



SCHOOL OF COMPUTATION,  
INFORMATION AND TECHNOLOGY —  
INFORMATICS

TECHNICAL UNIVERSITY OF MUNICH

Bachelor's Thesis in Informatics in Informatics

**From Hashtags to Ballot Boxes: A Close  
Look at the 2023 Turkish Election**

Efe Sener



SCHOOL OF COMPUTATION,  
INFORMATION AND TECHNOLOGY —  
INFORMATICS

TECHNICAL UNIVERSITY OF MUNICH

Bachelor's Thesis in Informatics in Informatics

**From Hashtags to Ballot Boxes: A Close  
Look at the 2023 Turkish Election**

**Von Hashtags zu Wahlentscheidungen: Ein  
umfassender Blick auf die Türkischen  
Wahlen 2023**

Author:	Efe Sener
Supervisor:	Prof. Dr. Georg Groh
Advisor:	Carolin Schuster
Submission Date:	15.03.2023

I confirm that this bachelor's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 15.03.2023

Efe Sener

## **Acknowledgments**

# Abstract

# Contents

<b>Acknowledgments</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	2
1.2 Research Questions . . . . .	3
<b>2 Related Work</b>	<b>4</b>
<b>3 Experiments</b>	<b>6</b>
3.1 The Dataset . . . . .	6
3.2 The Methodology . . . . .	8
3.2.1 The BERTopic Pipeline . . . . .	9
3.2.2 The Analysis Strategy . . . . .	10
<b>4 Results</b>	<b>13</b>
4.1 Detailed Analysis of Topic Categories . . . . .	16
4.1.1 Ruling Party Discourse Analysis . . . . .	16
4.1.2 Opposition Discourse Analysis . . . . .	18
4.1.3 Other Relevant Discourse Analysis . . . . .	21
4.2 Supplementary Visual Insights . . . . .	23
<b>5 Discussion</b>	<b>28</b>
5.1 Interpreting the results in a broader context . . . . .	28
5.2 Limitations . . . . .	28
5.3 Future Work . . . . .	28
<b>6 Conclusion</b>	<b>29</b>
6.1 Section . . . . .	29
6.1.1 Subsection . . . . .	29
<b>Abbreviations</b>	<b>32</b>

## *Contents*

---

<b>List of Figures</b>	<b>33</b>
<b>List of Tables</b>	<b>34</b>
<b>Bibliography</b>	<b>35</b>

# 1 Introduction

In recent years, governments and the public have realized the importance of social media, especially Twitter/X, which has a decisive role in mobilizing social and political activism (Uysal & Schroeder, 2019). Twitter has been instrumental in studying human behavior with social media data (Pfeffer et al., 2023), described as a digital social telescope by researchers in the social science field (Mejova et al., 2015). It has provided a somewhat free environment and guided social and political debates to gain new dimensions (Yerlikaya & Toker, 2020), where individual users can directly and publicly address comments to their representatives under conditions of anonymity (Theocharis et al., 2020). The robust rise in Twitter's popularity has stemmed from increasing accessibility to technology and affordability. Millions of people consume news from social media sites like Twitter (Anwar et al., 2021). In Turkey's case, Twitter began to be taken seriously after the unrest in the Middle East, especially after the Gezi Park protests in 2013 (Zaharna & Uysal, 2016), where Twitter was one of the most valuable media for protestor communication, given censorship (Ogan & Varol, 2017). In 2020, Turkey was ranked as the 10th most-used language on Twitter, with around 560 thousand tweets posted daily (Alshaabi et al., 2021).

This thesis aims to analyze the Twitter data, provided by Sabanci University (Najafi et al., 2022), to understand the Turkish Twitter discourse surrounding the May 2023 elections. Using innovative topic modeling techniques, this thesis will discover the most prevalent topics in Turkish Twitter between July 2022 and June 2023. It will uncover how these topics correlate with real-life events and how they reflect the election agendas of parties. This thesis will also compare the results with results observed in other countries. In a non-English-speaking country like Turkey, this thesis furthermore seeks to find solutions to the need for a more thorough and data-driven analysis of political discussions on Turkish Twitter.

In this chapter, the thesis starts by explaining the historical context and then continues to present the current political landscape. It demonstrates the importance of the May 2023 elections, emphasizes the significance of Twitter in Turkish politics, and deep dives into research questions. In the next chapter, the thesis examines various related works, asking similar questions and analyzing their results. After that, the thesis explains the Twitter dataset and used methodologies while collecting and analyzing the data. Next, the thesis deep dives into the analysis results, and later discusses the findings by



interpreting them, highlighting both the limitations and future work. The final section summarizes the results and its implications.

## 1.1 Background

It is crucial to examine Turkey's historical political context to understand the current complex political landscape and the May 2023 elections.

After the collapse of the Ottoman Empire, the Turkish Republic was declared in 1923. Some attempts were made, but the first multi-party elections were held in 1946. Until 1945, the Republican People's Party (CHP) was the only party in the parliament, and until 1950 it was the ruling party. The CHP was founded by Mustafa Kemal Atatürk, also the founder of the Turkish Republic.

With a multi-party system in a young republic, political power was now open to various groups. Different and new ideologies arose and started to organize politically (Rabasa & Larrabee, 2008). The military saw their role as the protector of the Republic and Atatürk's ideologies and overthrew the governments in 1960, 1971, and 1980. The 1980 military coup, which introduced a new constitution, was after a period of political fragmentation and civil instability in the 1970s.

During the 1970s, political Islamism started to emerge, which challenged the secularist nationalism and modernization ideologies of the CHP (Yilmaz & Bashirov, 2018). Changes in the political structure, the constitution, and civil liberties, major economic crises in 1994 and 2001 (Arđan, 2023) contributed to Islamic political groups' political influence and strength, to the emergence of new political players and parties like the Justice and Development Party (AKP) (Rabasa & Larrabee, 2008).

Since 2002, AKP has been in power in Turkey. Out of 15 elections, AKP just lost the local elections in 2019, in which the opposition coalition won more than four significant municipalities. Especially in Istanbul, the opposition won twice because the first election was canceled. For the May 2023 elections, the main opposition coalition was established from CHP, Good Party (İyiP), Felicity Party (SAADET), Democrat Party (DP), and two new parties were established out of AKP: Democracy and Progress Party (DEVA) and Future Party (GP) (Atila, 2022). Even though most of the polls favored the opposition in the May 2023 elections (Saç & Çoban, 2023), AKP has won the majority of the parliament and Recep Tayyip Erdoğan was elected in the kickoff elections for the third time as president, after serving two terms as president and two terms as prime minister since 2003.

## 1.2 Research Questions

This section introduces the research questions guiding this thesis, which are based on qualitative methods to analyze the Twitter discourse surrounding the May 2023 elections in Turkey.

The research questions are divided into two parts. The first part will cover the main research objective of this thesis, which is the analysis of the topic modeling results. The first question is as follows: “What were the most prevalent topics in Turkish Twitter discussions during the May 2023 elections?”. This question is necessary to understand the main topics of the May 2023 elections discussed in social media.

The next question is “How do real-life events during the election period correlate with shifts in discussion topics on Twitter, and in what ways do these shifts mirror political movements?”. This question focuses on the reflection of real-life events and political movements in Twitter discussions.

The third question is about parties and their election agendas: “How do the Twitter discussions about the ruling party and the opposition during the election lead-up reflect and compare to their respective election agendas and public statements?”. This question is essential to understand the reflection of the election agendas of the parties and the differences between them on Twitter.

With these questions in mind, the second part of the research questions covers the comparison of the results of the topic modeling with other research, where a similar approach was used for different countries. The main question is as follows: “How do the key themes, content, and engagement levels in the Turkish Twitter discourse surrounding the May 2023 elections compare with those observed in the past elections in other countries?”.

## 2 Related Work

The recent advances in Natural Language Processing (NLP) and easy access to open-source models allow researchers to study text data by performing sentiment and emotional analysis, topic modeling, semantic search, and many more. Large language models by OpenAI considerably explain how fast the NLP field develops.

In this thesis, topic modeling is performed on massive text data. Topic modeling is an unsupervised tool that helps extract the underlying themes from the given text data. There are several topic modeling approaches, and this thesis focuses on neural topic modeling. Unlike conventional models like Latent Dirichlet Allocation (LDA), a generative probabilistic model introduced by Blei et al. (2003), neural topic models have been used in important NLP tasks, including text generation, document summarisation, and translation, fields to which conventional topic models are complex to apply (Zhao et al., 2021). This thesis uses the neural topic model BERTopic, introduced by Grootendorst (2022), which is explained in detail in Chapter 3.

A tremendous number of studies have applied topic modeling in their research. In the political science field, Ilyas et al. (2020) performed topic modeling using LDA to discover daily discussion topics on Twitter about Brexit and to find out whether the topics discussed on Twitter were representative of actual events taking place, aligning with the second research question of this thesis. They found out that their model was representative of the actual events. Kaiser et al. (2020) used a structural topic model (STM), similar to LDA, to analyze the right media coverage during the 2016 US elections. The analysis shows that a media outlet is identified between the extreme far-right and mainstream right by finding out that they cover extreme and conservative topics. For the 2020 US elections, Anwar et al. (2021) applied topic modeling using BERTopic on pro-Trump tweets to analyze the most mentioned words for each topic and how frequent the topics were, aligning with the first research question of this thesis. Gritto (2022) applied BERTopic along with other German BERT models on Twitter data from German politicians and analyzed their results, aligning with the third research question of this thesis. She discovered that using BERTopic with the Sentence-BERT (SBERT) model yielded more valuable and significant topics. On the other hand, Contreras et al. (2022) used both LDA and BERTopic on Spanish Panamanian parliamentary proceedings. The research suggests that both models perform well with long multilingual political texts despite the small dataset.

It is essential to mention that according to the available literature, few studies apply topic modeling to multilingual political data. For the Turkish language, since the introduction of BERTurk by Schweter (2020), which is based on the BERT model by Devlin et al. (2019) trained on Turkish dataset, the Turkish NLP community is getting bigger and bigger day by day. Recently, a new model called TurkishBERTweet trained on the Turkish Twitter dataset was presented by the same team<sup>1</sup> that released the public social media dataset #Secim2023<sup>2</sup> (Najafi & Varol, 2023). The team has used TurkishBERTweet to conduct daily sentiment analysis and various other analyses on the #Secim2023 dataset, which will be discussed later.

This thesis will build upon the mentioned research and conduct one of the first neural topic modeling researches on a massive political Turkish language dataset using BERTopic. As mentioned in previous research, BERTopic yields more valuable and significant topics than other topic models, which is why this thesis will use that model. Since Najafi and Varol (2023) and also Najafi et al. (2022) analyze the same dataset as this thesis, but with different approaches and questions, this thesis will also use their results while answering the research questions.

---

<sup>1</sup>Center of Excellence in Data Analytics, Sabanci University, Turkey

<sup>2</sup>Najafi et al., 2022.

## 3 Experiments

This thesis uses the BERTopic model to apply topic modeling on the #Secim2023 dataset. Before diving into the results and discussion, this chapter explains the dataset, how the tweet hydration<sup>1</sup> is performed on the tweets from the dataset, how BERTopic and neural topic modeling works generally.

### 3.1 The Dataset

The dataset published by Najafi et al. (2022) consists of tweet IDs collected daily between July 2022 and June 2023, a total of around 250 million tweets. The frequency of the collected tweets is shown in Figure 3.1.

Due to Twitter’s Developer Agreement and Policy<sup>2</sup>, a public dataset can only include (tweet) IDs. In order to access all tweet information, they must be hydrated. Typically, a year before, a research group would have had access to Twitter Academic API<sup>3</sup> and used packages like Hydrator<sup>4</sup> to gather tweet information quickly. Unfortunately, after Elon Musk bought Twitter, Academic API was restricted and then shut down at the end of May 2023 (Calma, 2023), before the start of this thesis. Today, there are only paid options starting from 100\$ for 10,000 tweets per month, 0.3% of what was previously available for free access in a single day.

If one has tweet IDs, other methods exist to hydrate the tweets nowadays. All of the following methods use some embedded retrieval mechanism to gather the tweet information. The first method uses Twitter’s official page to retrieve embedded posts or videos given the tweet ID: <https://publish.twitter.com>. The second method, also used in this thesis, is implemented by React engineers in-house: <https://github.com/vercel/react-tweet>. One can have a JSON output with sufficient information for analysis by sending HTTP requests and tweet ID as a parameter.

As seen in Figure 3.1, the collected tweets (blue) are less than the total tweets in the dataset (orange). Out of 250 million tweets, only around 150 million tweets are collected. One of the main reasons for that is the deleted tweets. Since the hydration

---

<sup>1</sup>The process of retrieving a tweet’s complete information with only tweet ID.

<sup>2</sup><https://developer.twitter.com/en/developer-terms/agreement-and-policy>

<sup>3</sup><https://developer.twitter.com/en/use-cases/do-research/academic-research>

<sup>4</sup><https://github.com/DocNow/hydrator>

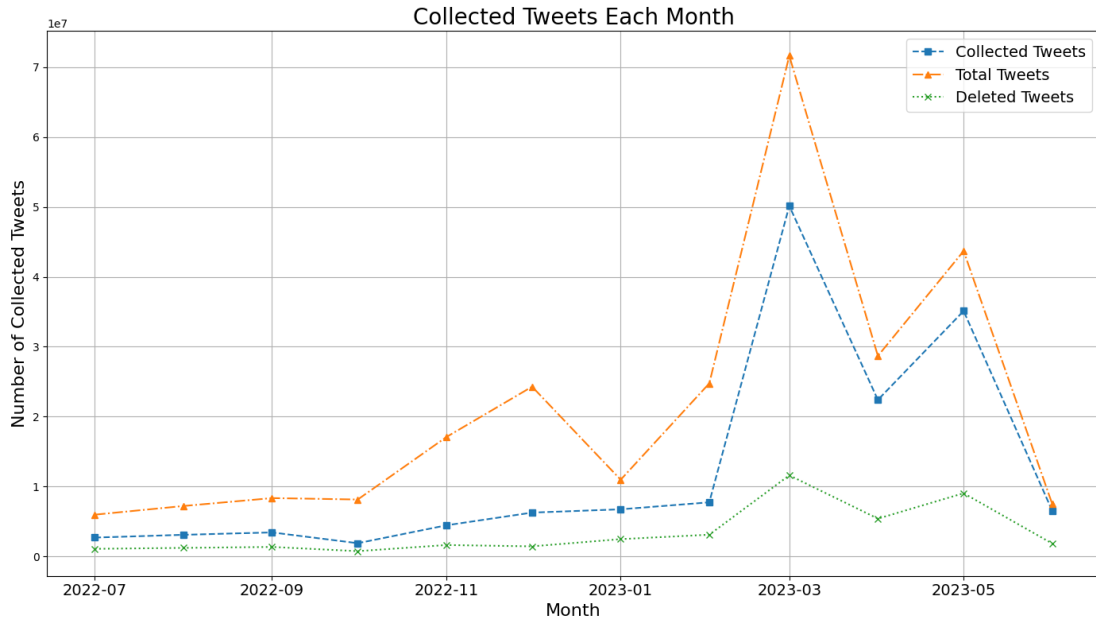


Figure 3.1: The graph displays the monthly count of collected tweets from the #Secim2023 database, spanning from July 2022 to June 2023, in increments of 10 million. The orange line displays the total number of tweets in the database, the blue line displays the total number of collected tweets from the database and the green line displays the deleted tweets on the month of hydration, around December 2023.

timeline for this thesis was between October 2023 and January 2024, and the tweets are between July 2022 and June 2023, there are approximately 50 million deleted tweets in the dataset (green).

There are several reasons why there are lots of deleted tweets. The main reason for the vast number is the deletion of highly interacted original tweet posts. According to Twitter, if the original tweet is deleted, the reposts to that tweet are also not available anymore<sup>5</sup>. The #Secim2023 dataset contains retweets, quotes, and replies in the majority. As discussed by Pfeffer et al. (2023) in their research, almost 80% of all tweets on Twitter refer to other tweets, with original tweets making up the rest.

Furthermore, the question of why people deleted their posts in the first place can be answered in two aspects. First, as mentioned in this New York Times article by Klosowski (2022), whether the person posting is a public figure or not is not essential:

<sup>5</sup><https://help.twitter.com/en/using-x/repost-faqs>

companies in the hiring process could run a social media background check, leading to rejection.

Secondly and even more alarming, the government can pull an old tweet out of context and use it against the person, leading to an arrest. The Turkish government's control over social media is widely recognized. It has instituted nationwide bans before and has arrested people accused of "provocative posts" continuously (Scott, 2023). Freedom House's 2023 report states that Turkey's global and internet freedom scores are classified as "not free" (Freedom House, 2023). These reasons could have eventually led to the deletion of many tweets after the election.

The gap between collected plus deleted tweets and total tweets lies under restricted rate limits by Twitter<sup>6</sup>. During hydration, every second or third response was empty, which led to second and third hydration batches of the missing tweets. There are around 50 million tweets that could not be collected. Due to time limitations and the lengthy duration of big data analysis, the hydration process resulted in the maximum feasible collection of tweets within the constraints.

## 3.2 The Methodology

As mentioned in the previous chapter, topic modeling is an unsupervised tool that helps extract the underlying themes from the given text data. BERTopic is a neural topic model, one of many topic modeling approaches. Due to time constraints and the time plan of this thesis, only the BERTopic model is used for topic modeling. Since several methods could be used, it is important to mention why BERTopic is used and why the others are not. Egger and Yu (2022) found out that for short and unstructured texts like Twitter data, BERTopic can extract contextual information, and it offers the most potential compared to different embedding-based topic models like Top2Vec. According to their research, BERTopic has high versatility and stability across domains and supports different topic modeling variations. Like other embedding-based topic models, it allows multilingual analysis, and there is no need for preprocessing of the original data. However, the embedding approach might cause too many topics and outliers in some cases, which makes the results more challenging to interpret and should be examined in detail. Some long documents could occasionally involve multiple topics, but in this approach, every document is assigned to a single topic, which could be a disadvantage.

---

<sup>6</sup><https://business.twitter.com/en/blog/update-on-tweets-limited-usage.html>

### 3.2.1 The BERTopic Pipeline

According to Grootendorst (2022), BERTopic generates topic representations in six steps, shown in Figure 3.2. First, without preprocessing, each document must be embedded using a pre-trained model. In this thesis, the SBERT model is used, introduced by Reimers and Gurevych (2019). SBERT modifies the BERT model and derives semantically meaningful sentence embeddings, also from multilingual documents, allowing tasks like clustering or information retrieval via semantic search. SBERT also allows the selection of various pre-trained multilingual models supporting more than 50 languages<sup>7</sup>. This thesis uses the paraphrase-multilingual-MiniLM-L12-v2 model, which supports Turkish and is the fastest and one of the best performers among other multilingual models (Reimers & Gurevych, 2020). This part of the pipeline allows to do chunk embeddings, saving the results and using them later, making the big data analysis easier with restricted hardware availabilities.

Secondly, the dimensionality of these resulting embeddings is reduced to optimize the clustering process by the Uniform Manifold Approximation and Projection (UMAP) algorithm, which plays a massive role in big data analysis (McInnes et al., 2020). Afterward, these low-dimension embeddings are clustered using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) technique, and the resulting clusters consist of semantically similar documents. This step allows unrelated documents to be assigned as noise or outliers, which improves the result, and both of these steps can be influenced very much by changing the parameters of the algorithms.

Fourthly, each cluster is tokenized using a Vectorizer like CountVectorizer<sup>9</sup>. Together with the weighting of these tokens, where a custom class-based variation of the Term Frequency – Inverse Document Frequency (c-TF-IDF) algorithm is used, they are responsible for creating the topic representations. Like the previous step, this step also allows room to play with various parameters to tune the model, which affects the results considerably.

At last comes the topic representation, where the topics can be fine-tuned using

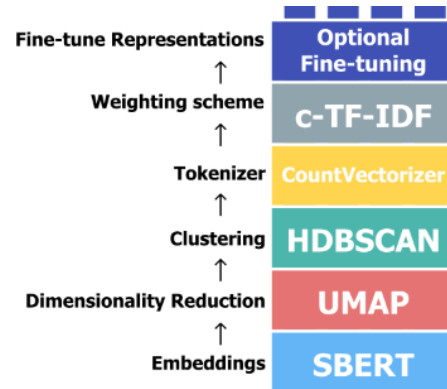


Figure 3.2: Default BERTopic Algorithm<sup>8</sup>

<sup>7</sup>[https://www.sbert.net/docs/pretrained\\_models.html](https://www.sbert.net/docs/pretrained_models.html)

<sup>8</sup><https://maartengr.github.io/BERTopic/algorithm/algorithm.html>

<sup>9</sup>[https://maartengr.github.io/BERTopic/getting\\_started/vectorizers/vectorizers.html](https://maartengr.github.io/BERTopic/getting_started/vectorizers/vectorizers.html)



various methods. This thesis uses KeyBERTInspired and Maximal Marginal Relevance (MMR) models<sup>10</sup>, which can be easily imported from the BERTopic library. They leverage the c-TF-IDF algorithm and weights keywords to represent the related topics. This thesis also leverages the power of LLMs by OpenAI, specifically GPT-4 Turbo, to better represent the resulting topics. One can also leverage other open-source LLMs, but most only support English or a few other languages.

#### 3.2.2 The Analysis Strategy

For hydration, this thesis used the *Social5* server provided by Research Group Social Computing<sup>11</sup>, which had sufficient hardware capabilities to run several scripts in parallel, and each of them was able to leverage multithreading, speeding up the process.

The result of the hydration was around 40 GB, which makes the analysis in one batch impossible. For this reason, hydrated tweet information was divided into smaller batches to make the analysis more manageable later. Before the topic modeling analysis, the resulting data is cleaned by extracting relevant information. That means only extracted tweets that include some text. These are then converted into lists and pulled into the analysis environment.

In the next step, to speed up the big data analysis process, this thesis used Google Colab<sup>12</sup> Pro that brings A100, V100, or T4 GPUs to usage, making the process considerably faster.

As mentioned before, because the data is big, it does not fit into the BERTopic model in one batch due to hardware restrictions. To tackle that issue, 2.5 million tweets are randomly sampled in proportion to the total number of collected tweets monthly, which makes up 1% of the total number of tweets in the #Secim2023 dataset.

The strategy used in this thesis was to first train the model with the sample tweets to find relevant topics in Turkish political discourse on Twitter. After that, all the remaining tweets were assigned a topic using the trained model without any need to train an additional model. This strategy also allows the analysis to be done in batches to make it smoother and more manageable.

The hydration, preprocessing, and analysis scripts mentioned are open-source and can be accessed on GitHub<sup>13</sup>. The analysis of this thesis is mainly based on two notebooks written by Maarten Grootendorst, father of BERTopic. The first is called

---

<sup>10</sup>[https://maartengr.github.io/BERTopic/getting\\_started/representation/representation.html](https://maartengr.github.io/BERTopic/getting_started/representation/representation.html)

<sup>11</sup><https://www.soc.cit.tum.de>

<sup>12</sup><https://colab.google>

<sup>13</sup><https://github.com/EfeSenerr/Thesis>

“Best Practices” and aims to get the best results<sup>14</sup>. In contrast, the second one is called “Big Data” and focuses on analyzing big data efficiently by optimizing and leveraging the power of the GPU<sup>15</sup>. The parameters used in the models are also inspired by them, resulting in minimal changes to optimize the results.

The changes also followed Grootendorst’s recommendations<sup>16</sup>. However, while making minimal changes in the parameters, the results can be highly affected by these changes. In this thesis, two approaches are made, resulting in two relatively different results while being similar in most topics. The results of this process are presented and discussed thoroughly in the following chapters.

In short, the following parameters changed from their default value are discussed.

In UMAP, `n_neighbors` is set to 25 from the default 15 to obtain a more global view of the embedding structure, resulting in larger clusters. Other parameters in UMAP followed the default values.

The `min_cluster_size` parameter in HDBSCAN is among the most critical parameters. It is set to 100 from 10 as a default to increase the minimum size of a cluster, thereby reducing the number of clusters the model generates. In the first analysis, this parameter was set to 20, and after getting too many topics, the parameter was increased to 100. For a massive dataset like #Secim2023, it could make sense to increase this parameter even more, but due to time restrictions, this thesis could not test this additionally. The `min_samples` parameter is usually automatically set to the value of `min_cluster_size`, but in this case, it is left in 20 to reduce the number of outliers generated. In the second analysis, the `prediction_data` parameter is set to True to predict new points later.

For passing out to CountVecorizer while training the model, one can prepare the vocabulary of the dataset used in training beforehand so that the tokenizer does not need to do the calculations later. It also reduces the necessary RAM during the training. This thesis creates a vocabulary of words and parses them such that they need to appear at least ten times in the data while removing the stopwords using Turkish stopwords from the NLTK library.

KeyBERT and MMR are used as the representation models of the topics. GPT-4 Turbo is later used for only the top 50 topics to reduce the costs constructed from it.

All these parameters mentioned above are passed to BERTopic along with its sampled text data and pre-computed embeddings. One can also play with some parameters of the BERTopic, but this thesis did not use any additional parameters. Some parameters do not play a critical role. For instance, the `min_topic_size` parameter in BERTopic is automatically set to the value of `min_cluster_size`. After several hours, this process

---

<sup>14</sup>[https://maartengr.github.io/BERTopic/getting\\_started/best\\_practices/best\\_practices.html](https://maartengr.github.io/BERTopic/getting_started/best_practices/best_practices.html)

<sup>15</sup>[https://huggingface.co/MaartenGr/BERTopic\\_Wikipedia](https://huggingface.co/MaartenGr/BERTopic_Wikipedia)

<sup>16</sup>[https://maartengr.github.io/BERTopic/getting\\_started/parameter%20tuning/parametertuning.html](https://maartengr.github.io/BERTopic/getting_started/parameter%20tuning/parametertuning.html)

resulted in topics that are the baseline of all topics in Turkish political discourse during the May 2023 elections. The resulting collected tweet data is then, in sequence, embedded and “transformed” by the trained topic model, which assigns topics to individual tweets. These results are demonstrated in the following chapter.

## 4 Results

As explained in the previous chapter, this thesis follows several topic modeling approaches to obtain the most representative result that can be analyzed. To emphasize the approaches again, the first approach follows Martin’s general recommendations and does not change any arguments. This BERTopic approach delivers around 7500 topics, which is too much to analyze and make sense of the results.

The main reason behind the mass number of topics is `min_cluster_size` remaining at 20, allowing around five thousand topics with less than five thousand tweets. The other main reason is the diversity of the analyzed data. To determine the topics, sample data of 1% of all collected tweets is used from a dataset that contains more than 200,000 daily tweets. When manually looking at the data, one can realize that even though the size is small, the topic can be related to a real-life event. Nevertheless, on the other hand, most of the topics can be combined or be in the outlier category.

With the experiences from the first approach, the second approach tries to minimize the cluster size of the topics to obtain a better result. When the `min_cluster_size` remained at 100, around 1000 topics were returned from the topic model. Similar to the first approach, this approach also covered the most significant topics. Because the cluster sizes are bigger with this approach, the topics are more challenging to label and categorize in one area. Although one can realize the similarities between these approaches in the top eight representative words, the representative tweets differ. These representative tweets are used in OpenAI prompt to label the topics, so the quality of these tweets is of high importance. In this approach, most representative tweets include three similar long texts with a high proportion of hashtags and mentions. That is also why almost half of the top 20 topics are demands or requests from the government, where many users and bots spam almost similar tweets to put the topic in trends, promoting politicians, parties, and specific agendas (Najafi et al., 2022).

After considering these arguments, this thesis mainly analyzes the results of the first approach, where one can realize topic trends consisting of specific topics that make comparing and analyzing the results considerably more straightforward.

It is essential to mention that outliers are not mentioned via graphs but are remembered during the results. Although some approaches in this tried to redistribute outliers, outliers exist for a reason. As Grootendorst notes, outliers are to be expected, and pushing the output to have no outliers may not correctly represent the data.

One disadvantage of following the default pipeline of BERTopic is that when the model is first trained, the biggest cluster is usually the outliers. However, when training the model and transforming the rest of the data, the outlier cluster is considered a usual topic cluster like any other. For instance, during one of the approaches, the model had more tweets in the outlier cluster while training than after transforming all the collected data, which might, on the one hand, cause lots of meaningless topics and, on the other hand, not accurately represent the data. There are several new approaches that BERTopic supports that overcome that issue. One can try *Online Topic Modeling* to incrementally train the model, or *Merge Multiple Fitted Models* approach, where multiple models are trained and, in the end, merged so that no information is lost. These approaches are discussed in Section 5.3.

The top results of the topic modeling are presented in Table 4.1, ordered by the number of tweets in the clusters. The Table 4.1 shows the topic labels on the left and their respective topic's eight representative words on the right. Although it would be hard to translate the representative words, the labels are labeled and then translated by GPT-4 Turbo for better interpretation. One can look at the GitHub repository<sup>1</sup> for the original results of the model for the top 50 topics. Since OpenAI models are not cost-free, they have been only used for the top 50 topics.

The top 15 topics represent various themes around Turkish political discourse. These topics include general political discussions, praise and criticism of both ruling and opposition parties, wishes and demands from the government, and the earthquake disaster.

Before diving deeply into the topics and understanding the big picture, clarifying some topics for better context would be beneficial. Topics like *Political Discussions*, *Political Discussions & Critism*, *Critism of Political Figures*,

*Praise of Political Figures* are more general and include tweets with positive and negative sentiments towards both the ruling and the opposition parties. One can realize that the representative words of the first topic include Twitter user names, which are active accounts that share news mainly about the elections. A high number of replies and a high level of interaction with their tweets made this category the most noteworthy.

However, topics like *Opposition Candidates*, *Opposition Coalition & Criticism*, *Opposition Figures & Preferences* and *Erdogan's Election Chance* focus on and express opinions about specific target groups, either the ruling party or the opposition parties and their candidates.

It is hard not to notice the fourth biggest cluster *Religious Wishes*, which includes tweets that cover aspects of prayer and Islam. While most of the representative sentences

---

<sup>1</sup>[https://github.com/EfeSenerr/Thesis\\_paper](https://github.com/EfeSenerr/Thesis_paper)

Topic Label	Representative Words
Political Discussions	aysedogan1955, yirmiucderece, hassa61, furkancerkes, sec-imtr2023, cenginyurt52, aykiricomtr, pushholder
Opposition Candidates	oy, kılıçdaroğlu, oyları, seçmeni, oylar, vermem, aday, seçimde
Erdogan & Nationalism	türk, turkey, türkiye, yüzyılı, türkiyeyüzyılı, stanbul, mil-liyetçisi, cumhuriyeti
Religious Wishes	allah, versin, eylesin, müslüman, razı, mekanı, namaz, dini
Political Discussions & Criticisms	cır, terketmek, denilince, uyruklu, ayrılmış, muharrem, vezir_yuce, bahtiyar_ergn
Criticism of Political Figures	siyaset, siyasetin, siyasetiniz, siyasetçi, siyaseti, batsın, siyasete, siyasette
Opposition Coalition & Criticism	erbakancayazan, örneğinde, yolunuzdan, iddialar, atıldılar, seçicez, ihtimalde, omurgasızlıktır
Treason Accusations	ülkeyi, ülkeye, ülke, ülkenin, ülkede, yağdanlıkları, unutturacaksınız, sığınmayın
Opposition Figures & Preferences	eu, me, que, aq, pra, amk, tudo, dedem
Retirement System	emeklilik, kısmi, kademeli, emekli, emeklilikte, prim, 5000
Earthquake & Demands	neticesiz, uzunca, deprem, depremin, süredir, atanma, yumuşak, depremde
Provocation & Discussions	kızmaz, ahhh, vura, yanarız, soğancı, mitink, ar-sibjk1903bjk, baskan
Praise of Political Figures	başkanım, başkan, başbakan, başkanın, cumhurbaşkanım, selo, mehmetfatihser5, mehmetersoy57
Teachers & Demands	öğretmen, öğretmenler, ataması, ilave, ücretli, öğretmenlere, cumhuriyetimizin, öğretmenlerin
Erdogan's Election Chance	rdoğan, erdoğan, erdoğanı, egemenlere, turda, tura, vereceğim, oyu

Table 4.1: Result of the topic modeling with outlier redistribution. The top eight representative words from MMR model for each of the top 15 topics with their respective topic label translated into English.

show tweets expressing condolences to various political figures who have lost their relatives, many of these tweets also include phrases of support and sympathy, such as ‘*God bless you*’, underlining solidarity.

The label *Treason Accusations* might sound absurd. The main reason for that is, again, the representative tweets that include some accusations against the political figures, blaming them with extremes like committing treason or being a terrorist. As one of the largest clusters, topics like this illustrate the polarization within the Turkish political landscape, which has been prominent over the years (Çevik, 2018). Another topic in this table, *Provocation & Discussions*, also highlights the polarization level in the Turkish

Twitter political discourse. The representative tweets are aggressive and show a high level of emotions.

*Erdogan & Nationalism* topic and its representative words show a significant degree of nationalist ideas. The label itself includes Erdogan because of the representative tweets mentioning Erdogan and the tenth representative word *Erdogan*, which is not seen in the table. However, it is essential to mention that both the opposition and ruling party coalition include parties based on Turkish nationalism that played a significant role in May 2023 elections, parties being IyiP and Nationalist Movement Party (MHP).

Although it is more explicit in other topic modeling strategies that this thesis tried, one can also see the volume of topics like *Teachers & Demands* and *Earthquake & Demands* in Table 4.1. The role of Twitter in the latter topic is crucial. Because of the earthquake on 6 February, all controversial telecommunication lines were disrupted, where Twitter has been widely used for rescue operations (Çevik & Aksoy, 2023).

All these mentioned topics are deeply analyzed in the following section, with graphs supporting understanding of the topics and their trends.

## 4.1 Detailed Analysis of Topic Categories

For simplicity and better understanding purposes, this thesis analyzes the topic modeling results in three categories: topics related to the ruling party visualized in Figure 4.1, the opposition visualized in Figure 4.2, and the remaining topics visualized in Figure 4.3.

These graphs visualize the selected topics in a normalized monthly distribution. Normalization has been done considering the top 50 topics. Otherwise, the interpretation of topics and their trends was almost impossible. One can not realize any trends in a frequency graph because the #Secim2023 dataset is not collected equally between months, demonstrated in Figure 3.1.

In all these graphs, the biggest cluster *Political Discussions* is shown as the baseline in a straight blue line to compare these different graphs and analyze the trends more smoothly.

### 4.1.1 Ruling Party Discourse Analysis

Beginning the analysis with the first graph 4.1, it covers the monthly distribution of the ruling party and Erdogan-related topics. It is essential to acknowledge the baseline topic *Political Discussions*, keeping its importance in the months to elections, massively increasing after the start of 2023 up to more than 10% of the top 50 tweets. That is normal, considering that in the months leading up to the elections, almost all the news was political, and rallies around Turkey were happening daily. Undecided opposition

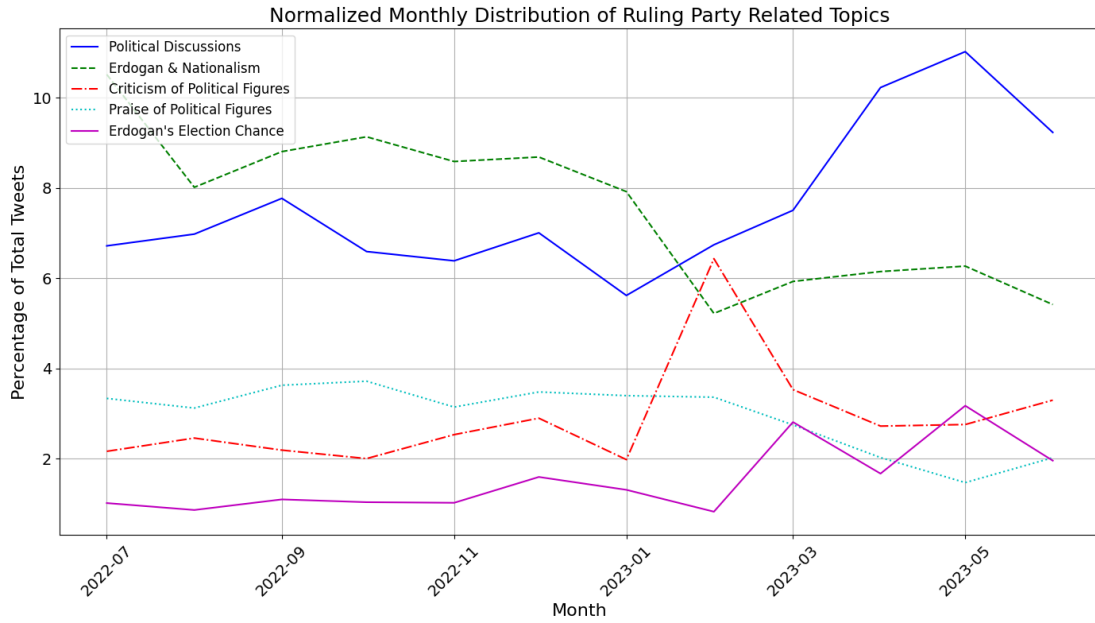


Figure 4.1: Percentage of ruling party related topics in the top 50 topics. The blue line represents the biggest cluster and is the baseline, and the other lines represent topics related to Erdogan and the ruling party.

candidacy, hot topics like the economic crisis and migration policy, and uncertainties in domestic and foreign policies led social media to a political discussion hub.

Having the baseline topic covered, the next topic with a green dotted line covers the topic *Erdogan & Nationalism*. As mentioned previously, it is essential not to converge Turkish Nationalist ideas with Erdogan and the AKP regime. There are nationalism-based parties in both the opposition and the ruling coalition. The opposition coalition's name is even 'Nation Alliance', in other words, the 'Table of Six'.

Turkish Nationalism has always been one of the base ideas of Turkish politics since the republic's foundation. The reason for that is now the opposition and once the founding party, CHP, who laid Turkey's six founding fundamental ideologies called the 'Six Arrows'. One of the *arrows* represents Nationalism. However, as Hotelling's model suggests, throughout the years, because CHP's ideology moved more toward the left wing, parties like AKP or newly established parties like ZP started to fill the empty right-wing spots (Caramani, 2020). Representative tweets showing support for different political figures using nationalist ideas also underline this aspect.

The migration policy was one of the hottest topics of the May 2023 elections. The government's attitude of welcoming more than six million refugees while getting



funding from the European Union for this purpose sparked lots of tensions amongst different ideologies and parties. In 2021, a new party called the Victory Party (ZP) was formed from the ranks of the opposition. It then rapidly increased its popularity as a mainly single-issue party focused on expelling refugees back to their countries (Esen, 2022a). ZP formed the second opposition coalition against Erdogan, and their candidate received more than 5% of the votes.

These events highlight why Nationalism has been one of the most trending topics during the months up to the election. However, in the new year, more specifically after the 6 February earthquake, other topics gained more significance, which can also be seen in the Figure 4.1.

One of the topics that gained relatively more importance is the *Criticism of Political Figures*, visualized as a red dotted and straight line in Figure 4.1. On the other hand, the topic *Praise of Political Figures*, visualized as a turquoise dotted line, has lost volume in time. The main explanation for that is the 6 February earthquake, which immediately changed the course of the elections by changing everybody's attention towards the southeast of Turkey.

The following reasons, also underlined by Çevik and Aksoy (2023), can explain the change in the trends of these topics, where criticism outweighed the praise of political figures on Twitter. The system lacked law enforcement until the earthquake, where the contractors of the building constructions sought to maximize their profits without taking the necessary precautions. The weakened state capacity led to the slowness of emergency response during the catastrophe. The military was missing, and there were no plans against this kind of emergency, or in other words, if there was a plan, it was not executed thoroughly. NGOs and volunteers organized themselves and worked together for days after the earthquake. Twitter was always the central platform for organizations and cry for help. Nonetheless, the government slowed down Twitter to prevent 'disinformation' from spreading, which worsened this process.

The last topic in this graph is *Erdogan's Election Chance*, visualized as a pink straight line. It can be seen that this topic remained relatively low and reached the bottom rock during the month of the earthquake, and in March, it increased a lot and outranked the other two topics in the election month. The reasons for this increase will be discussed thoroughly in Figure 4.2. However, in short, the lack of collaboration and not being decisive about a candidate among the opposition ranks make this increase feasible. Polls also support this argument (Çevik & Aksoy, 2023).

#### 4.1.2 Opposition Discourse Analysis

The analysis continues with the other side of the table, Figure 4.2, which covers the monthly distribution of the opposition-related topics. As a reminder, *Political Discussions*

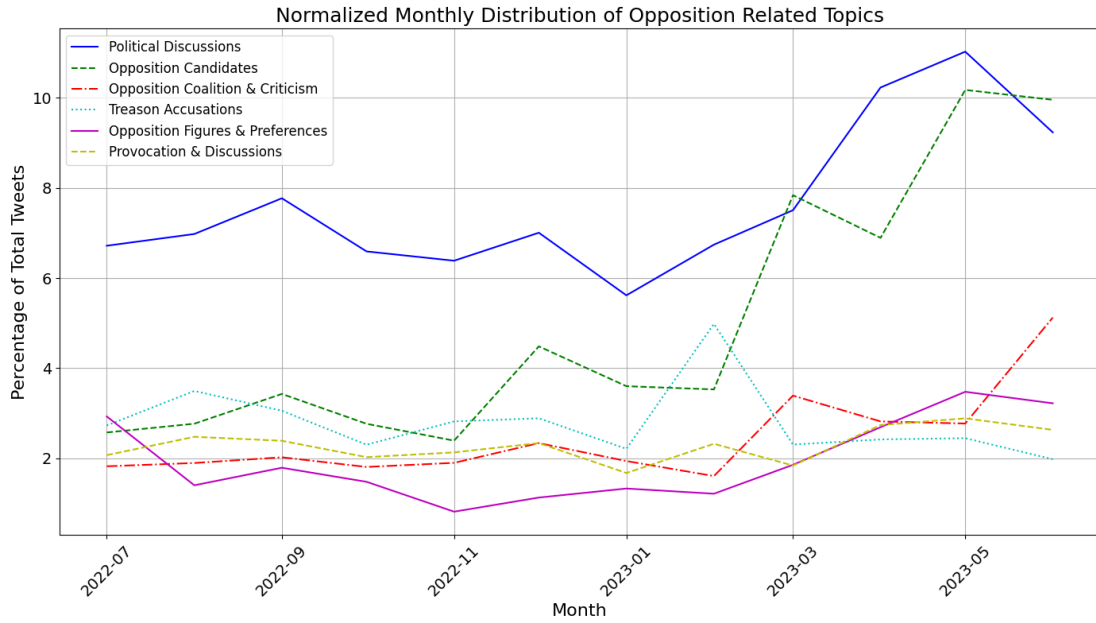


Figure 4.2: Percentage of opposition related topics in the top 50 topics. The blue line represents the biggest cluster and is the baseline, and the other lines represent topics related to opposition parties and their candidates.

topic can also be found in this graph, which is used as a baseline for comparing the graphs.

Continuing with the second topic, *Opposition Candidates* visualized as a green line in the graph. While the representative words cover topics like vote, the main opposition coalition Nation Alliance’s candidate Kılıçdaroğlu, and elections, the representative tweets are about people’s opinions about who they would vote for in the opposition ranks. For instance, some of them state that ‘*they would not vote for Kılıçdaroğlu, who does not have any victory against Erdogan for all these years.*’, or ‘*they would only vote for Mansur Yavaş*’, who is the mayor of the capital city Ankara.

It is essential to present some context here. The main opposition coalition, Nation Alliance, a.k.a. ‘Table of Six’, consists of six parties that signed a joint manifesto in 2022 outlining plans to abolish the presidential system and restore the rule of law and civil liberties by returning to the previous parliamentary system (Esen, 2022b). The coalition wanted to achieve this goal by selecting a joint presidential candidate. Like most people in the opposition ranks, (Esen, 2022b) saw selecting a joint presidential candidate as the most demanding task. Since the president holds immense powers in the Turkish presidential election system, the parties were uncertain whether to hand that authority

to one individual, so they sought to concentrate more on the alliance than the candidate. Due to these reasons, the coalition pushed this matter further back and could only start to discuss these matters after the earthquake, which is easily remarkable on Figure 4.2.

In March 2023, the majority of parties in the coalition then agreed on Kılıçdaroğlu, who was the leader of the main opposition party CHP and also played an essential role in the establishment of this coalition. After the agreement, the second largest opposition party İYİP also considered that although Kılıçdaroğlu's collaborative personality and focus on unity was an asset for a democratic transition, but stood behind the idea that this could weaken his electoral prospects against Erdoğan. As noteworthy evidence, one could show the most recent polls showing that he follows the popular mayors of Ankara, Mansur Yavaş; Istanbul, Ekrem İmamoğlu, and even the İYİP leader Meral Akşener (Esen, 2022b).

The trending level of the topic also aligns with the social movements at that time (Gültekin, 2023). Just before the earthquake, on 5 February, Kılıçdaroğlu's candidacy topic sparked again. There were campaigns against Kılıçdaroğlu on the streets and also on social media with a slogan 'Do not be a candidate Kılıçdaroğlu.'. People were holding pan cards with the same slogan in front of the CHP general headquarters and sharing it on Twitter (Ulaş, 2023). Some influencers even planned to march to the headquarters the next day to prevent his candidacy. That is one of the main reasons why the earthquake on 6 February changed and deeply affected the discourse of the May 2023 elections.

İYİP considered Kılıçdaroğlu lacked sufficient public appeal to win the elections. That is why Akşener opposed his candidacy and supported the mayors of Ankara and Istanbul. In short, after several days, the parties agreed on the vice presidency of these two mayors and continued the election campaign as a united front. All these events resulted in the topic *Opposition Candidates* naming itself as the most articulated topic in March on Twitter, surpassing the baseline. Even after the election month, the topic continued to be relevant, beating the baseline again. The discussion rallied then around whether Kılıçdaroğlu was 'the right candidate'. Several months after the elections, Kılıçdaroğlu lost the in-party elections after leading the main opposition party for 13 years (Gündoğan et al., 2023).

The next trending topic, *Opposition Coalition & Criticism* visualized as a red straight and dotted line, follows a similar tendency, gaining importance in March and continuing to grow even after the elections. However, this topic mainly focuses on direct criticism against the opposition parties. For example, the representative tweets in this cluster complain about Ali Babacan, the leader of DEVA, called him a puppet whose congressmen candidates were entering the elections in the same list as CHP. On the other hand, some tweets refer to Kılıçdaroğlu as a traitor to his own party for this very same reason.

The following topic, *Treason Accusations* visualized as a turquoise dotted line, shows accusations against those who did not support the same ideology as the writer of these Tweets. Blaming political figures with extremes like committing treason or being a terrorist is a standard tongue that brings the polarization in the Turkish political discourse to light. One can observe that this topic gained importance in February. The reason can be easily tied to the earthquake, such as people accusing the government of lacking law enforcement or not taking the necessary measures for the earthquake.

Represented as a pink straight line, *Opposition Figures & Preferences* highlights a mixture of tweets that include swear words, foreign words and sentences, and discussions around opposition figures. One can understand the abbreviations of Turkish swearwords in the representative words column in Figure 4.2, and it is likely that the BERTopic topic model could not make any sense of the words and merged the topic with foreign words and sentences. Another point the topic model had difficulties understanding is the word '*dedem*', translated to '*my grandfather*'. Grandfather nickname has been used by the younger generation on social media and overall to refer to the opposition candidate Kılıçdaroğlu. The nickname itself has more of a positive sentiment, emphasizing his kindness and, on the other hand, his age. Although this topic beat all other topics than the baseline topic in July 2022, it started to gain importance when the elections were getting closer and peaked during election month, possibly related to unexpected election results, including comments about the opposition.

The topic, *Provocation & Discussions*, marked by a yellow dotted line, indicates the polarization level in the Turkish Twitter political discourse, as stated before. The representative tweets are aggressive and show negative sentiments, with some of the tweets accusing political figures of being provocateurs. The trend of the topic continues steadily, increasing slightly just before the election month.

#### 4.1.3 Other Relevant Discourse Analysis

The analysis proceeds with the last graph, as presented in Figure 4.3, showcasing the monthly trends of other than ruling party or opposition-related topics. As the baseline, *Political Discussions* topic can be seen as the blue straight line in the graph. One can realize that multiple topics have crossed the baseline topic over the months up to the election.

Kicking things off with the topic *Religious Wishes*, visualized as a green line, the topic contains tweets that cover aspects of prayer and Islam. It features representative words like '*Allah*', '*müslüman*', and '*din*', which translates to '*Muslim*' and '*religion*'. Some representative sentences exhibit tweets praising political figures from different ranks, saying '*God bless you*'. Although the topic remained at the same level as the baseline and peaked during the month of the earthquake, afterward, it kept its importance while

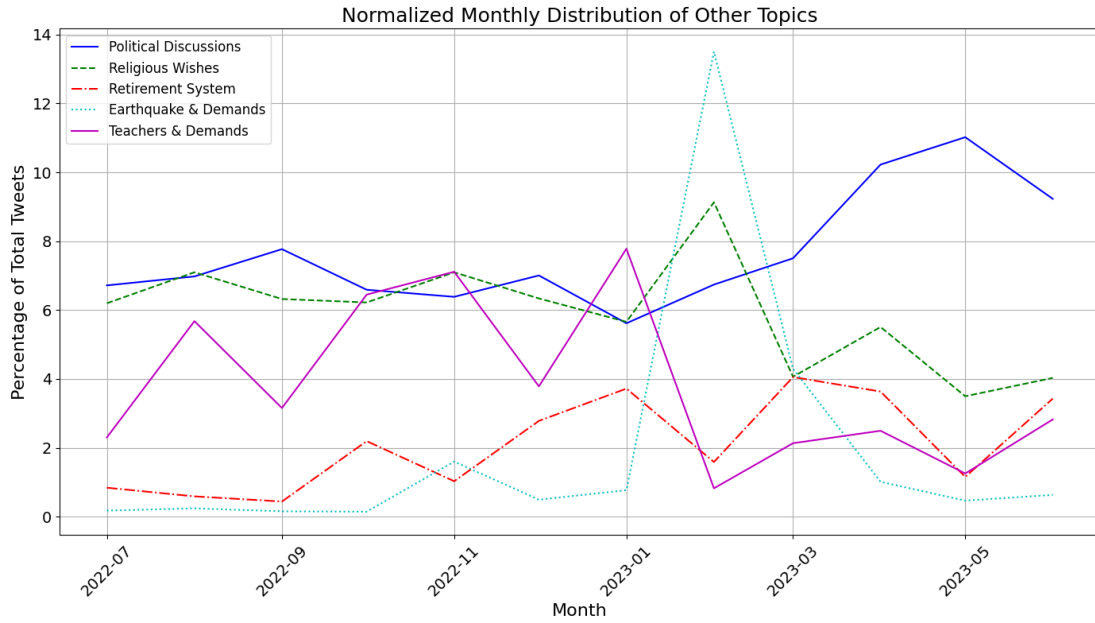


Figure 4.3: Percentage of other related topics in the top 50 topics. The blue line represents the biggest cluster and is the baseline, and the other lines represent topics like demands from the government and the earthquake.

falling in volume, leaving room for political discussions.

Illustrated with a red straight and dotted line, the *Retirement System* topic demonstrates tweets around the retirement system policies. Before the elections, one of the government's promises was to renew the last social security legislation in 1999. At the beginning of March, the government passed new legislation that softens existing rules on retirement age and retirement eligibility requirements, allowing almost more than 2 million workers to retire immediately (Erdem & Bişgin, 2023).

With the enactment of new legislation and a massive increase in the minimum wage, the government increased its spending massively, which resulted in increased public opinion just before the elections (Çevik & Aksoy, 2023). One can recognize that the topic peaked in the month of enactment, slightly lowering in volume afterward.

The topic *Earthquake & Demands* is self-explanatory, highlighting the earthquake on 6 February. In the earthquake, more than 53,000 people lost their lives, which created a social trauma and changed the course of the elections. Although the government mishandled the earthquake, as Çevik et al. (2023) shares, the opposition failed to get the votes from the region, and Erdoğan maintained his popularity in the region, where he received more than 70% the votes in some provinces (Michaelson & Narlı, 2023). As

mentioned previously, the role of social media in the earthquake is noteworthy. It is remarkable that the topic peaks massively in February. However, it is very interesting, and maybe also not expected, that it lost almost all of its volume in just two months, making just around 1% of the tweets afterward.

The last topic in the examination is *Teachers & Demands*. The representative sentences exist of sentences like '*We demand 100 thousand additional teacher positions.*'. One of the representative words '*cumhuriyet*', translated to '*republic*', points out that tweets exist of a rhyme, calling for an additional 100 thousand teachers at the 100. anniversary of the Turkish Republic. The massive volume of this topic can not be unseen, almost beating the baseline topic two times. It is essential to mention that, as Pfeffer et al. (2023) also points out, Twitter has been used through bots and different means of coordinated automated activities to show influence and power in the public discourse by making the relevant topic trending and keeping it before eyes.

Regardless, this topic aligns with the agenda of the Nation Alliance, where the coalition promised to create new positions for the teachers in one of their rallies (Haber, 2023).

## 4.2 Supplementary Visual Insights

This chapter presents additional visual representations of the topic model results.

The first visualization is inspired by Maarten Grootendorst, who visualized the topic model results from millions of Wikipedia pages in his Huggingface repository<sup>2</sup>. Like this thesis, he also used BERTopic for big data analysis and visualized the results in the end.

The Figure 4.4 presents the top 50 topics, which have been reduced to two-dimensional embedding space using UMAP. It pictures different clusters with different colors with respect to their sizes and does not contain any outliers. The white background could not be seen if the figure would have included outliers. The topics are illustrated by their top three representative words. One can realize that their top three representative words can differ from the ones from the previous table, Table 4.1.

This is because a different BERTopic pipeline has been used for this visualization. This pipeline includes pre-computing of required parameters of BERTopic, which are generally computed during the fitting procedure. Different from the training pipeline of the base topic model explained in Subsection 3.2.1, the first step starts with UMAP, which has a parameter `n_components` of two instead of five. That leverages the power of GPU by using the cuML library and reduces the topics to two-dimensional space.

---

<sup>2</sup>[https://huggingface.co/MaartenGr/BERTopic\\_Wikipedia](https://huggingface.co/MaartenGr/BERTopic_Wikipedia)

<sup>3</sup>[https://huggingface.co/MaartenGr/BERTopic\\_Wikipedia](https://huggingface.co/MaartenGr/BERTopic_Wikipedia)

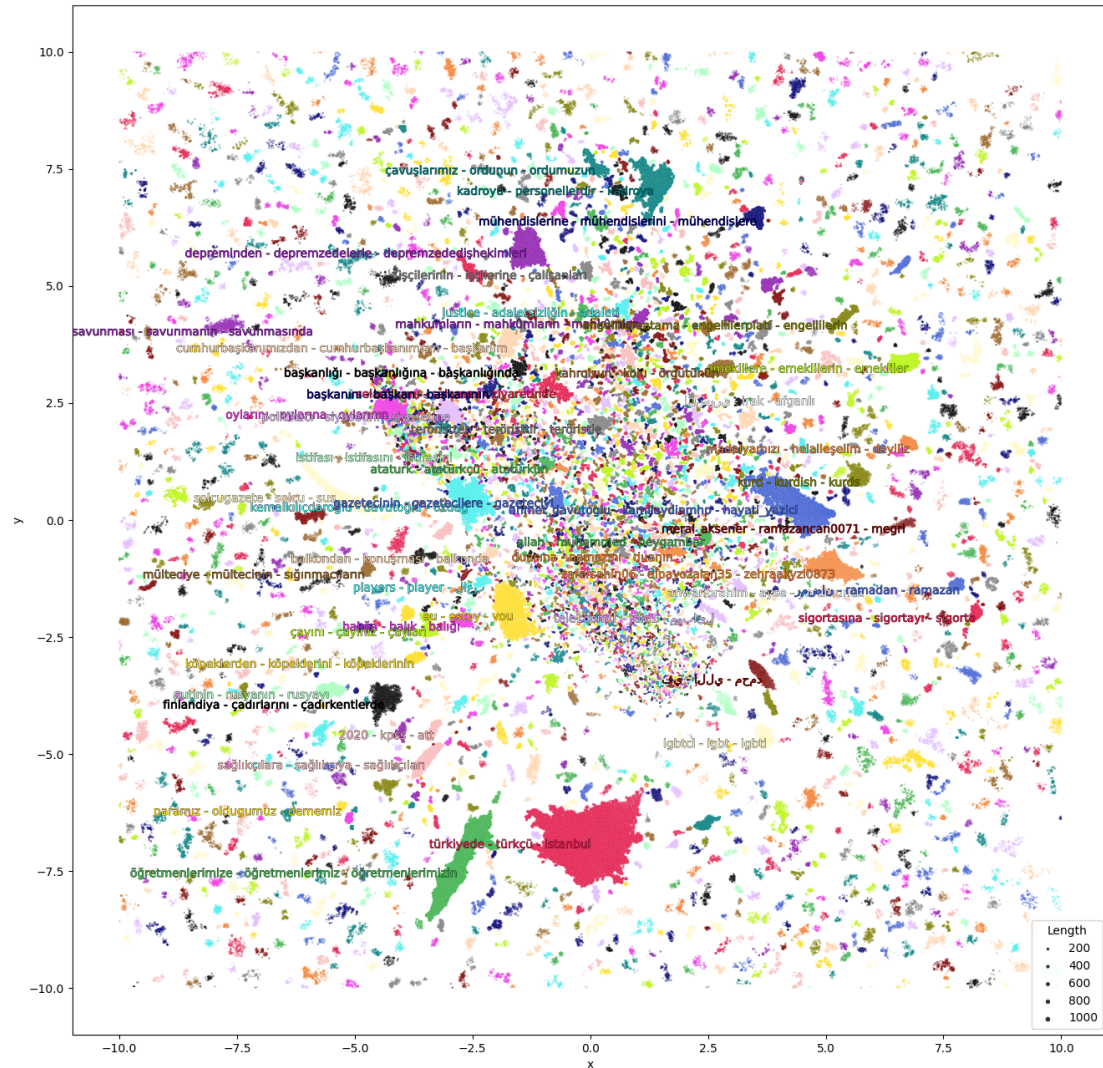


Figure 4.4: The top 50 topics visualized and reduced to 2-dimensional space using cuML’s UMAP, following Maarten Grootendorst’s guide.<sup>3</sup>

Afterward, HDBSCAN uses the two-dimensional reduced embeddings to pre-compute the clusters of the topics from scratch. These results can be fed to BERTopic to perform manual topic modeling<sup>4</sup>. If the model already has topic labels, it can use those to label them using various representation models.

With embeddings, reduced dimensionality and clusters on hand allow the speeding

<sup>4</sup>[https://maartengr.github.io/BERTopic/getting\\_started/manual/manual.html](https://maartengr.github.io/BERTopic/getting_started/manual/manual.html)

up of the fitting process. Additionally, Grootendorst presents a trick to get better results. He states that if one directly gives BERTopic the reduced embeddings, BERTopic would use those to create topic vectors, which could be better. For better results, he suggests giving the model the full embeddings and creating a custom dimensionality reduction class that will return the reduced embeddings. Further details can be found in the thesis GitHub repository.

The results of this pipeline are then used for visualizing the Figure 4.4, with the usage of `seaborn` and `matplotlib` library.

Some of the topics are relatively straightforward to pick up. The enormous red cluster on the bottom of the figure, titled as *'türkiyede – türkçü – istanbul'*, visualizes the topic GPT-4 Turbo labeled as *Erdogan & Nationalism* from the default pipeline. Next to that topic to the left, visualized as a green pipeline, is the topic *Teachers & Demands*.

The middle of this figure is considerably chaotic. Almost all of the clusters in the middle are related to the election. The topics from the political figures, votes, major Twitter accounts, and the resignation of known figures are covered. At the top of the figure, the cluster related to the earthquake can be seen as purple, titled as *'depreminde – depremlilerle – depremliler'*.

The following visualization, Figure 4.5, is about the hierarchical clustering of the top 50 topics of the sampled data and has been created using BERTopic function `visualize_hierarchy()`. The fitted BERTopic model has been used, and the parameter `top_n_topics` has been set to 50, which returns that figure. A ward linkage function performs the hierarchical clustering based on the cosine distance matrix between topic embeddings. Since the GPT-4 Turbo labeling has been done after fitting the model, and only for the top 50 topics, the labels in the figure show default labels after the fitting. One can look at the English labels of each top 50 topic in the GitHub repository of the thesis.

If the Figure 4.5 is deeply analyzed, the topics are divided into different colors, which are then connected together to build the total representation. The colors present 13 main subjects among the top 50 topic representations.

Beginning the exploration with the top topics, the green cluster represents the demands of military sergeants willing to be permanently employed. A batch of sergeants who worked on a limited contract in the state institutions for almost 36 years sought permanent employment.

The subsequent topic, visualized as red, addresses the demands of young people endorsing their internships to be recognized as the start date for their insurance, a critical factor determining the start of retirement eligibility.

The following three topics, visualized as turquoise, brown, and green, represent the concerns of agriculture engineers and their work appointments, the retirement system, teachers, and their demands, respectively. The last topic in the top cluster



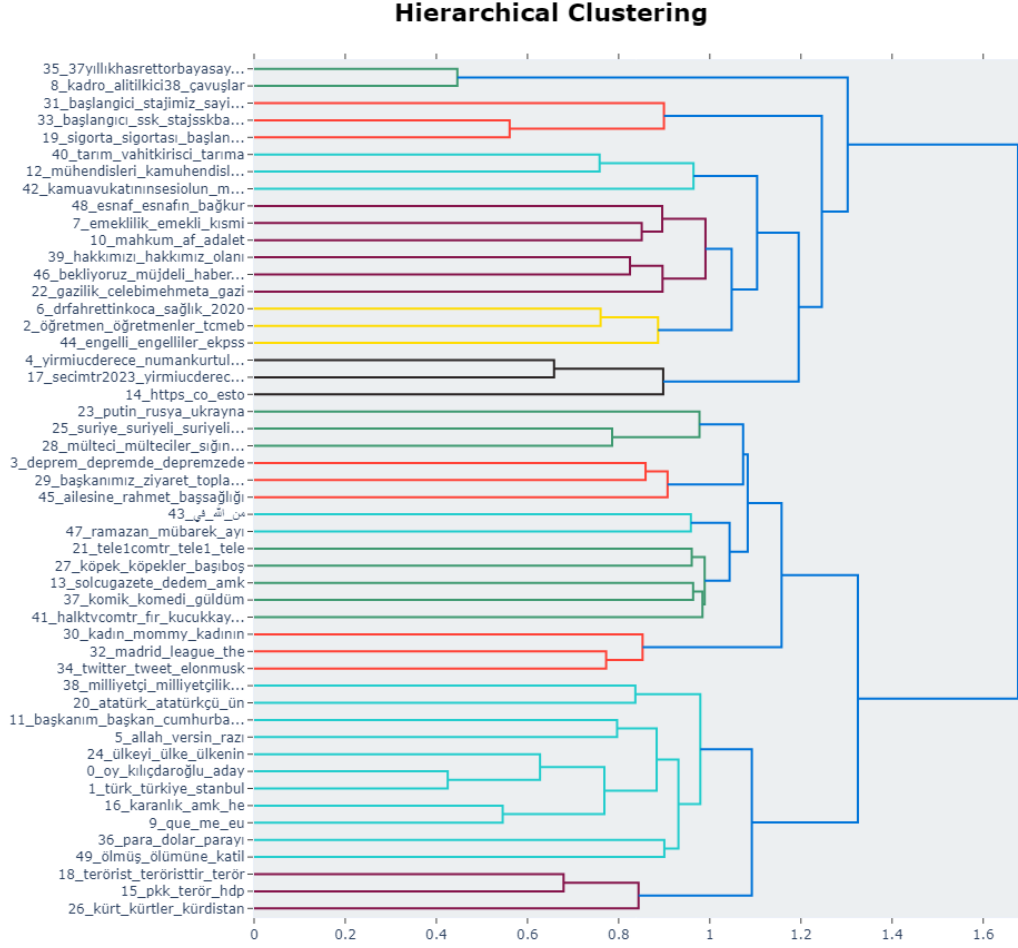


Figure 4.5: Hierarchical clustering of the top 50 topics using BERTopic functions.

covers general political discussions on Twitter.

Commencing with the bottom cluster, the green cluster illustrates topics related to foreigners, refugees, and specifically, people from Ukraine and Syria in separate subclusters. The volume of the topic related to Ukraine has stayed at a high level since the beginning of the dataset till the beginning of 2023, and the importance of the migration policy in the May 2023 elections was discussed in the previous section.

The succeeding two topic clusters, visualized as red and turquoise, demonstrate topics related to the earthquake, Arabic texts and Ramadan, respectively.

The next two clusters, colored as green and red, contain eight different partially political topics that are challenging to generalize. One topic that is considerably interesting to point out would be the topic titled as *27\_köpek\_köpekler\_başiboş*. It refers to problems caused by free-roaming dogs on the streets of Turkey, including increased dog attacks, hygiene concerns, safety risks, and traffic accidents, worsened by uncontrolled abandonment and the expansion of these animals in public spaces.

Another topic to highlight is the topic with *30\_kadın\_mommy\_kadının*. *kadın* can be translated to *woman* and both *kadın* and *mommy* refers to the İyİP leader Meral Akşener. Most of the young people on Twitter refer to her as a mom, underlining her warmth and sympathy.

The topic titled as *34\_twitter\_tweet\_elonmusk* contains representative words referring to Elon Musk's acquisition of Twitter, which started around April and concluded around October, the timeframe which this dataset also partially covers.

The second last cluster, portrayed as turquoise, covers various political topics from Nationalism to Atatürk, from political figures to presidential candidates, from religious wishes to treason accusations, from exchange rates to other economic topics.

The last cluster, illustrated as brown, covers topics related to terrorism and the Kurdish community. One can realize the topic titled as *15\_pkk\_terör\_hdp*, which includes in its top three representative words terrorism and Peoples' Democratic Party (HDP). HDP received around 9% of the votes in the May 2023 elections, with a significant portion of their support coming from the Kurdish community. The AKP government and other nationalists accuse HDP that it is linked to the banned Kurdistan Workers' Party (PKK), which the European Union also lists as a terrorist organization (Esen, 2022b), which we can also see from the topic representation being in the top 50 topics and its' representative tweets.

To remind it again, the extensive Turkish results and graphs, including the code, can be found on the GitHub repository of the thesis.

## 5 Discussion

The last chapter focused on results derived from the BERTopic topic model and analyzed them in detail. This chapter focuses on interpreting these results in a broader context and discusses this thesis's limitations and future work.

### 5.1 Interpreting the results in a broader context

x

### 5.2 Limitations

Citation test (Lamport, 1994).

### 5.3 Future Work

Citation test (Lamport, 1994).

# 6 Conclusion

## 6.1 Section

Citation test (Lamport, 1994).

Acronyms must be added in `main.tex` and are referenced using macros. The first occurrence is automatically replaced with the long version of the acronym, while all subsequent usages use the abbreviation.

E.g. `\ac{TUM}`, `\ac{TUM}`  $\Rightarrow$  Technical University of Munich (TUM), TUM

For more details, see the documentation of the acronym package<sup>1</sup>.

### 6.1.1 Subsection

See Table 6.1, Figure 6.1, Figure 6.2, Figure 6.3.

Table 6.1: An example for a simple table.

A	B	C	D
1	2	1	2
2	3	2	3

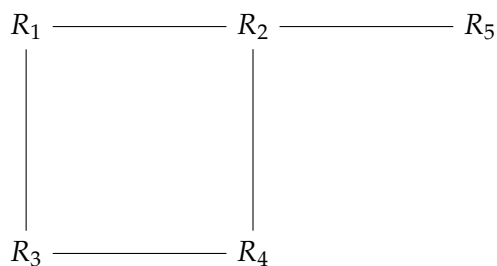


Figure 6.1: An example for a simple drawing.

---

<sup>1</sup><https://ctan.org/pkg/acronym>

Topic Label	Representative Words
Elections & Candidates	oy, kılıçdaroğlu, aday, seçim, erdoğan, tayyip, seçimi, istifa
Erdogan & Political Developments	türk, türkiye, stanbul, yüzyılı, turkey, türkiyeyüzyılı, ankara, cumhuriyeti
Political Agenda & Demands	yirmiucderece, numankurtulmus, aysedogan1955, cenginyurt52, herkesicinchip, meral_aksener, secimtr2023, hassa61
Religious Wishes & Political Figures	allah, versin, razı, eylesin, müslüman, etsin, din, rabbim,
Teachers & Demands	öğretmen, öğretmenler, tcmeb, ataması, prof_mahmutozer, 100, bin, kpss
Earthquake & Demands	deprem, depremde, depremlerde, depremin, yapıyayı, yumuşak, depremden, müstakil
Nation & Country	ülkeyi, ülke, ülkenin, ülkeye, ülkede, millet, milletin, milleti
Leading Figures	başkanım, başkan, cumhurbaşkanım, mansuryavas06, ekrem_imamoglu, başbakan, cumhurbaşkanı, sayın
Fair Trial & Amnesty Demands	mahkum, af, adalet, genelaf, 77, adil, adli, mahkumlar
Mixed Emotions	que, me, eu, não, no, dedem, aq, amk
Civil Servants & Demands	sırtlayan, hafızası, kurumların, yükünü, kamunun, 3600ekgösterge, devletine, umudumuz
Retirement System	emeklilik, emekli, kısmi, kademeli, prim, 5000, yaş, zorunlu
Reserves Demands	degildi, tercihimiz, dileğimiz, milletvekilim, etmektir, yedek, talebimiz, mehmetfatihser5
Medical Secretaries & Demands	drfahrettinkoca, sağlık, 2020, tıbbi, sağlıkçılar, sekreterlik, sağlıkçı, yönetimi,
Election News & Comments	secimtr2023, yirmiucderece, cumhuriyetgzt, vekilince, https, co, ozan_blk07, furkancerkes

Table 6.2: Result of the topic modeling with reduced topic distribution. The top eight representative words from MMR model for each of the top 15 topics with their respective topic label translated into English.

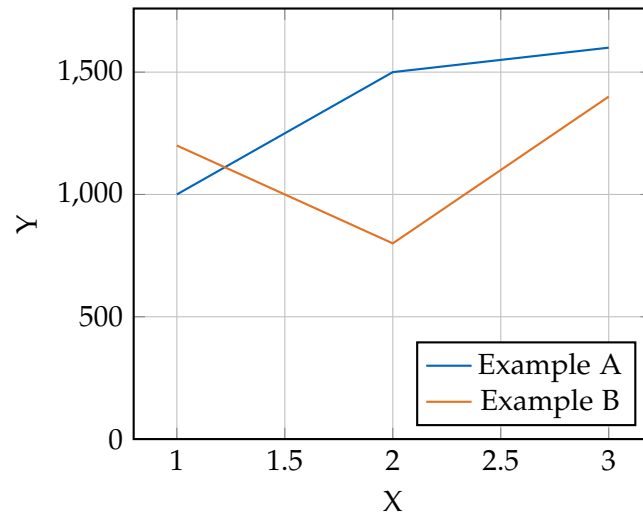


Figure 6.2: An example for a simple plot.

```
SELECT * FROM tbl WHERE tbl.str = "str"
```

Figure 6.3: An example for a source code listing.

# Abbreviations

**TUM** Technical University of Munich

**CHP** Republican People's Party

**AKP** Justice and Development Party

**MHP** Nationalist Movement Party

**HDP** Peoples' Democratic Party

**DEVA** Democracy and Progress Party

**IyiP** Good Party

**ZP** Victory Party

**SAADET** Felicity Party

**DP** Democrat Party

**DEVA** Democracy and Progress Party

**GP** Future Party

**NLP** Natural Language Processing

**PKK** Kurdistan Workers' Party

## List of Figures

3.1	Collected Tweets each month . . . . .	7
3.2	Default BERTopic Algorithm . . . . .	9
4.1	Normalized monthly distribution of ruling party related topics . . . . .	17
4.2	Normalized monthly distribution of opposition related topics . . . . .	19
4.3	Normalized monthly distribution of other related topics . . . . .	22
4.4	The top 50 topics visualized and reduced to two dimensional embedding space using cuML's UMAP . . . . .	24
4.5	Hierarchical Clustering . . . . .	26
6.1	Example drawing . . . . .	29
6.2	Example plot . . . . .	31
6.3	Example listing . . . . .	31



## List of Tables

4.1	Result of the topic modeling with labels and representative words. . . .	15
6.1	Example table . . . . .	29
6.2	Result of the topic modeling with labels and representative words. . . .	30

# Bibliography

- Alshaabi, T., Dewhurst, D. R., Minot, J. R., Arnold, M. V., Adams, J. L., Danforth, C. M., & Dodds, P. S. (2021). The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009-2020. *EPJ data science*, 10(1), 15. <https://doi.org/10.1140/epjds/s13688-021-00271-0>
- Anwar, A., Ilyas, H., Yaqub, U., & Zaman, S. (2021). Analyzing QAnon on twitter in context of US elections 2020: Analysis of user messages and profiles using VADER and BERT topic modeling. *DG.O2021: The 22nd Annual International Conference on Digital Government Research*, 82–88. <https://doi.org/10.1145/3463677.3463718>
- Ardan, M. (2023). 1994 Financial Crisis in Turkey. In B. Açıkgöz (Ed.), *Black Swan: Economic Crises, Volume II* (pp. 95–126). Springer Nature. [https://doi.org/10.1007/978-981-99-2318-2\\_7](https://doi.org/10.1007/978-981-99-2318-2_7)
- Atila, S. (2022). 3 kasım 2002’den bugüne akp ve erdoğan’ın 20 yıllık seçim tarihi. Retrieved February 3, 2024, from <https://medyascope.tv/2022/11/03/3-kasim-2002den-bugune-akp-ve-erdoganin-20-yillik-secim-tarihi/>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://doi.org/10.5555/944919.944937>
- Calma, J. (2023). Twitter just closed the book on academic research. Retrieved February 7, 2024, from <https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research>
- Caramani, D. (2020). *Comparative Politics*. Oxford University Press.
- Çevik, S. (2018). The Future of Opposition in Turkey. Overcoming Identity Politics Is the Key for Success. Retrieved February 25, 2024, from <https://www.swp-berlin.org/publikation/the-future-of-opposition-in-turkey>
- Çevik, S., & Aksoy, H. A. (2023). Political and economic implications of the Turkish earthquakes. <https://doi.org/10.18449/2023C19>
- Çevik, S., Aksoy, H. A., & Aydın, Y. (2023). The Opposition Alliance in Turkey: A Viable Alternative to Erdoğan? <https://doi.org/10.18449/2023C42>
- Contreras, K., Verbel, G., Sanchez, J., & Sanchez-Galan, J. E. (2022). Using topic modelling for analyzing panamanian parliamentary proceedings with neural and

- statistical methods. 2022 *IEEE 40th Central America and Panama Convention (CONCAPAN)*, 1–6. <https://doi.org/10.1109/CONCAPAN48024.2022.9997766>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7. <https://doi.org/10.3389/fsoc.2022.886498>
- Erdem, E., & Bişgin, E. (2023). Turkey - Changes to Minimum Retirement Age Requirement. Retrieved February 27, 2024, from <https://kpmg.com/xx/en/home/insights/2023/06/flash-alert-2023-121.html>
- Esen, B. (2022a). Post-2023 election scenarios in turkey. <https://doi.org/10.18449/2022C55>
- Esen, B. (2022b). The Opposition Alliance in Turkey: A Viable Alternative to Erdoğan? <https://doi.org/10.18449/2022C52v02>
- Freedom House. (2023). *Turkey* (tech. rep.). Retrieved February 20, 2024, from <https://freedomhouse.org/country/turkey/freedom-world/2023>
- Gritto, A. (2022). *Application of neural topic models to twitter data from German politicians (bat)*. Ludwig-Maximilians-Universität München. <https://doi.org/10.5282/ubm/epub.92617>
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. <https://doi.org/10.48550/arxiv.2203.05794>
- Gültekin, L. (2023). Kılıçdaroğlu'nun adaylığını niçin riskli buluyorum? - Diken. Retrieved February 28, 2024, from <https://www.diken.com.tr/kilicdaroglunun-adayligini-nicin-riskli-buluyorum/>
- Gündoğan, B., Yıldızalp, M., & Tosun, M. (2023). CHP'nin yeni genel başkanı Özgür Özel oldu. Retrieved February 28, 2024, from <https://www.aa.com.tr/tr/politika/chpnin-yeni-genel-baskani-ozgur-ozel-oldu-/3044068>
- Haber, T. (2023). Kılıçdaroğlu: Cumhuriyet'in 100. yılında 100 bin öğretmen ataması yapacağız. Retrieved February 28, 2024, from <https://www.trthaber.com/haber/gundem/kilicdaroglu-cumhuriyetin-100-yilinda-100-bin-ogretmen-atamasi-yapacagiz-765768.html>
- Ilyas, S. H. W., Soomro, Z. T., Anwar, A., Shahzad, H., & Yaqub, U. (2020). Analyzing brexit's impact using sentiment analysis and topic modeling on twitter discussion. *The 21st Annual International Conference on Digital Government Research*, 1–6. <https://doi.org/10.1145/3396956.3396973>

- Kaiser, J., Rauchfleisch, A., & Bourassa, N. (2020). Connecting the (far-)right dots: A topic modeling and hyperlink analysis of (far-)right media coverage during the US elections 2016. *Digital Journalism*, 8(3), 422–441. <https://doi.org/10.1080/21670811.2019.1682629>
- Klosowski, T. (2022). Why You should Delete (All) your tweets. Retrieved February 20, 2024, from <https://www.nytimes.com/wirecutter/blog/why-you-should-delete-your-tweets/>
- Lamport, L. (1994). *Latex : A documentation preparation system user's guide and reference manual*. Addison-Wesley Professional.
- McInnes, L., Healy, J., & Melville, J. (2020, September 17). UMAP: Uniform manifold approximation and projection for dimension reduction. <https://doi.org/10.48550/arXiv.1802.03426>
- Mejova, Y., Weber, I., & Macy, M. W. (Eds.). (2015). *Twitter: A digital socioscope*. Cambridge University Press. <https://doi.org/10.1017/CBO9781316182635>
- Michaelson, R., & Narlı, D. B. (2023). 'He works hard': Voters in Turkey's quake zone backing Erdoğan in runoff. Retrieved February 27, 2024, from <https://www.theguardian.com/world/2023/may/26/turkey-quake-zone-voters-backing-erdogan-in-runoff>
- Najafi, A., Mugurtay, N., Demirci, E., Demirkiran, S., Karadeniz, H. A., & Varol, O. (2022). #Secim2023: First Public Dataset for Studying Turkish General Election. Retrieved October 28, 2023, from <http://arxiv.org/abs/2211.13121>
- Najafi, A., & Varol, O. (2023). TurkishBERTweet: Fast and Reliable Large Language Model for Social Media Analysis. <https://doi.org/10.48550/arXiv.2311.18063>
- Ogan, C., & Varol, O. (2017). What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during gezi park. *Information, Communication & Society*, 20(8), 1220–1238. <https://doi.org/10.1080/1369118X.2016.1229006>
- Pfeffer, J., Matter, D., Jaidka, K., Varol, O., Mashhadi, A., Lasser, J., Assenmacher, D., Wu, S., Yang, D., Brantner, C., Romero, D. M., Otterbacher, J., Schwemmer, C., Joseph, K., Garcia, D., & Morstatter, F. (2023, April 11). Just another day on twitter: A complete 24 hours of twitter data. <http://arxiv.org/abs/2301.11429>
- Rabasa, A., & Larrabee, F. S. (2008). *The Rise of Political Islam in Turkey* (1st ed.). RAND Corporation. <https://www.jstor.org/stable/10.7249/mg726osd>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Retrieved November 7, 2023, from <http://arxiv.org/abs/1908.10084>
- Reimers, N., & Gurevych, I. (2020, October 5). Making monolingual sentence embeddings multilingual using knowledge distillation. <https://doi.org/10.48550/arXiv.2004.09813>

- Saç, E., & Çoban, T. (2023). Seçim sonuçlarından geriye bakmak: Anketler neden, nasıl yanıldı? Retrieved July 24, 2023, from <https://teyit.org/dosya/secim-sonuclarindan-geriye-bakmak-anketler-neden-nasil-yanildi>
- Schweter, S. (2020). BERTurk - BERT models for Turkish. <https://doi.org/10.5281/zenodo.3770924>
- Scott, M. (2023). How Turkey's Erdoğan uses social media to cling onto power. Retrieved February 3, 2024, from <https://www.politico.eu/article/recep-tayyip-erdogan-elon-musk-twitter-turkey-elections-social-media-power/>
- Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on twitter. *SAGE Open*, 10(2), 2158244020919447. <https://doi.org/10.1177/2158244020919447>
- Ulaş. (2023). CHP'li Killik'in, parti önünde 'Kılıçdaroğlu aday olmasın' pankartı açan gence tepkisi tartışma yarattı. Retrieved February 28, 2024, from <https://onedio.com/haber/chp-li-killik-in-parti-onunde-kilicdaroglu-aday-olmasin-pankarti-acan-gence-tepkisi-tartisma-yaratti-1126170>
- Uysal, N., & Schroeder, J. (2019). Turkey's twitter public diplomacy: Towards a "new" cult of personality. *Public Relations Review*, 45(5), 101837. <https://doi.org/10.1016/j.pubrev.2019.101837>
- Yerlikaya, T., & Toker, S. (2020). Social media and fake news in the post-truth era: The manipulation of politics in the election process. *Insight Turkey*, 177–196. <https://doi.org/10.25253/99.2020222.11>
- Yilmaz, I., & Bashirov, G. (2018). The AKP after 15 years: Emergence of erdoganism in turkey. *Third World Quarterly*, 39(9), 1812–1830. <https://doi.org/10.1080/01436597.2018.1447371>
- Zaharna, R. S., & Uysal, N. (2016). Going for the jugular in public diplomacy: How adversarial publics using social media are challenging state legitimacy. *Public Relations Review*, 42(1), 109–119. <https://doi.org/10.1016/j.pubrev.2015.07.006>
- Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021, February 28). Topic modelling meets deep neural networks: A survey. <https://doi.org/10.48550/arXiv.2103.00498>