, allowing to adjust various different parameters. Since topic modeling is an unsupervised algorithm, multiple iterations of experimentation are required to identify optimal parameters, a process known as fine-tuning. The topic modeling results are summarized in Table 4. Full list of generated topics can be seen in Appendix A. Additionally, Dynamic Topic Modeling results are presented in Section 3.1.2.

3.1.1 Static Topic Modeling

Summary of BERTopic parameters is presented in Table 5. The presented parameters were selected by searching for the best parameters combination in the provided search space. The model with overall highest evaluation metrics were selected and the metric is shown in **Value** column. In the table, an additional algorithm is provided, which was not mentioned earlier - **Maximal Marginal Relevance** or MMR [11]. MMR is a technique designed to balance relevance and diversity in keyword selection. It achieves this by evaluating how closely each keyword aligns with the main topic of the document while also considering how distinct it is from the keywords that have already been selected. This approach aims to ensure that the chosen keywords are not only relevant to the document but also diverse, attempting to reduce redundancy and provide a broader representation of the documents content.

Table 4. *A summary of topics and narratives extracted by BERTopic in 2024 Lithuanian parliamentary election context.*

ID	Count	Keywords	Narrative
1	1142	zemaitaitis, dawn, nemunas, nemunas dawn, remigijus zemaitaitis, remigijus, dawn nemunas, leader nemunas, leader nemunas dawn, zemaitaitis leader	The political landscape in Lithuania surrounding the ruling coalition including the ”Nemunas Dawn” party, led by Remigijus Žemaitaitis, amidst controversies regarding his leadership, candidacy nominations for key ministries, and public protests against alleged anti-Semitic rhetoric. Key figures include Prime Minister-designate Gintautas Paluckas and Agnė Širinskienė, with implications for coalition dynamics and international criticism of the partnership due to legal and ethical concerns.
2	651	karbauskis, peasants, skvernelis, lvzs, saulius skvernelis, saulius, union, greens, peasants greens, ramunas	This topic discusses the political dynamics and coalition negotiations surrounding the Lithuanian Peasants and Greens Union (LVŽS), led by Ramūnas Karbauskis, and the Democratic Union ”For the Sake of Lithuania,” headed by Saulius Skvernelis, in the context of the recent Seimas elections. It highlights the electoral performance of these parties, their leadership positions, and the complexity of forming a governing majority, addressing issues of political responsibility, coalition stability, and the implications of past political cultures on future partnerships.
3	494	gintautas, paluckas, gintautas paluckas, prime, prime minister, minister gintautas, prime minister gintautas, gintautas paluck, mr paluckas	The candidacy and appointment of Gintautas Paluckas as Prime Minister of Lithuania, including coalition negotiations and the formation of a new government cabinet, alongside discussions on policy proposals, budget considerations, and ministerial appointments, amidst the backdrop of the social democratic political landscape and recent electoral outcomes.

ID	Count	Keywords	Narrative
4	372	blinkevicuite, vilija, vilija blinkeviciute, chairwoman, prime minister, leader social, leader social democrats, prime, blinkevicuite chairwoman	The topic centers around Vilija Blinkevičiūtė, the chairwoman of the Lithuanian Social Democratic Party (LSDP), and her decision not to pursue the position of Prime Minister following the party's electoral success in the Seimas elections. It highlights the political dynamics of coalition building with other parties, including the Democratic Union "For the Sake of Lithuania," and reflects on the implications of her leadership decisions, party strategies, public sentiment, and challenges within the political landscape of Lithuania. The discussions also touch upon themes of trust and reliability in political commitments, as well as reactions from various political figures and parties regarding the coalition and government formation.
5	277	kasciunas, laurynas kasciunas, laurynas, defense, minister national, national defense, minister national defense, defense laurynas, defense laurynas kasciunas, national defense laurynas	The role and statements of Laurynas Kasčiūnas, Minister of National Defense of Lithuania, regarding military developments, defense cooperation with European countries, and support for Ukraine's defense industry during the context of emerging defense challenges in Lithuania and ongoing regional tensions. The discussions include air defense systems, procurement of military ammunition, and collaborative agreements with Germany and Ukraine, reflecting broader national security concerns and the political landscape related to the upcoming Seimas elections.
6	256	cmilytenielsen, viktorija, viktorija cmilytenielsen, seimas viktorija, speaker seimas viktorija, seimas viktorija cmilytenielsen, speaker, speaker seimas, liberal, liberal movement	Discussions surrounding the leadership and actions of Viktorija Čmilytė-Nielsen, Chairwoman of the Liberal Movement and Speaker of the Seimas, in relation to political accountability, coalition dynamics, and parliamentary elections in Lithuania, including responses to other political parties and significant figures like Eugenijus Gentvilas.

Table 5. *Summary of BERTopic parameters.*

Algorithm	Parameter	Value	Search Space
UMAP	`n_neighbors`	15	5, 15, 25, 50, 100
	`n_components`	5	2, 5, 10, 20
	`metric`	cosine	euclidean, cosine
HDBSCAN	`min_cluster_size`	50	10, 50, 100, 200
	`metric`	euclidean	euclidean, l2
	`cluster_selection_method`	eom	eom
CountVectorizer	`stop_words`	english	english
	`ngram_range`	(1, 3)	(1, 1), (1, 2), (1, 3)
	`min_df (min_doc_freq)`	2	1, 2, 5
ClassTfidfTransformer	`reduce_frequent_words`	True	True, False
OpenAI Representation Model	`model`	gpt-4o-mini	gpt-4o-mini
	`nr_docs`	20	10, 15, 20
	`diversity`	0.3	from 0.1 to 1 by 0.1
Maximal Marginal Relevance	`diversity`	0.4	from 0.1 to 1 by 0.1

The analysis revealed the controversies and public reactions surrounding Remigijus Žemaitaitis and the ”Nemunas Dawn” party, which was the most popular topic with 1142 documents. Allegations of anti-Semitic rhetoric by Žemaitaitis ignited public protests and international criticism, raising questions about coalition stability and the ethical dimensions of political partnerships.

The model also emphasized the role of defense and security in political discourse. Laurynas Kasčiūnas, the Minister of National Defense, emerged as a key figure in 277 documents, with discussions focusing on military cooperation and regional security in the context of geopolitical challenges.

The leadership and actions of Viktorija Čmilytė-Nielsen, Speaker of the Seimas and leader of the Liberal Movement, were also explored, focusing on her influence on political accountability, coalition dynamics, and the parliamentary elections in Lithuania. The topic of her consists of 256 documents.

In addition to these key topics, 2636 documents were clustered as outliers, indicating content that did not align closely with any specific topic identified by the model.

Table 6. *BERTopic evaluation results.*

NPMI	Topic Diversity	UMass	CTC_{TA}	$CTC_{Intrusion}$	CTC_{Rating}
-0.2484	0.9333	-14.2050	3.8824	0.9915	3.8824

According to the evaluation scores, the BERTopic model effectively clustered the dataset and provided valuable context for narrative construction. The high *Topic Diversity* metric indicates that the clusters were well-differentiated, with extracted keywords and narratives distinctly representing various politicians or parties during the elections. However, the extracted keywords were suboptimal and lacked sufficient detail to fully capture the topical structure of specific clusters. This limitation is evident in the low *UMass* and *NPMI* scores. The evaluation scores are presented in Table 6.

3.1.2 Dynamic Topic Modeling

The dynamics of narrative popularity over time are illustrated in Figure 7. This section was modeled by using Dynamic BERTopic configuration with same parameters as in Table 5. A trend can be

observed—increasing the volume of documents as election dates approach. Following the elections, media interest in the topic remains elevated for a short period, as expected. However, approximately one month after the elections, a significant decline is observed, eventually reaching an all-time low in frequency.

Table 7. *Dynamic BERTopic evaluation results.*

NPMI	Topic Diversity	UMass	CTC_{TA}	$CTC_{Intrusion}$	CTC_{Rating}
0.0075	0.0050	−11.0928	3.1727	0.9966	3.8252

The evaluation metrics for Dynamic BERTopic are presented in Table 7. Compared to the static BERTopic evaluation results, the performance of dynamic topic modeling declined in terms of the *Topic Diversity* metric. This decline can be attributed to the fact that dynamic topic modeling generated a total of 469 topics over the dataset's time period, which lowered the metrics' statistical significance. Additionally, it is expected that documents within each time slice were clustered into similar groups, resulting in recurring topics across time slices. Because of these reasons, the overall topic diversity was significantly reduced compared to the static topic model. Despite this, the dynamically generated topics showed similar performance to the statically generated topics across other evaluation metrics.

An example of how a narrative evolves over time is illustrated in Figure 8. The analysis begins on 2024-09-25 and focuses on the most prominent discovered narrative, which revolves around Remigijus Žemaitaitis and the political party "Nemunas Dawn". Models with architectures like BERTopic not only uncover topics within text corpora, but also track the evolution of these topics over time. This example effectively demonstrates the ability of such models to construct a coherent and linear narrative focused on a single recurring theme.

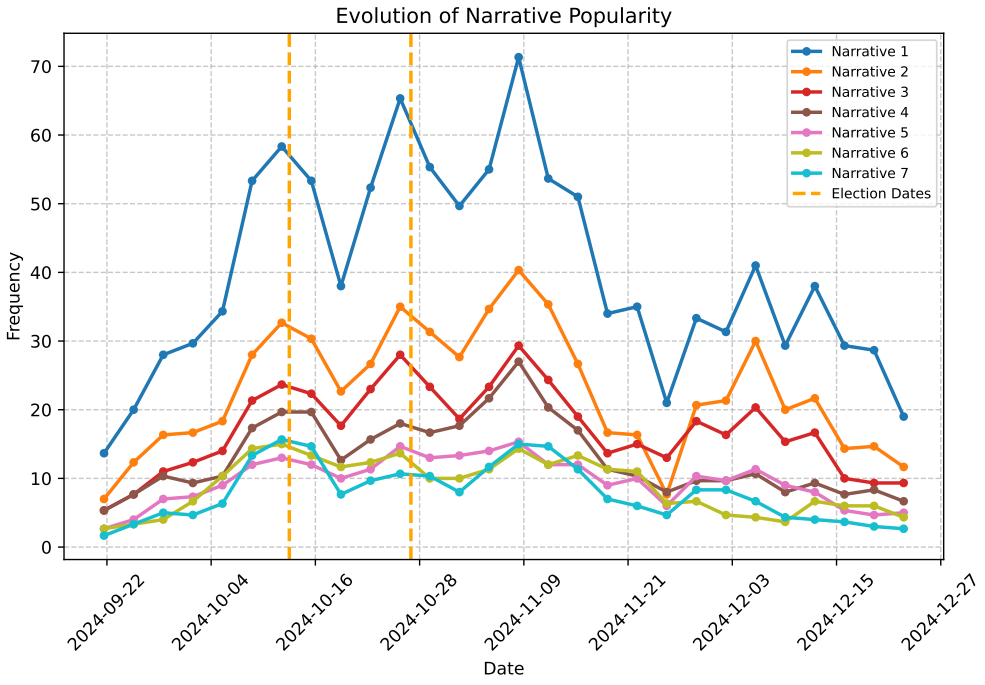

Figure 7. *Evolution of Narrative Popularity over time, modeled by BERTopic.*

Figure 8. *Narrative evolution of most frequent narrative discovered by BERTopic.*

3.1.3 Conclusions

In conclusion, the BERTopic analysis provided valuable insights into the political discussions and events surrounding the 2024 Lithuanian parliamentary elections. By using both static and dynamic topic modeling, the study highlighted key themes, such as the controversies involving the ”Nemunas Dawn” party and its leader Remigijus Žemaitaitis, decisions by political leaders like Gintautas Paluckas and Vilija Blinkevičiūtė, and the focus on defense and security led by Laurynas Kasčiūnas. The static model did a good job grouping related topics, while the dynamic model showed how these topics changed over time. Although there were some challenges with keyword selection, the model proved helpful in understanding and following important political narratives.

3.2 ANTM Results

The second model used in this study was ANTM. Unlike the BERTopic model, ANTM is considerably less popular and is not actively maintained, although the original research paper [45] claims it outperforms BERTopic. Although the implementation is publicly available, modifications to the original model specification were necessary to successfully run the model and perform the intended analysis. Unlike BERTopic, which supports multiple types of topic modeling, ANTM is designed for dynamic topic modeling. As a result, this section focuses on narrative modeling and evaluation within that scope. The parameters used for ANTM model are presented in Table 8. The parameter selection was performed in the same manner as in BERTopic case.

Table 8. *Summary of ANTM parameters.*

Algorithm	**Parameter**	**Value**	**Search Space**
ANTM	`window_size`	6	2, 4, 6
	`overlap`	2	2, 3, 4
HDBSCAN	`min_cluster_size`	10	10, 25, 50
	`metric`	euclidean	euclidean
	`cluster_selection_method`	eom	eom
CountVectorizer	`stop_words`	english	english
OpenAI Representation Model	`model`	gpt-4o-mini	gpt-4o-mini
	`nr_docs`	20	10, 20, 30

ANTM modeled dynamics of narrative popularity are illustrated in Figure 9. Compared to the modeled dynamics of the BERTopic model, it is difficult to understand the actual changes in popularity of certain narratives. For example, the most popular narrative is about the Minister of National Defense Laurynas Kasčiūnas, which has clearly had a spike in popularity on election date and in the aftermath of elections. However, this narrative has two clear and distinct spikes throughout the timeline, while at other dates, the frequency of it was relatively low. This may suggest that ANTM, as a model, has a poor modeling of the temporal aspect of the narrative. Using sliding window and AlignedUMAP method, the model seems to not adjust clusters at every time step, instead re-clusters all documents once again. The narrative stays the same considering thematic aspect but loses temporal connection between time steps.

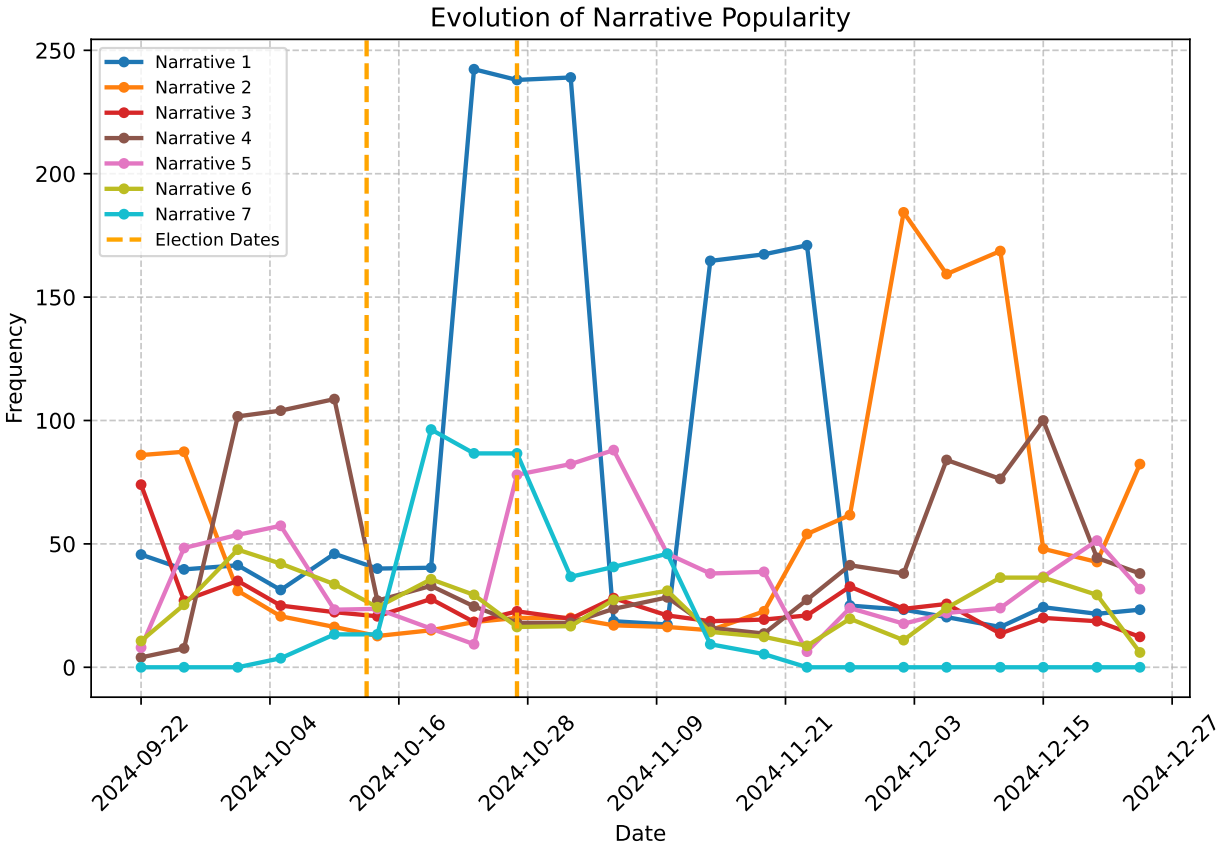

Figure 9. *Evolution of Narrative Popularity over time, modeled by ANTM.*

The evaluation results for ANTM are presented in Table 9. Compared to the previous dynamic model evaluation for BERTopic, what really stands out is the massive increase of *Topic Diversity*. It seems that ANTM has modeled much more diverse narratives over time, which means that keywords in each narrative are more diverse when going over time steps, compared to dynamic BERTopic. Other metrics, such as CTC_{TA}, $CTC_{Intrusion}$ has also improved, but the difference is not that significant compared to *Topic Diversity*. ANTM performed worse than BERTopic when compared based on *NPMI, UMass* and CTC_{Rating}.

Table 9. *ANTM evaluation results.*

NPMI	Topic Diversity	UMass	CTC_{TA}	$CTC_{Intrusion}$	CTC_{Rating}
-0.3455	0.5014	-16.1049	3.7636	0.9976	3.7212

However, just looking at those evaluation metrics and selecting models based on that is not sufficient. Since topic modeling is considered unsupervised learning, the expected output is arbitrary and cannot always be evaluated based on the evaluation metrics. Evaluation should also depend on the expert evaluation, whose task should be to look at the output of models and evaluate whether the generated narratives are coherent, have the same topical aspect through each time step.

Figure 10. *Narrative evolution of most frequent narrative discovered by ANTM.*

The example of narrative evolution modeled by ANTM is illustrated in Figure 10. By examining the initial topic, we can expect that the overall narrative focuses on national defense, collaboration between Germany and Lithuania, and ministers of defense. This expectation is supported by the keywords associated with this topic: *fundamental, attend, commander, minister, mayor, level, eastern, recently, partnership, strategic, NATO, director, bilateral, army, defense, Pistorius, German, national, defense, Laurynas*. This theme remains consistent through steps 1 and 2 of narrative evolution. However, after step 3, the topic shifts to the controversies surrounding R. Žemaitaitis. Interestingly, when reviewing the keywords for step 3, neither R. Žemaitaitis nor his party appears among them. Instead, the keywords include: *majority, Seimas, received, public, may, possibility, criminal, entered, long, inciting, information, scandal, network, June, year, last, Facebook, published, nationally, divisive*.

This shift highlights a limitation of the ANTM technique for dynamic topic modeling. Specifically, it demonstrates the potential for the overall narrative theme to change unexpectedly after a modeling step or for the keywords to be insufficient to capture the actual topic. Addressing these issues requires improving the topic modeling algorithm to ensure better continuity between time steps and improving the keyword extraction algorithm. The latter issue, however, can be partially mitigated by incorporating large language models (LLMs) into dynamic topic modeling, which is a theme of this study. By providing flawed keywords and representative documents of the document cluster, LLMs can summarize the topic more effectively, identify the relevant entities (R. Žemaitaitis), and contextualize the cluster documents, therefore mitigating the flaws in keyword extraction.

3.3 Section Summary

The aim of this study was to explore methods for identifying time-varying topics within textual datasets, otherwise known as narrative search. In particular, a dataset consisting of various text snippets in the context of 2024 Lithuanian parliamentary elections. The methodologies employed were BERTopic static and dynamic model, as well as ANTM dynamic topic model, to extract narratives and evaluate their capabilities. In summary, these are the key insights and findings:

1. **Dataset characteristics**.

    - The data for this study was collected between September and December 2024. It included data from various Lithuanian media sources, such as websites, press, social media, TV, radio and podcasts, with texts translated into English to ease the process of analysis. It included only sentences with relevant keywords to reduce the amount of data needed to translate and embed.
    - The first round of elections which took place on the 13th of October attracted more media coverage then the second round, which took place on the 27th of October. We can say that this is due to the nature of the public interest since second round focuses more on individual politicians, not the parties. The dataset was gathered by looking for most recognizable entities, not for all participators of election, so naturally the number of documents decreased.

2. **BERTopic model insights**.

    - BERTopic is a fairly flexible model which supports both static and dynamic topic modeling. Using static modeling technique effectively clustered and highlighted key political narratives as controversies involving R. Žemaitaitis and his party, coalition negotiations, national defense discussions.
    - Dynamic topic modeling successfully captured the evolution of narratives over time, highlighting the change of popularity in certain narratives, change of inner narrative topics.
    - The challenges associated with using BERTopic include suboptimal keyword extraction and limitations in representing complex narratives, which are reflected by evaluation metrics such as NPMI and Topic Diversity.

3. **ANTM model insights**.

    - ANTM provided a more diverse set of narratives, improving metrics such as *Topic Diversity*. However, the handling of temporal connections and thematic coherence was weaker compared to BERTopic.
    - Narrative evolution modeled by ANTM occasionally deviates from thematic consistency, with keywords failing to capture context of topic. Due to this, a fragmentation of narratives were observed in the dynamic topic modeling.

4. **Model comparison**.

    - BERTopic showed great performance in clustering and representing static topics, while ANTM demonstrated a broader dynamic topic range. However, ANTM struggled to maintain coherent narratives across time steps.
    - Both methods explored in the study benefited from using LLMs to generate narratives, but keywords extraction is an aspect which could be improved on both models.

5. **LLMs and Topic Modeling**

    - The goal of this study was to propose a method of how to extract narratives from a temporal dataset using Large Language Models (LLMs). This work demonstrated that LLMs can not only enhance information retrieval from topic models but also serve as an effective evaluation tool for assessing topic coherence and narrative quality.
    - LLMs were generally able to capture the broader context provided by representative documents in the clustered dataset, thereby expanding upon the limited information offered by keywords alone. This approach combines extracted keywords with representative documents to generate comprehensible narratives that evolve over time.

To finalize the findings, the evaluation results presented in Table 10 show that both BERTopic and ANTM efficiently performed the task of topic modeling on the selected dataset, with comparable results across most evaluation metrics. This proves that topic modeling, as an unsupervised method, requires researchers or analysts to understand the context of dataset to select the most suitable model. Despite their differences, both models offered valuable insights into temporal evolution of narratives within the dataset.

Table 10. *Dynamic topic modeling evaluation results.*

Model	**NPMI**	**Topic Diversity**	**UMass**	**CTC_{TA}**	**$CTC_{Intrusion}$**	**CTC_{Rating}**
ANTM	-0.3455	**0.5014**	-16.1049	**3.7636**	**0.9976**	3.7212
BERTopic	**0.0075**	0.0050	$-$**11.0928**	3.1727	0.9966	**3.8252**

In addition to these comparisons, this study shows how the integration of Large Language Models (LLMs) into the topic modeling enhances extraction and evaluation. LLMs excelled in bridging the gaps of keyword based topic modeling by generating coherent, human readable narratives that capture context of evolving themes over time. These findings shows the potential for further research into combining traditional natural language processing techniques with LLMs capabilities.

Discussion and Conclusions

The conducted study presented a process for extracting and analyzing topics varying in time, or as referenced in this work, narratives, from textual datasets using Dynamic Topic Modeling (DTM) in combination with Large Language Models (LLMs). The dataset focused on 2024 Lithuanian parliamentary elections, which was used to perform research and compare two algorithmic topic modeling techniques - BERTopic and ANTM. Also, this study demonstrated how LLMs can enhance narrative extraction and evaluation. The aim of this section is to combine the main findings, evaluate their importance, discuss the study's limitations and possible paths for further research.

3.4 Key Findings

In this work, the use of algorithmic topic modeling approaches which use embeddings and clustering techniques, were examined. While traditional probabilistic models like LDA [7] treat documents as combinations of topics, algorithmic models such as BERTopic [20] or ANTM [45] that were analyzed in this work, transform text into numerical representations, called embeddings, which allow analysis of the semantic context in documents. This research demonstrated that the algorithmic topic modeling approach enables effective extraction of topics and their evolution over time.

Regarding model performance, both BERTopic and ANTM successfully identified meaningful narratives in the Lithuanian parliamentary election dataset, however, there were some differences. **BERTopic** demonstrated better performance in maintaining temporal coherence of narratives, with a more consistent representation of topics across time steps. Its ability to model narrative evolution was seen in capturing the controversies surrounding R. Žemaitaitis and the "Nemunas Dawn" party, showing how this narrative developed from initial mentions of coalition negotiations to more specific criticisms regarding anti-Semitism implications. BERTopic excelled in $NPMI$, $UMass$, and CTC_{Rating} metrics, suggesting that the extracted topics were more interpretable and coherent.

ANTM produced more diverse topics (as showed by its higher Topic Diversity score of 0.5014 compared to BERTopic's 0.0050), performed better on alignment metric CTC_{TA}, and that topics extracted by ANTM had less intruder words, as shown by $CTC_{Intrusion}$ metric. However, it struggled with maintaining thematic structure across time steps, occasionally shifting narratives, as seen in the example where a defense focused narrative unexpectedly transitioned to the controversies surrounding Žemaitaitis.

These findings show that the choice of the model should depend on the specific analytic needs and the final decision should be made by the context of the research. When narrative coherence over time is preferred, BERTopic may be the option, since it demonstrated better performance on keeping the alignment of topics over time. On the other hand, when a more diverse set of narratives is needed, ANTM might be preferred.

3.5 Integration of Large Language Models

The next thing this work experimented with is the integration of LLMs into topic modeling pipeline. The idea of utilizing LLMs for the purpose of topic modeling tasks were explored. It was based on the concept that traditional topic representation as keywords might be hard to understand, but using LLMs in combination with the keywords and the representative documents extracted from the clustered dataset might produce more human readable topics.

Firstly, the LLM integration was explored in the context of evaluation. The introduction of contextualized topic coherence metrics (CTC_{TA}, $CTC_{Intrusion}$, CTC_{Rating}) using LLMs provides a more nuanced evaluation of topic quality than traditional metrics alone. This usage of contextualized topic coherence metrics, compared to traditional metrics, allows to utilize the context understanding of LLMs, making it more comparable with human judgment.

Then, the goal was to enhance narrative extraction. By providing LLMs with both keywords and representative documents from clusters, the experiments showed how they can generate comprehensive, context aware narrative descriptions that overcome limitations in keyword extraction. This approach proved particularly important when extracted keywords were insufficient to capture complex political narratives or when the keywords alone were not sufficient enough to understand the complex nature of the documents. By using LLMs to generate such topic names and descriptions, the context of each cluster and the narratives in the dataset becomes more understandable.

3.6 Lithuanian Political Landscape

The analysis performed by using topic models revealed some insights of Lithuanian political landscape during the 2024 parliamentary elections. From the dataset alone, it was possible to see a clear pattern of media interest growing around election dates, which particularly reached its peak during the first round on October 13, followed by gradual decline in attention that reached lowest levels in December. During the election period, the most popular narratives centered on controversial figures and events, with focus on R. Žemaitaitis and accusations of anti-Semitism which was significantly covered by media. Defense and security emerged as high popularity narratives, especially in discussions regarding Minister of National Defense L. Kasčiūnas, showing the wider context of geopolitics which was and remains and important theme of Lithuanian politics.

3.7 Limitations

Despite all the important findings and results of this work, there were several limitations which should be acknowledged. Solving these limitations could potentially improve the results in this study and future research.

Translation effects should be taken into consideration. The dataset consisted of texts translated from Lithuanian to English, which may have resulted in inaccuracies in the actual semantic meaning of texts. Translation involves some degree of interpretation, and having in mind the context of selected dataset, which consisted of texts during Lithuanian parliamentary elections, some of the

political terminology, cultural background may have lost its true meaning. This limitation could have been avoided if the selected embedding model would have been trained on actual Lithuanian data. However, due to the large volume of texts analyzed and the dataset having already been translated and embedded, the decision was made not to employ a separate model.

The sentence level segmentation approach might also have its own limitations. To manage the amount of documents needed to analyze, the dataset consisted only of sentences with relevant political entities mentioned, not full texts. This data segmentation might have potentially compromised the contextual semantic relationships between sentences and decreased the clustering quality. Document level analysis could have been more effective alternative in this case.

Next, the dataset consisted of documents which were filtered based on specific political entities, selected by analysts who conducted the actual data gathering. This filtering might have excluded relevant texts that do not directly mention the political entities, but are relevant to Lithuanian parliamentary elections. A better approach and a possibility for future research would have been the use of topic models to identify relevant texts that have the semantic context suitable for this research.

Finally, the selected evaluation strategy can also be considered a limitation. The evaluation metrics used in this research were of two types: automated and LLM-based. LLM-based metrics imitate human assessment, trying to bridge the gap between automated and human evaluation. However, they might maintain some degree of subjectivity and not align with human analyst judgment in specific domains, such as politics. To have better evaluation results or even try to find correlation between human and LLM judgment, the results could be evaluated by political experts or domain specialists familiar with Lithuanian political discourse to assess the validity of the identified narratives.

3.8 Future Research

With limitations of this study assessed, it is easier to explore possible ways for future research. Some of the aspects might have been repeating from previous sections, but this section should consolidate and discuss them all together.

Using language-specific embedding models (in this case, Lithuanian) instead of relying on general purpose embedding models would allow analysis of texts in their original language without translation to English. This approach could potentially improve clustering quality since the texts would maintain their original contextual meaning. The usage of such models would also enable an interesting comparison between the results of this research, where no specific Lithuanian language model was used, and results that could be obtained with language specific models.

This research was specifically based on two approaches used by both examined models: dimensionality reduction and clustering methods. However, dimensionality reduction of embeddings before clustering might lead to loss of important contextual information. Perhaps using an approach which does not require reducing the embeddings dimension, such as community detection algorithms, could result in better clustering quality. Also, the clustering algorithms used in this research were advanced. The exploration of simpler clustering methods could be an interesting option for future research.

End-to-end LLM integration into topic modeling could lead to even more simplification. As LLMs are constantly advancing, increasing their context windows and allowing for more context in the input, the topic modeling pipeline could be completely transformed. LLMs might be provided with a full cluster's documents in order to extract topics, which would render the keyword extraction step in topic modeling unnecessary. It would be interesting to explore how LLMs could handle the full dataset of texts, potentially removing the dimensionality reduction and clustering steps from the topic modeling pipeline entirely.

3.9 Results

This section presents the key results of the study. The results summarize both methodological advances and practical discoveries from applying dynamic topic modeling techniques to selected textual dataset.

1. Extensive literature review was conducted on various topic modeling techniques. It revealed the shift from traditional probabilistic topic modeling approaches like LDA to algorithmic methods which uses embeddings, dimensionality reduction and clustering techniques. This literature review created strong theoretical foundation for the combination of dimensionality reduction, clustering algorithms, embeddings and large language models to extract meaningful narratives from textual dataset.

2. The 2024 Lithuanian parliamentary election dataset was selected and analyzed as a corpus to investigate evolving narratives. It consisted of 7385 documents from diverse media sources. The conducted analysis showed media coverage change overtime and the topical analysis revealed how narratives changed over time in the Lithuanian political landscape.

3. New contextualized topic coherence metric was introduced, called Contextualized Topic Coherence – Topic Alignment (CTC_{TA}). This metric is designed to asses the alignment between extracted topic keywords and narratives generated using LLM. The aim of CTC_{TA} is to measure how well algorithmic keyword extraction captures the thematic structure that LLMs identify in document clusters.

4. This study compared traditional automated topic coherence metrics (e.g., NPMI, UMass) with novel contextualized topic coherence metrics ($CTC_{Intrusion}$, CTC_{Rating}, CTC_{TA}) that leverage LLMs. The contextualized metrics provide a more nuanced evaluation of topic quality, bridging the gap between algorithmic measures and human judgment.

3.10 Conclusions

This study has proposed and validated a process based on topic modeling to identify time-varying topics in a textual dataset, utilizing machine learning techniques and Large Language Models (LLMs). The following conclusions can be made:

1. This research demonstrates that dynamic topic modeling can capture and track the evolution of narratives over time. The literature review and methodology shows that combining Large Language Models with algorithmic topic modeling techniques transforms simple topic keyword extraction into dynamic narrative discovery, providing tools that help not only to understand what topics exist, but how stories evolve and merge over time.

2. The comparative evaluation of BERTopic and ANTM showed their differences in strengths when performing topic modeling tasks. BERTopic was better in maintaining temporal coherence of narratives across time steps, while ANTM produced more diverse topic representations with less intruder words. This analysis provided a systematic process for selecting topic models based on specific research goals - when better narrative consistency is needed, BERTopic should be selected, and when more diverse topics is the goal - ANTM is preferred.

3. The proposed topic identification and tracking process (see Table 3) successfully captured the evolution of narratives throughout the election period in Lithuanian. For example, it was shown how discussions regarding R. Žemaitaitis changed from initial coalition negotiations to more specific criticism regarding anti-Semitism.

4. Automated topic coherence metrics as $NPMI$ or $UMass$ are not sufficient for evaluating topic models enhanced with Large Language Models. The research shows that contextualized topic coherence metrics using LLMs ($CTC_{Intrusion}$, CTC_{Rating}, CTC_{TA}) provide deeper topic quality assessment that better aligns with human judgment.

5. Comparative evaluation showed that BERTopic outperformed ANTM on coherence and interpretability metrics: $NPMI$ (0.0075 vs. -0.3455), $UMass$ (-11.0928 vs. -16.1049), and CTC_{Rating} (3.8252 vs. 3.7212). Meanwhile, ANTM scored better on topic diversity and semantic clarity: Topic Diversity (0.5014 vs. 0.0050), $CTC_{Intrusion}$ (0.9976 vs. 0.9966), and CTC_{TA} (3.7636 vs. 3.1727). These differences suggest choosing BERTopic for coherent topic evolution and ANTM for broader thematic variety.

6. Finally, the integration of LLMs improved both narrative extraction and evaluation, addressing limitations of keyword based topic modeling techniques. Using the contextual knowledge of LLMs, the proposed topic extraction process generates comprehensive narrative names and descriptions that encapsulate the complex political context. Without the integration of LLMs and relying solely on the keyword based approach, it would be more difficult to understand the underlying political background of the dataset. Overall, this approach showed how NLP methods can be improved with generative language models to improve analysis of information space.

Despite limitations related to translation, sentence level segmentation, data filtering and evaluation subjectivity, this research provides potential directions for future work in dynamic topic modeling. Improvements to the examined approach include using language-specific embedding model, adjusting or completely skipping dimensionality reduction and clustering steps and the implementation of fully end-to-end LLM based topic modeling pipeline.

As the amount of available information continues to grow and expand, approaches that can effectively extract, track, and analyze narratives over time are becoming increasingly valuable. This work contributes to addressing the challenge of information overload by exploring methods to identify coherent storylines within large textual datasets, improving our ability to navigate and understand complex information space.

References

[1] A. Akbik, D. Blythe, R. Vollgraf. “Contextual String Embeddings for Sequence Labeling.” In: *Proceedings of the 27th International Conference on Computational Linguistics*. Edited by E. M. Bender, L. Derczynski, P. Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pages 1638–1649. URL: https://aclanthology.org/C18-1139.

[2] R. Albalawi, T. H. Yeap, M. Benyoucef. “Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis.” In: *Frontiers in Artificial Intelligence* 3 (2020). ISSN: 2624-8212. https://doi.org/10.3389/frai.2020.00042.

[3] N. Aletras, M. Stevenson. “Evaluating Topic Coherence Using Distributional Semantics.” In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Edited by A. Koller, K. Erk. Potsdam, Germany: Association for Computational Linguistics, 2013, pages 13–22. URL: https://aclanthology.org/W13-0102.

[4] M. S. Asyaky, R. Mandala. “Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP.” In: *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. 2021, pages 1–6. https://doi.org/10.1109/ICAICTA53211.2021.9640285.

[5] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, E. Artemova, E. Tutubalina, G. Chowell. “A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration.” In: *Epidemiologia* 2.3 (2021), pages 315–324. ISSN: 2673-3986. https://doi.org/10.3390/epidemiologia2030024.

[6] D. Blei, L. Carin, D. Dunson. “Probabilistic Topic Models.” In: *IEEE Signal Processing Magazine* 27.6 (2010), pages 55–65. https://doi.org/10.1109/MSP.2010.938079.

[7] D. Blei, A. Ng, M. Jordan. “Latent Dirichlet Allocation.” In: *Advances in Neural Information Processing Systems*. Edited by T. Dietterich, S. Becker, Z. Ghahramani. Volume 14. MIT Press, 2001. URL: https://proceedings.neurips.cc/paper_files/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf.

[8] D. M. Blei, J. D. Lafferty. “Dynamic topic models.” In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML ’06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pages 113–120. ISBN: 1595933832. https://doi.org/10.1145/1143844.1143859.

[9] R. Bohn, J. Short. “Info Capacity| Measuring Consumer Information.” In: *International Journal of Communication* 6.0 (2012). ISSN: 1932-8036. URL: https://ijoc.org/index.php/ijoc/article/view/1566.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, et al. “Language Models are Few-Shot Learners.” In: *Advances in Neural Information Processing Systems*. Edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin. Volume 33. Curran Associates, Inc., 2020, pages 1877–

1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[11] J. Carbonell, J. 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. Melbourne, Australia: Association for Computing Machinery, 1998, pages 335–336. ISBN: 1581130155. https://doi.org/10.1145/290941.291025.

[12] E. Chen, E. Ferrara. “Tweets in Time of Conflict: A Public Dataset Tracking the Twitter Discourse on the War between Ukraine and Russia.” In: *Proceedings of the International AAAI Conference on Web and Social Media* 17.1 (2023), pages 1006–1013. https://doi.org/10.1609/icwsm.v17i1.22208.

[13] K. Chen, Z. He, K. Burghardt, J. Zhang, K. Lerman. “IsamasRed: A Public Dataset Tracking Reddit Discussions on Israel-Hamas Conflict.” In: *Proceedings of the International AAAI Conference on Web and Social Media* 18.1 (2024), pages 1900–1912. https://doi.org/10.1609/icwsm.v18i1.31434.

[14] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. 2024. URL: https://arxiv.org/abs/2403.04132.

[15] C. B. Clement, M. Bierbaum, K. P. O’Keeffe, A. A. Alemi. *On the Use of ArXiv as a Dataset*. 2019. URL: https://arxiv.org/abs/1905.00075.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Edited by J. Burstein, C. Doran, T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423.

[17] A. B. Dieng, F. J. R. Ruiz, D. M. Blei. “Topic Modeling in Embedding Spaces.” In: *Transactions of the Association for Computational Linguistics* 8 (2020), pages 439–453. ISSN: 2307-387X. https://doi.org/10.1162/tacl_a_00325.

[18] M. Ester, H.-P. Kriegel, J. Sander, X. Xu. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD’96. Portland, Oregon: AAAI Press, 1996, pages 226–231.

[19] B. Forrester, S. Ghajar-Khosravi, S. Waldman. “Machine Learning-Enabled Narrative Search in the Information Environment.” In: *15th NATO OR&A Conference*. 2021, pages 18–20. URL: https://cradpdf.drdc-rddc.gc.ca/PDFS/unc422/p816453_A1b.pdf.

[20] M. Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. 2022. URL: https://arxiv.org/abs/2203.05794.

[21] M. Günther, J. Ong, I. Mohr, A. Abdessalem, et al. *Jina Embeddings 2: 8192-Token General-Purpose Text Embeddings for Long Documents*. 2024. URL: https://arxiv.org/abs/2310.19923.

[22] H. W. A. Hanley, Z. Durumeric. "Partial Mobilization: Tracking Multilingual Information Flows amongst Russian Media Outlets and Telegram." In: *Proceedings of the International AAAI Conference on Web and Social Media* 18.1 (2024), pages 528–541. https://doi.org/10.1609/icwsm.v18i1.31332.

[23] H. W. A. Hanley, D. Kumar, Z. Durumeric. "Happenstance: Utilizing Semantic Search to Track Russian State Media Narratives about the Russo-Ukrainian War on Reddit." In: *Proceedings of the International AAAI Conference on Web and Social Media* 17.1 (2023), pages 327–338. https://doi.org/10.1609/icwsm.v17i1.22149.

[24] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt. "Measuring Massive Multitask Language Understanding." In: *arXiv e-prints*, arXiv:2009.03300 (2020), arXiv:2009.03300. https://doi.org/10.48550/arXiv.2009.03300.

[25] F. Heppell, K. Bontcheva, C. Scarton. ""Analysing State-Backed Propaganda Websites: a New Dataset and Linguistic Study"." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Edited by H. Bouamor, J. Pino, K. Bali. Singapore: Association for Computational Linguistics, 2023, pages 5729–5741. https://doi.org/10.18653/v1/2023.emnlp-main.349.

[26] F. Heppell, K. Bontcheva, C. Scarton. "Analysing State-Backed Propaganda Websites: a New Dataset and Linguistic Study." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Edited by H. Bouamor, J. Pino, K. Bali. Singapore: Association for Computational Linguistics, 2023, pages 5729–5741. https://doi.org/10.18653/v1/2023.emnlp-main.349.

[27] A. Hoyle, P. Goel, A. Hian-Cheong, D. Peskov, J. Boyd-Graber, P. Resnik. "Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence." In: *Advances in Neural Information Processing Systems*. Edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan. Volume 34. Curran Associates, Inc., 2021, pages 2018–2033. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/0f83556a305d789b1d71815e8ea4f4b0-Paper.pdf.

[28] H. Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), page 417. https://doi.org/10.1037/h0071325.

[29] M. T. Islam, J. W. Fleischer. *Manifold-aligned Neighbor Embedding*. 2022. URL: https://arxiv.org/abs/2205.11257.

[30] T. Joachims. "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization." In: *Proceedings of the Fourteenth International Conference on Machine Learning*. ICML '97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pages 143–151. ISBN: 1558604863.

[31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig. “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.” In: *ACM Comput. Surv.* 55.9 (2023). ISSN: 0360-0300. https://doi.org/10.1145/3560815.

[32] L. van der Maaten, G. Hinton. “Visualizing Data using t-SNE.” In: *Journal of Machine Learning Research* 9.86 (2008), pages 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[33] N. Maslej, L. Fattorini, R. Perrault, V. Parli, et al. *The AI Index 2024 Annual Report*. Technical report. Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2024. URL: https://aiindex.stanford.edu/report/.

[34] L. McInnes, J. Healy. “Accelerated Hierarchical Density Based Clustering.” In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2017, pages 33–42. https://doi.org/10.1109/ICDMW.2017.12.

[35] L. McInnes, J. Healy, J. Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” In: *arXiv e-prints*, arXiv:1802.03426 (2018), arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426.

[36] T. Mikolov, K. Chen, G. Corrado, J. Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013. URL: https://arxiv.org/abs/1301.3781.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean. “Distributed Representations of Words and Phrases and their Compositionality.” In: *Advances in Neural Information Processing Systems*. Edited by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger. Volume 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[38] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum. “Optimizing Semantic Coherence in Topic Models.” In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edited by R. Barzilay, M. Johnson. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011, pages 262–272. URL: https://aclanthology.org/D11-1024.

[39] N. Muennighoff, N. Tazi, L. Magne, N. Reimers. “MTEB: Massive Text Embedding Benchmark.” In: *arXiv preprint arXiv:2210.07316* (2022). https://doi.org/10.48550/ARXIV.2210.07316.

[40] A. Narayan, B. Berger, H. Cho. “Assessing Single-Cell Transcriptomic Variability through Density-Preserving Data Visualization.” In: *Nature Biotechnology* (2021). https://doi.org/10.1038/s41587-020-00801-7.

[41] G. B. Newby. “Metric multidimensional information space.” In: *TREC*. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. 1996.

[42] J. Pennington, R. Socher, C. Manning. “GloVe: Global Vectors for Word Representation.” In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edited by A. Moschitti, B. Pang, W. Daelemans. Doha, Qatar: Association for Computational Linguistics, 2014, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162.

[43] O. Press, N. A. Smith, M. Lewis. *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation*. 2022. URL: https://arxiv.org/abs/2108.12409.

[44] H. Rahimi, D. Mimno, J. Hoover, H. Naacke, C. Constantin, B. Amann. “Contextualized Topic Coherence Metrics.” In: *Findings of the Association for Computational Linguistics: EACL 2024*. Edited by Y. Graham, M. Purver. St. Julian’s, Malta: Association for Computational Linguistics, 2024, pages 1760–1773. URL: https://aclanthology.org/2024.findings-eacl.123/.

[45] H. Rahimi, H. Naacke, C. Constantin, B. Amann. *ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics*. 2023. URL: https://arxiv.org/abs/2302.01501.

[46] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang. “SQuAD: 100,000+ Questions for Machine Comprehension of Text.” In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Edited by J. Su, K. Duh, X. Carreras. Austin, Texas: Association for Computational Linguistics, 2016, pages 2383–2392. https://doi.org/10.18653/v1/D16-1264.

[47] K. Ravi, A. Ernesto Vela, R. Ewetz. “Classifying the Ideological Orientation of User-Submitted Texts in Social Media.” In: *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2022, pages 413–418. https://doi.org/10.1109/ICMLA55696.2022.00066.

[48] M. Röder, A. Both, 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. Shanghai, China: Association for Computing Machinery, 2015, pages 399–408. ISBN: 9781450333177. https://doi.org/10.1145/2684822.2685324.

[49] M. Steyvers, T. Griffiths. “Probabilistic topic models.” In: *Handbook of latent semantic analysis* 427.7 (2007), pages 424–440.

[50] G. Team. *Gemini: A Family of Highly Capable Multimodal Models*. 2024. URL: https://arxiv.org/abs/2312.11805.

[51] L. Thompson, D. Mimno. *Topic Modeling with Contextualized Word Representation Clusters*. 2020. URL: https://arxiv.org/abs/2010.12626.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. “Attention is All you Need.” In: *Advances in Neural Information Processing Systems*. Edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett. Volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[53] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman. “SuperGLUE: a stickier benchmark for general-purpose language understanding systems.” In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[54] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt. *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. 2023. URL: https://arxiv.org/abs/2302.11382.

[55] X. Wu, T. Nguyen, A. T. Luu. “A Survey on Neural Topic Models: Methods, Applications, and Challenges.” In: (2023). https://doi.org/10.21203/rs.3.rs-3049182/v1.

Appendix A. Full list of generated topics by BERTopic

Table 1. *Full list of topics and narratives extracted by BERTopic in 2024 Lithuanian parliamentary election context.*

Count	Keywords	Narrative
1142	zemaitaitis, dawn, nemunas, nemunas dawn, remigijus zemaitaitis, remigijus, dawn nemunas, leader nemunas, leader nemunas dawn, zemaitaitis leader	The political landscape in Lithuania surrounding the ruling coalition including the ”Nemunas Dawn” party, led by Remigijus Žemaitaitis, amidst controversies regarding his leadership, candidacy nominations for key ministries, and public protests against alleged anti-Semitic rhetoric. Key figures include Prime Minister-designate Gintautas Paluckas and Agnė Širinskienė, with implications for coalition dynamics and international criticism of the partnership due to legal and ethical concerns.
651	karbauskis, peasants, skvernelis, lvzs, saulius skvernelis, saulius, union, greens, peasants greens, ramunas	This topic discusses the political dynamics and coalition negotiations surrounding the Lithuanian Peasants and Greens Union (LVŽS), led by Ramūnas Karbauskis, and the Democratic Union ”For the Sake of Lithuania,” headed by Saulius Skvernelis, in the context of the recent Seimas elections. It highlights the electoral performance of these parties, their leadership positions, and the complexity of forming a governing majority, addressing issues of political responsibility, coalition stability, and the implications of past political cultures on future partnerships.
494	gintautas, paluckas, gintautas paluckas, prime, prime minister, minister gintautas, prime minister gintautas, gintautas paluck, mr paluckas	The candidacy and appointment of Gintautas Paluckas as Prime Minister of Lithuania, including coalition negotiations and the formation of a new government cabinet, alongside discussions on policy proposals, budget considerations, and ministerial appointments, amidst the backdrop of the social democratic political landscape and recent electoral outcomes.

Count	Keywords	Narrative
372	blinkevicuite, vilija, vilija blinkeviciute, chairwoman„ prime minister, leader social, leader social democrats, prime, blinkevicuite chairwoman	The topic centers around Vilija Blinkevičiūtė, the chairwoman of the Lithuanian Social Democratic Party (LSDP), and her decision not to pursue the position of Prime Minister following the party's electoral success in the Seimas elections. It highlights the political dynamics of coalition building with other parties, including the Democratic Union "For the Sake of Lithuania," and reflects on the implications of her leadership decisions, party strategies, public sentiment, and challenges within the political landscape of Lithuania. The discussions also touch upon themes of trust and reliability in political commitments, as well as reactions from various political figures and parties regarding the coalition and government formation.
277	kasciunas, laurynas kasciunas, laurynas, defense, minister national, national defense, minister national defense, defense laurynas, defense laurynas kasciunas, national defense laurynas	The role and statements of Laurynas Kasčiūnas, Minister of National Defense of Lithuania, regarding military developments, defense cooperation with European countries, and support for Ukraine's defense industry during the context of emerging defense challenges in Lithuania and ongoing regional tensions. The discussions include air defense systems, procurement of military ammunition, and collaborative agreements with Germany and Ukraine, reflecting broader national security concerns and the political landscape related to the upcoming Seimas elections.
256	cmilytenielsen, viktorija, viktorija cmilytenielsen, seimas viktorija, speaker seimas viktorija, seimas viktorija cmilytenielsen, speaker, speaker seimas, liberal, liberal movement	Discussions surrounding the leadership and actions of Viktorija Čmilytė-Nielsen, Chairwoman of the Liberal Movement and Speaker of the Seimas, in relation to political accountability, coalition dynamics, and parliamentary elections in Lithuania, including responses to other political parties and significant figures like Eugenijus Gentvilas.

Count	Keywords	Narrative
226	simonyte, ingrida simonyte, ingrida, minister ingrida simonyte, minister ingrida, prime minister ingrida, prime minister, prime, outgoing, minister	The evolving political landscape in Lithuania, focusing on outgoing Prime Minister Ingrida Šimonytė and her potential role in future elections, government continuity, political challenges, and the implications of ongoing scandals involving conservative party members. The discussion includes sentiments around leadership, the government's performance, and the dynamics within the conservative ranks, especially concerning the upcoming parliamentary elections.
217	lsdp, party lsdp, democratic party lsdp, lithuanian social democratic, lithuanian social, democratic, social democratic party, democratic party, social democratic, union sake lithuania	Formation of a governing coalition in Lithuania involving the Lithuanian Social Democratic Party (LSDP), the Democratic Union "For the Sake of Lithuania", and the "Nemunas Dawn" party, amidst controversies regarding leadership actions and public protests. Key figures include Gintautas Paluckas and Remigijus Žemaitaitis, with implications for left-wing values and political stability in the Seimas following the recent parliamentary elections.
211	ir, lietuvos, kad, nemuno, vardan, vardan lietuvos, nemuno aura, bus, laisvs, konservatoriai	The dataset revolves around the upcoming 2024 Lithuanian parliamentary elections, highlighting significant political parties such as the Democratic Union "For the Sake of Lithuania", the Freedom Party, and the Lithuanian Social Democratic Party (LSDP). Key figures include Remigijus Žemaitaitis, leader of the "Nemunas Aušra" party, who faces controversy due to allegations of anti-Semitic remarks. The data reveals ongoing coalition discussions, electoral strategies, and party performances, reflecting a diverse political landscape and the challenges faced by different political factions, including the conservatives and social democrats. The sentiment indicates tensions regarding coalition formations and the responses to voter concerns.

Count	Keywords	Narrative
180	savickas, economy, finance, sabutis, ministry, economy innovation, linge, lukas savickas, innovation, social security	The topic focuses on the recent developments in Lithuania's parliamentary elections, highlighting the appointments and proposals of various ministers, including Lukas Savickas as Minister of Economy and Innovation and Eugenijus Sabutis as Minister of Transport. It explores the implications of budgetary decisions, tax reforms, and the government's response to economic challenges, reflecting the political dynamics among parties and their leaders, such as Šarūnas Birutis, Mindaugas Lingė, and Inga Ruginienė. The discussions emphasize the importance of innovation, social security, and financial policy in shaping the country's economic future, while revealing underlying tensions regarding funding and governance.
154	photo, lrt, bns, bns photo, peleckis, peleckis bns, stacevicius lrt, stacevicius, paulius peleckis bns, paulius peleckis	Lithuanian Parliamentary Elections 2023 - Coverage of candidates Gintautas Paluckas, Saulius Skvernelis, Remigijus Žemaitaitis, Vilija Blinkevičiūtė, and others, highlighting key events, coalition agreements, election outcomes, and controversies. Photographic documentation by Paulius Peleckis, Lukas April, J. Stacevicius, and V. Raupelis emphasizes the electoral landscape and political dynamics within Lithuania.
153	freedom party, freedom, armonaite, raskevicius, ausrine armonaite, ausrine, world, freemasons, zirmunai, world lithuanian	The recent elections in Lithuania highlight the challenges faced by the Freedom Party, led by Aušrinė Armonaitė, as it struggled to gain voter support and failed to secure a parliamentary mandate. Key figures include Tomas Vytautas Raskevičius and Morgana Danielė, who contested in various single-member constituencies. The elections featured a competitive political landscape, with traditional parties like the Conservatives and Social Democrats performing strongly, while controversies emerged around alleged electoral irregularities linked to the ”Freemasons.” The necessity for the Freedom Party to redefine its strategy and leadership following its electoral defeat is a significant aspect of the ongoing political discourse.

Count	Keywords	Narrative
142	unionlithuanian, homeland unionlithuanian, unionlithuanian christian, homeland, homeland unionlithuanian christian, christian democrats tslkd, democrats tslkd, unionlithuanian christian democrats, christian, christian democrats	Analysis of the recent changes in leadership within the Homeland Union-Lithuanian Christian Democrats (TS-LKD) following the unsuccessful Seimas elections, focusing on Gabrielius Landsbergis' resignation, the reactions from party members, and the implications for future political strategies and coalitions in Lithuania.
99	landsbergis, gabrielius landsbergis, gabrielius, foreign minister, minister gabrielius, minister gabrielius landsbergis, foreign minister gabrielius, foreign, diplomacy, minister foreign affairs	The role of Foreign Minister Gabrielius Landsbergis in promoting Lithuania's foreign affairs and support for Ukraine during the ongoing geopolitical tensions, including his engagements with EU diplomacy, discussions on sanctions against Russia, and efforts to strengthen bilateral ties with the United States and other countries amidst the backdrop of Lithuania's parliamentary elections.
64	saulius skvernelis, saulius, teodoras biliunas bns, biliunas bns, teodoras, vilija blinkevicuite saulius, blinkevicuite saulius, blinkevicuite saulius skvernelis, skvernelis, saulius skvernelis photo	Overview of the Lithuanian Parliament Elections featuring key figures such as Saulius Skvernelis, Vilija Blinkevičiūtė, Eduardas Vaitkus, and Mindaugas Puidokas, highlighting coalition formations, electoral strategies, and public sentiments regarding health reforms and electoral conduct.

Count	Keywords	Narrative
57	zemaitaitis remigijus, remigijus zemaitaitis, remigijus, zemaitaitis remigijus zemaitaitis, zemaitaitis, mikutavicius, lithuanian youth, help, experienced, youth	Discussions surrounding the recent political activities and controversies in Lithuania, focusing on Remigijus Žemaitaitis and his influence in the parliamentary elections. The topic includes sentiments expressed by various youth organizations regarding government coalitions, public protests, property tax debates, and concerns over political integrity, amidst a backdrop of changing public sentiment toward the Social Democrats and discussions on progressive taxation and working conditions.
54	chairman tslkd, tslkd, february, election chairman tslkd, election chairman, held, chairman, leader elected, election, did want	Election and Leadership Transition in the Conservative Party TS-LKD, featuring the resignation of Chairman Gabrielius Landsbergis, the interim leadership of Radvilė Morkūnaitė-Mikulėnienė, and the upcoming election for a new party leader scheduled for February. The situation reflects internal party dynamics and the desire for unity amidst changing leadership.