Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

Language Technology

BACHELORTHESIS

Analyzing Political Discourse on Online Platforms using Topic Modelling

Till N. Schaland

Field of Study: Software-System-Entwicklung
Matriculation No.: 7416419
1st Examiner: Dr. Seid Muhie Yimam, Universität Hamburg
2nd Examiner: Prof. Dr. Chris Biemann, Universität Hamburg

Language Technology
Department of Informatics
Faculty of Mathematics, Informatics and Natural Sciences

Universität Hamburg
Hamburg, Germany

A thesis submitted for the degree of

Bachelor of Science (B. Sc.)

Analyzing Political Discourse on Online Platforms using Topic Modelling

Bachelor's Thesis submitted by: Till N. Schaland

Date of Submission: 28.10.2025

Supervisor(s):
Dr. Seid Muhie Yimam, Universität Hamburg

Committee:
1st Examiner: Dr. Seid Muhie Yimam, Universität Hamburg
2nd Examiner: Prof. Dr. Chris Biemann, Universität Hamburg

Universität Hamburg, Hamburg, Germany
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Language Technology

Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Bachelorstudiengang Software-System-Entwicklung selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe. Sofern im Zuge der Erstellung der vorliegenden Abschlussarbeit generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten Wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

28.10.2025, Oldenburg

Date

T. Schaland

Signature

(Till N. Schaland)

Abstract

In recent years, online platforms have increasingly been used by different individuals across all age groups and social classes. They have evolved into influential spaces for political discourse and public opinion placements, especially during election periods. This vast amount of textual data cannot be analyzed manually to gain insights into public opinions, public discourses, and highly discussed themes. Topic modeling offers a promising approach to analyze what themes / topics are being discussed by the public and how the public sentiment around these topics are. This thesis compares traditional and modern approaches to topic modeling for analyzing political discourse on online platforms with an specific focus of the German election period in 2025.

A dataset of over 380,000 online posts was collected and processed from multiple online sources provided by the *KIFürDemokratie* project. Four topic modeling methods were evaluated: traditional models such as LSA and LDA, and modern approaches using transformers like BERTopic and TopicGPT. Hyperparameter optimization was applied using Bayesian optimization provided by the *OCTIS* framework. The hyperparameter optimization uses a composite objective function combining topic coherence (C_v) and topic diversity (TD) to ensure that the models are optimized to ensure both coherent and diverse topics.

The results show that BERTopic and TopicGPT outperform traditional models in both quantitative and qualitative evaluations. BERTopic achieved the best scores in both quantitative metrics and in the qualitative evaluation. Dynamic topic modeling with BERTopic revealed distinct temporal patterns, including spikes in discourse intensity around key election events and the emergence of new topics reflecting how the public discourse shifts. Due to its ability to capture fast-paced, coherent, and interpretable topics, a more advanced variant of the dynamic topic modeling approach was developed and integrated into the *KIFürDemokratie* project's analytical dashboard. It is trained continuously on new daily online posts to also identify new emerging topics which may arise in the future.

Overall, this study demonstrates that modern topic modeling approaches provide a more powerful alternative to traditional models for analyzing online content.

Contents

1 Introduction

This chapter will first cover the motivation of this thesis and the importance of natural language processing, with a specific focus on topic modeling. Following the motivation, we will discuss the problem that this thesis aims to address and solve, as well as the research question it aims to answer.

1.1 Motivation

Although online platforms such as Facebook, Twitter, and Instagram were initially used for personal interactions among family and friends, in the past decade, they have evolved into influential spaces for political discourse and public opinion placement. In recent years, online platforms have been increasingly used for political discourse, particularly during elections.

In Germany, recent events, such as the dissolution of the governing coalition and recent developments in the German and European economies, have further polarized public opinion. This division is increasingly visible on online platforms, particularly those where politicians and citizens leverage them to publish content and distribute their viewpoints on different topics. The project KIFürDemokratie[1] highlights the transformative role of AI in policymaking. The analysis of online data aims to bridge the gap between policymakers and citizens to strengthen the democratic process. By applying artificial intelligence (AI) to analyze online content, voters' concerns can be made more accessible for policymakers, enabling them to better understand public discourse. This thesis aims to leverage AI techniques to address the increasing volume and complexity of political discourse during elections.

Online platforms are used extensively by users for political discourse and opinion formation. This vast amount of data cannot be analyzed by individuals alone for policy making and analyzing voters' concerns. However, Natural Language Processing (NLP) has advanced significantly due to theoretical advancements such as the transformer architecture and technical advances such as the incredible improvement in the utilization of Graphics Processing Units (GPUs) in deep learning architectures. These developments

1. Für Demokratie e.V.

made it possible to handle large textual datasets and extract richer semantic information from text. By applying NLP, these data can be used to gain insights into the voters' interests and concerns and to understand how the public responds during the election period in Germany. Topic modeling is a method for analyzing different topics, patterns, and hidden structures within large collections of textual data. This makes it useful for studying how the frequency of the topic changes over time and capturing changes in the content of different topics within the election period.

1.2 Problem Statement

Existing research and studies focus primarily on static topic modeling, analyzing how political topics are discussed on social media. Such approaches can discover key topics within a corpus but fail to capture how these topics evolve over time. Most existing work concentrated on analyzing social media platforms such as Twitter or Instagram during election campaigns. Therefore, these studies offer a narrow view of political communication online. However, political discourses happen on various online platforms, including news comment sections, blogs, and online forums. Moreover, these studies only focus on static topic modeling and do not incorporate how topics evolve and how real-world political events and news influence them.

Topic modeling provides a promising approach for discovering topics in large textual data. Traditional topic modeling approaches, such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), are commonly used. However, transformer-based approaches, such as BERTopic and topic modeling approaches leveraging large language models (LLMs), can better capture semantics in textual data and, therefore, offer promising alternatives.

This thesis aims to address this gap by comparing traditional and transformer-based topic models. This thesis assesses the performance of these methods in modeling static topics across online platforms and provides an analysis of how political discourses change and are affected by real-world political events and news.

1.2.1 Research Questions

The problem, mentioned in Section 1.2, opens up several research questions. This thesis compares traditional topic modeling techniques with newer techniques based on small language models (SLM) and LLM to capture topics in political discourse on online platforms.

Since most analyses on social media platforms perform static topic modeling on generated content, this thesis will furthermore aim to answer the question of how topic modeling can be used to analyze temporal content during election periods, especially during the German election period until the early parliamentary elections on 23 February 2025. This opens up the first research question, which can be formulated as follows.

1. How do traditional machine learning approaches compare to transformer-based approaches for topic modeling on politically related data on online platforms?

This question enables a comparative evaluation of traditional methods, such as LDA and LSA, SLM-based frameworks such as BERTopic, and more recent LLM-driven

methods like TopicGPT, assessing their ability to capture and interpret political topics on online platforms.

2. **How do political topics evolve over time across online platforms, and what insights can be drawn about the dynamics of public discourse?**

This research question aims to analyze temporal trends in topic importance and discover trends in how political discourse evolves over time across online platforms.

Answering these questions in this thesis will contribute to developing more effective tools for researchers, policymakers, and journalists to analyze how public discourses are changed and influenced by real-world events and news. Helping them to understand public concerns better and gain insights into the dynamics of discourses on online platforms.

1.3 Structure of this Thesis

Chapter 2 provides an overview of the theoretical knowledge required to apply topic modeling to textual data. First, topic modeling is defined and the problem it tries to solve is addressed, followed by an introduction to early topic models and their conceptual foundation. Subsequently, the foundation of neural networks in NLP, starting with the early developments of neural networks in NLP, is introduced, beginning with the concept of distributed representations, followed by the rise of the transformer architecture and its influence on representations of words and documents. Following this, neural topic models are introduced, forming the state-of-the-art topic modeling approaches. Lastly, general applications of topic modeling are explored, specifically focusing on applying topic models to online content and political discourses. Chapter 3 introduces the approach applied for this thesis. This includes collecting and preprocessing the data, developing a framework for applying topic modeling to the collected data, and optimizing and evaluating these models. Lastly, the results of the topic models will be discussed in Chapter 4, highlighting the advantages and disadvantages of each model in the context of analyzing political discourses on online platforms. In addition, the ability of BERTopic in the temporal analysis will be analyzed, evaluating whether it can track topic dynamics and how they are influenced by real-world politics and events.

2
Background and Related Work

2.1 Background

This chapter covers the theoretical knowledge needed for this thesis. First, Section 2.1.1 and Section 2.1.2 will introduce topic modeling and the development from traditional machine learning models. Secondly, an overview of recent advancements in natural language processing is given, starting with the introduction of distributed representations, the transformer architecture, and BERT to model advanced state-of-the-art approaches such as large language models (Section 2.1.3). Furthermore, in Section 2.1.4, an introduction to novel topic modeling techniques is given, which makes use of the transformer to perform topic modeling. Lastly, Section 2.1.6 introduces dynamic topic modeling, covering the first dynamic topic modeling (DTM) model, dynamic latent Dirichlet allocation, and ending with BERTopic for dynamic topic modeling.

2.1.1 Topic Modeling

Topic Modeling is a technique in Natural Language Processing that aims to reveal hidden structures, also called latent topics, in large collections of documents. It is an unsupervised learning technique that identifies topics, represented as groups of words that co-occur frequently together. Topic Modeling was initially introduced by Deerwester et al. (1990) using Latent Semantic Indexing, also known as Latent Semantic Analysis, which was primarily used for information retrieval. Another method was introduced by Blei et al. (2003), which lays out the fundamental probabilistic framework behind topic modeling. It is widely recognized as a foundational work in topic modeling and will be explained in Section 2.1.2. Latent Dirichlet Allocation (LDA) is a generative, probabilistic model. In contrast to Latent Semantic Analysis, which uses singular value decomposition to reveal latent topics, it is based on Bayesian methods. For this approach, the authors assume that documents are made out of topics, and topics are made out of words. Both approaches work on Document-Term-Matrices, namely, bag-of-words (BOW) or term frequency inverse document frequency (TF-IDF) to represent documents. Both methods ignore the order in which words appear in the document and only

focus on the occurrences of words per document. With the rise of neural networks, topic modeling techniques have evolved over time by using vector representations of documents using representation learning. BERTopic (Grootendorst, 2022) is an approach that maps documents to vector representations, performs dimensionality reduction, and then clusters the reduced representations to discover topics.

2.1.2 Traditional Topic Modeling

As outlined in Section 2.1.1, topic modeling is an unsupervised machine learning technique to discover latent topics from large amounts of unstructured text. While introducing the general concept of topic modeling, this section will delve deeper into these traditional methods that have set the foundation of topic modeling. Specifically, the two most influential methods are latent semantic analysis (LSA) and latent Dirichlet allocation (LDA).

Latent Semantic Analysis (LSA)

LSA was initially introduced by Dumais et al. (1988) under the name latent semantic indexing (LSI), addressing deficiencies of keyword retrieval methods, and was primarily used for information retrieval. However, LSA has been widely used in topic modeling, because it identifies latent structures in text corpora by applying singular value decomposition (SVD) (Deerwester et al., 1990). LSI uses singular-value decomposition on a matrix constructed from the text corpora. The matrix is constructed as a term-document matrix where each row is a term and each column represents a document. The key idea of LSA is that by applying SVD to the term-document matrix, it can identify a latent structure of the data that captures the similarity of words, not only by co-occurrences but also by the association of words across multiple documents. Thus, words that appear in the same context across multiple documents share a latent meaning and therefore are located in the same semantic structure.

As described in Deerwester et al. (1990), singular value decomposition assumes that any rectangular matrix ($t \times d$) of terms and documents, X, can be decomposed into the product of three other matrices:

$$X = T_0 S_0 D_0', \tag{2.1}$$

such that T_0 and D_0 have orthonormal columns and S_0 is diagonal. The resulting matrix T_0 has the dimension of $t \times r$ and represents the terms in the reduced semantic space, whereas the matrix D_0 is of rank $r \times d$ and represents the documents in the reduced semantic space. The diagonal matrix S_0 contains the singular values in descending order and represents the importance of each latent dimension. Thus, when applying LSA to extract, e.g., k topics, the decomposed matrix is truncated based on the importance of the singular values by keeping the k largest singular values of the S_0 matrix and keeping the corresponding column in T_0 and D_0. This truncation of the matrices results in the following approximation of X:

$$\hat{X} = T_{0,k} S_{0,k} D_{0,k}. \tag{2.2}$$

Since LSA relies solely on linear algebra, the approach has various limitations. (Thomas K Landauer and Laham, 1998p. 35) explicitly mentions that SVD suffers from bad

computational efficiency. This constraint makes the application of LSA on large text corpora impractical.

Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model introduced by Blei et al. (2003) that can reveal latent topics in text corpora. LDA assumes that each document is generated by a mixture of topics, and a topic is defined as a distribution over words. Specifically, LDA assumes the following generative process for each document **w** in a corpus ***D***:

1. Choose $N \sim Poisson(\lambda)$.
2. Choose $\theta \sim Dir(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial conditioned on the topic z_n.

This generative model can be visualized using graphical models, as shown in Figure 2.1. The boxes in the graphical model represent 'plates', meaning that these are repetitions of the process within the plate. As described above, the model draws a topic distribution from a Dirichlet distribution for each document w in corpus D. This corresponds to the outer place in the graphical model. The inner plate represents the iterative process of sampling a word for each number of words in a document based on the chosen topic of the current word z_n and the word-topic distribution within β.

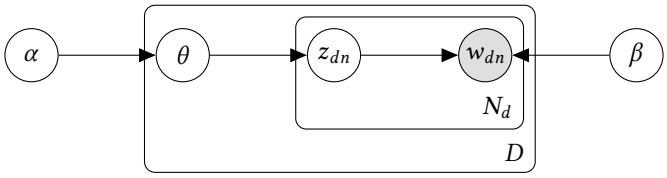

Figure 2.1: Graphical model of Latent Dirichlet Allocation. Adapted from Blei et al. (2003)

Blei et al. (2003) define the word-topic distribution as a $k \times V$ matrix β, where each element β_{ij} represents the probability of word j given topic i. Where k is the dimensionality of the Dirichlet distribution, that is, the number of topics, and V is the vocabulary size. The topic distribution θ for each document is drawn from a Dirichlet distribution and defines the topic mixtures for a given document.

The Dirichlet distribution is a probability distribution and can be used to model the Multinomial distribution. Furthermore, the distribution is supported by a $(K-1)$-Simplex, whose property is that the sum of the probabilities of the k-vector, which lies in the $(k-1)$-simplex, equals 1. The Dirichlet distribution is controlled by α. All Dirichlet distributions can be broadly categorized into three categories: the cases $\alpha < 1$, $\alpha = 1$, and $\alpha > 1$.

For $\alpha < 1$ the probabilities of the distribution are more dense on the edges and corners, leading to a more sparse probability distribution. In the case that $\alpha = 1$, the distribution is similar to a uniform distribution. In the latter case ($\alpha > 1$), the distribution is more concentrated towards the center, which leads to more balanced topic proportions. Because the Dirichlet distribution is a conjugate prior of the Multinomial distribution, we can think of it as a distribution over distributions.

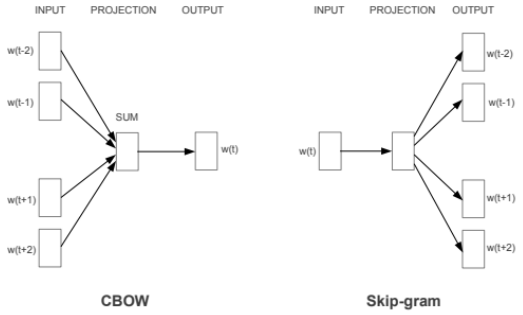

Figure 2.2: Overview of Continuous Bag-of-Words and Skip-gram model architecture. Source: Mikolov et al. (2013).

2.1.3 Neural Networks in Natural Language Processing

In the early stages of NLP development, the field mostly relied on traditional methods such as BoW and TF-IDF representation of documents. These methods do not capture the semantics or syntactic structures of language. In contrast, neural network-based approaches provided a way of including these features in the representations by directly learning these from large corpora of text, leading to representations that are sensitive to context. As the field progresses, more advanced architectures arise using new techniques such as the attention mechanism, which are included in modern architectures such as the transformer architecture.

The following subsections will describe how neural networks are applied in NLP, beginning with the foundational concept of word embeddings, then focusing on the transformer architecture and its impact on document and word representation, and concluding with the most recent developments of the transformer architecture and large language models.

Distributed Representation

In the early years of natural language processing, text was represented as BoW. BoW represents documents as a vector, where each index represents a word and the value represents the number of occurrences in the document. Although these approaches could represent document similarity, they do not provide similarity between single words. The work by Mikolov et al. (2013) describes this as: ”Many current NLP systems and techniques treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary”. In his paper, Mikolov et al. (2013) introduced a method for representing a word as a densely distributed representation. In this work, Mikolov et al. (2013) introduced two architectures and training tasks that allowed him to learn these representations directly from the data: the continuous bag-of-words (CBOW) and skip-gram models (see Figure 2.2).

The CBOWs training objective is to predict a given word using its context words. The second architecture introduced is the continuous skip-gram model, similar to the CBOW model. However, instead of predicting a word given the context, the training objective is to predict the context words given a single word as input. The resulting representations from the learned architecture offered dense vectors compared to BoW representations that only represented these as sparse representations, which usually

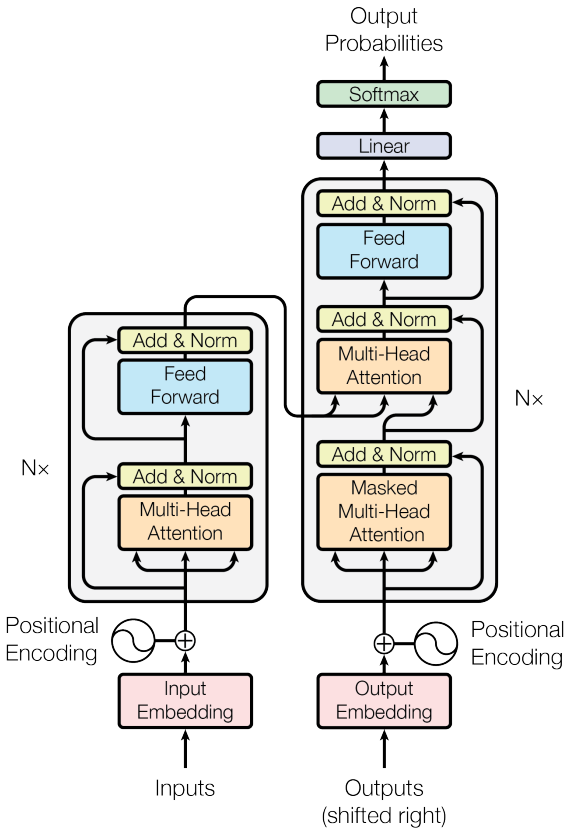

Figure 2.3: The Transformer-model architecture. Source: Vaswani et al. (2017).

results in a high dimensionality (length of vocabulary) and a lack of semantic information encoded into these vectors.

These dense representations captured semantic information, such as similarities or analogies. Although densely distributed representations improved the performance of various NLP applications, these static representations were insensitive to context, meaning that the representation of words did not change within the context in which they appeared. For example, these densely distributed representations do not capture the difference between the fruit ”Apple” and the Company ”Apple” (Mikolov et al., 2013).

This context insensitivity limits the model's ability to capture language nuances, especially in tasks requiring deeper semantic understandings. These limitations of the distributed representations led to the need for contextualized word representations. While research based on Word2Vec, such as GloVe (Pennington et al., 2014), tried to enhance traditional word embeddings by including global co-occurrence patterns, the need for contextualized word embeddings remained.

Transformer Architecture

The Transformer architecture introduced in the paper ”Attention is All You Need” by Vaswani et al. (2017) is specifically designed to address the limitations of context-insensitive distributed representations. Figure 2.3 provides an overview of the Transformer model, which follows the encoder-decoder architecture that is already commonly used in various NLP tasks such as machine translation. Unlike previous approaches, such as Recurrent Neural Networks (RNN) (Rumelhart et al., 1986) and Long-Short-

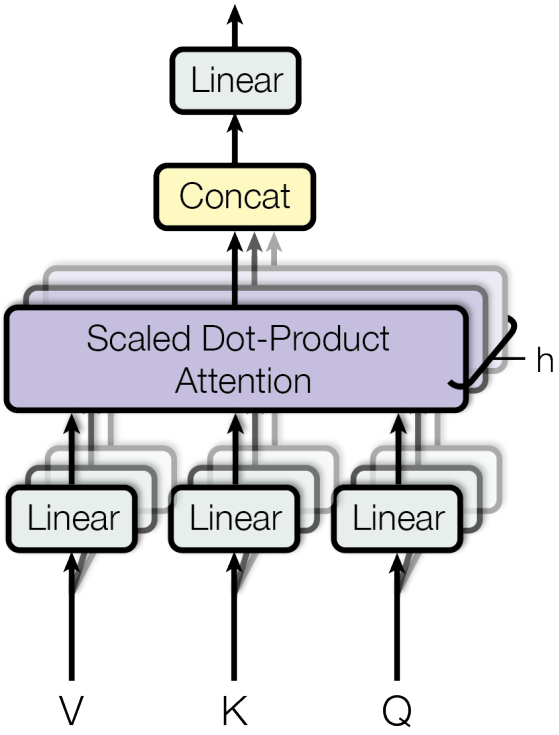

Figure 2.4: Multihead-Attention using Scaled Dot-Product Attention. Source: Vaswani et al. (2017).

Term-Memory (LSTMs) (Hochreiter and Schmidhuber, 1997), which process textual data sequentially, this architecture utilizes parallelism through the ”Attention” mechanism to model semantic relationships between inputs. Although ”Attention” was already used in earlier architectures, in combination with RNNs and LSTMs, these suffered from the computational complexity introduced by sequential modeling of the input data. The transformer architecture solves this issue by modeling the input data only using ”Attention”, eliminating the drawbacks of sequential processing.

The encoder-decoder architecture first encodes a given input $(x_1, ..., x_n)$ into a vector representation, which captures contextual and semantical information about the sentence. The decoder decodes these vector representations into an output sequence $(y_1, ..., y_n)$ (Vaswani et al., 2017).

Attention, the core of the transformer architecture, uses the scaled dot product to find relationships between the input sequences. Figure 2.4 illustrates the Multihead Attention mechanism based on scaled dot-product attention. Specifically, they use self-attention, which models relationships within the input sequence. Scaled dot product attention calculates these relationships by calculating the dot product of the queries and keys. Furthermore, they divide the result by the square root of d_k; this helps to prevent the model from suffering from the so-called ”vanishing gradient problem”(Vaswani et al., 2017). Finally, the scaled dot product is fed into a softmax layer to obtain the weights on the values, which are again multiplied against the corresponding values. Due to simultaneous computation, this can be formulated as a matrix formulation as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.3}$$

The work by Vaswani et al. (2017) suggests that instead of applying a single attention mechanism, the model will benefit from performing multiple attention mechanisms. These attention mechanisms are called heads and are calculated using $Attention$. Furthermore, Vaswani et al. (2017) suggest first projecting each query, key, and value into a subspace of the original embedding by multiplying the inputs using weight matrices $W^Q, W^K \in \mathbb{R}^{d_{model} \times d_k}$, and $W^K \in \mathbb{R}^{d_{model} \times d_v}$ specific to each head, leading to the following equation:

$$head_i = Attention(QW_i^Q, K_i^K, V_i^V) \tag{2.4}$$

The final Multi-Head Attention layer is then defined as:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2.5}$$

Whereas $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ is the final layer that projects the output of the Multi-Head Attention to the final output dimension d_{model}.

Finally, the output of Multi-Head Attention is passed through a so-called ”position-wise” fully connected layer. ”Position-wise” means the feed-forward network is applied to each input token separately and identically. This means that for each input x in the input sequence, the FNN is defined as:

$$FNN(x) = ReLu(xW_1 + b_1)W_2 + b2. \tag{2.6}$$

Bidirectional Encoder Representations from Transformers

Bidirectional Representations from Transformers (BERT) is a model for representing documents using the formerly introduced Transformer architecture. BERT consists of multiple stacked Transformer encoders, as introduced by Vaswani et al. (2017).
However, Devlin et al. (2019) slightly changed the representations from input and output to distinguish sequences in the input using segmentation embeddings. These embeddings aim to make BERT able to perform various tasks, such as named entity recognition and question-answering tasks, usually in the form of a question and an answer. The author clarifies that the input "sentence" can be of arbitrary length and may also consist of multiple sentences or paragraphs (Devlin et al., 2019).

The term *bidirectional* in the model's name is not reflected in the actual architecture but rather in the pretraining objective of the model. Unlike other models that use pretraining, it does not use traditional left-to-right or right-to-left training, but performs two-stage pretraining with different objectives (Devlin et al., 2019).

The first pretraining task, masked language modeling (MLM), randomly masks some input tokens. The objective of these pretraining tasks is to try to predict the masked tokens, given context, both from left and right of the word that is being predicted. However, since in the fine-tuning phase of the model, where it is tailored to a specific downstream task, the ”[MASK]” token is not present (Devlin et al., 2019). The authors solve this mismatch by utilizing a probabilistic approach to the MLM phase. Instead of always replacing the masked word with the ”[MASK]” token, the authors suggest the

following: "If the i-th token is chosen, we replace the i-th token with (1) the [MASK] token 80% of the time (2) a random token 10% of the time (3) the unchanged i-th token 10% of the time." (Devlin et al., 2019).

Devlin et al. (2019) states: "Many important downstream tasks, such as Question Answering (QA) and Natural Language Inference (NLI), are based on understanding the relationship between two sentences, which is not directly captured by language modeling" (Devlin et al., 2019). This led to the second pretraining task – next sentence prediction (NSP), which is basically just a binary classification tasked with two sentences. The data for this task consists of two input sentences, in which half of the time the actual next sentence is used, and the other half of the time a random sentence from the corpus is chosen. The objective of the model is then to predict whether sentence B is actually the next sentence of sentence A.

Although this model architecture is primarily used for classification tasks, question-answering, and language understanding tasks such as named entity recognition, some experiments showed that the network was able to produce document/sentence level representations by using the by using the [CLS] tokens output or by averaging the word vectors. However, these representations did not capture textual similarity well and were limited in their ability to create meaningful representations for whole documents (Reimers and Gurevych, 2019; Devlin et al., 2019).

Siamese BERT-Networks

Sentence-BERT is a modification of the BERT network described in Section 2.1.3, which uses Siamese and triplet network structures to create meaningful sentence embeddings (Reimers and Gurevych, 2019). Basically, Siamese networks use two identical subnetworks with shared weights, meaning that both inputs are processed by the same model. The resulting representations are then compared or combined for further tasks, such as classification or similarity estimation.

Reimers and Gurevych (2019) suggested different fine-tuning objectives for the BERT networks:

1. **Classification Objective Function**

$$o = softmax(W_t(u, v, |u - v|)) \tag{2.7}$$

2. **Regression Objective Function**

$$cosine\_sim(u, v) = \frac{U * V}{||U||||V||} \tag{2.8}$$

3. **Triplet Objective Function**

$$max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0) \tag{2.9}$$

In triplet networks, the data to train the model consists of three inputs, namely: an anchor a, a positive sentence p, and a negative sentence n. The task is to minimize the distances between a and p, while maximizing the distance between a and n. This is done by minimizing the triplet objective function $max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0)$.

Large Language Models

Large Language Models (LLMs) are a family of generative neural networks that describe language models that are scaled to billions of parameters. However, with the emergence of the large language model, a shift in the training processes of these models occurs. They are not only pretrained using traditional language modeling tasks, but also fine-tuned using instruction tuning and reinforcement learning by human feedback (RLHF). In this section, the focus is set on the Generative Pretrained Transformer (GPT).

GPT is a decoder-only transformer that mostly follows the original transformer architecture with slight variations. One of these is that, instead of using two masked multi-head attention layers, they employ a single masked multi-head attention layer following a single feedforward neural network. Radford and Narasimhan (2018) used unsupervised pretraining to model the standard language to maximize the likelihood of the corpus $U = \{u_1, ..., u_n\}$ given its context tokens with the parameters of the networks Θ. Mathematically, this is formulated as follows:

$$L_i(U) = \sum \log P(u_i | u_{i-k}, ..., u_{i-1}; \Theta) \tag{2.10}$$

After pretraining, the model was fine-tuned using supervised learning for specific downstream tasks. In earlier models, such as GPT, this was applied to tasks including summarization, question-answering, classification, and similarity. In subsequent years, Radford et al. (2019) demonstrated that by simply scaling the model's architecture to larger parameters, GPT-2 was able to achieve stronger performance without task-specific fine-tuning, showing the ability of zero-shot and one-shot tasks. Zero-shot tasks can be modeled as $p(output|input, task)$. Later Brown et al. (2020) introduced GPT-3, extending these findings by scaling the models even further by increasing the model and training dataset size. GPT-3 was able to show strong performance using zero-shot, one-shot, and few-shot tasks on various benchmarks, which they introduce as "in-context learning"(Brown et al., 2020). In-context learning means that a language model is capable of performing unseen tasks, without fine-tuning, by providing instructions without updating the model's parameters.

2.1.4 Transformer-Based Topic Modeling

Although traditional topic models, such as LDA or LSA, were widely adopted for analyzing larger corpora of documents and are still commonly used in various applications, their ability is limited. Due to the representation of documents using one-hot encoding, TF-IDF, these models lack the ability to understand the semantics of the documents and words, thus limiting the ability to create meaningful topics. With the subsequent advancement of neural networks in natural language processing, a new research direction is emerging to improve topic modeling using neural networks. Based on these research directions, several neural topic models have been introduced, including BERTopic, a framework for neural topic modeling.

Semantic Topic Modeling using BERTopic

BERTopic, introduced by Grootendorst (2022), is a topic modeling framework that uses distributed representations and clustering to determine latent topics in large corpora of text. While traditional approaches use statistical/probabilistic methods that rely solely

on co-occurrences of words, BERTopic uses a fundamentally different approach for finding latent topics. The fundamental concept of BERTopic aligns with Grootendorst (2022) assumption that "documents containing the same topic are semantically similar" (Grootendorst, 2022, p. 2). BERTopic formulates topic modeling as a clustering task and extends the approach by using a modified version of TF-IDF – class-based TF-IDF (C-TF-IDF) – to extract topic representations (Grootendorst, 2022). BERTopic performs topic modeling through a multi-step framework that can be broadly categorized into three steps: document representation, clustering, and topic representation.

Distributed Representation The first step in BERTopic is to convert each document into a numerical vector representation. This is done using pretrained transformer models, such as Sentence-Transformers, which generate dense embeddings that capture the semantic meaning of a text. Unlike traditional approaches based on bag-of-words or TF-IDF, these embeddings consider the context of words within sentences, allowing semantically similar documents to be represented by vectors that are close to each other in the embedding space. This property is crucial for BERTopic, as it assumes that documents belonging to the same topic will also be close together in this semantic space.

Dimensionality Reduction using UMAP The second step in BERTopic involves clustering the document embeddings. However, high-dimensional data often suffer from the so-called curse of dimensionality, which can make it difficult to identify meaningful clusters(Grootendorst, 2022). To address this issue, BERTopic applies a dimensionality reduction method before clustering. Specifically, it uses the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes et al., 2018), which can reduce high-dimensional data while preserving important local and global structures.

UMAP is a nonlinear dimensionality reduction algorithm based on manifold learning. The main idea is that high-dimensional data often lie on a manifold, and UMAP tries to find a low-dimensional representation that keeps the same overall structure of the high-dimensional manifold (McInnes et al., 2018). The algorithm can be divided into two main stages: first, it constructs a weighted graph of the data in the high-dimensional space, and second, it learns a low-dimensional representation that best preserves the high-dimensional structure or manifold.

High-dimensional graph construction. UMAP starts by building a k-nearest-neighbor graph from the dataset $X = \{x_1, \dots, x_N\}$ using a distance metric d. For each data point x_i, it finds the k nearest neighbors and defines the connection strength between x_i and x_j as (McInnes et al., 2018):

$$w(x_i, x_j) = \exp\left(-\frac{\max(0,\ d(x_i, x_j) - \rho_i)}{\sigma_i}\right), \tag{2.11}$$

where ρ_i ensures that every point is connected to at least one neighbor, and σ_i adjusts how distances are scaled around x_i. This creates a directed, weighted graph where closer points have higher connection weights.

Graph symmetrization. Since the graph is directed, UMAP combines edges from both directions to create a symmetric graph using the so-called *fuzzy union* (McInnes et al., 2018):

$$w_{ij} = w(x_i, x_j) + w(x_j, x_i) - w(x_i, x_j)\, w(x_j, x_i). \tag{2.12}$$

This results in a single undirected graph that represents the overall relationships between data points in the high-dimensional space.

Low-dimensional embedding. In the second stage, UMAP learns a low-dimensional embedding $Y = \{y_1, \dots, y_N\}$ that preserves the structure of the high-dimensional graph. To do this, it constructs a similar fuzzy graph W' in the low-dimensional space and minimizes the difference between both graphs using the following cross-entropy loss (McInnes et al., 2018):

$$C = \sum_{i<j} \left[w_{ij} \log \frac{w_{ij}}{w'_{ij}} + (1 - w_{ij}) \log \frac{1 - w_{ij}}{1 - w'_{ij}} \right]. \tag{2.13}$$

This loss is optimized using stochastic gradient descent until the distances between the points in the low-dimensional space reflect those in the original data as closely as possible.

Hierarchical Density-Based Spatial Clustering of Applications with Noise After reducing the dimensionality of the document embeddings, BERTopic uses Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2013) to group semantically similar documents. HDBSCAN is an extension of the well-known DBSCAN (Ester et al., 1996) algorithm that can find clusters of varying densities and automatically detect noise points. Unlike DBSCAN, which produces a single flat clustering, HDBSCAN builds a hierarchy of clusters and then extracts the most stable ones.

The key idea behind HDBSCAN is to measure how densely data points are packed together. It does this by defining a core distance for each point x_p, which is the distance to its m_{pts}-th nearest neighbor. Based on this, it calculates the mutual reachability distance between two points x_p and x_q as

$$d_{mreach}(x_p, x_q) = \max\{d_{core}(x_p), d_{core}(x_q), d(x_p, x_q)\}, \tag{2.14}$$

where d is a chosen distance metric (Campello et al., 2013). This definition smooths out variations in local density and makes clustering more robust.

Using these distances, HDBSCAN builds a weighted graph where data points are vertices and edge weights correspond to mutual reachability distances. From this graph, a minimum spanning tree is constructed to represent how clusters merge as the distance threshold increases. By analyzing this tree, HDBSCAN identifies clusters that remain stable over a wide range of distance thresholds-these are considered the most meaningful clusters.

The stability of a cluster C_i is defined as the sum of the persistence of its points over different density levels:

$$S(C_i) = \sum_{x_j \in C_i} \left(\lambda_{max}(x_j, C_i) - \lambda_{min}(C_i) \right), \tag{2.15}$$

where $\lambda_{min}(C_i)$ and $\lambda_{max}(x_j, C_i)$ represent the minimum and maximum density levels at which a cluster and its members exist. Intuitively, clusters that remain stable across a large range of densities have higher $S(C_i)$ values and are therefore preferred.

This hierarchical approach allows HDBSCAN to automatically determine the number of clusters and handle datasets with varying densities, which makes it well-suited for text embeddings. In the context of BERTopic, this ensures that semantically similar documents form coherent clusters while outliers or irrelevant texts are filtered out.

Class-based TF-IDF Lastly, BERTopic extracts topic representations from the formed clusters, using C-TF-IDF to calculate a topic-word distribution matrix $W_{t,c}$ as follows:

$$W_{t,c} = tf_{t,c} * \log(1 + \frac{A}{tf_t}) \tag{2.16}$$

C-TF-IDF models the importance of words within a topic, in contrast to all other topics (Grootendorst, 2022). The importance of words for a given topic is calculated by treating the documents belonging to a specific topic as a single document. The C-TF-IDF then calculates the frequency of a term t in a class c and multiplies it by the class's inverse frequency. The frequency of the inversion class is calculated taking the logarithm of the fraction $1 + \frac{A}{tf_t}$, where A denotes the average number of words per class and tf_t is the frequency of a term t in all documents, instead of focusing only on a specific class. The final topic representations can then be easily extracted from the topic-word distribution matrix $W_{t,c}$ by using the top k words with the highest weight for each class.

2.1.5 Prompting Large Language Models for Topic Modeling

With the latest advancements of large language models, new approaches for topic modeling are being discovered to identify latent topics in documents by prompting large language models. Since large language models are trained on vast amounts of text corpora, they have proven to show impressive results on various tasks without further fine-tuning. During the pretraining phase of LLMs, they learn the semantics of words and whole documents, making them especially useful for unique domains.

Parallel and Sequential Prompting Approaches

Doi et al. (2024) introduced a study on LLM-based topic modeling, focusing specifically on short texts. The authors distinguish two primary approaches for performing topic modeling with LLMs: (1) parallel prompting and (2) sequential prompting (Doi et al., 2024).

Parallel Prompting. In the parallel prompting approach, each document is processed independently using specific instructions to extract latent topics. Although Doi et al. (2024) did not evaluate the computational time required for a full corpus, this method is generally expected to be faster, as it does not depend on the outputs of previous prompts.

Sequential Prompting. In contrast, the sequential prompting approach builds on prior outputs: each prompt includes the topics identified so far (Doi et al., 2024). One key advantage of sequential prompting is its ability to produce less redundant and more coherent topic sets, since the LLM can choose between existing topics or generate new ones only when necessary. However, this approach becomes less scalable for very large corpora, as the context length of the prompts can grow indefinitely, increasing computational costs.

Prompt-Based Framework for Topic Modeling

Pham et al. (2024) proposed a prompt-based framework for topic modeling that adopts the sequential prompting strategy. The framework consists of two stages: Topic Generation and Topic Assignment. Each step consists of multiple sub-steps to ensure the generation of coherent and non-redundant topics and to prevent potential hallucinations from the LLM. The first step consists of the initial topic generation, followed by a refinement step in which the topics are further refined to ensure a coherent and non-redundant topic list (Pham et al., 2024).

Topic Generation. In the initial topic generation step, the framework generates a new topic for a given topic and a set of topics to identify an existing topic or generate a new topic. Pham et al. describes this process as: ”Given a document d from the corpus and a set of example topics S, the model is instructed to assign d to an existing topic in S or generate a new topic that better describes d and add it to S”(Pham et al., 2024).

The generated topic list S is subsequently refined by first generating topic embeddings using Sentence-Transformers to identify potential pairs of similar topics. The identified topic pairs are then provided as input to the LLM, which is instructed to merge topics that address identical topics. Additionally, infrequent topics are removed from S based on a predefined frequency threshold.

Topic Assignment. In the second stage, each document d in the corpus is assigned a topic from the list of refined topics in S. Each document d in the corpora is assigned to one or more topics. Lastly, the authors of Pham et al. (2024) suggest performing self-correction to address invalid topic assignments and hallucinations(Pham et al., 2024). During the self-correction step, the assigned topics are compared against the topic list S, if the assigned topics are not within the topic list or contain "None" / "Error", the document is again prompted to assign a new topic from S.

2.1.6 Dynamic (Temporal) Topic Modeling

Dynamic Topic Modeling (DTM) was first introduced by Blei and Lafferty (2006). Blei and Lafferty states: "The themes in a document collection evolve over time, and it is of interest to explicitly model the dynamics of underlying topics." (Blei and Lafferty, 2006), expressing the need to analyze the evolution of the topic of temporal documents. The groundwork for dynamic topic modeling was laid by Blei and Lafferty (2006), by proposing DTMs, which laid the foundations of dynamic topic modeling by extending latent Dirichlet allocation to capture evolving topics. In subsequent years, various researchers introduced additional topic models, including BERTopic by Grootendorst (2022), which was primarily a framework for topic modeling using State-of-the-Art transformers that includes a technique to perform dynamic topic modeling. In the following subsections, the first dynamic topic models are introduced, an extension of the formerly known Latent Dirichlet Allocation, before describing the process of how BERTopic achieves DTM.

Dynamic Latent Dirichlet Allocation

A naive approach to analyzing temporal documents is to apply LDA to each time slice independently. However, this approach would make the transition of topics between time slices abrupt and nonsmooth, since LDA assumes that documents are drawn exchangeably from the same set of topics(Blei and Lafferty, 2006).

To overcome this disadvantage, Blei and Lafferty (2006) assumes that a topic in a slice t evolves from topics in the previous time slice $t-1$. Blei and Lafferty (2006) achieves this by assuming that the topic-word distribution for a given time slice t and a topic k is influenced by the previous topic-word distribution $\beta_{t-1,k}$ that changes slightly using Gaussian noise (Blei and Lafferty, 2006), mathematically:

$$\beta_{t,k} \mid \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I). \tag{2.17}$$

Similarly, the document-topic proportions for time slice t are also drawn conditioned on the previous topic proportions of time slice $t-1$ given by:

$$\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I). \tag{2.18}$$

The final DTM assumes the following generative process:

1. For each time slice t of a sequential corpus:
   (a) Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
   (b) Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$
   (c) For each document:
      i. Draw $\eta \sim \mathcal{N}(\alpha_t, \alpha^2 I)$
      ii. For each word:
         A. Draw $z \sim Mult(\pi(\eta))$.
         B. Draw $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.

With this approach, for each time slice t, the topic-word distribution as well as the topic-document distribution changes slightly for each time slice, allowing the model to produce a smooth topic representation for each time slice that depends on previous time slices.

Dynamic Topic Modeling using BERTopic

Since BERTopic uses distributed representations of documents, the resulting clusters share a diverse vocabulary, neglecting the effect of wordings in different time stamps, because they share the same semantics and thus are semantically similar to each other. Grootendorst (2022) describes this effect as: "Here, we assume that the temporal nature of topics should not influence the creation of global topics. The same topic might appear across different times, albeit possibly represented differently." This assumption leads to a simple but effective approach to perform dynamic topic modeling using distributed representation and the formerly introduced C-TF-IDF(Grootendorst, 2022).

BERTopic is fitted on the entire corpus, ignoring the temporal nature of the corpus, as done in static topic modeling described in Section 2.1.4. Afterwards, for each timestamp

i, the C-TF-IDF is adjusted to calculate the term frequency for the given timestamp multiplied by the IDF value of the whole corpus:

$$W_{t,c,i} = tf_{t,c,i} * \log(1 + \frac{A}{tf_t}) \tag{2.19}$$

However, even though this approach would model the evolution of topics across different timestamps, the approach is equivalent to the naive approach as described in 2.1.6, leading to non-smooth representations. To incorporate the dependency of topic representations between time slices, similar to how DTM Blei and Lafferty (2006) works, Grootendorst suggests using the C-TF-IDF matrices of each timestamp. BERTopic does this by first normalizing the vectors of the C-TF-IDF for all topics across all timestamps using the L1-norm. Suppose $v_{t,i}$ is the vector representing a topic t at timestamp i, the normalized vector is then calculated by:

$$v_{t,i}^{norm} = \frac{v_{t,i}}{\|v_{t,i}\|_1}. \tag{2.20}$$

$\|v_{t,i}\|_1$ represents the L1-norm of the topic vector at timestamp i and is calculated as follows:

$$\text{L1-norm}(v_{t,i}) = \sum_j |v_{t,i,j}|, \tag{2.21}$$

where $v_{t,i,j}$ corresponds to the j-th component in the vector. Lastly, the representations are smoothed by taking the average of a topic vector for a given timestamp i with the topic vector of the previous timestamp $i - 1$.

This approach of dynamic topic modeling using distributed representations makes it very effective and computationally lightweight, since the dynamic topic modeling only uses C-TF-IDF for creating temporal topic representation.

2.2 Related Work

The rapid growth of social media platforms has transformed political communication and created new opportunities for analyzing public political discourse during electoral campaigns. Topic modeling has emerged as an essential technique for extracting meaningful topics from the vast amounts of unstructured textual data available on these platforms. Applying topic modeling to social media content enables researchers and decision-makers to better understand public opinion, political narratives, and reactions to policy decisions. The evolution from traditional methods like LDA and LSA to neural approaches like BERTopic and large language model-based methods presents both new opportunities and challenges, particularly in the context of political research.

This section reviews existing literature on social media political discourse analysis and comparisons of topic modeling techniques, identifying key gaps addressed by this thesis: the predominant focus on political figures rather than public discourse, and the limited comparative evaluation of traditional versus neural topic modeling approaches specifically within the political domain.

2.2.1 A Comparison Between Traditional and Neural Topic Models

Egger and Yu (2022) compared four topic modeling algorithms – LDA, Non-negative Matrix Factorization (NMF), Top2Vec and BERTopic – applied to tweets discussing COVID-19. Specifically, they analyze how these methods handle short and noisy textual data from social media. The study demonstrates the superior performance of neural topic modeling techniques, such as BERTopic and Top2Vec, over traditional methods such as LDA and NMF. In addition, Egger and Yu (2022) points out that achieving optimal results with LDA often requires careful selection of hyperparameters, including the prior assumption of document-topic distribution, topic-word distribution, and the number of topics. In particular, Egger and Yu (2022) highlights the ability of neural topic modeling techniques to capture nuances and context-rich topics better than traditional topic modeling techniques.

Kaur and Wallace (2024) also compared traditional topic modeling techniques like LDA and NMF against BERTopic, highlighting the limitations of LDA and NMF on social media data. They conducted interviews and evaluations with 12 qualitative researchers providing feedback on usability, interpretability, and insights. Their study showed that BERTopic outperformed LDA and NMF in providing detailed and coherent topics.

2.2.2 Political Discourse Analysis on Social Media

Achmann and Wolff (2023) applied BERTopic on social media data from Instagram related to the 2021 German Federal Election Campaign to gain insights into the political communication on these platforms. Their dataset included posts and stories from various accounts of major political parties and official candidates. They not only performed topic modeling on textual data, but also on images extracted in the data collection step using Optical Character Recognition (OCR). The results of their study show that most topics contained policy-related content, whereas the majority of stories focus on documentation of campaign rallies and events. Furthermore, BERTopic uncovered several policy-related topics such as climate change, renewable energies, education, and social issues.

Similarly, Hellwig et al. (2024) analyzed tweets during the German federal election of 2021 using BERTopic. Their study aimed at analyzing major topics discussed during the election and how they differ from topics of the public mentioning politicians. A key challenge was to capture not only text in tweets but also the textual content embedded in images shared on Twitter. They extended the dataset by extracting text from images using OCR. Their findings highlight that COVID-19, climate policy, and financial policy were central themes, while event-driven issues such as Afghanistan or Israel peaked at specific moments during the election period. Their results also reported that LDA produced fewer meaningful topics compared to BERTopic, as it fails to reveal distinct topics in the corpora. Their results highlight the advantages of BERTopic for handling short and noisy social media texts.

Unlike prior studies that either focus on politicians' communication (Achmann and Wolff, 2023) or compare politicians and the public (Hellwig et al., 2024), this thesis focuses on public discourse across platforms. Moreover, by systematically comparing

both traditional approaches (LSA, LDA) and transformer-based methods (BERTopic, TopicGPT), it addresses the limited comparative evaluations of topic modeling techniques within the political domain.

3 Methodology

The previous chapters provided an in-depth introduction to topic modeling and its application to the analysis of large document collections. The previous chapters introduced LSA and LDA, two of the first models to achieve significant success in automated document analysis, as well as BERTopic, a state-of-the-art framework based on transformer architectures. BERTopic extends traditional topic modeling by incorporating not only word co-occurrences but also the semantic relationships between words and sentences. Furthermore, previous chapters discussed related applications of topic modeling in social media analysis and political discourse research. Building on this foundation, this chapter outlines the experimental setup used to compare different topic modeling approaches in the analysis of political discourse on social media. The following sections describe the data collection and preprocessing pipeline, the training and optimization of the models, and the evaluation metrics used for model comparison.

3.1 Data Collection

The dataset used in this thesis was obtained from the Institut für Management- und Wirtschaftsforschung (IMWF)[1] as part of the collaborative research project KIFürDemokratie[2], conducted jointly with the University of Hamburg. The IMWF is an institute for management and economic research that collects data on thousands of companies, brands, and political figures on over 438 million websites (IMWF Institute, 2024).

Within the KIFürDemokratie project (Veliz et al., 2025), data were filtered using predefined signal words to identify politically relevant content, with a specific focus on right-wing extremism and populism. Data are collected through multiple feed URLs, which are automatically processed and stored in the project database. A daily script extracts the latest content from each feed and updates the database accordingly.

After filtering and storing the data, the resulting dataset includes various online sources such as web blogs, closed groups, microblogs, video portals, and press-related

1. https://www.imwf.de/
2. https://ki-fuer-demokratie.de/

Platform	Type of Source	Platform	Type of Source
Blog	Weblog	Online	News site
Facebook	Closed Group	Online	Popular Press
X / Twitter	Microblog	YouTube	Video Portal
Mastodon	Microblog	Online	Daily Newspaper (regional)
Bluesky	Microblog	Online	Daily Newspaper (national)
Forum	Forum	Online	Program TV/Radio
Online	Party, Club, Association	Telegram	Closed Group
Online	Press Service	Online	Trade and Professional Press
TikTok	Closed Group	Online	Business/Company
Instagram	Photo Portal	Online	Weekly Newspaper/Magazine
Online	News Agency	Online	Advertising Journal
Reddit	Forum	Threads	Microblog
Newsletter	Newsletter	Online	Consumer Portal
Online	Sunday Paper	Online	Customer Magazine
LinkedIn	Closed Group	Reader Comments	Reader Comments
Online	Wiki		

Table 3.1: List of Internet Sources and Their Types

platforms. Table 3.1 provides an overview of the data sources and their corresponding types.

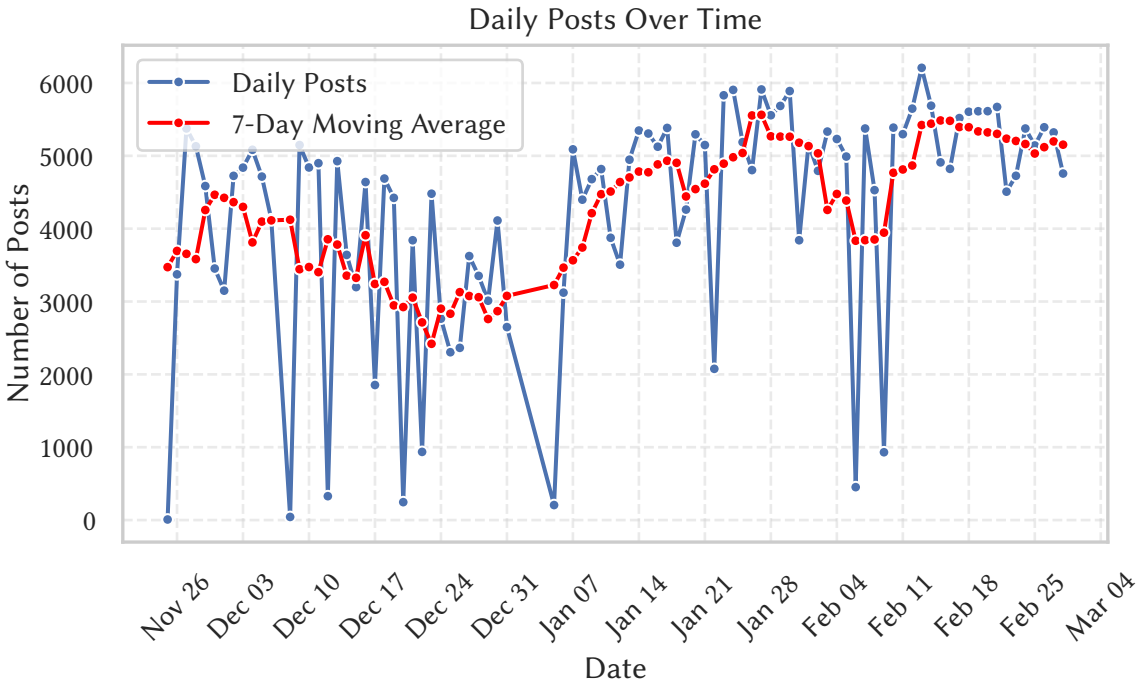

Figure 3.1: Number of Posts per day

The dataset was further restricted to November 25, 2024, to March 1, 2025. This filtering step resulted in a dataset containing 387,874 posts from ten different types of sources

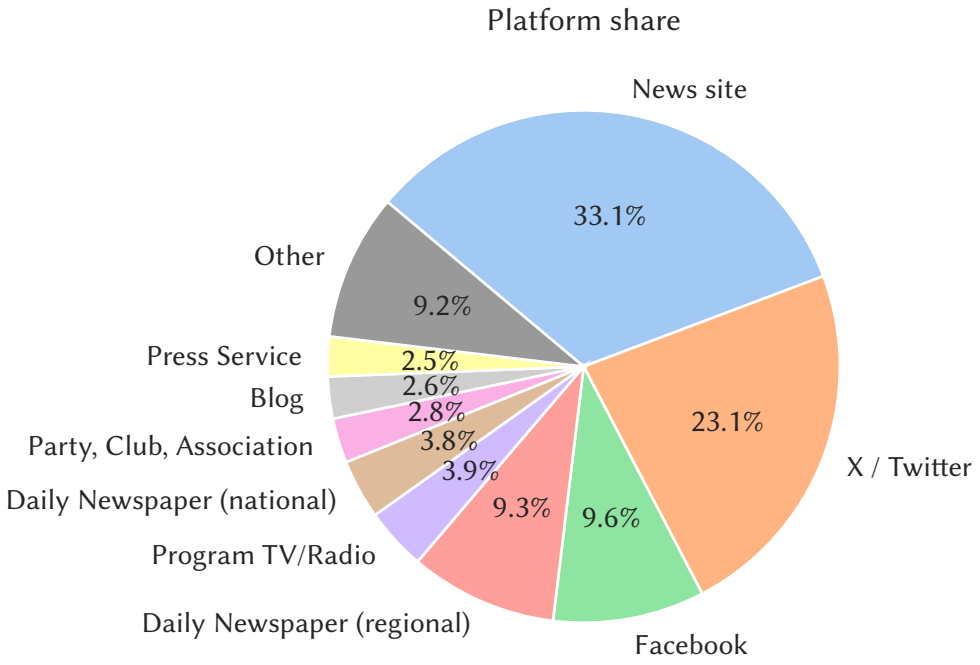

Figure 3.2: The individual share of platforms.

across 33 platforms. The dataset comprises 8,682 channels, with **Twitter** and **Facebook** being the most represented channels, followed by **de.dayfr.com**, **ausschreibungen-deutschland.de**, and **MSN Deutschland**.

Figure 3.2 shows the distribution of platforms in the dataset. It reveals that approximately 90.8% of all posts originate from the nine most active channels, with News sites accounting for 33.1% and Twitter accounting for 23.1% of the total content. Other microblogging platforms, such as Bluesky and Mastodon, are also represented to a lesser extent and are included in the 'Other' category.

Between November 25, 2024, and March 1, 2025, an average of 4,017 posts were extracted per day. Figure 3.1 illustrates the daily extraction volume, showing significant variation, with some days yielding minimal or no new content.

3.2 Data Preprocessing

Because the raw scraped data contain considerable noise, preprocessing was required before it could be used in any of the topic modeling techniques. Upon inspection, the data included raw HTML snippets and other artifacts that could negatively affect the resulting topic representations. To address this, a two-stage preprocessing approach was implemented, consisting of (1) general text cleaning and (2) model-specific preprocessing.

All preprocessing steps were implemented in Python, which offers a comprehensive ecosystem of libraries for natural language processing and data handling. The first stage removed general noise, such as HTML tags (e.g., `<div>`, `<em>`), line breaks, tabs, and redundant spaces. Personal information, such as user mentions, URLs, and phone numbers, was removed, as these elements do not contribute to the identification of

Dataset Version	Size of Dataset
Raw Dataset	387,874
Deduplicated Dataset	310,868
Deduplication after the Basic Preprocessing	301,783
Traditional Dataset	296,856
BERT Dataset	301,159

Table 3.2: Dataset Size across Different Datasets

latent topics. These general preprocessing steps were applied consistently across all models (LDA, LSA, and BERTopic).

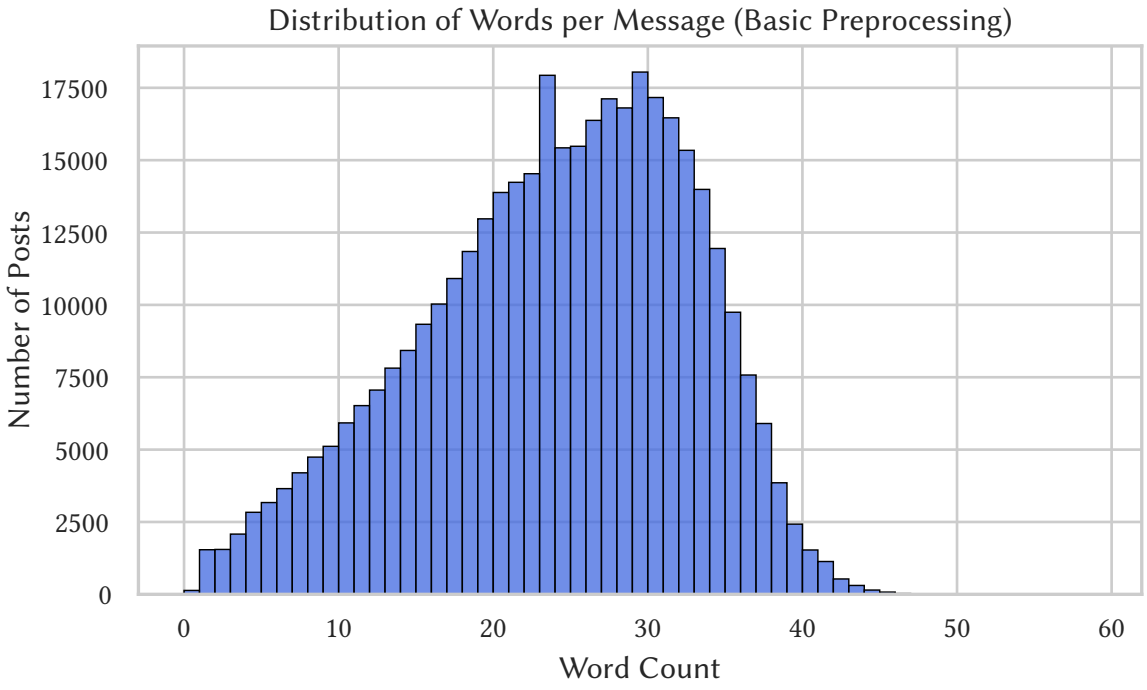

Figure 3.3: Distribution of the number of words across the dataset.

Figure 3.3 illustrates the distribution of document lengths in the dataset. Most documents contain between one and fifty words, which aligns with typically social media posts and comments that are short.

The second preprocessing stage was model-specific. For LDA and LSA, which rely on document-term matrices, additional cleaning steps were performed, including stop-word and emoji removal. In contrast, BERTopic leverages document embeddings and therefore requires less preprocessing, as it can process full sentences without removing stop words or punctuation.

Due to the substantial amount of duplicate content across different sources, a deduplication step was also conducted, reducing the corpus by approximately 22.2% to a total of 301,783 documents.

Finally, because hashtags often contain valuable semantic information, they were preserved by removing only the '#' prefix. This ensured that hashtags contributed meaningfully to topic identification without being discarded as noise.

3.3 Topic Models used for Political Discourse Analysis

This section describes the topic modeling techniques applied in this thesis to analyze political discourse on social media. It introduces LSA, LDA, and BERTopic as representative models for static topic modeling, followed by TopicGPT, a novel prompting-based approach that leverages LLMs for topic extraction and assignment. Finally, this section outlines the process of determining optimal model parameters and briefly introduces dynamic topic modeling to capture temporal changes in topic relevance and representation during the German election period.

3.3.1 Latent Semantic Analysis (LSA)

LSA (Deerwester et al., 1990) is one of the earliest approaches for discovering latent topics in large corpora of text. It applies SVD to decompose the document-word matrix into three smaller matrices that capture the underlying structure of term co-occurrences. In contrast to BERTopic, which automatically infers the number of topics, LSA requires the number of topics k to be specified in advance, as the reduced matrix must be truncated to rank k. Since k is unknown in the context of political discourse analysis, it must be estimated during the optimization process.

Two additional parameters, `power_iterations` and `extra_samples`, improve the numerical accuracy of the decomposition. The `power_iterations` parameter controls how many refinement steps are used to approximate the singular value decomposition, where higher values yield a more accurate low-rank representation. The `extra_samples` parameter determines the size of the intermediate subspace before truncation, allowing more robust estimation of latent structures.

3.3.2 Latent Dirichlet Allocation (LDA)

LDA (Blei et al., 2003) is the most widely used probabilistic topic model. It assumes that each document is represented as a mixture of topics θ drawn from a Dirichlet distribution with parameter α, and that each word is drawn from a topic-specific distribution over the vocabulary (β), which itself is drawn from a Dirichlet prior with parameter η.

The hyperparameters α and η serve as priors on the document-topic and topic-word distribution, respectively, and both must be defined or estimated during model optimization. As with LSA, the number of topics k must be predetermined, since it determines the dimensionality of the latent space. Because k is unknown beforehand, it is included as part of the hyperparameter search.

3.3.3 BERTopic

BERTopic (Grootendorst, 2022) represents a transformer-based framework that performs topic modeling using document embeddings and density-based clustering. Unlike LDA

or LSA, the number of topics is not specified beforehand but emerges naturally from the clustering process.

BERTopic includes several components: document embeddings, dimensionality reduction using UMAP, HDBSCAN, and topic representation using C-TF-IDF. Each stage introduces parameters that influence model performance.

For UMAP, the most relevant parameters are `n_neighbors`, `min_dist`, and `n_components`. The `n_neighbors` parameter controls the balance between local and global structure. Lower values capture fine-grained local relationships, while high values preserve more global relationships. The `min_dist` parameter determines how tightly points are packed in the low-dimensional space. Lower values yield denser clusters, while higher values produce more dispersed clusters. Finally, `n_components` specifies the dimensionality of the low-dimensional representation.

For HDBSCAN, the most important parameters are `min_cluster_size` and `min_samples`. The `min_cluster_size` parameter defines the minimum number of points required to form a cluster. Increasing this value yields fewer but larger clusters. `min_samples` specifies how conservative the clustering is. Higher values result in more points being treated as noise.

Because these parameters interact closely, they must be tuned jointly to achieve optimal topic separation and stability.

3.3.4 TopicGPT

TopicGPT (Pham et al., 2024) is a prompt-based framework that uses LLMs to generate and assign topics without additional training. The model prompts an LLM to extract topic descriptions directly from text and then to assign documents to these topics. By default, TopicGPT outputs only topic labels or titles rather than word-based representations, which complicates direct comparison with traditional topic models such as LSA, LDA, and BERTopic.

To ensure comparability, the TopicGPT framework was extended with an additional step based on the C-TF-IDF method introduced by Grootendorst (2022). This step computes word-level topic representations by aggregating term frequencies across documents assigned to each topic and weighting them by their inverse document frequencies. The resulting representations capture the most informative words per topic, enabling consistent evaluation alongside other models.

TopicGPT exposes two primary parameters: the similarity threshold used during topic refinement, which controls the merging of semantically similar topics, and the temperature parameter, which affects the diversity of topic generation. However, due to the computational cost of prompting large language models, TopicGPT was not optimized using the OCTIS framework. Instead, it was applied with default LLM parameters and a fixed topic-embedding distance threshold.

Finding optimal hyperparameters is a crucial step in topic modeling, as they strongly influence the interpretability and coherence of discovered topics. The following section describes the optimization strategy applied in this thesis to determine the best parameter configurations for each model.

3.4 Hyperparameter Optimization

A crucial step in topic modeling is selecting optimal hyperparameters for generating coherent and meaningful topics. Because topic modeling is an unsupervised task, optimal configurations cannot be inferred directly from labeled data. This thesis employs Bayesian optimization using the Optimizing and Comparing Topic Models Is Simple (OCTIS) framework (Terragni et al., 2021) to automatically determine the best model configurations. The following section outlines how Bayesian optimization was applied and summarizes the search space defined for each topic model.

3.4.1 OCTIS Framework

OCTIS (Terragni et al., 2021) is a framework designed to facilitate the training, comparison, and optimization of topic models in a unified environment. It automates common experimental steps, including preprocessing, model training, evaluation, and hyperparameter tuning. The framework supports a wide range of topic modeling algorithms. OCTIS provides several evaluation metrics, including topic coherence, diversity, and classification-based scores, and supports optimization through Bayesian optimization. In this thesis, OCTIS is used for training and hyperparameter optimization of LDA, LSA, and BERTopic.

Integration of BERTopic into OCTIS. Since the OCTIS framework does not natively support the BERTopic model, a custom OCTIS-compatible implementation was developed as part of this thesis. The model class was created by inheriting from the OCTIS abstract class `AbstractModel`. This integration allows BERTopic to be trained, evaluated, and optimized using the same workflow as the other models, ensuring consistent comparison across all topic modeling approaches.

3.4.2 Bayesian Optimization in OCTIS

Bayesian optimization iteratively explores a predefined search space to maximize an objective function $f(x)$, where x represents a candidate hyperparameter configuration. A surrogate model approximates the true objective function and provides a confidence estimate of the function, which is used to select the next point to be evaluated. This is done by an acquisition function a_n, that decides which input x should be chosen next to evaluate the model by solving $x_{n+1} = \underset{x \in A}{\operatorname{argmax}}\, a_n(x)$ (Snoek et al., 2012). A commonly used acquisition function is the *Expected Improvement* function (Snoek et al., 2012).

The objective function combines topic coherence and diversity to ensure that optimized models generate interpretable yet distinct topics (see Section 3.4.4).

3.4.3 Hyperparameter Search Space

The hyperparameters to include in the search space are selected specifically for each model.

LDA. For Latent Dirichlet Allocation, the hyperparameter search space includes the number of topics `num_topics`, the document-topic prior α, and the topic-word prior η.

The number of topics varies between 50 and 300 to capture different levels of typical granularity. For the priors, α is tested in three configurations, `symmetric`, `asymmetric`, and `auto`, while *eta* is evaluated in its `symmetric` and `auto` settings. When setting either α or η to `symmetric`, the probabilities are set as:

$$\alpha_i = \frac{1.0}{\text{num\_topics}}. \tag{3.1}$$

The `asymmetric` setting uses a fixed normalized asymmetric prior, defined as:

$$\alpha_i = \frac{1.0}{\text{i} + \sqrt{\text{num\_topics}}}. \tag{3.2}$$

The `auto` setting, on the other hand, learns an asymmetric prior from the corpus.

Because the dataset consists mainly of short social-media posts (1-50 words per document; see Section 3.2), documents are expected to contain only one or two dominant topics. Setting low or asymmetric prior values encourages sparsity in the topic-document and topic-word distributions, resulting in higher probabilities for a few relevant topics per document and near-zero probabilities for others, as desired for short-text topic modeling.

LSA. For LSA, the only hyperparameter included in the search space is the number of topics, `num_topics`, which determines the rank k of the reduced document-term matrix. The parameter varies between 50 and 300 to identify the level of dimensionality that yielded the most coherence and diverse topic structure. The model is trained using the one-pass algorithm (Halko et al., 2011) with a fixed `chunk_size` of 20,000, two `power_iterations`, and 100 `extra_samples` to ensure numerical stability and efficient computation. This setup balances accuracy and runtime while preserving the latent semantic structure of short, sparse social media documents.

BERTopic. BERTopic is a framework that is built on distributed representations obtained by using transformers. The topics found by BERTopic are primarily formed by the underlying clustering algorithm HDBSCAN, which is applied to the reduced representations obtained by UMAP. UMAP offers several parameters for performing the dimensionality reduction, which are primarily controlled by the `n_neighbors`, `min_dist`, and `n_components` parameters. Since this experiment aims to analyze the political discourse on social media platforms, focusing on recent events, the dimensionality reduction should focus more on local structures than global structures. This leads to relatively low `n_neighbors` to look at when performing UMAP and is set to a range of 2 - 50, which results in more fine-grained structures. The dimensionality on which UMAP projects the original space should be based on the context of HDBSCAN, which uses the reduced representations. McInnes et al. (2017) of HDBSCAN state that the performance decreases in high-dimensional data. Furthermore, they clarify that HDBSCAN performs well on dimensions of up to 100, so the search space of the `n_components` parameter for UMAP is set to a range of 2 to 50 dimensions. The `min_dist` parameter in UMAP controls how tightly similar points are packed together. This parameter search space is set to a range of 0.001 to 0.1 because the embeddings of the documents will most likely be relatively close in the embedding space, since most documents focus on political topics. For HDBSCAN, both `min_cluster_size`

and `min_samples` are included in the search space, since both parameters significantly influence the resulting clusters. The parameter `min_cluster_size` controls the group size that should be considered a cluster. The relevant clusters to analyze political discourses should consist of at least a meaningful number of points to avoid focusing on irrelevant topics discussed on social media. Similarly, the groupings should not be categorized too broadly to capture topics within the domain of politics. Thus, the parameter is set to a range of 50 to 350 points to consider only groups with a meaningful size. To control the number of outliers generated for each grouping, the `min_samples` parameters are set to a relatively small range of 1 to 50 points. For embedding the documents, the multilingual model `paraphrase-multilingual-MiniLM-L12-v2`[3] was used, which has shown strong performance on German text clustering tasks.

TopicGPT. Due to the computational cost of prompting large language models, TopicGPT was not optimized through OCTIS. Instead, two LLMs, `Mistral-7B-Instruct`[4] and `GPT-4o-mini`[5] were compared directly, following the methodology of Pham et al. (2024). `Mistral-7B-Instruct` balances performance and computational efficiency, representing the current state of open-source instruction-tuned models. `GPT-4o-mini` offers a comparison with a frontier commercial model, enabling assessment of performance differences between open-source and proprietary solutions in the context of political discourse topic modeling.

3.4.4 Optimization Metrics

Since topic modeling is an unsupervised learning technique, classification-based metrics are not applicable for this task. Instead, topic quality is assessed either through human judgment or by using intrinsic measures such as topic coherence and diversity. A quantitative objective is required for Bayesian optimization, which evaluates how well a given set of hyperparameters produce coherent and distinct topics.

Topic Coherence. Topic coherence measures the interpretability of a topic by measuring the semantic similarity between its most probable words. Röder et al. (2015) introduced C_v as a topic coherence metric, which combines the indirect cosine measure, the direct confirmation measure, normalized pointwise mutual information (NPMI), and the Boolean sliding window. C_v segments the top words $V^{(t)}$ using one-set segmentation S^{one}_{set}. Where each word $V^{(t)}$ is paired with all other most probable words in t. Mathematically, this means:

$$S^{one}_{set} = \{(W', W^*) | W' = \{v_i^{(t)}\}; v_i^{(t)} \in V^{(t)}; W^* = V^{(t)}\} \tag{3.3}$$

Additionally, instead of estimating the probability using Boolean documents (the number of documents in which the word appears), C_v uses the Boolean sliding window (P_{sw}), which counts the number of appearances of a word per sliding window. Finally, Röder et al. (2015) calculates an indirect confirmation measure instead of a direct confirmation measure. The indirect confirmation measure assumes that words v_m and v_l support each other even if they do not frequently appear together. To capture

3. https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
4. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
5. https://openai.com/de-DE/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

the hidden support, Blei and Lafferty (2006) proposed to calculate the support for other words in the corpus, which leads to a strong correlation between these direct confirmation measures. Thus, semantically supported words should have a similar vector of confirmation measures to all other words. Mathematically, this can be formulated as:

$$\vec{v}_{m_{nlr},\gamma}(W') = \{\sum_{w_i \in W'} m_{nlr}(w_i, w_j)^{\gamma}\}_{j=1,...,|w|} \tag{3.4}$$

,

where m_{nlr} is the normalized log-ratio measure (NPMI) and is defined as:

$$m_{nlr} = \frac{m_{lr}(S_i)}{-\log(P(W', W^*) + \epsilon)} \tag{3.5}$$

and m_{lr} is the log-ratio measure (PMI) defined as:

$$m_{lr} = log\frac{P(W', W^*) + \epsilon}{P(W') * P(W^*)}. \tag{3.6}$$

Röder et al. (2015) named the resulting vectors for a given word set pair from $S_i = (V^{(t)'}, V^{(t)} *)$ context vectors. These context vectors are then used to calculate an indirect confirmation using cosine similarity. Finally, the confirmations of all subsets S_i of the topic t are aggregated to a single coherence score using the arithmetic mean (σ_a). The final coherence score for the whole model is simply the mean of all topic coherence scores.

Topic Diversity. To address this limitation, additional evaluation metrics are considered during the Bayesian optimization process. In particular, topic diversity, introduced by Dieng et al. (2020), measures how distinct the generated topics are. It is defined as the proportion of unique words among the most probable words of all topics:

$$TD(T) = \frac{|\bigcup_{k=1}^{K} W_k|}{K \cdot N} \tag{3.7}$$

where $T = \{t_1, ..., t_k\}$ denotes the set of generated topics, N is the number of top words considered per topic, and W_k is the set of top N words for topic t_k. A higher value of $TD(T)$ indicates greater diversity, meaning that the topics cover a broader range of words.

Objective Function. To optimize coherence and diversity, the Bayesian optimization process maximizes the product of coherence and diversity:

$$f(T) = C_v(T) \times TD(T). \tag{3.8}$$

This objective function ensures that the selected model configuration produces topics that are both semantically meaningful and non-overlapping.

3.4.5 Selection of the best Parameters

For each model, the best hyperparameter configuration is selected as the one that achieves the highest value of the objective function $f(T)$. Formally:

$$\theta^* = \arg\max_x f(T(\theta)) \tag{3.9}$$

where θ represents a candidate hyperparameter set and $T(\theta)$ the corresponding generated topics.

This approach ensures that each topic model is evaluated using its most optimal configuration before being compared against other models, providing a fair basis for subsequent analysis of political discourse.

3.5 Dynamic Topic Modeling

The second research question of this thesis is:

> How do political topics evolve over time across online platforms, and what insights can be drawn about the dynamics of public discourse?

To answer this question, dynamic topic modeling is applied to the dataset to analyze topic evolution over the course of the German election. Unlike static models such as LDA and LSA, BERTopic provides a built-in approach for dynamic topic modeling, making it suitable for this study. While classical DTMs based on LDA (Blei and Lafferty, 2006) can also model topic evolution, LSA lacks an inherent temporal component and would require manual alignment of topics across time slices, introducing potential inconsistencies. Even though DTM based on LDA is capable of modeling temporal shifts in topics, the model is not considered for the analysis of topic evolution in this thesis. Therefore, BERTopic is chosen as the primary approach for dynamic topic modeling in this thesis.

BERTopic assumes that topics remain semantically stable across timestamps, meaning that while topic representations may shift over time, their underlying meaning remains consistent (Grootendorst, 2022). These assumptions allow for meaningful tracking of topics through the election period.

3.5.1 BERTopic for Dynamic Topic Modeling

This subsection describes the implementation of BERTopic's dynamic topic modeling pipeline, including temporal segmentation of the dataset and the creation of time-dependent topic representations.

Temporal Segmentation of the Dataset

To analyze topic evolution, the dataset was divided into temporal segments based on predefined time intervals. These segments can be based on:

- Fixed time windows (e.g., daily, weekly, or monthly segments).
- Key political events (e.g., major debates, polling days).

In this thesis, daily segmentation was chosen to reflect the fast-paced nature of social media, where political discussions often shift rapidly in response to unfolding events. Using daily segments ensures that the topic modeling captures short-term dynamics, such as sudden spikes in attention to particular issues, emerging discourses, or shifts in public sentiment. Segmenting the dataset into weekly or event-based windows could smooth over these rapid changes, potentially obscuring important insights about how discourse evolves in response to ongoing events.

Implementation of Dynamic Topic Modeling

BERTopic's dynamic topic modeling method extracts topic representations at each timestamp without retraining the entire model several times (see Section 2.1.6). For this, the hyperparameters optimized via Bayesian optimization are leveraged, as explained in Section 3.4, to train the topic model without temporal aspects. Once the model is fitted on the non-temporal data, the corresponding timestamps for each document are extracted. Next, the temporal topic representation is calculated using the BERTopic model's method `topics_over_time`, which calculates C-TF-IDF for each timestamp.

BERTopic fine-tunes the topic representation for each timestamp using *evolutionary tuning* and *global tuning*. Evolutionary tuning smooths the topic representation obtained by C-TF-IDF for each timestamp t by averaging the representation at the timestamp t with the representation of the previous timestamp $t-1$. Similarly, global tuning of the topic representation for each timestamp t is achieved by averaging the result of a topic representation at timestamp t with the global topic representation, which was calculated by C-TF-IDF on each cluster without the temporal aspect of the data. Both techniques can be applied independently or in combination. When both are applied, *evolutionary tuning* is performed first, followed by *global tuning*, to enhance the temporal topic representation.

4
Results and Evaluation

This chapter presents and evaluates the results of the topic modeling experiments described in Chapter 3. The results are organized into three main sections addressing the research objectives. Section 4.1 reports the results of the hyperparameter optimization for LSA, LDA, and BERTopic, identifying the configurations that maximize topic coherence and diversity. Section 4.2 provides a systematic performance comparison across all models, including TopicGPT, using quantitative metrics and qualitative manual inspection of topic quality, directly addressing the first research question regarding the most effective topic modeling approach for political discourse analysis. Section 4.4 applies BERTopic in a dynamic setting to analyze the temporal evolution of political discourse throughout the election period, addressing the second research question concerning topic dynamics over time.

4.1 Model Optimization Results

The first phase of the experimental evaluation focused on identifying the optimal hyperparameter configurations for LDA, LSA, and BERTopic. Each model was optimized using Bayesian optimization implemented in the `OCTIS` framework (Terragni et al., 2021). The objective function combined topic coherence C_v and topic diversity TD, ensuring that the optimized models produced semantically coherent and distinct topics.

Bayesian optimization was applied for 25 iterations per model, with each hyperparameter configuration evaluated over three independent runs to account for stochastic variations. The mean, median, and standard deviation of the objective function were recorded for each iteration. The following subsections summarize the optimization process and report the best-performing configurations for each model based on the final objective function score.

4.1.1 LSA Optimization Results

The first model tested in the optimization process was LSA. This subsection presents the results of the Bayesian optimization applied to LSA as described in Section 4.1.

Figure 4.1 shows the optimization objective, as defined in Section 4.1. In the first iterations, the optimization objective score increased slowly and steadily to a local maximum. Afterwards, the objective function converged at around 0.07 before reaching its maximum at iteration 22. Both topic coherence and diversity followed a similar trend as the objective function reached its maximum at the same iteration.

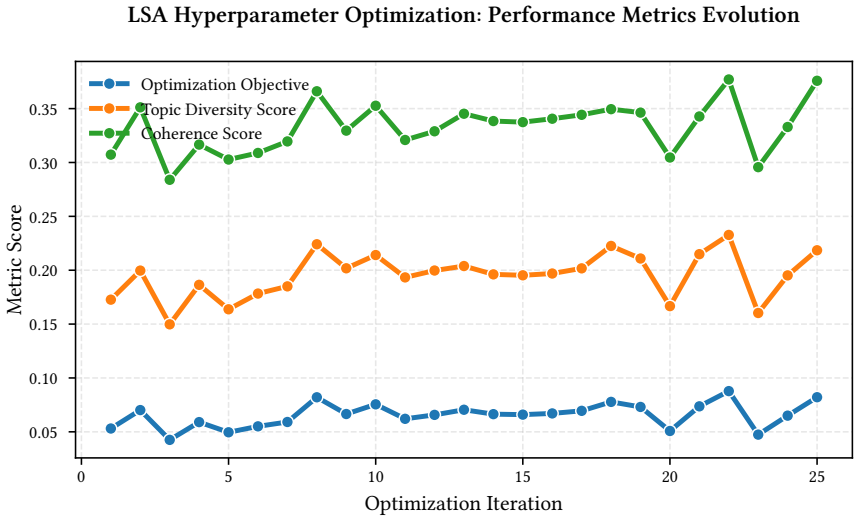

Figure 4.1: Bayesian Optimization score across iterations. The score is the product of topic diversity and coherence (C_v). For each iteration, a unique hyperparameter configuration is evaluated by running three separate models with the same settings and the final score is the mean of these runs.

Table A.1 lists all iterations with their mean, median, and standard deviation values, as well as corresponding coherence and diversity scores. The best results were achieved in iterations 8, 22, and 25, with mean scores of 0.07, 0.07, and 0.088. The very small standard deviations (all below 0.005) show that LSA produces very stable results across runs. This is to be expected since LSA is a deterministic model, and only small deviations occur due to the optimized implementation used by Gensim (Řehůřek and Sojka, 2010). Figure 4.2 visualizes the deviations of each iteration's runs, clearly showing the stability of LSA.

Overall, LSA achieved a low topic coherence and even lower topic diversity scores. Based on the results, the hyperparameter from iteration 22 was chosen for the model comparison as it achieved the highest mean score of the objective function with very low standard deviation.

4.1.2 LDA Optimization Results

After evaluating the performance of LSA, the next model to be optimized was LDA, which uses a probabilistic approach to discover latent topics. The search space for the Bayesian optimization was previously described in Section 3.4.3.

Figure 4.3 visualizes the trajectory of the optimization objective function. In contrast to LSA, the objective score increases rapidly in the early iterations and does not stabilize until after the 18th iteration, where it reaches an optimal space. However, topic coherence and diversity scores fluctuate strongly during the optimization process across all iterations. This fluctuation in topic coherence and diversity most likely arose from

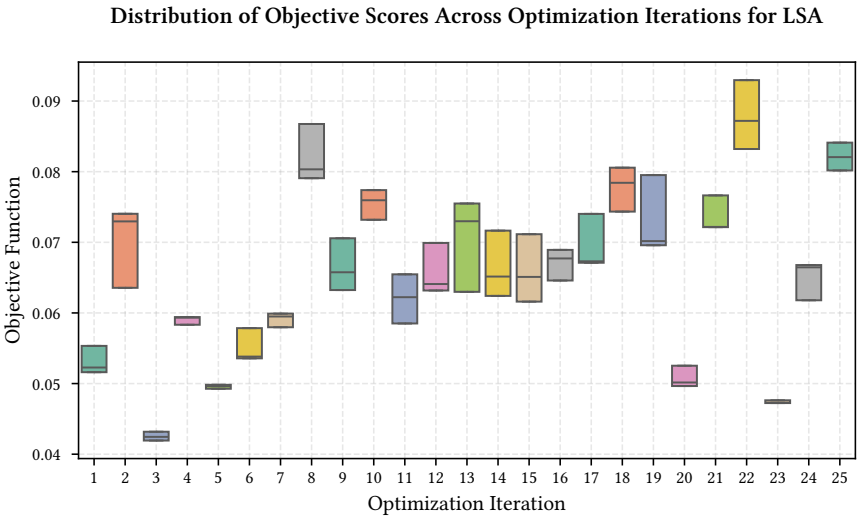

Figure 4.2: Distribution of objective function values across optimization iterations for LSA. Each box plot represents the distribution of objective scores obtained during a specific iteration, showing the median (center line), interquartile range(box), and range (whiskers).

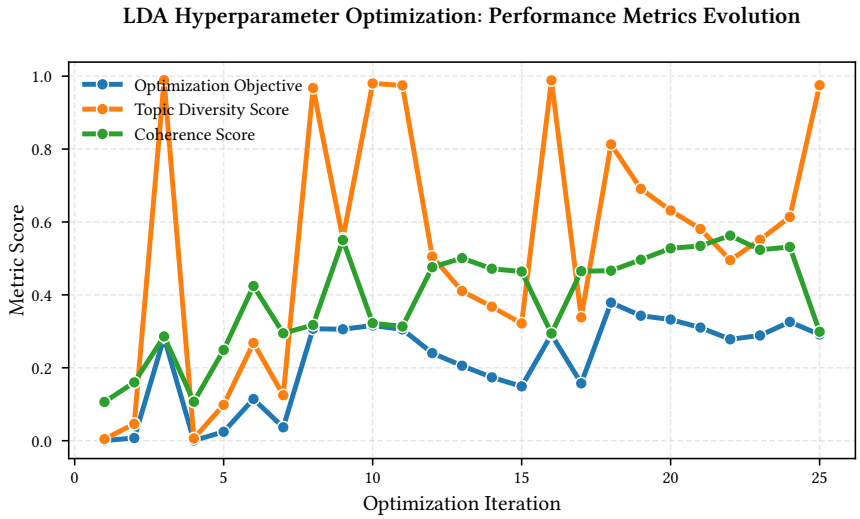

Figure 4.3: Bayesian Optimization score across iterations. The score is the product of topic diversity and coherence (C_v). For each iteration, a unique hyperparameter configuration is evaluated by running three separate models with the same settings and the final score is the mean of these runs.

the more complex search space compared to LSA's optimization process. Notably, topic coherence and diversity did not increase or decrease simultaneously but demonstrated a trade-off between both metrics. This differs from LSA optimization trajectories, where both metrics increased or decreased simultaneously.

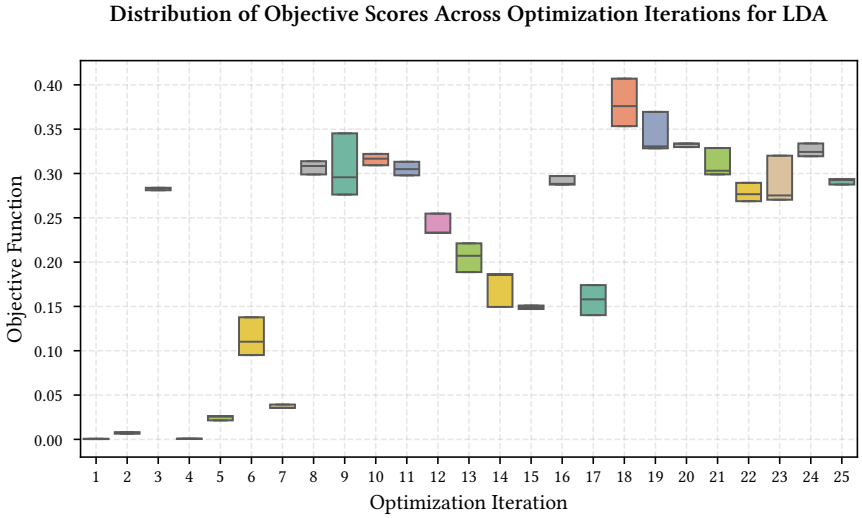

Figure 4.4: Distribution of objective function values across optimization iterations for LDA. Each box plot represents the distribution of objective scores obtained during a specific iteration, showing the median (center line), interquartile range(box), and range (whiskers).

Even though LDA is a probabilistic model due to its generative assumption of documents being a mixture of topics, the distribution of objective scores across runs in an iteration is relatively small, similar to LSA. The best result was achieved in iteration 18 with a mean score of 0.379 and a standard deviation of only 0.022. The optimal parameters for LDA are found to be 139 topics with α set to symmetric and η set to auto. Detailed results for all iterations are provided in Appendix B.1.

4.1.3 BERTopic Optimization Results

BERTopic showed a similar optimization trajectory to LSA, where it did not increase significantly across iterations. However, BERTopic suffered from similar spikes in the optimization trajectory as LDA does, as seen in Figure 4.5. In the early iterations, the objective function slightly decreases. Afterwards, the performance improves rapidly and reaches its highest value at iteration 6 before decreasing to similar scores as before the spike. Beyond this point, the objective score remained mostly even with occasional spikes.

As seen in Table C.1, BERTopic achieves the highest score at iteration 6 with a mean objective score of 0.639. However, this iteration also had the highest standard deviation with a value of 0.145. This suggests a highly unstable and suboptimal topic model. Figure 4.6 demonstrates the highly fluctuating scores within an iteration. Iterations 11, 16, and 20 share similar observations, indicating hyperparameter settings resulting in unstable topic models with highly varying topic diversity and topic coherence scores.

To ensure stable and reproducible results, the hyperparameters with high standard deviation scores are ignored, and instead, the next best hyperparameter is selected

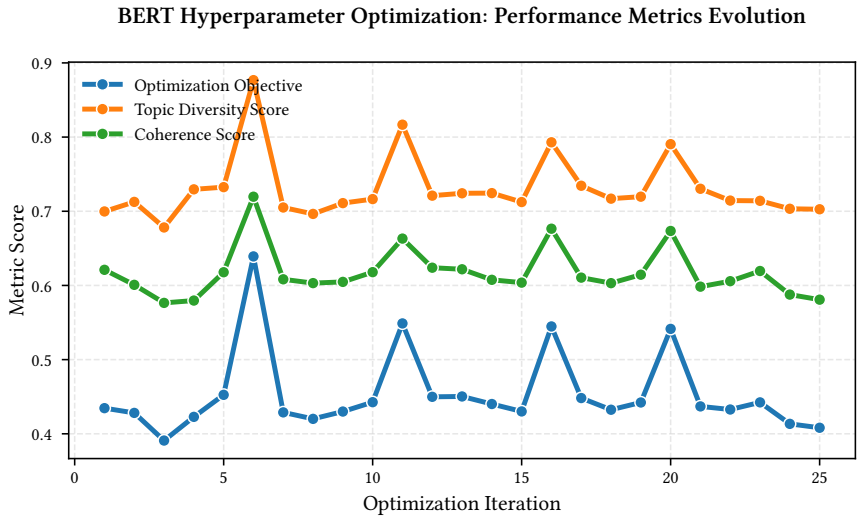

Figure 4.5: Bayesian Optimization score across iterations. The score is the product of topic diversity and coherence (C_v). For each iteration, a unique hyperparameter configuration is evaluated by running three separate models with the same settings and the final score is the mean of these runs.

for further experimentation and comparison. This led to the hyperparameter settings found in iteration 5 being the best performing model with a mean score of 0.453 and a standard deviation of 0.006. The hyperparameters for this iteration were found as follows: `min_cluster_size` being set to 314 with a `min_dist` of 0.055, `min_samples` of 27, `n_components` of 44, and `n_neighbors` of 34.

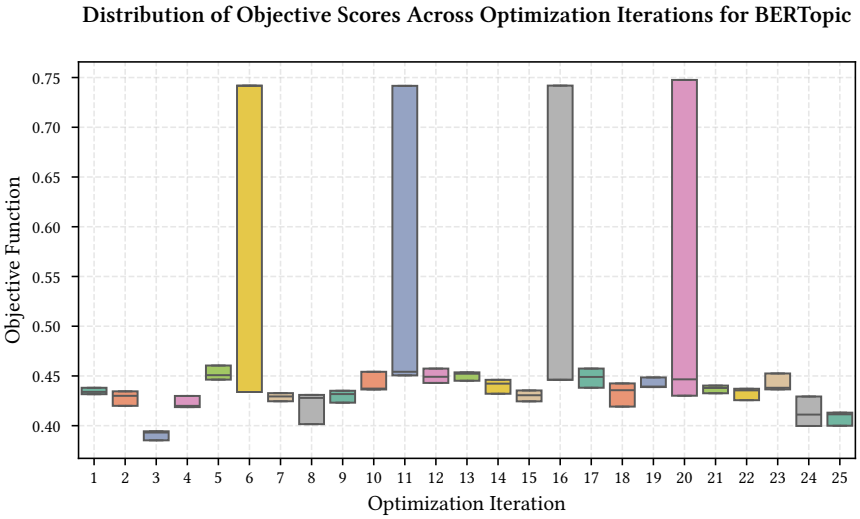

Figure 4.6: Distribution of objective function values across optimization iterations for BERTopic. Each box plot represents the distribution of objective scores obtained during a specific iteration, showing the median (center line), interquartile range(box), and range (whiskers).

4.1.4 TopicGPT Results

Since TopicGPT is a framework that applies Large Language Models for topic modeling, this approach was not optimized using Bayesian Optimization. Instead, following the original paper's experimental setup, two different LLMs were evaluated. The choice of *Mistral-7B-Instruct* and *GPT-4o-mini* was inspired by the original paper to enable direct comparison with their findings (Pham et al., 2024). Table 4.1 summarizes the performance metrics for both models.

The choice of LLM impacted both the topic generation behavior and the final model performance. *Mistral-7B-Instruct* generated 1080 topics during the initial generation phase, which were subsequently refined to 13 core topics, before expanding to 16 topics after subtopic generation. This behavior aligns with observations by Pham et al. (2024), who noted that smaller open-source LLMs (in their case, Mistral-7B-Instruct) tend to generate many redundant topics rather than selecting from existing topics. The final 16 topics captured broad themes, including politically related and non-politically related topics. It scored a mean objective score of 0.35 with a topic coherence score of 0.51 and a topic diversity score of 0.68.

In contrast, *GPT-4o-mini* was more efficient in the initial topic generation phase, generating 39 topics, which were refined to 17 topics, and resulted in 62 distinct topics. This model achieved a higher mean score of 0.40, with topic coherence and diversity values of 0.54 and 0.74, respectively. The larger set of final topics suggests that *GPT-4o-mini* captured more fine-grained distinctions within political discourse with a higher coherence and diversity score.

Model	# Topics	Objective Function	Coherence C_v	Topic Diversity	Error Rate	Hallucination Rate
Mistral-7B-Instruct	16	0.35107	0.51533	0.68125	0.2729 %	0.7444 %
GPT-4o-mini	62	0.404890	0.540425	0.749206	0.0620 %	1.2408 %

Table 4.1: Evaluation of the TopicGPT framework with Mistral7B-Instruct and GPT-4o-mini, showing the number of topics, objective function, coherence (C_v), topic diversity (TD), and error/hallucination rate.

While *Mistral-7B-Instruct* had an error rate approximately five times higher than *GPT-4o-mini*, the latter showed a higher hallucination rate, about twice that of *Mistral-7B-Instruct*. These metrics refer to cases where the model either produced invalid topic assignments (errors) or generated non-existent topics (hallucinations).

Since this thesis focuses on analyzing political discourse during the German election, the *GPT-4o-mini* model was used for further analysis and comparison between LSA, LDA, and BERTopic because it was able to capture more specific and politically relevant topics.

In summary, each model was optimized or evaluated to identify the hyperparameter settings that achieved the best balance between topic coherence, diversity, and stability. LSA and LDA showed stable performance with moderate coherence values, while BERTopic achieved the highest overall objective scores but also unstable parameter settings. For TopicGPT, *GPT-4o-mini* was selected as the most suitable LLM, as it generated more coherent and politically relevant topics than *Mistral-7B-Instruct*. The following section compares their results in terms of topic quality, interpretability, and suitability for analyzing political discourse.

4.2 Model Evaluation

This section compares the performance and output quality of the optimized topic models identified in Section 4.1. Specifically, it contrasts the results of LSA, LDA, BERTopic, and TopicGPT using both quantitative metrics, such as topic coherence and diversity, and qualitative assessments of topic quality and interpretability.

4.2.1 Quantitative Evaluation

Model	Best Iteration	# Topics	Coherence C_v	Topic Diversity
LSA	22	57	0.377	0.232
LDA	18	135	0.483	0.830
BERTopic	4	99	0.615	0.726
TopicGPT(GPT-4o-mini)	-	63	0.540	0.749

Table 4.2: Comparison of the best-performing topic models after hyperparameter optimization, showing the number number fo topics, topic coherence, and topic diversity.

Table 4.2 summarizes the performance of the best hyperparameter for each model. Among the traditional approaches, LDA achieved the highest topic diversity of 0.83, indicating that it discovered a broad range of distinct topics. However, its coherence score of 0.483 remained notably lower than that of BERTopic. This suggests that although LDA topics are well separated, they are less semantically consistent. BERTopic achieved the best overall balance between topic coherence and diversity. LSA, by contrast, produced considerably lower scores on both coherence and diversity metrics. TopicGPT (GPT-4o-mini) performed comparably to BERTopic, achieving high diversity and slightly lower coherence. This indicates that large language models can be applied as competitive zero-shot or few-shot topic models.

4.2.2 Qualitative Evaluation

Beyond quantitative metrics, a qualitative inspection of the discovered topics by each model is conducted. This subsection aims to evaluate and compare the coherence, interpretability, and ability to capture relevant political topics during the election period. Each model's discovered topics are inspected and interpreted by also inspecting representative docs for top topics.

Latent Semantic Analysis (LSA). While LDA produces clearly defined, discrete topics that documents can be assigned to, LSA works fundamentally differently (Blei et al., 2003). LSA identifies latent axes that capture the maximum variance in the data, resulting in axes that represent semantic spectra rather than distinct topics (Deerwester et al., 1990). Each component typically reflects a continuum of related topics or stances rather than a single topic (Deerwester et al., 1990). The list of all LSA axes with their highest positive and negative terms is provided in AppendixA, Table A.2. For example, one dominant axis (axis 1) reflected a moral-ideological polarization or contrast. The positive

side's most influential terms were *heuchler*, *spd*, *grüne*, *rechtstaat*, *unglaubwürdig*, and *cdu*, while the negative terms included *afd*, and *remigration*. Documents with high positive scores predominantly expressed disapproval and blame towards established political parties (see Quote 4.2.2).

> *"Rechtsstaat anywhere Grüne Heuchler Von A bis Z."*
> *"Niemals CDU CSU ihr Heuchler sPD Grüne Linke BSW alle weg ihr seid alles Heuchler."*

In contrast, documents with high negative scores focused on ideological advocacy like the AfD's Remigration agenda as seen in Quote 4.2.2.

> *Alice Weidel macht den Begriff der Remigration ganz nach dem Vorbild in Österreich Und wenn das dann Remigration heißt dann heißt das eben Remigration so Weidel auf dem AfD Parteitag.*
> *Remigration und das sofort Trump AfD.*

Another prominent axis (axis 2) combines earlier oppositions of Topic 1, instead of revealing new topics, and contrasts them with civic and institutional words. These combined negative poles are defined by *heuchler*, *remigration*, *afd*, combining moral accusations and right-wing rhetoric. The positive contrast is defined by terms such as *terroristen*, *rechtstaat*, and *waffenstillstand*. The positive pole of the topic mainly discusses short documents about terrorists or constitutional states (see Quote 4.2.2).

> *Deutschland ist immer noch eine Rechtsstaat.*
> *Wo ist Deutschland noch ein Rechtsstaat.*
> *Dieser Rechtsstaat ist wirklich am Ende.*
> *Aber nicht die Terroristen Die dürfen nicht mehr zurück.*

Axis 11 contrasted documents primarily about the AfD and left-green parties with migration policies, especially in the context of Friedrich Merz. The positive pole contains documents both pro- and anti-AfD, while the negative pole is defined by terms like *merz*, *verschärfung*, and *grenzkontrollen*. The negative-scoring documents discussed mostly strengthening border control, mentioning Merz and the CDU. Most likely, these documents referred to the proposal of CDU in the Bundestag about stricter border controls, which was highly discussed in the public[1].

Another interesting axis focused on contrasting left- and right-wing extremism. The positive pole was defined by terms like *linksextremisten*, *sicherheitsrisiko*, *cdu*, and *mitte*. The documents assigned to the positive pole of the axis were mostly about accusing people of being left extremists or concerns about security risks. On the other hand, the negative pole is defined by terms such as *rechtsextremisten*, *spd*, and *asylbewerber*. Documents with negative scores mostly accused people of being right-wing extremists.

Overall, LSA reveals ideological structures within the corpus rather than coherent topics. While axis 1 captures a somewhat meaningful opposition in the discourse, the resulting axis often overlaps and recombines terms from other axes rather than identifying new topics. This redundancy illustrates that LSA is primarily developed for information retrieval. LSA topics were highly defined by one or two terms with a high score, leading to documents being highly loaded towards these axes when the term is contained in the document. Consequently, individual axes are difficult to interpret as topics, and their boundaries remain unclear.

1. https://www.tagesschau.de/inland/innenpolitik/migration-antrag-union-100.html

Latent Dirichlet Allocation (LDA). LDA discovered several political topics that mention specific political parties such as *Die Grünen*, *CDU*, and *AFD*, but also topics about international conflicts and geopolitical issues, including the Israel-Gaza conflict, the Russia-Ukraine war, and US-related topics. The model was also able to discover economic topics, including energy politics, inflation, and taxes, but also socially related topics like welfare policy, pandemic-related topics, and public protest topics. A detailed table of all LDA topics with their representative words is provided in Appendix B, Table B.2.

Topics 67 and 69 were both related to migration and remigration. They include terms such as *remigration*, *fordert*, *afd*, or *einwanderung*, *illegale*, and *einwanderer*. Topic 69, which focuses on the right-wing slogan and AfD-aligned policy keyword *remigration*, covers a discourse about remigration, especially around AfD. Topic 67 combined discussions about migration and illegal immigration from multiple perspectives. Other topics focus on the topic of border controls in general, as well as the Bundestag proposal by CDU/CSU about strengthening the border controls after multiple attacks in late 2024 and early 2025.

Other topics that LDA identified included foreign policies, including US-related topics around the US president Donald Trump, focusing specifically on migration. International conflicts such as the Israel-Gaza conflict or the Ukraine-Russia conflict also formed topics in the model.

Notably, LDA identified a similar topic focusing on hypocrisy as LSA did. This topic was exclusively represented by documents that covered the term *heuchler* and were usually short, containing only a few words.

Other topics focused on left- and right-wing extremism, either in the context of Germany (topics 109 and 58), but also right-wing extremism in the context of Israel.

In summary, the LDA model produced interpretable and diverse topics, including ideological, security, and migration-related topics. Compared to LSA, whose topics were not easily interpretable and mostly non-coherent, LDA's topics were mostly coherent and interpretable, with minor exceptions. Furthermore, LDA's topics were often characterized by a few dominant terms that strongly defined the topic. This was most likely due to the nature of the documents, which were overall very short. Despite this limitation, LDA was able to discover political topics discussed on social media effectively.

BERTopic. Unlike LDA or LSA, which rely on word co-occurrences, BERTopic leverages transformer-based representations that capture semantic relationships. This allows BERTopic to identify similar documents even when they do not share similar words within a document. This makes it especially useful for short documents, which LSA and LDA struggle with, as seen in the previous paragraphs.

BERTopic discovered *99* distinct topics, including a topic (-1) where outliers are assigned to. A complete Table of all BERTopic topics with their representative terms is provided in Appendix C, Table C.2. BERTopic assigned 145,924 documents as outliers into topic -1, which represented roughly 48% of all documents. Although one of the advantages of BERTopic is the identification of outliers, 48% seems like a very high number of outliers in the dataset; some outliers are expected and desired due to the nature of documents in social media, containing extremely short documents and possible outliers.

BERTopic discovered a broad range of political and social themes during the German election period, ranging from international conflicts such as the Ukraine-Russia conflict,

the Syria conflict, and the Israel-Gaza conflict in the Middle East. Germany-related topics, including a topic about tax burden, migration politics, security risks, and a topic about the constitutional state, were discovered as well.

The Ukraine-Russia conflict, with a size of 11,396 documents, was found to be the most discussed topic out of all topics identified by BERTopic. The discussion was mainly about requesting a ceasefire, which is highlighted by the terms that represent the topic (*ukraine*, *russland*, *waffenstillstand*, and *putin*). The most representative documents of the topic (as shown in Quote 4.2.2) underline that the topic's terms are coherent and interpretable, and underline the discussion around the Ukraine-Russia conflict.

> *Trump sagte weiter, er hoffe auf einen baldigen Waffenstillstand zwischen Russland und der Ukraine.*
> *Chrupalla: Waffenstillstand in Ukraine sinnvoll AfD Atomkraft Russland Ukraine Waffenstillstand.*
> *Weder Russland noch die Ukraine sind derzeit zu einem Waffenstillstand bereit.*

Similar to LSA and LDA, BERTopic identified a topic around hypocrisy, represented by terms like *heuchelei*, *unglaubwürdig*, *sauerei*, and *schäbig*. Again, the representative documents of the topic confirm the coherence of the terms and interpretability.

> *Heuchler!!!!!!*
> *Alles gewollt ihr Heuchler.*
> *HEUCHLER!!!*

Interestingly, BERTopic also identified a prominent topic about sports (*gegenwehr*, *unzufriedenheit*, *fans*). The discourse around this topic mainly focused on the dissatisfaction of their football clubs and general discussions around football clubs and games.

Overall, BERTopic demonstrated its strength in extracting semantically meaningful and interpretable topics even from short, informal texts. While the high outlier rate suggested challenges inherent to social media data, the model's ability to reveal both major political debates and everyday social conversations highlighted its robustness and versatility in topic discovery.

TopicGPT (GPT-4o-mini). TopicGPT generated 39 topics after the first generation, and after refinement, 17 topics remained. The final generation, which produced subtopics from the refined topics, yielded 66 topics, including four duplicated topics, leaving 63 distinct topics. Table D.1 in Appendix D provides a complete overview of all TopicGPT topics with their representative terms and generated labels. The framework covered topics across a wide range of political, social, and economic topics. These topics included taxation, tax burden, and tax policy, religiously related topics, social issues such as extremism, violence, human rights, and migration, and war, terrorism, conflicts, and peace. The 10 most prominent topics generated by TopicGPT are Politics, Social Issues, Immigration and Migration, War, Security, Tax Burden, Terrorism, Human Rights, Remigration, and Culture.

The topic *Politics* included 78,253 documents. Terms that represented this Topic based on *C-TF-IDF* included *deutschland*, *grünen*, *demokratie*, *cdu*, *spd*, and *afd*. It grouped documents about political parties or politics in general, mostly with a negative sentiment (as shown in Quote 4.2.2).

> *Totalversagen der Politik.*
> *Hobby Politiker.*
> *Eine bürgerfeindliche Partei. HabeckKannEsNicht.*

Another highly prominent topic was *Social Issues* (topic 15). It discussed social issues in general, ranging from dissatisfaction with social systems to social imbalances (as shown in Quote 4.2.2).

> *Dort stehen auch die Sozialsysteme aufgrund fehlender finanzieller Mittel vor großen Herausforderungen.*
> *Unzufriedenheit, Spaltung und Frust bestimmen die Diskussionen in den Foren, auf den Rängen und in den sozialen Medien.*
> *Alle Bemühungen des Landkreises nach mehr Klima- und Umweltschutz würden mit den Füßen getreten und die Politik einmal mehr unglaubwürdig.*

The topic was represented by terms like *verschwörungstheorie*, *sozialsysteme*, *soziale*, *umverteilung*, *rechtsextremisten*, *ungleichheit*, and *armutszeugnis*, highlighting the focus on social issues/imbalances in society. These terms also demonstrated the coherence and interpretability of the topic terms with the representative documents, as well as with the Label generated by the `GPT-4o-mini`.

Other topics, such as *War* (topic 50), focused on international conflicts mostly related to the Israel-Gaza and Ukraine-Russia conflicts.

4.3 Cross-Model Comparison

This section focuses on comparing the findings of the qualitative evaluation presented in Section 4.2.2.

The results from Section 4.2.2 showed a clear difference between traditional and transformer-based approaches. *LSA* and *LDA* followed the expected behavior of classical models on short and noisy documents: they rely primarily on co-occurrences, which limited their ability to capture coherent and diverse topics. Especially in the dataset used in this thesis, traditional models showed that topics are mostly defined by one or two terms that highly define the topic, grouping topics mostly based on these terms, rather than on coherent and meaningful topics.

Among the traditional models, *LDA* performed substantially better than *LSA*, producing more coherent and interpretable topics that corresponded to political topics such as migration, extremism, and foreign policies. However, *LDA* often broke political themes up into multiple overlapping topics. In contrast, *LSA* primarily identified ideological axes rather than discrete topics. This behavior made *LSA* more difficult to interpret and also formed axes that were rather non-informative and noisy.

Transformer-based models, on the other hand, demonstrated clear advantages. *BERTopic* produced coherent and diverse topics that clearly correspond to political themes currently discussed in society and on social media. While *BERTopic* produced a high proportion of outliers of around 48% of all documents, which at first seemed very high, it could also be interpreted as an advantage over other topic models, since online datasets including social media, news comments, or blogs naturally contain a lot of noisy or extremely short documents that did not necessarily contribute to coherent

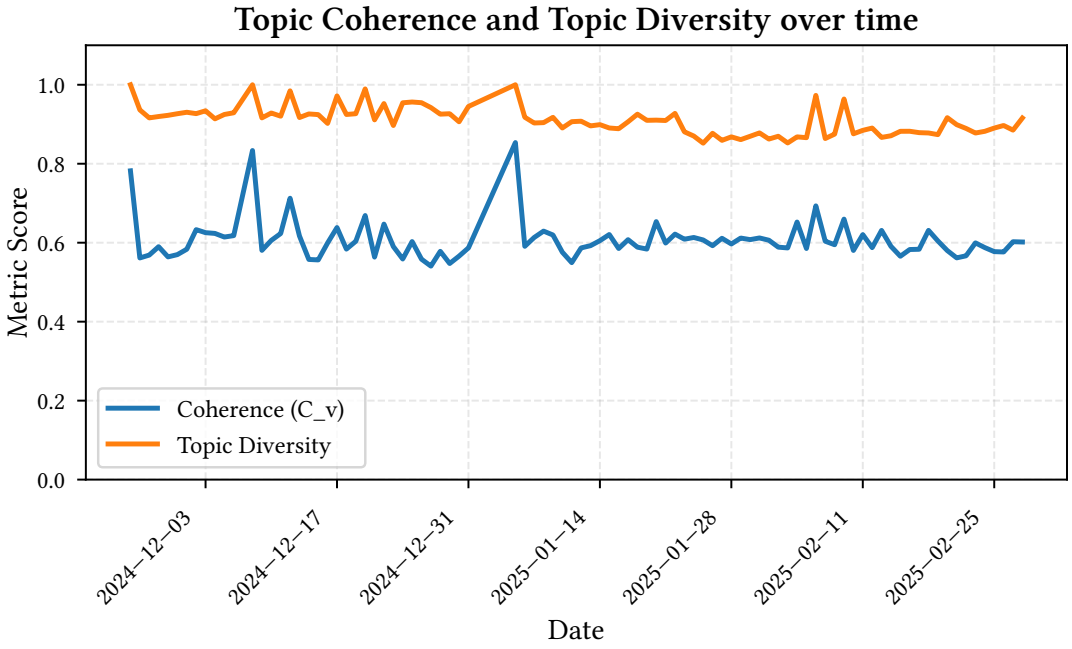

Figure 4.7: Topic coherence and diversity of BERTopic during the german election period per day.

topics. By finding these outliers, *BERTopic* filters irrelevant documents, which likely results in more coherent and diverse topic representations.

TopicGPT showed strong potential as a complementary approach, particularly for qualitative analysis, where human-readable labels are valuable. However, one huge disadvantage is the fact that *TopicGPT* relies on LLMs to generate topics and assign documents to these generated topics, resulting in intensive costs associated with this framework.

4.4 Temporal Analysis

DTM was applied using the optimized BERTopic parameters identified in Section 4.1. The corpus was segmented into daily time slices spanning from 25 November 2024 to 1 March 2025, covering the German election period. For each time segment, BERTopic independently generates topic representations using C-TF-IDF, and then uses the global C-TF-IDF and C-TF-IDF from the previous time segment $t - 1$ to smooth the topic representations. This allows BERTopic to (1) find the topic frequency at each time segment t and (2) how the representation changes over the election period. Topic trends were quantified using topic coherence (C_v) and diversity (TD) to assess temporal stability.

4.4.1 Temporal Analysis

Figure 4.7 shows the evolution of topic coherence and diversity across all 91 daily segments. Both topic coherence and diversity remained relatively constant over time, with minor fluctuations near days with low document counts, as visible at the beginning of the timeline. This indicates that the topics discovered by BERTopic are robust over

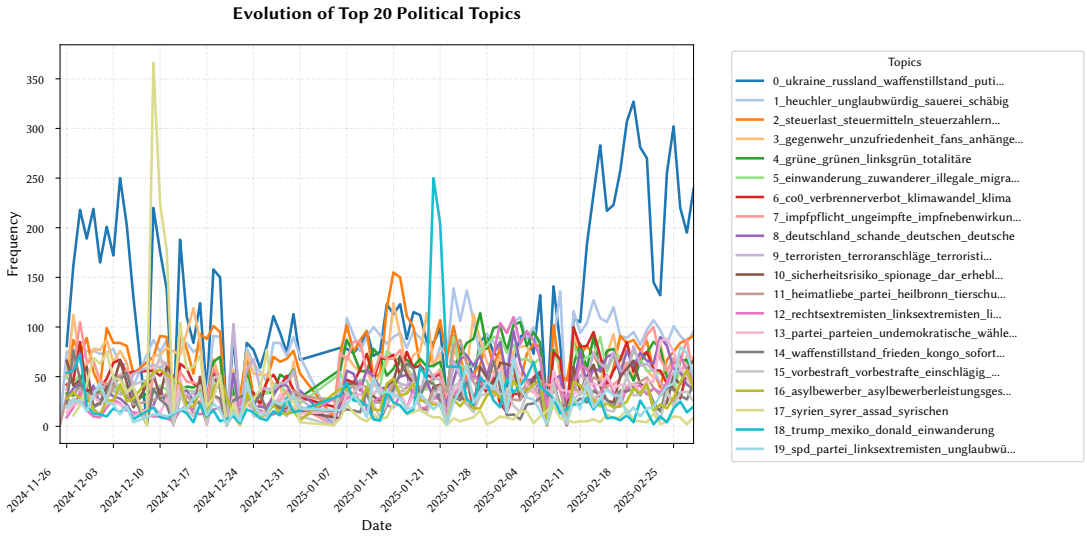

Figure 4.8: Evolution of the top 20 BERTopic topics during the german election period.

time and that short-term fluctuations in the number of documents do not degrade the quality of the model.

Two local spikes in coherence (on 7th December 2024 and 5th January 2025) correspond to a sudden drop in the number of documents for those dates (cf. Figure 3.1).

Figure 4.8 visualizes the frequency of the 20 most prominent topics across the German election period. Several topics show distinct temporal patterns reflecting real-world political and social events. The Ukraine-Russia conflict and the fall of the Assad regime in Syria are two of the most discussed topics on online platforms, as seen in Figure 4.8. The graph clearly shows that these topics were discussed frequently on social media, with a peak frequency of around 330 documents and 360 documents, respectively.

4.4.2 Event-Driven Topic Shifts

This subsection examines how major events and news developments influenced the evolution of topics identified by BERTopic. To this end, several widely discussed societal events were selected to evaluate whether BERTopic's dynamic topic modeling approach could (1) accurately capture shifts in public discourse and (2) reveal how these discourses evolved over time.

Ukraine-Russia Conflict. The graph also shows how the discourse itself shifted over time, influenced by events such as the court case regarding Marsalek.[2] On November 28, 2024, a court case was conducted regarding Marsalek, a former Wirecard board member, regarding potential espionage for Russia. At this date, the terms that influenced the topics most were *russland*, *ukraine*, *marsalek*, *waffenstillstand*, and *bulgaren*. However, on February 25, 2025, the same topic focuses on the meeting of French President Macron and US President Donald Trump, which discussed the future of the Russia-Ukraine war. Accordingly, Topic's two most representative terms changed to *ukraine*, *waffenstillstand*, *macron*, *russland*, and *putin*.

2. https://www.tagesschau.de/investigativ/br-recherche/marsalek-prozess-100.html

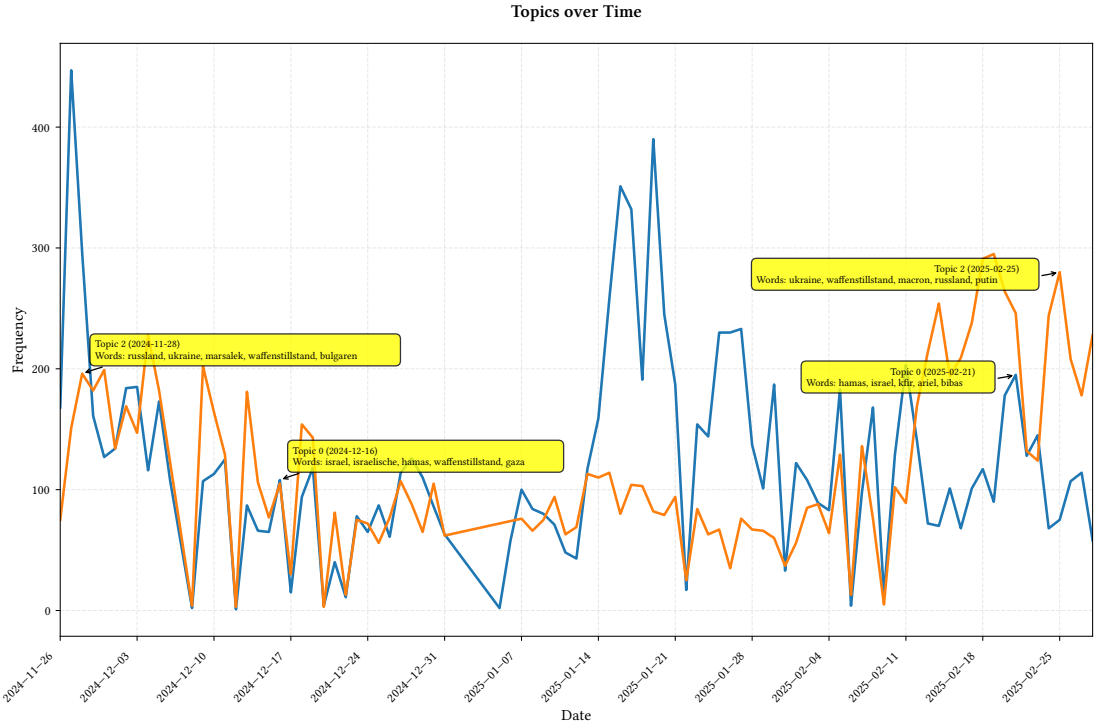

Figure 4.9: Temporal evolution of key discussion from late 2024 to early 2025. Each line represents a distinct topic, with annotate points at certain points in time listing the most representative words at that time.

Hamas–Israel Conflict. The exact figure also captures discourse shifts regarding the Israel–Gaza war. A pronounced spike occurs between February 21–26, 2025, when the IDF stated that *Ariel* (4) and *Kfir* (ca. 9–10 Monate) *Bibas* in Hamas captivity had been killed, contradicting earlier Hamas claims they died in airstrikes.[3] Concurrently, Israeli authorities reported that a body Hamas had handed over and labeled as *Shiri Bibas* did not match her DNA,[4] further intensifying public debate. Topic two reflected this window by shifting its top influence terms from a broader ceasefire frame toward more specific markers: on February 25, 2025, a higher relevance for the terms *hamas*, *israel*, *kfir*, *ariel*, and *bibas* was selected by BERTopic.

Recent Incidents in Germany. Figure 4.10 shows how several topics evolved in relation to recent incidents in Germany. Topic 46 and Topic 98 mainly described discussions about the Magdeburg incident on December 20, 2024.[5]

Months after the incident in Magdeburg another incident occurred in Aschaffenburg as reported by Der Spiegel.[6] Topics 20, 23, and 98 were related to the Aschaffenburg attack. For both events, the number of documents assigned to these topics clearly spiked on the day of the attack and in the following days.

3. https://www.timesofisrael.com/liveblog_entry/netanyahu-says-murderers-of-ariel-and-kfir-bibas-do-not-deserve-to-wal
4. https://www.juedische-allgemeine.de/israel/es-ist-nicht-shiri-bibas/
5. https://www.spiegel.de/panorama/justiz/magdeburg-news-autofahrer-faehrt-in-menschenmenge-behoerden-gehen-von-
6. https://www.spiegel.de/panorama/aschaffenburg-das-ist-ueber-den-messerangriff-mit-zwei-toten-bekannt-a-d0ed0363-0

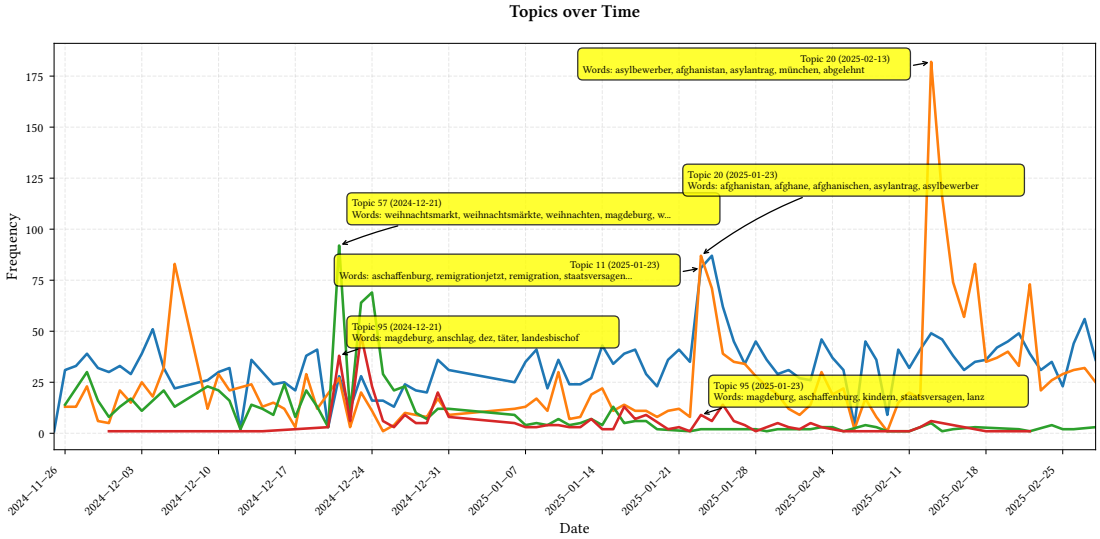

Figure 4.10: Evolution of topics regarding attacks in Magdeburg and Aschaffenburg.

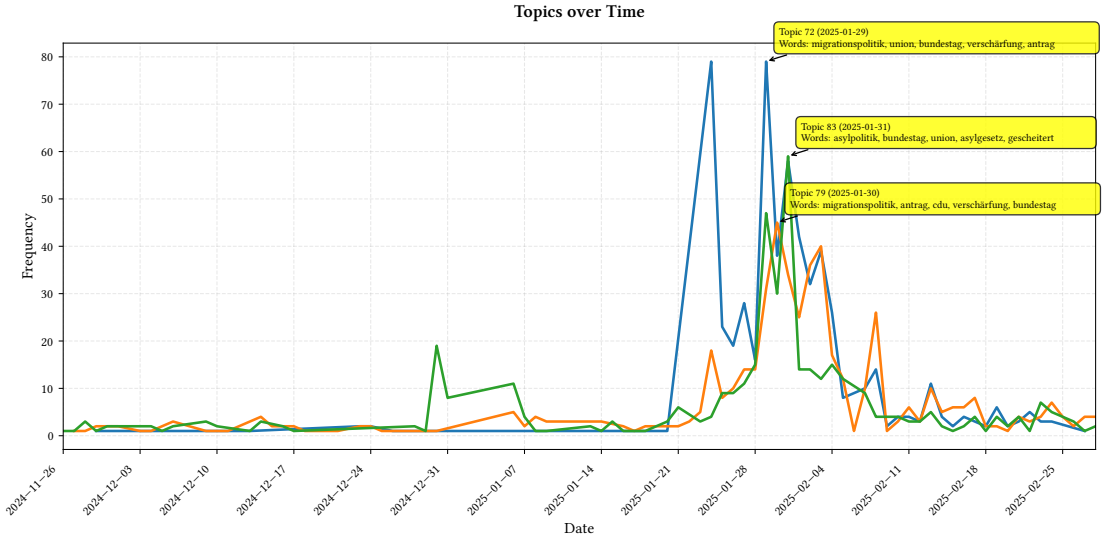

Figure 4.11: Evolution of Topics discussing the CDU proposal for strengthening the border controls and migration politics.

CDU Bundestag proposal regarding migration politics. Following these incidents in Magdeburg and Aschaffenburg, the politicians responded by strengthening the migration policies. The German party CDU proposed a five-point plan for strengthening the migration policies, including border controls, which had to be voted on in the Bundestag [7]. The proposal was highly discussed in the public and by the news, both positively and negatively. The public discussion was also captured by BERTopic, showing a high frequency increase in related topics (as shown in 4.10). The topic terms shifted from general terms to more incident-related terms like, clearly showing how the discourse focused on these incidents.

7. https://www.tagesschau.de/inland/innenpolitik/migration-antrag-union-100.html

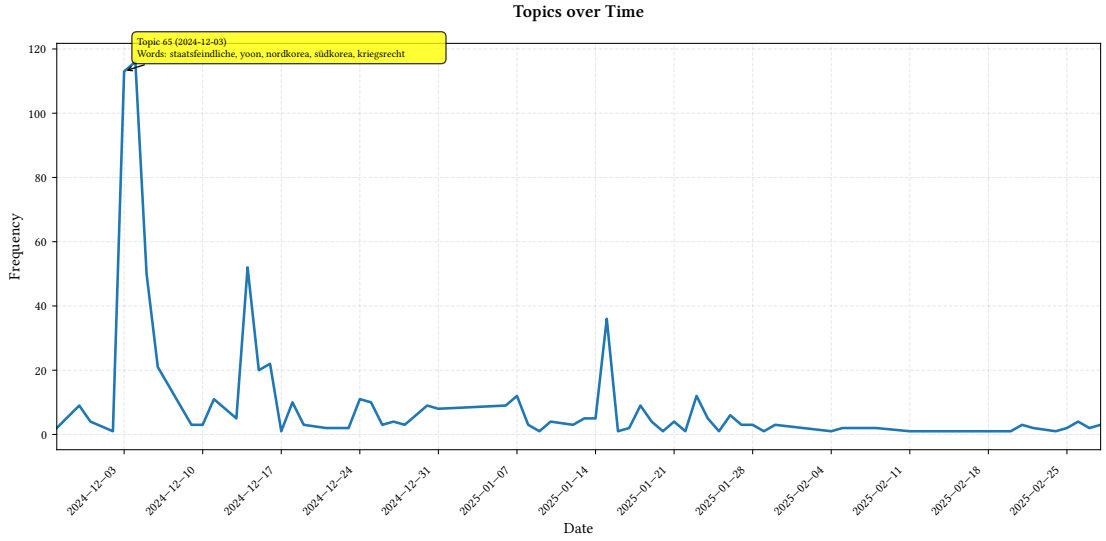

Figure 4.12: Evolution of Topic concerning the declaration of martial law by President Yoon Suk-yeol.

South Korea: declaration of martial law. Another noticeable event captured by BERTopic was the declaration of martial law by the South Korean President Yoon Suk-yeol on December 3, 2024 [8]. In an address, Yoon accused the opposition Democratic Party of pro-North Korean activities and anti-state behavior. He then declared martial law. BERTopic identified a distinct topic related to this incident and showed a spike in the topic frequency containing terms such as *staatsfeindliche*, *yoon*, *nordkorea*, *südkorea*, and *kriegsrecht*. This reflected the public attention and media coverage surrounding this event and illustrated BERTopic's ability to detect sudden peaks in topic frequency.

4.4.3 Summary

Overall, the analysis showed that BERTopic was able to reflect real-world events through noticeable changes in topic frequency and representative terms. The identified spikes around major incidents and political debates show that BERTopic's dynamic topic modeling technique could capture short-term shifts in public discourse. At the same time, the results also show that topic boundaries can overlap when several related discussions occur in parallel, which should be considered when interpreting the model output.

4.5 Chapter Summary

This chapter presents and evaluates the results of the topic modeling experiments. The optimization results show that BERTopic achieves the highest balance between topic coherence and diversity, followed by TopicGPT, while LDA and LSA perform less effectively on short social media texts. The application of BERTopic in temporal analysis shows that the model can capture event-driven changes in political discourse over time. Overall, the results indicate that transformer-based approaches are more

8. https://www.tagesschau.de/ausland/asien/suedkorea-praesident-kriegsrecht-100.html

suitable for analyzing short text commonly found on online platforms, addressing both research questions formulated in Chapter 3.

However, as mentioned before, a crucial step in all topic models discussed in this thesis requires a human-in-the-loop to evaluate topic models. While TopicGPT demonstrated potential in topic modeling, it is inherently too expensive to perform topic modeling on large corpora of text. A promising future research direction would be to assess whether and how LLMs can be used to automatically interpret and label topics discovered by LSA, LDA, and BERTopic and investigate the potential of LLMs to act as a human-in-the-loop replacement for topic analysis.

5 Conclusion and Future Work

5.1 Conclusion

This thesis compared the application of traditional and transformer-based topic modeling methods for analyzing political discourse on online platforms. The thesis aimed to answer two research questions: (1) How do traditional machine learning approaches compare to transformer-based approaches for topic modeling on politically related data on online platforms, and (2) How do political topics evolve over time across online platforms, and what insights can be drawn about the dynamics of public discourse?

The thesis compared traditional topic models such as LSA and LDA with transformer-based approaches like BERTopic and TopicGPT. All models were trained on social media data collected by the IMWF Institute in the research project *KIFürDemokratie*. A two-stage preprocessing pipeline was responsible for first reducing artifacts, such as HTML tags, user mentions, and personal information, and the second step included model-specific preprocessing for each model. Each model's optimal hyperparameters were found using Bayesian optimization provided by *OCTIS* and evaluated based on the product of topic coherence and diversity.

Both LSA and LDA demonstrated inferior results compared to transformer-based approaches like BERTopic and TopicGPT. The documents were mostly grouped based on single words due to the sparsity problem of the data. Therefore, traditional approaches showed less interpretable and redundant topics, making the interpretability and coherence less ideal. BERTopic achieved the highest score in both topic coherence and diversity, outperforming traditional methods such as LSA and LDA. The qualitative evaluation confirmed these findings, making the topics discovered by BERTopic more coherent and interpretable than traditional models. The topics also showed real-world political themes commonly discussed in the public. TopicGPT demonstrated potential for generating interpretable topics but suffers from high associated costs due to the use of LLMs.

While the `OCTIS` framework was helpful for systematic and reproducible hyperparameter optimization, this may be more suited for models that provide a clearer quantitative function compared to using only topic coherence and diversity. In use

cases such as political discourse analysis, quantitative optimization alone does not necessarily align with interpretive or journalistic goals. In such applications, a more qualitative and manual model selection process, focusing on the relevance, coherence, and narrative fit of the discovered topics, might yield results that better support domain experts in their analysis.

Additionally, BERTopic was applied for the temporal analysis of topics within the dataset by using the dynamic topic modeling functionality provided by BERTopic. In the temporal analysis, BERTopic was applied to analyze how political discourse evolves over time. The analysis showed how topic frequency shifts over time, with newly emerging topics that were mostly discussed on that date. Furthermore, it shows how topics themselves evolve through changes in the most representative words associated with topics. These temporal patterns showed how the political discourse on social media shifts and was influenced by real-world politics and news.

However, despite this reflection, the results still demonstrate that traditional models such as LDA and LSA struggle with extremely short and noisy data like social media comments, due to the sparsity of such text. BERTopic, by contrast, did not suffer from this limitation because it leverages contextual embeddings that capture semantic meaning even with short documents. Therefore, transformer-based approaches remain superior for short-text topic modeling, both in terms of topic quality and temporal stability.

5.2 Future Work

Future research and development could extend this work in several directions. From a methodological perspective, transformer-based approaches such as BERTopic could benefit from fine-tuned embedding models specifically trained on political or social media-related data. Such fine-tuning may improve the model's ability to distinguish political topics and differences in stance, sentiment, or framing, thereby increasing topic coherence and interpretability. For traditional models like LDA and LSA, future work could focus on mitigating the sparsity problem inherent in short social media comments. One promising approach would be to enrich each comment with additional contextual information, such as the text of the original post or article it responds to. This could provide more complete document representations and potentially improve the models' capability to identify coherent and diverse topics.

From a practical perspective, this thesis was conducted as part of a collaborative project with journalists who aim to continuously analyze online discourse. Building on this work's findings, a prototype pipeline has been developed that automatically trains new topic models daily using the most recent data. Each daily model is merged with the topic model from the previous day by comparing topic embeddings and identifying semantically similar topics. If the similarity between two topics exceeds a predefined threshold, the topics are merged; otherwise, new topics are added dynamically. This enables continuous topic discovery and adaptation over time, allowing journalists to monitor emerging discussions that earlier models did not capture. Due to the domain's sensitivity, the application has not been published publicly. However, researchers, policymakers, or journalists interested in testing the application can request access by contacting Till N. Schaland or the Hub of Computing and Data Science of University of Hamburg. Future work could further evaluate and optimize this merging strategy, for example, by experimenting with different similarity measures, dynamic thresholds, or

hierarchical topic alignment techniques to improve long-term consistency in dynamic topic modeling.

Appendices

A
LSA.

Table A.2: All topics identified by LSA with topic IDs and the highest scored positive and negative words.

Topic ID	Positive Top Words	Negative Top Words
0	afd, remigration, spd, deutschland, grüne, cdu, heuchler, rechtsstaat, merz, grünen	
1	heuchler, spd, grüne, rechtsstaat, terroristen, unglaubwürdig, grünen, cdu	remigration, afd
2	terroristen, rechtsstaat, deutschland, abgeschoben, waffenstillstand, schande, israel	heuchler, remigration, afd
3	unglaubwürdig, spd, afd, cdu	terroristen, waffenstillstand, abgeschoben, heuchler, israel, remigration
4	unglaubwürdig, terroristen, waffenstillstand, israel, hamas	rechtsstaat, deutschland, abgeschoben, schande, demokratie
5	rechtsstaat, terroristen, demokratie, grüne, waffenstillstand, israel	abgeschoben, unglaubwürdig, straftäter, deutschland
6	afd, verschärfung, bundestag, migrationspolitik, merz, union	unglaubwürdig, rechtsstaat, remigration, abgeschoben
7	rechtsstaat, abgeschoben, demokratie, afd, verschärfung, migrationspolitik	deutschland, schande, unglaubwürdig, partei
8	terroristen, grüne, spd	waffenstillstand, ukraine, unglaubwürdig, trump, rechtsstaat, heuchler, russland

Continued on next page

Topic ID	Positive Top Words	Negative Top Words
9	spd, grüne, waffenstillstand	terroristen, afd, rechtsstaat, bundestag, unglaubwürdig, verschärfung, deutschland
10	waffenstillstand, afd, schande, abgeschoben, israel	einwanderung, illegale, grenzkontrollen, migration, land
11	afd, linksgrün, einwanderung, grüne	merz, verschärfung, spd, grenzkontrollen, friedrich, remigration
12	spd, grünen, partei, linksgrün, prozent	grüne, einwanderung, verschärfung, deutschland, migrationspolitik
13	einwanderung, spd, illegale, afd, cdu	linksgrün, grüne, unzufriedenheit, rechtsextremisten, asylbewerber
14	partei, unzufriedenheit, prozent, grüne, rechtsextremisten	linksgrün, cdu, einwanderung, deutschland, merz
15	partei, schande, linksgrün, verschärfung, bundestag, einwanderung	afd, deutschland, prozent, grenzkontrollen
16	straftaten, grenzkontrollen, terroristische, gwb, partei, terroristischen, aktivitäten, zusammenhang, ss, israel	
17	verschärfung, straftaten, terroristische, gwb, spd, terroristischen	grenzkontrollen, merz, partei, prozent
18	prozent, unzufriedenheit, anhängern, verschärfung, linksgrün	rechtsextremisten, grenzkontrollen, linksextremisten, land, afd
19	grenzkontrollen, verschärfung, bundestag, spd	merz, friedrich, straftaten, sicherheitspolitik, trump, rechtsextremisten
20	cdu, israel, einwanderung, merz, armutszeugnis	trump, grenzkontrollen, donald, ukraine, spd
21	cdu, unzufriedenheit, csu	linksextremisten, rechtsextremisten, prozent, grünen, merz, einwanderung, mitte
22	rechtsextremisten, linksextremisten, cdu, mitte, trump, csu	schäbig, afd, einfach, merz
23	schäbig, einfach, rechtsextremisten, trump, prozent, armutszeugnis	unzufriedenheit, sicherheitspolitik, sicherheitsrisiko, grünen
24	impfpflicht, sicherheitsrisiko, gestimmt, fpö, bundestag	unzufriedenheit, armutszeugnis, land, politik, linksextremisten
25	armutszeugnis, land, asylbewerber, prozent, partei, sicherheitsrisiko	unzufriedenheit, schäbig, linksextremisten, rechtsextremisten
26	armutszeugnis, trump, donald, impfpflicht, israel	schäbig, ukraine, linksextremisten, sicherheitspolitik, asylbewerber
27	armutszeugnis, linksextremisten, impfpflicht, ukraine, grünen	sicherheitsrisiko, fpö, rechtsextremisten, övp, israel
28	armutszeugnis, rechtsextremisten, sicherheitsrisiko, prozent, ukraine, cdu, grenzkontrollen	asylbewerber, trump, impfpflicht

Continued on next page

Topic ID	Positive Top Words	Negative Top Words
29	linksextremisten, sicherheitsrisiko, cdu, sauerei, mitte, trump, land	rechtsextremisten, spd, asylbewerber
30	sicherheitspolitik, grünen, umverteilung, israel, ungleichheit, außen	unzufriedenheit, land, sicherheitsrisiko, ukraine
31	land, grünen, umverteilung, rechtsextremisten, schande, sauerei	armutszeugnis, partei, linksextremisten, sicherheitspolitik
32	sicherheitspolitik, land, cdu, schäbig, demokratie, csu	sauerei, vereinigung, kriminelle, merz
33	sauerei, csu, cdu	vereinigung, land, kriminelle, linke, armut, tod, bringen
34	umverteilung, ungleichheit, merz, unten	sicherheitspolitik, sauerei, land, bundestag, außen, grünen
35	grünen, anhängern, verantwortungslos	sauerei, spd, land, sicherheitspolitik, demokratie, ungleichheit, schande
36	umverteilung, asylbewerber, unzufriedenheit, bundestag, trump, sicherheitspolitik	spionage, verschwörungstheorie, aschaffenburg, unterdrückung
37	aschaffenburg, verantwortungslos, politik, magdeburg, mitte, sauerei	verschwörungstheorie, armutszeugnis, unterdrückung, demokratie
38	bundestag, unwürdig, schande, verantwortungslos, spionage, politik, verschwörungstheorie	verschärfung, partei, anhängern
39	verantwortungslos, unwürdig, politik, mitte	aschaffenburg, anhängern, umverteilung, schande, armutszeugnis, magdeburg
40	verantwortungslos, spionage, anhängern, unterdrückung, totalitäre	verschwörungstheorie, aschaffenburg, politik, waffenstillstand, grünen
41	ungleichheit, demokratie, unwürdig, unterdrückung, bundestag, linke, soziale	verschwörungstheorie, umverteilung, spionage
42	anhängern, ungleichheit	illegale, totalitäre, verbrennerverbot, verschwörung, grünen, einwanderer, armutszeugnis, unterdrückung
43	umverteilung, mitte, verbrennerverbot, linke, unwürdig, demokratie	ungleichheit, soziale, linksextremisten, schande
44	spionage, ungleichheit, russland, volksverräter	verheerende, verschwörung, unterdrückung, zwangsgebühren, einfach, ukraine
45	usa, israel, unwürdig, verbrennerverbot, anhängern	fpö, parteien, rechtsextremer, verschwörung, linke
46	verschwörung, verantwortungslos, verbrennerverbot, anhängern	verheerende, unterdrückung, totalitäre, grünen, politik, bundestag

Continued on next page

Topic ID	Positive Top Words	Negative Top Words
47	verschwörung, mitte, unwürdig	regierung, politik, bundestag, fpö, ukraine, unterdrückung, migration
48	illegale, linke, demokratie, einwanderer, verantwortungslos, usa, verheerende	verschwörung, regierung, unwürdig
49	opera, news, sehen, linke, verschwörung, illegale	cdulinkebswspdgrünearmuttod, russland, vereinigung, politik
50	verbrennerverbot, cdulinkebswspdgrünearmuttod, kriminelle, unterdrückung, totalitäre, mitte	linke, israel, russland, terroristische
51	usa, regierung, terroristische, verschwörung, illegale	verbrennerverbot, opera, israel, gwb, sehen
52	terroristische, gwb, tätigkeit, entrichtung, ss	straftaten, terroristischen, verbrennerverbot, zusammenhang, aktivitäten
53	terroristische, zwangsgebühren, verantwortungslos, schande, verbrennerverbot, organisation	gwb, verschwörung, ssss, israel
54	verbrennerverbot, schande	nutzungsbedingungen, threadanzeige, einblenden, russland, gwb, ssss, entfernen, beitrag
55	nutzungsbedingungen, threadanzeige, einblenden	russland, entfernen, beitrag, geiseln, anzeigeanzeige, asylbewerbernachrichtenag, gwb
56	entfernen, beitrag, anzeigeanzeige, asylbewerbernachrichtenag, kürzung, hisbollah, sozialleistungen	geiseln, ukraine, hamas

#	Time	Median	Mean	Std. Dev.	# topics	Topic Diversity	Coherence
1	562.751	0.052	0.053	0.002	172	0.172	0.304
2	432.797	0.073	0.070	0.005	97	0.202	0.354
3	873.311	0.042	0.043	0.001	280	0.151	0.284
4	492.270	0.059	0.059	0.000	140	0.186	0.317
5	566.337	0.050	0.050	0.000	182	0.164	0.303
6	501.850	0.054	0.055	0.002	160	0.180	0.311
7	415.321	0.059	0.059	0.001	129	0.186	0.320
8	289.865	0.080	0.082	0.003	65	0.222	0.367
9	414.874	0.066	0.067	0.003	112	0.199	0.330
10	311.148	0.076	0.076	0.002	71	0.214	0.355
11	443.886	0.062	0.062	0.003	120	0.194	0.320
12	428.172	0.064	0.066	0.003	116	0.196	0.328
13	369.652	0.073	0.070	0.005	86	0.210	0.347
14	446.741	0.065	0.066	0.004	104	0.192	0.339
15	458.087	0.065	0.066	0.004	99	0.195	0.334
16	401.380	0.068	0.067	0.002	98	0.197	0.342
17	409.987	0.067	0.069	0.003	96	0.199	0.343
18	361.173	0.078	0.078	0.003	74	0.220	0.351
19	435.058	0.070	0.073	0.005	92	0.207	0.342
20	597.403	0.050	0.051	0.001	174	0.166	0.303
21	439.642	0.072	0.074	0.002	85	0.214	0.345
22	339.706	0.087	0.088	0.004	57	0.232	0.377
23	690.625	0.047	0.047	0.000	230	0.160	0.297
24	436.413	0.066	0.065	0.002	118	0.197	0.337
25	491.585	0.082	0.082	0.002	63	0.219	0.375

Table A.1: Bayesian optimization results for LSI, showing all iterations with evaluated hyperparameter, objective function statistics (mean, median, standard deviation), and corresponding topic coherence and diversity scores.

B
LDA.

Table B.2: All topics identified by LDA with topic IDs and the representative words.

Topic ID	Top Words
0	weder, scharia, usw, daten, tiktok, zb, herrscht, falle, kürzlich, reichen
1	frage, braucht, unseren, undemokratische, deutscher, wählt, gründe, bekämpfung, demonstrationen, drittel
2	kinder, finden, wort, neuer, müsste, terror, rente, toten, bundesrepublik, gleiche
3	parteien, politischen, genau, mitte, rechts, politik, links, regeln, heutigen, unzufriedenheit
4	fast, völlig, stunden, feigheit, bevor, zurückweisungen, fand, stoppen, beginnt, sitzen
5	steuerlast, montag, volk, senken, truppen, einkommen, inhalte, assad, rentner, vorher
6	lassen, sicherheit, reden, donnerstag, suchen, investitionen, essen, bessere, staatsbürgerschaft, frankfurt
7	liegen, unten, nationale, dortmund, rein, begonnen, beobachtet, machten, nutzung, künftige
8	gewalttäter, bevölkerung, kritik, angesichts, mutmaßliche, bereit, namen, psychisch, schweizer, unzufriedenheit
9	erklärt, kriminelle, vorwürfe, innenministerin, acab, entsprechend, änderung, mitgliedern, entspricht, erfüllen
10	verschwörung, impfpflicht, geschichte, jahres, form, kontrolle, islamistische, erzählt, thomas, haltung
11	brandmauer, informationen, infrastruktur, waffenruhe, zerstört, klimawandel, führung, bedroht, gesamten, erlebt
12	sprechen, schließlich, halt, plan, aussagen, zeigte, geplant, retten, trifft, reihe
13	prozent, spd, grünen, bereich, asylpolitik, beitrag, einzige, integration, bedrohung, zuwanderung
14	bundestagswahl, wahlkampf, situation, forderung, koalition, angekündigt, parlament, schaden, erreichen, verschärfung

Continued on next page

Topic ID	Top Words
15	thema, selbstverteidigung, diskussion, gefängnis, bezug, nötig, bestätigt, propaganda, sachen, russlands
16	knapp, international, flucht, voll, bereichen, auschwitz, co, gäste, geboren, weist
17	zahl, auswirkungen, haft, ergebnis, taten, öffentlich, aktiv, gesundheit, unabhängig, journalisten
18	unterdrückung, seite, bundeskanzler, bezahlkarte, baerbock, menschenrechte, vertreter, kämpft, christoph, verantwortlichen
19	zeigt, letzten, jährige, blick, könne, unzufriedenheit, hoffe, möglichkeiten, bewegung, lösen
20	herausforderungen, sexualstraftaten, schön, fühlen, vielmehr, massiven, pandemie, vertreibung, amerikanische, kontext
21	liegt, linken, junge, ort, zahlreiche, dringend, vergessen, besteht, details, risiko
22	trump, zuwanderer, donald, präsident, usa, haus, beenden, hoffnung, tausende, chance
23	per, punkt, abend, gesamte, entwickelt, niedersachsen, verbrecher, meinungen, matthias, vorstellen
24	fall, lage, auslöschung, gegeben, forderte, schützen, handeln, wirtschaftliche, wolle, moskau
25	debatte, leistungen, raum, gespräch, betroffen, sicherheitsbehörden, post, gemeinden, verfassung, müller
26	regierung, scheiterns, rolle, sprach, christian, spielen, lindner, tragen, linksextremistischen, erklärung
27	februar, deutlich, unzufriedenheit, hieß, deutliche, gespräche, zerstörung, vorschläge, irak, bisherigen
28	halten, schweiz, berichtet, trumps, team, person, erwarten, new, ministerpräsident, starken
29	frieden, afghanistan, islamistischen, waffenstillstand, unkontrollierte, dauerhaften, beamte, afp, festnahme, interessiert
30	musk, nächsten, elon, armut, glaube, bund, gestern, leitkultur, sabotage, geplante
31	aschaffenburg, männer, anträge, tödlichen, bedingungen, klingt, messerattacke, gehalten, versprochen, äußert
32	terroranschläge, hintergrund, liebe, frankreich, schließen, ziele, reaktion, innere, ums, deutschland
33	sagen, krieg, führte, teile, arbeitet, eingeführt, gegenteil, falls, job, behauptet
34	berlin, brandanschlag, verschwörungstheorie, fremdenfeindlichkeit, sorgen, rassismus, rechtsextreme, martin, diskutiert, preis
35	verheerende, führen, angeles, verlassen, kämpfen, feuer, brände, abschieben, übernehmen, voller
36	entscheidung, anschläge, geprägt, verdacht, hervor, nix, gescheitert, absolute, zusätzliche, greift
37	abgeschoben, hohen, deutschland, wagenknecht, massive, teilte, sahra, gesellschaftlichen, szene, reich
38	betont, verhalten, botschaft, fazit, veränderung, prüfung, anna, kanton, abschluss, zunehmenden

Continued on next page

Topic ID	Top Words
39	quelle, maßnahmen, straftaten, vorbestraften, märz, einschlägig, anzeigen, schwere, körperverletzung, bekannten
40	angaben, findet, länder, inkompetenten, steuern, globalen, siehe, gezogen, amtszeit, globale
41	klar, bringen, obwohl, bsw, verloren, vereinigung, thüringen, inneren, beteiligung, richtigen
42	sicherheitspolitik, außen, lässt, meinung, sozialen, spielt, archivbild, teilweise, nähe, beste
43	politische, sozialsysteme, erhalten, bislang, wahlen, täglich, jugendliche, ausbeutung, deutschland, rumänien
44	gruppe, genommen, erreicht, äußerte, israelische, berichten, gegner, linie, betroffenen, echte
45	partei, grüne, linke, wahl, spd, handelt, deutschlands, dennoch, deutschland, auto
46	grund, asylgesuche, kritisierte, demokraten, waffenexporte, demonstrieren, besuch, jugendlichen, frank, besondere
47	hält, betonte, dr, amt, mitglied, lügen, militärische, missbraucht, formen, real
48	jährigen, fordern, klare, kommenden, gefunden, verschiedenen, mithilfe, terrorismus, höheren, saison
49	euro, millionen, migranten, widerstand, beendet, eindruck, dollar, ursache, häufiger, unterstützer
50	flüchtlingspolitik, totalitäre, mögliche, straße, folge, längst, polizisten, familien, wahrheit, guten
51	trotz, magdeburg, kampf, fest, weltweit, stellte, klima, ideologie, solingen, titel
52	januar, dürfte, monat, antifa, finanzierung, mörder, terroristen, instagram, begrenzung, flughafen
53	frau, weidel, video, ländern, alice, justiz, vertrauen, gesetz, ändern, öffentlichen
54	gilt, kosten, faeser, insgesamt, offenbar, beginn, konsequenzen, plant, staatskosten, kontrollen
55	spionage, verurteilt, ausland, zeitung, drohnen, spur, anlass, zählt, aufgenommen, haftstrafe
56	ungleichheit, verantwortungslos, soziale, freiheit, flüchtlinge, bringt, schaffen, gerechtigkeit, trainer, april
57	wirtschaft, weiterhin, derzeit, bildung, schuld, stimme, peter, leisten, tv, jahrzehnten
58	linksextremisten, gazastreifen, angriff, region, sachsen, bewohner, wohlstand, wachstum, oktober, festgehalten
59	waffenstillstand, ukraine, russland, putin, möglichen, krieg, bürgergeld, usa, schreibt, sanktionen
60	asylbewerber, geben, abschiebungen, zusammenarbeit, aktuelle, abschiebung, bitte, internationalen, weise, unternehmensgewinne
61	leute, nehmen, tod, pläne, gefordert, kauf, entscheiden, deutsch, freund, last
62	geld, art, stärken, vorbestrafter, zusammenhang, bezahlen, solle, ss, unterwegs, ansicht
63	welt, unwürdig, spiel, merkel, michael, schule, härte, angela, projekte, deutschland
64	eu, rahmen, heimat, hinweis, entschieden, wahrscheinlich, ukrainekrieg, außerhalb, kirche, sorgte

Continued on next page

Topic ID	Top Words
65	gegenwehr, frauen, gewalt, opfer, täter, the, einsatz, of, gestellt, film
66	hinterhältig, jungen, läuft, begangen, einreisen, hinsichtlich, verstanden, einstellung, drängen, entwurf
67	einwanderung, illegale, themen, einwanderer, habeck, wähler, grenze, illegaler, robert, wien
68	lesen, pro, nato, podcast, kriminalität, möglicherweise, entwicklungen, beamten, erkennen, wichtigsten
69	remigration, fordert, afd, zustimmung, wächst, höhere, offensichtlich, interessen, aufgabe, verbindung
70	sogar, lösung, freie, mexiko, volksverräter, zölle, umwelt, kanada, umsetzen, vorerst
71	ebenfalls, gesetze, rest, wolfgang, verbessern, wider, gewonnen, unzufriedenheit, schmidt, zufrieden
72	terrorisiert, nutzen, fokus, bürokratie, gezeigt, kind, verfolgt, niemals, bilder, faschisten
73	merz, cdu, fdp, friedrich, zwangsgebühren, brauchen, afd, bedeutet, cducsu, nächste
74	sieht, politiker, probleme, sicher, bundesregierung, glaubt, anstatt, hielt, gesetzentwurf, eh
75	mord, rücken, aufmerksam, verstehe, moderne, beleuchtet, datenschutz, show, dorf, kurzen
76	heuchler, eigentlich, zuvor, bundespolizei, passiert, woche, gerne, abgelehnt, letzte, arbeitsmarkt
77	scheint, gekommen, zumindest, starke, denke, solange, via, steigen, ausgesetzt, schneller
78	dpa, wichtig, gemeinsam, asyl, masseneinwanderung, sorgt, entwicklung, energie, demokratische, schulen
79	grenzkontrollen, grenzen, personen, finde, unterstützen, deutschland, dauerhafte, deutschen, tut, bundeswehr
80	bleibt, china, größte, stelle, sorge, inflation, wahlprogramm, fallen, eingesetzt, anklage
81	deportation, biden, bezeichnete, joe, jüdischen, süden, veränderungen, meldungen, konfrontiert, angeblichen
82	israel, terroristen, hamas, waffenstillstand, geiseln, gaza, monaten, hand, bekommt, freilassung
83	artikel, sehen, setzt, news, rede, direkt, begriff, erstmals, sehe, gesellschaftliche
84	afd, bundestag, verschärfung, union, migrationspolitik, stimmen, antrag, mittwoch, hilfe, deutschen
85	unglaubwürdig, sauerei, gesellschaft, sofort, min, setzen, tatsächlich, bayern, staatsversagen, richter
86	sicherheitsrisiko, mehrere, eltern, verantwortlich, kamen, ehemaligen, verhaftet, interessant, dar, gefährliche
87	terroristische, spricht, staaten, bezeichnet, organisation, eingestuft, weiterer, reagiert, gehe, aktivitäten
88	straftäter, weiß, ukrainische, gefährder, ausreisepflichtige, brauche, entlassen, töten, menge, freiwillig
89	vorbestraft, stellen, demokratischen, forderungen, sogenannten, schritt, ziehen, verfassungsbruch, verfassungswidrigen, ermordet

Continued on next page

Topic ID	Top Words
90	zeigen, zahlen, gericht, somit, alter, symbolbild, abschaffung, zweifel, alexander, erfahrungen
91	november, dienstag, rückkehr, trägt, norden, falsche, opfern, bedenken, kämpfe, geschlossen
92	schande, könnten, eher, asylantrag, stark, markus, idee, moment, flüchtlingen, meist
93	erklärte, armee, israelischen, terroristen, experten, landkreis, syrischen, aktion, spö, verbot
94	bekommen, europäische, illegalen, hamburg, mutmaßlichen, unterstützung, anhänger, spiegel, verein, wirtschaftspolitik
95	anschlag, absolut, verübt, ki, eindeutig, grenzmauer, verteilung, ermordung, kalifornien, dänemark
96	mehrheit, müsse, sonntag, gebe, alternative, fehler, worte, vater, schafft, deutschland
97	steuerzahlern, csu, heimatliebe, paar, platz, verfassungsschutz, konsequent, extrem, verdachtsfall, jährlich
98	angst, waffen, is, angeblich, sucht, wiener, gewaltausbrüchen, fragt, amtseinführung, kreuz
99	europa, europäischen, uspräsident, washington, link, kennen, ausländische, fachkräfte, usa, hauptstadt
100	beispielsweise, künftig, vergleich, fällen, altparteien, wirkt, auftrag, alleine, erleben, ddr
101	folgen, mehrfach, zuletzt, verheerende, mutter, gesehen, homophob, massiv, beiträge, daniel
102	hohe, hysterisch, schrieb, unterstützt, sicht, inzwischen, internationale, gelten, nennt, häufig
103	heißt, steuermitteln, jan, versucht, druck, finanziert, vorgehen, falsch, angriffen, umfrage
104	anzeige, weiterlesen, erwartet, nacht, corona, amerikanischen, tief, total, traurig, schließung
105	münchen, freitag, opposition, kultur, antwort, protest, westen, heraus, fakten, punkte
106	foto, minuten, aufgrund, samstag, veröffentlicht, gründen, überfall, wohnraum, unzufriedenheit, versuche
107	all, angriffe, skrupelloser, müssten, streit, erinnern, zerstören, ausgerechnet, änderungen, erhebliche
108	österreich, fpö, asylbewerbern, aktuell, övp, verhindern, kickl, nannte, polen, länger
109	rechtsextremisten, wählen, unserem, gebracht, umgang, land, demnach, herbert, deutschland, sitzt
110	polizei, gefährder, medien, behörden, abstimmung, totalversagen, festgenommen, laut, mindestens, vorgeworfen
111	notwendig, münchner, vance, folgt, respekt, sicherheitskonferenz, abzuschieben, innen, baby, toleranz
112	landes, gruppen, monate, soldaten, kräfte, erinnert, zugang, dinge, hunderte, institutionen
113	schutz, gleichzeitig, innerhalb, meisten, jähriger, schwer, urteil, hass, verletzt, bietet
114	rechtsextremer, israels, vorbestrafte, russischen, interview, ausländer, iran, mitarbeiter, jüngsten, darstellen
115	syrien, stehen, dezember, zukunft, russische, hisbollah, libanon, türkei, militär, europas

Continued on next page

Topic ID	Top Words
116	führt, massenmord, juden, verfahren, nazis, mannheim, minderheiten, bilden, befreien, partie
117	unverschämtheit, arbeit, wirtschaftsministeriums, sowohl, zufolge, meinungsfreiheit, möglichkeit, de, stunde, studie
118	rechtsstaat, demokratie, leider, politisch, zeiten, ermittlungen, raus, laufen, sieg, erfahren
119	kritisiert, nimmt, antisemitismus, andreas, programm, wert, genannt, scharf, fc, legen
120	linksgrün, bild, schäbig, ungeimpfte, staatsanwaltschaft, italien, messer, alt, waldbrände, kümmern
121	schnell, gehört, einfluss, to, fördern, extremisten, ngos, fehlende, fürchten, gelernt
122	zudem, bleiben, stellt, gefahr, dar, strafe, ermittler, staates, steigende, erhöht
123	verhandlungen, frei, hamasterroristen, kraft, bzw, vorfall, waffenstillstand, kaufen, treten, händen
124	bürger, unternehmen, fehlentwicklungen, plötzlich, bieten, bekämpfen, finanzielle, wiederum, zahlreichen, luft
125	scholz, verbrennerverbot, nachrichten, kanzler, olaf, sozialverbände, bedeutung, gewählt, zunehmend, kindern
126	familie, milliarden, nrw, mitglieder, erfolg, bürgermeister, solidarität, inhalt, geplanten, juli
127	umverteilung, bürgerrechte, warnt, geführt, fans, vergangenheit, richtung, hingegen, stärker, mehreren
128	teilen, neu, egal, rechtsextremen, komplett, ließ, brandenburg, feiern, zusätzlich, palästinensischen
129	leben, stadt, ziel, erneut, arbeiten, zeichen, klimaschutz, neues, denken, verhindert
130	deutsche, einfach, stand, staat, droht, gehören, system, wären, söder, demo
131	migration, getötet, asylgesetz, realität, weihnachtsmarkt, hilft, bezahlt, offene, irgendwie, fehlen
132	landesgrenzen, hinaus, wochen, treffen, herr, hinweg, alten, präsidenten, selenskyj, alte
133	organisationen, etc, nämlich, sport, ereignisse, befindet, david, gefährden, bedrohungen, indes
134	armutszeugnis, fragen, problem, aktuellen, oktober, september, verantwortung, anfang, helfen, wirft

#	Time (s)	Median	Mean	Std Dev	Alpha	Eta	# Topics	Topic Diversity	Coherence
1	2094.680	0.000	0.000	0.000	asymmetric	symmetric	296	0.003	0.106
2	1776.383	0.008	0.007	0.001	asymmetric	auto	194	0.049	0.160
3	1084.539	0.284	0.283	0.001	symmetric	symmetric	122	0.989	0.287
4	2006.410	0.001	0.001	0.000	auto	symmetric	264	0.008	0.107
5	1367.076	0.026	0.024	0.002	auto	symmetric	174	0.097	0.239
6	1262.827	0.110	0.114	0.018	auto	symmetric	156	0.253	0.435
7	1570.936	0.036	0.037	0.002	auto	auto	168	0.124	0.289
8	886.925	0.308	0.307	0.006	auto	symmetric	85	0.969	0.318
9	1198.112	0.296	0.306	0.029	auto	symmetric	144	0.561	0.540
10	1302.113	0.317	0.316	0.005	symmetric	auto	127	0.981	0.326
11	1309.428	0.305	0.305	0.006	auto	auto	126	0.976	0.312
12	1127.898	0.233	0.240	0.010	asymmetric	symmetric	144	0.505	0.462
13	1271.331	0.207	0.206	0.013	auto	auto	148	0.421	0.500
14	1231.373	0.185	0.174	0.017	symmetric	symmetric	151	0.377	0.486
15	1481.654	0.149	0.149	0.002	auto	auto	153	0.324	0.460
16	1291.098	0.288	0.291	0.004	auto	auto	124	0.988	0.292
17	1449.013	0.158	0.157	0.014	auto	auto	151	0.329	0.464
18	1338.262	0.376	0.379	0.022	symmetric	auto	135	0.830	0.483
19	1331.005	0.331	0.343	0.019	symmetric	auto	139	0.692	0.502
20	1270.743	0.334	0.332	0.002	symmetric	auto	141	0.630	0.529
21	1379.281	0.303	0.310	0.013	symmetric	auto	142	0.579	0.526
22	1424.564	0.277	0.278	0.009	symmetric	auto	146	0.503	0.575
23	1400.551	0.275	0.289	0.022	symmetric	auto	143	0.543	0.528
24	1156.922	0.324	0.326	0.006	symmetric	symmetric	142	0.620	0.520
25	959.121	0.293	0.291	0.003	symmetric	symmetric	102	0.975	0.300

Table B.1: Bayesian optimization results for LDA, showing all iterations with evaluated hyperparameter, objective function statistics (mean, median, standard deviation), and corresponding topic coherence and diversity scores.

C
BERTopic

Table C.2: All topics identified by BERTopic with topic IDs and the representative words.

Topic	Top Words
-1	afd, deutschland, abgeschoben, verschärfung, rechtsstaat, spd, cdu, remigration, terroristen, einwanderung
0	ukraine, russland, waffenstillstand, putin, russischen, russische, trump, selenskyj, krieg, ukrainische
1	heuchler, unglaubwürdig, sauerei, schäbig, unverschämtheit, verantwortungslos, abgeschoben, einfach, unwürdig, hinterhältig
2	steuerlast, steuermitteln, steuerzahlern, zwangsgebühren, senken, euro, sozialsysteme, finanziert, geld, steuergerechtigkeit
3	gegenwehr, unzufriedenheit, fans, anhängern, spiel, spieler, trainer, mannschaft, fc, saison
4	grüne, linksgrün, grünen, totalitäre, grün, sekte, heuchler, linksgrüne, linke, partei
5	einwanderung, zuwanderer, illegale, migration, sozialsysteme, masseneinwanderung, einwanderer, zuwanderung, migranten, deportation
6	co0, verbrennerverbot, klimawandel, klima, klimaschutz, umverteilung, energiewende, klimakrise, steuer, ungleichheit
7	impfpflicht, ungeimpfte, impfnebenwirkungen, corona, impfung, allgemeine, gestimmt, einrichtungsbezogene, impfungen, covid
8	deutschland, schande, deutschen, deutsche, partei, rechtsstaat, spd, grüne, antideutsche, grünen
9	terroristen, terroranschläge, terroristische, rebellen, anschläge, terror, land, zivilisten, lassen, geiseln
10	sicherheitsrisiko, spionage, dar, erhebliches, darstellen, sabotage, cyberangriffe, daten, unternehmen, anhängern
11	heimatliebe, partei, heilbronn, tierschutz, aschaffenburg, basisdemokratische, arbeit, freie, remigration, deutschlandabernormal

Continued on next page

Topic	Top Words
12	rechtsextremisten, linksextremisten, linksextremistische, mitte, linksextremistischen, rechtsextremer, linksextremistisch, antifa, extremisten, bundestag
13	partei, parteien, undemokratische, wähler, demokratie, politiker, undemokratischen, unzufriedenheit, unglaubwürdig, schande
14	waffenstillstand, frieden, kongo, sofortigen, weltfriedenstag, verhandlungen, dauerhaften, friedensverhandlungen, geiseln, kraft
15	vorbestraft, vorbestrafte, einschlägig, mehrfach, vorbestraften, bewährung, angeklagte, verurteilt, jährige, gericht
16	asylbewerber, asylbewerberleistungsgesetz, bezahlkarte, leistungen, asylbewerbern, asylantrag, erhalten, sgb, asylbewerberinnen, asylgesetz
17	syrien, syrer, assad, syrischen, syrische, hts, aleppo, alassad, syriens, sturz
18	trump, mexiko, donald, einwanderung, einwanderer, illegale, trumps, kanada, zölle, usa
19	spd, partei, linksextremisten, unglaubwürdig, heuchler, umverteilung, antifa, klingbeil, schäbig, wählen
20	afghanistan, asylbewerber, asylantrag, afghanen, afghane, abgelehnt, abgeschoben, taliban, jähriger, afghanischer
21	deutschland, steuerlast, deutschen, deutsche, ungleichheit, armutszeugnis, wirtschaft, steuerzahlern, euro, hohe
22	waffenstillstand, hamas, gaza, israel, geiseln, gazastreifen, freilassung, israels, israelischen, waffenruhe
23	cdu, csu, cducsu, unglaubwürdig, linksgrün, brandmauer, spd, heuchler, grünen, fdp
24	hamas, hamasterroristen, terroristen, oktober, geiseln, israel, gazastreifen, verschleppt, gaza, entführt
25	rechtsstaat, vertrauen, justiz, gesetze, funktioniert, unserem, leben, gesetz, unseren, funktionierenden
26	remigration, heißt, afd, heißen, rettet, millionenfache, weidel, begriff, wort, remigrationjetzt
27	israel, israels, terroristen, israelische, armee, antiisraelische, palästinenser, auslöschung, israelischen, palästinensischen
28	afd, verdachtsfall, afdanhängern, partei, rechtsextremer, rechtsextremisten, systemgegner, verfassungsschutz, parteien, wählen
29	islam, scharia, muslime, islamischen, islamisten, islamische, terroristen, islamistische, muslimische, muslimischen
30	trump, donald, trumps, anhängern, trumpanhängern, biden, washington, urpräsident, präsident, joe
31	china, spionage, chinesische, tiktok, chinesischen, usa, chinas, peking, bundesanwaltschaft, deepseek
32	grenzkontrollen, kontrollen, binnengrenzen, grenzen, schengenraum, niederlande, deutschen, luxemburg, september, deutschland
33	ungleichheit, soziale, umverteilung, unten, oxfam, sozialer, armut, sozialen, reich, wachsende
34	österreich, kickl, fpö, österreicher, wien, övp, österreichs, österreichischen, herbert, österreichische

Continued on next page

Topic	Top Words
35	merz, friedrich, migrationspolitik, verschärfung, unionskanzlerkandidat, bundestag, cduchef, anträge, cdukanzlerkandidat, afd
36	unzufriedenheit, scheiterns, gefühl, stress, führen, beziehung, gefühle, führt, alltag, leben
37	deutschland, einwanderung, zuwanderer, deutschen, deutsche, arbeitsmarkt, migranten, migration, braucht, remigration
38	hisbollah, libanon, israel, waffenstillstand, libanesischen, israelische, armee, südlibanon, waffenruhe, libanesische
39	eu, verbrennerverbot, europa, eukommission, europäischen, brüssel, regulierungswahn, europäische, kommission, euverbrennerverbot
40	musk, elon, musks, tesla, geste, trumpanhängern, trump, techmilliardär, anhängern, herz
41	frauen, sexualstraftaten, gewalt, unterdrückung, mädchen, männer, genderideologie, ungleichheit, opfer, frau
42	brandanschlag, feuer, brand, brandstiftung, nacht, polizei, feuerwehr, asylbewerberunterkunft, verübt, flammen
43	nazis, nazi, faschismus, hitler, faschisten, nationalsozialismus, totalitäre, rechtsextremisten, hitlers, faschistischen
44	verschwörungstheorie, verschwörung, theorie, verschwörungstheorien, realität, bestätigt, ne, wahrheit, abgetan, galt
45	unternehmensgewinne, banken, inflation, aktien, fed, anleger, verschärfung, investoren, ezb, wachstum
46	selbstverteidigung, techniken, selbstbehauptung, frauen, kampfsport, kampfkunst, karate, fitness, waffe, waffen
47	sicherheitspolitik, außen, europa, sicherheitskonferenz, europäische, europäischen, münchner, eu, europas, usa
48	kirche, kirchen, papst, christlichen, katholischen, jesus, religion, christen, franziskus, christliche
49	grenzkontrollen, dauerhafte, zurückweisungen, grenzen, binnengrenzen, grenze, zurückweisung, gültige, kontrollen, bundespolizei
50	medien, zwangsgebühren, journalismus, medienmanipulation, journalisten, öffentlichrechtlichen, sozialen, fake, media, propaganda
51	grüne, spd, grünen, einwanderung, masseneinwanderung, illegale, migration, migrationspolitik, unkontrollierte, grün
52	merkel, angela, merkels, flüchtlingspolitik, cdu, merz, frau, kanzlerin, afd, masseneinwanderung
53	verschwörung, netflix, spionage, the, film, action, doves, black, staffel, thriller
54	deutschland, asylbewerber, asylantrag, deutschen, flüchtlinge, zahl, asylgesuche, asylbewerbern, flüchtlingspolitik, asyl
55	merz, friedrich, herr, unglaubwürdig, cdu, rechtsextremisten, afd, brandmauer, union, kanzler
56	eltern, kinder, mutter, hinterhältig, frau, vater, hysterisch, verantwortungslos, kind, edith

Continued on next page

Topic	Top Words
57	weihnachtsmarkt, weihnachten, weihnachtsmärkte, weihnachtsmärkten, magdeburg, weihnachtsbaum, anschlag, magdeburger, anhängern, dezember
58	unzufriedenheit, kunden, unternehmen, job, beschäftigten, mitarbeiter, arbeitgeber, prozent, gehalt, beschäftigte
59	spd, grünen, grüne, linksextremisten, linke, fdp, linken, parteien, linksextremistischen, linksgrün
60	armutszeugnis, absolutes, land, echtes, gesellschaft, einfach, tafeln, lebensmittel, echt, welt
61	bundespolizei, grenzkontrollen, rahmen, waidhaus, kleve, autobahn, wiedereingeführten, ots, jährigen, a0
62	patienten, ärzte, apotheken, arzt, unzufriedenheit, ungeimpfte, krankenhaus, gesundheitssystem, pharma, demenz
63	juden, deportation, auschwitz, massenmord, jüdinnen, jüdischen, roma, sinti, nazis, holocaust
64	angeles, brände, verheerende, waldbrände, wüten, kalifornien, feuer, flammen, südkalifornien, häuser
65	yoon, staatsfeindliche, südkorea, nordkorea, kriegsrecht, opposition, verfassungsbruch, kräfte, wirft, yeol
66	türkei, pkk, kurden, türkische, erdogan, kurdischen, sdf, türkischen, kurdische, ypg
67	remigration, begriff, wort, unwort, wahlprogramm, afd, jahres, riesa, umstrittene, kampfbegriff
68	tiere, hund, hunde, katzen, tier, tieren, sauerei, tierheim, katze, klauenseuche
69	georgescu, rumänien, calin, georgien, bukarest, rechtsextremer, kandidat, rumänische, stichwahl, rumänischen
70	straftaten, ss, terroristischen, aktivitäten, gwb, zusammenhang, terroristische, ausschlussgründe, ssss, vereinigungen
71	homophob, beleidigt, queere, rassistisch, kaulitz, homosexuelle, queer, schwulen, bill, schwule
72	migrationspolitik, union, bundestag, verschärfung, antrag, anträge, afd, mittwoch, unionsfraktion, fünfpunkteplan
73	totalitäre, unterdrückung, meinungsdiktatur, meinungsfreiheit, systeme, regime, diktatur, demokratie, freiheit, kontrolle
74	sicherheitspolitik, außen, hochdynamischen, außenpolitischen, entscheidungen, bereich, sicherheit, sepos, veränderungen, spezifisch
75	rechtsstaat, demokratie, bürgerrechte, demokratischen, demokratischer, verfassung, freiheit, demokratische, verteidigt, menschenrechte
76	habeck, robert, habecks, wirtschaftsminister, kanzlerkandidat, sozialsysteme, wirtschaftsministeriums, gigantischste, energiewendeirrsinn, wirtschaftszerstörung
77	terroristen, deutschland, terroranschläge, deutsche, islamistische, deutschen, anschläge, österreich, münchen, terroristische
78	fdp, fpö, spö, övp, unglaubwürdig, neos, scheiterns, partei, kickl, verhandlungen
79	migrationspolitik, cdu, verschärfung, antrag, bundestag, csu, cducsu, afd, mittwoch, durchgesetzt

Continued on next page

Topic	Top Words
80	flüchtlingspolitik, wirtschaftsflüchtlinge, flüchtlinge, humane, flüchtlingen, humanitär, thema, integration, geflüchteten, flüchtlingsstrom
81	scholz, olaf, bundeskanzler, spd, kanzler, scheiterns, unverschämtheit, schröder, ampelregierung, fdp
82	dresden, rechtsextremisten, potsdam, rechtsextremer, aufmarsch, demonstration, tausende, demonstrieren, samstag, tausend
83	asylpolitik, bundestag, union, antrag, verschärfung, stimmen, cdu, afd, mehrheit, asylgesetz
84	juden, antisemitismus, jüdische, jüdischen, antisemitische, jüdinnen, auslöschung, massenmord, judenhass, jüdischer
85	vereinigung, kriminelle, armut, tod, bringen, bsw, cdulinkebswspdgrünearmuttod, rechtsstaat, linke, cdu
86	drohnen, spionage, theorien, gesichtet, militärübungen, geheime, marl, außerirdischem, drohne, reichen
87	schule, schulen, lehrer, kindergarten, schüler, schulleitung, kinder, eltern, wiederkehr, unzufriedenheit
88	rassismus, fremdenfeindlichkeit, diskriminierung, antisemitismus, unterdrückung, weiße, form, hass, rassistische, ausgrenzung
89	gwb, entrichtung, tätigkeit, gewerblichen, einstellung, zahlung, sozialversicherungsbeiträge, ssss, steuern, verpflichtungen
90	schweiz, schweizer, svp, amherd, zürich, sicherheitsrisiko, bundesrätin, rechtsstaat, rücktritt, sicherheitspolitik
91	energie, energiepreise, sozialsysteme, masseneinwanderung, wirtschaft, energiewende, einwanderung, energiepolitik, infrastruktur, kaputt
92	orf, zwangsgebühren, orfzwangsgebühren, abschaffen, fpö, zwangsgebührensender, sender, berichterstattung, zahlen, finanziert
93	weidel, alice, kanzlerkandidatin, remigration, begriff, riesa, afdchefin, afdkanzlerkandidatin, afdparteitag, afdspitzenkandidatin
94	österreich, österreicher, zuwanderer, wien, fpö, österreichischen, österreichische, österreichs, spö, österreichern
95	magdeburg, anschlag, staatsversagen, aschaffenburg, täter, taleb, attentäter, faeser, opfer, gefährderkategorie
96	prozent, anhängern, spdanhängern, bsw, afdanhängern, cducsu, fdp, union, unionsanhängern, bevorzugen
97	eu, europäischen, europa, asylbewerbern, europäische, migration, einwanderung, euasylrechts, asylbewerber, außengrenzen

#	Time	Median	Mean	Std. Dev.	min cluster size	min dist	min samples	n components	n neighbors	Topic Diversity	Coherence
1	1895.728	0.434	0.435	0.003	196	0.069	49	18	30	0.699	0.622
2	2220.518	0.430	0.428	0.006	97	0.049	49	13	44	0.717	0.599
3	3499.795	0.393	0.391	0.004	115	0.100	1	23	13	0.675	0.574
4	2400.911	0.420	0.423	0.005	98	0.099	8	10	45	0.730	0.581
5	2905.533	0.451	0.453	0.006	314	0.055	27	44	34	0.726	0.615
6	1686.055	0.742	0.639	0.145	180	0.097	35	22	25	0.925	0.742
7	2500.695	0.429	0.429	0.003	202	0.070	13	35	40	0.706	0.608
8	2409.330	0.428	0.420	0.013	227	0.095	32	21	4	0.702	0.605
9	2046.283	0.432	0.430	0.005	178	0.081	49	22	11	0.713	0.606
10	2390.336	0.437	0.443	0.008	184	0.100	35	31	19	0.715	0.617
11	2928.617	0.454	0.549	0.136	331	0.097	38	47	24	0.728	0.624
12	1380.063	0.449	0.450	0.006	239	0.098	40	10	25	0.720	0.627
13	3791.857	0.452	0.450	0.004	78	0.097	48	43	27	0.724	0.624
14	3435.179	0.442	0.440	0.006	88	0.098	35	41	25	0.722	0.604
15	2111.898	0.431	0.430	0.004	181	0.008	35	22	23	0.713	0.604
16	1527.778	0.446	0.545	0.139	156	0.097	39	13	28	0.728	0.615
17	3055.827	0.449	0.448	0.008	265	0.097	36	48	27	0.734	0.615
18	2075.587	0.436	0.432	0.010	180	0.061	8	24	25	0.718	0.607
19	1937.384	0.439	0.442	0.004	224	0.097	32	21	10	0.722	0.614
20	1865.274	0.447	0.541	0.146	149	0.097	33	23	46	0.733	0.609
21	2317.230	0.438	0.437	0.003	86	0.097	29	11	27	0.731	0.595
22	2631.670	0.435	0.433	0.005	167	0.097	37	38	25	0.716	0.607
23	2069.204	0.438	0.442	0.007	212	0.097	48	26	27	0.709	0.619
24	2239.604	0.411	0.413	0.012	144	0.097	35	14	7	0.693	0.593
25	2592.691	0.411	0.408	0.006	186	0.097	7	23	6	0.703	0.579

Table C.1: Bayesian optimization results for BERTopic, showing all iterations with evaluated hyperparameter, objective function statistics (mean, median, standard deviation), and corresponding topic coherence and diversity scores.

D
TopicGPT

Table D.1: All topics identified by TopicGPT(GPT-4o-mini) with topic IDs and representative words calculated by C-TF-IDF and the Topic Label generated by GPT-4o-mini.

Topic ID	Topic Name	Top Words
0	Taxation	bastian, greenpeace, wirtschaftsexperte, neuwirth, superreiche, steuerzahlern, besteuern, umverteilung, steuermitteln, steuergerechtigkeit
1	Tax Burden	sozialrechtliche, verstoß, gewerblichen, entrichtung, einstellung, tätigkeit, verpflichtungen, zahlung, steuern, steuerlast
2	Tax Policy	einstellung, tätigkeit, gewerblichen, entrichtung, steuern, bzw, zwingende, fakultative, ausschlussgründe, gwb
3	Crafts	tolles, heimat, freizeitanzeige, hinaus, landesgrenzen, puzzle, geschenk, sauerei, heimatliebe, anhängern
4	DIY Decorations	spülmittel, siphon, einweicht, geschleudert, ladungssicherung, befestigt, dekorationen, bremsmanöver, fahrzeuginneren, christbaumkugeln
5	Baking Techniques	segen, fluch, opa, fahrradstraßen, sauerei, enkelin, penis, backt, plätzchen, weihnachtsbäckerei
6	Music	gegenwehr, hinaus, fans, landesgrenzen, band, widerstand, unterdrückung, anhängern, musik, heuchler

Continued on next page

Topic ID	Topic Name	Top Words
7	Music as Resistance	adel, musik, saus, forest, braus, sherwood, landvolk, celtic, widerstand, unterdrückung
8	Electronic Music	date, 00, author, publish, syndicated, ra, musik, techno, source, elektronischen
9	Global Music Distribution	waffenexporte, position, militärische, legt, russland, handel, globalen, acts, rohstoffe, china
10	Religion	papst, brandanschlag, christen, islam, anhängern, kirchen, religion, synagoge, kirche, scharia
11	Church Authority	sed, revolution, gehörte, friedlichen, beendete, protagonisten, regimekritikern, prägenden, schärfsten, theologe
12	Religious Identity	beiträge, social, media, johannes, aktiv, islamische, beschuldigte, religiös, postete, athlet
13	Religious Conflicts	antisemitischen, schmierereien, islam, religiöse, verübt, religiösen, brandanschlag, oldenburg, scharia, synagoge
14	Secularism	gesellschaft, säkularismus, zustände, friedlichen, staat, religion, tolerieren, rechtsstaates, säkulare, säkularen
15	Social Issues	verschwörungstheorie, sozialsysteme, soziale, umverteilung, rechtsextremisten, deutschland, fremdenfeindlichkeit, ungleichheit, armutszeugnis, unzufriedenheit
16	Extremism	parteien, rechts, grüne, spd, extremismus, extremisten, afd, linksextremisten, rechtsextremer, rechtsextremisten
17	Violence	täter, vorbestraft, einsatzkräfte, tötungsdelikte, gewaltausbrüchen, gewalttaten, gewaltexzesse, gewalt, polizei, gewalttäter
18	Human Rights	freiheit, sexualstraftaten, meinungsfreiheit, homophob, rechtsstaat, menschenrechte, vorbestraft, frauen, bürgerrechte, unterdrückung
19	Migration	inflation, atomausstieg, wirtschaftszerstörung, altersarmut, massenmigration, sozialsysteme, asylbewerber, migranten, zuwanderer, migration
20	Social Inequality	euro, wachsende, prozent, umverteilung, einkommen, armut, sozialer, sozialen, soziale, ungleichheit

Continued on next page

Topic ID	Topic Name	Top Words
21	Natural Disasters	umschließt, aufhebt, reisewarnung, epidemien, naturkatastrophen, feuer, waldbrände, brände, angeles, verheerende
22	Floods	malaga, andalusiens, september, horwood, valencia, ufer, schäden, verheerende, überschwemmungen, hochwasser
23	Drought	dpa, wüstenbildung, gehört, archivbild, überschwemmungen, klimawandels, schmitt, dürren, mirjam, dürre
24	Tsunami	stromdisaster, salzwasser, copyright, content, date, author, publish, syndicated, source, tsunami
25	Technology	wirtschaftsministeriums, bankensystem, unterstützungsmöglichkeiten, verschärfung, technologie, bitcoin, sicherheitsrisiko, digitalen, ki, verbrennerverbot
26	Artificial Intelligence	gemeinwohlorientiert, überwachungsmaßnahmen, künstlichen, chatbot, menschheit, chips, künstlicher, künstliche, intelligenz, ki
27	Software Alternatives	support, xp, alternativen, betriebssystem, microsoft, nutzer, unzufriedenheit, linux, software, windows
28	Cybersecurity	phishing, unternehmen, windows, software, passwörter, cyberangriffe, daten, cyber, spionage, sicherheitsrisiko
29	Entertainment	serie, app, netflix, live, kanal, bundesliga, frag, wwg1wga, the, verschwörung
30	Video Games	netflix, angriffe, verheerende, spiels, spiel, fallout, konami, meistern, spieler, action
31	Streaming Services	zwangsgebühren, doves, staffel, kritiken, action, prosieben, back, serie, streaming, netflix
32	Economics	wirtschaftliche, wachstum, inflation, us, aktien, unternehmen, verschärfung, wirtschaft, wirtschaftsministeriums, unternehmensgewinne
33	Corporate Profits	erweisen, aktie, dividenden, einkommensströme, ausgeschüttet, aktien, unternehmen, gewinne, aktionäre, unternehmensgewinne
34	Security	vereinigungen, terroristischen, sicherheit, aktivitäten, gefährder, terroristische, sicherheitspolitik, straftaten, sicherheitsrisiko, spionage

Continued on next page

Topic ID	Topic Name	Top Words
35	Terrorism	zahlung, terroristen, ausschlussgründe, steuern, gewerblichen, einstellung, entrichtung, tätigkeit, gwb, terroristische
36	Espionage	russland, heimliche, geheimnisvolle, tiefgarage, verdachts, krahs, mitarbeiter, krah, china, spionage
37	Politics	deutschland, grünen, demokratie, anhängern, grüne, partei, cdu, spd, afd, rechtsstaat
38	Political Accountability	demokratie, fdp, politischen, partei, cdu, unglaubwürdig, undemokratische, verfassungsbruch, spd, rechtsstaat
39	Foreign Relations	präsident, gemeinsamkeiten, eu, russland, ukraine, usa, trump, us, außen, sicherheitspolitik
40	Governance and Democracy	eu, cdu, undemokratischen, demokratische, gewaltenteilung, parteien, demokratischen, undemokratische, demokratie, rechtsstaat
41	Women's Rights	gesetzeshüter, gesetzesinitiative, gesetzesinitiativen, gesetzeskonform, gesetzeskonformen, gesetzeslage, gesetzeslücke, gesetzeslücken, gesetzesnovellierung, 6steckborn
42	Indigenous Rights	bezeichnet, bevölkerung, unterdrückung, geprägt, invasion, gräueltaten, survival, folgezeit, day, indigenen
43	Political Repression	bürgerrechte, kräfte, verfassungsmäßige, politische, regierungsarbeit, staatsfeindliche, gegner, politischen, unterdrückung, rechtsstaat
44	Freedom of Expression	rechtsstaat, meinung, pressefreiheit, menschenrechte, demokratie, zensur, meinungsäußerung, redefreiheit, unterdrückung, meinungsfreiheit
45	Rights of the Accused	angeklagten, prüfung, anhörung, angeklagte, betroffene, kaution, gilt, vorbestraft, rechtsstaat, unschuldsvermutung
46	Culture	unwürdig, unglaubwürdig, unzufriedenheit, verschwörung, hinaus, kultur, anhängern, heimatliebe, landesgrenzen, leitkultur
47	Cultural Identity	wissenschaftfeindlich, vorausschauendes, anmer, anpassender, scheib, wendiger, linke, wendehals, leitkultur, liensberger

Continued on next page

Topic ID	Topic Name	Top Words
48	Urban-Rural Divide	gesetzeshüter, gesetzesinitiative, gesetzesinitiativen, gesetzeskonform, gesetzeskonformen, gesetzeslage, gesetzeslücke, gesetzeslücken, gesetzesnovellierung, 6steckborn
49	Cultural Education	düsseldorf, präsentiert, ausstellung, vielfach, berger, allee, mildere, fotografie, stadtmuseum, fotodokumentarische
50	War	waffenstillstand, israels, israelische, terroristische, geiseln, oktober, armee, israel, hamas, terroristen
51	Ceasefire	putin, krieg, frieden, geiseln, russland, gaza, verhandlungen, israel, ukraine, waffenstillstand
52	Peace and Conflict	hamas, sicherheitsgarantien, krieg, gaza, russland, verhandlungen, israel, frieden, ukraine, waffenstillstand
53	Stability	schone, krieg, diene, ausgehandelt, konflikt, russland, übereinkommen, stabilität, ukraine, waffenstillstand
54	Health and Wellness	covid, unzufriedenheit, verheerende, masern, corona, impfung, gesundheit, impfnebenwirkungen, ungeimpfte, impfpflicht
55	Preventive Health	finanzierungsmix, cornelia, warn, strukturierte, pflegefall, corbyn, kontrollierbaren, cordoba, präventionsstrategie, corona
56	Infectious Diseases	kriege, naturkatastrophen, abhängig, umständen, ausspricht, auswärtige, reisewarnung, aufhebt, umschließt, epidemien
57	Chronic Conditions	gesundheitssystem, eigenverantwortung, gesellschaftliches, chronische, gezieltes, chronischen, chronischer, vir, erkrankungen, patientenmanagement
58	Immigration and Migration	merz, migration, afd, bundestag, illegale, grenzkontrollen, verschärfung, asylbewerber, migrationspolitik, einwanderung
59	Internal Displacement	störung, vertriebene, ärztliche, atteste, posttraumatischen, belastungsstörung, wiedererleben, beeinträchtigenden, traumatischer, schlafproblemen
60	Remigration	deutschland, aschaffenburg, wahlprogramm, heißt, alice, wort, weidel, begriff, afd, remigration

Continued on next page

Topic ID	Topic Name	Top Words
61	Deportation	abschiebehaft, deutschland, asylbewerber, migranten, gefährder, abschiebung, straftäter, abschiebungen, deportation, abgeschoben
62	Outliers	wm, saison, league, anhängern, fc, spiel, mannschaft, trainer, unzufriedenheit, gegenwehr

References

Michael Achmann and Christian Wolff. 2023. Policy Issues vs. Documentation: Using BERTopic to Gain Insight in the Political Communication in Instagram Stories and Posts during the 2021 German Federal Election Campaign. *Digital Humanities in the Nordic and Baltic Countries Publications* 5, no. 1 (October): 11–28. (Cited on page 19).

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning,* 113–120. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery. (Cited on pages 16 sqq., 30 sq.).

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, no. null (March): 993–1022. (Cited on pages 4, 6, 25, 39).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems.* NIPS '20. Vancouver, BC, Canada: Curran Associates Inc. (Cited on page 12).

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining,* edited by Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, 160–172. Berlin, Heidelberg: Springer Berlin Heidelberg. (Cited on page 14).

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41 (6): 391–407. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9. (Cited on pages 4 sq., 25, 39).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 Jill Burstein, Christy Doran, and Thamar Solorio, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics, June. (Cited on pages 10 sq.).

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic Modeling in Embedding Spaces. Edited by Mark Johnson, Brian Roark, and Ani Nenkova. *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 8:439–453. (Cited on page 30).

Tomoki Doi, Masaru Isonuma, and Hitomi Yanaka. 2024. Topic Modeling for Short Texts with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop),* edited by Xiyan Fu and Eve Fleisig, 21–33. Bangkok, Thailand: Association for Computational Linguistics, August. (Cited on page 15).

S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 281–285. CHI '88. Washington, D.C., USA: Association for Computing Machinery. (Cited on page 5).

Roman Egger and Joanne Yu. 2022. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology* 7 (May). (Cited on page 19).

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. 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,* 226–231. KDD'96. Portland, Oregon: AAAI Press. (Cited on page 14).

KI für Demokratie e.V. KI Für Demokratie. https://ki-fuer-demokratie.de/. Accessed: 2025-08-01. (Cited on page 1).

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794,* (cited on pages 5, 12 sq., 15 sqq., 25 sq., 31).

N. Halko, P. G. Martinsson, and J. A. Tropp. 2011. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review* 53 (2): 217–288. eprint: https://doi.org/10.1137/090771806. (Cited on page 28).

Nils Constantin Hellwig, Jakob Fehle, Markus Bink, Thomas Schmidt, and Christian Wolff. 2024. Exploring Twitter discourse with BERTopic: topic modeling of tweets related to the major German parties during the 2021 German federal election. *International Journal of Speech Technology* 27, no. 4 (October): 901–921. (Cited on page 19).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* (Cambridge, MA, USA) 9, no. 8 (November): 1735–1780. (Cited on page 9).

IMWF Institute. 2024. Institute für Management- und Wirtschaftsforschung. Available at: https://www.imwf.de. (Cited on page 21).

Amandeep Kaur and James R Wallace. 2024. Moving Beyond LDA: A Comparison of Unsupervised Topic Modelling Techniques for Qualitative Data Analysis of Online Communities. *arXiv preprint arXiv:2412.14486,* (cited on page 19).

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2, no. 11 (March). (Cited on page 28).

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3 (29): 861. (Cited on pages 13 sq.).

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (Cited on pages 7 sq.).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* edited by Alessandro Moschitti, Bo Pang, and Walter Daelemans, 1532–1543. Doha, Qatar: Association for Computational Linguistics, October. (Cited on page 8).

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A Prompt-based Topic Modeling Framework. arXiv: 2311.01449 [cs.CL]. (Cited on pages 16, 26, 29, 38).

Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. (Cited on page 12).

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (Cited on page 12).

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora [in English]. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45–50. Http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May. (Cited on page 34).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 3982–3992. Hong Kong, China: Association for Computational Linguistics, November. (Cited on page 11).

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* 399–408. WSDM '15. Shanghai, China: Association for Computing Machinery. (Cited on pages 29 sq.).

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323:533–536. (Cited on page 8).

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical Bayesian optimization of machine learning algorithms. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2,* 2951–2959. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc. (Cited on page 27).

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and Optimizing Topic Models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations,* edited by Dimitra Gkatzia and Djamé Seddah, 263–270. Online: Association for Computational Linguistics, April. (Cited on pages 27, 33).

Peter W. Foltz Thomas K Landauer and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25 (2-3): 259–284. eprint: https://doi.org/10.1080/01638539809545028. (Cited on page 5).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems,* edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol. 30. Curran Associates, Inc. (Cited on pages 8 sqq.).

Rudy Alexandro Garrido Veliz, Till Nikolaus Schaland, Simon Bergmoser, Florian Horwege, Somya Bansal, Ritesh Nahar, Martin Semmann, Jörg Forthmann, and Seid Muhie Yimam. 2025. KI4Demokratie: An AI-Based Platform for Monitoring and Fostering Democratic Discourse. arXiv: 2506.09947 [cs.CY]. (Cited on page 21).