# BERT Model Analysis of Dutch Perceptions on Countries during Political Upheaval

Annet Prudon (*5910447*)

Master's Thesis
MSc Applied Data Science
Faculty of Science
Utrecht University

Supervised by Dr. E. J. Meijers, and Dr. M. J. Smit

28 June 2024

Utrecht University

**Abstract**

This thesis explores how perceptions of countries are shaped and specifically influenced by political upheaval. Perceptions of places are often not created by direct experiences but through media, art, news, and social discourse. These perceptions are dynamic, continuously evolving due to the narratives and discourses generated by a wide range of actors. To investigate the perception of regions and the impact of political upheaval, this study utilizes BERT models (BERTopic and robBERT v2) for topic modelling and sentiment analysis on parliamentary debate transcripts from the Dutch House of Representatives. The findings indicate that rapid changes, such as the Russian invasion of Ukraine, are more easily detectable in terms of topics and sentiment compared to gradual changes like Brexit. However, the results are preliminary, suggesting a need for further research with refined methodologies. Future work should include applying the analysis to the entire dataset and comparing it with Ukraine and UK filtered datasets. Enhancing data preprocessing steps is also essential to improve topic interpretability. Additionally, sentiment analysis should be conducted using a robBERT v2 model trained specifically on parliamentary debates. These improvements aim to strengthen the conclusions of this thesis and facilitate broader research on the impact of political upheaval in other countries.

# TABLE OF CONTENTS

# 1 Introduction

Since World War II there have been multiple initiatives to create an unified Europe, to fulfill the need for a cohesive European identity. The identity of a place and its people is a well-researched subject. However, most research overlooks the perspective of one country from another, which is essential for understanding the perception of other nations and fostering a sense of unity (Peng et al., 2020). There is significant interest in this subject because the perception of one country by another greatly influences its development in various fields, such as economic growth, migration trends, and tourism dynamics. Additionally, it impacts the willingness to provide aid during difficult times (Agostini and van Zomeren, 2021, Chayinska et al., 2017). These perceptions, however, are not static; they are particularly susceptible to transformation as a result of political upheaval (Ivanov et al., 2017). This sense of unity is crucial for supporting each other during difficult times, such as in the event of a conflict.

Moreover, human perception is always determined by the mental shortcuts made to save time in comprehending our surroundings. As a result, we humans form stereotypes about every subject, which influence our thoughts and actions. With a similar mechanism we humans try to grasp the world and perceive its varieties. Our perception of countries is often sculpted not by firsthand experiences, but through the lens of media, art, news, and social conversations. These perceptions are in constant change, continually reshaped by the evolving narratives and discourses from an extensive variety of actors.

To gain a better understanding of how these perceptions of other countries are formed, Natural Language Processing (NLP) can be of service. Over the past decade, NLP has taken a lift. This expansion is driven by the availability of extensive text-based resources and the development of various NLP models. As a result, a substantial number of research papers have been published, focusing on methods to extract and interpret textual information using NLP models. These studies explore different data sources and model architectures. The former varies from news articles to legal documents, and Twitter discussions to political speeches. The latter varies from logistic regression to decision trees, and neural networks to transformers. This research specifically uses BERT type models to analyse transcripts of political debates.

Regarding data sources, most research on measuring people's sentiment or opinion on specific subjects relies on Twitter data (Ruz et al., 2020, Chaudhry et al., 2021, Ilyas et al., 2020). However, since the take-over and reorganisation of X, formerly Twitter, the data is no longer accessible and individual expressions have become more extreme. Therefore, the source of a new, reliable, and well-documented text-based source is needed. In addition to social media data, political speeches and legislative documents are increasingly being analysed using NLP models (Abercrombie and Batista-Navarro, 2020, Müller-Hansen et al., 2021, Greene and Cross, 2017, Gillioz et al., 2020, Silva et al., 2021). These types of research mainly seek information about the political agenda and focus of the relevant government. This thesis focuses on the parliamentary debates of the House of Representatives in the Netherlands. Despite the availability of transcribed Dutch debates, which are publicly accessible through an API, there is limited existing research on this topic.

Regarding model architectures, there are numerous new techniques that use (slightly) different assumptions and methods to extract information from textual data. One of the most well-known and widely used models is BERT, a deep learning model introduced by Google in 2018. Consequently, it has undergone extensive testing and various model variations have been developed. Specifically, a topic modelling technique BERTopic (Grootendorst, 2022) and a large pre-trained general Dutch language model robBERT (Delobelle et al., 2020). The former can implement Dynamic Topic Modelling (DTM), which is required for this research due to the dependency between topics over time. The latter is trained on Dutch text and outperforms other sentiment analysis (BERT) models.

Therefore, this research aims to use the described BERT models to extract information from transcripts of parliamentary debates in the House of Representatives in the Netherlands. The goal is to gain insight into the evolution of public perception. Specifically, to gain a deeper understanding of how politicians discuss a particular country and how this discourse changes after political upheavals. This will be a first step towards obtaining knowledge about the imaging of a region using BERT models.

**Research question:** *What is the effect of political upheaval on the way of speaking about a country by analysing changes in discussed topics and sentiments through political debates?*

To be specific, this research focuses on the defining moments in Ukraine and United Kingdom (UK), respectively the Russian invasion of Ukraine and the Brexit. The former involves rapid change with a military operation where life and death are at stake. The latter is a slower process with a greater focus on financial and trade matters. Despite their inherent differences, both topics are highly relevant in the Netherlands and are therefore frequently discussed in the House of Representatives.

This research contributes to the knowledge gap in multiple ways. First, parliamentary debate transcripts are rarely used in scientific research, despite being easily accessible in the Netherlands. It is indeed a very different data source compared to Twitter data. In political debates, representatives speak with careful consideration and substantiated opinions, unlike the impulsive and extreme rhetoric often found on Twitter. Additionally, the House of Representatives should provide a good representation of the general public opinion since its members are democratically elected. Second, testing BERT models on Dutch textual data is beneficial because most NLP research primarily focuses on English texts. Especially since the two recently developed models used in this research (`BERTopic` and `robBERT`) are easily applicable and fitting for the Dutch language (more on this in the Methods section). Third, these BERT models are not often tested on parliamentary debate data. Fourth, this research helps to take the first steps in understanding the perception of regions and the impact of political upheaval on these perceptions.

This thesis is further organized as follows. In the next section, Data, the data and its source are discussed. Thereafter, in Preprocessing the preprocessing of the data is described. In Methods the implementation of the BERT models, and evaluation measures are discussed. Then, the results of the models are presented in Results, after which the implications and limitations of this research are discussed in Discussion. The

overall conclusion is given in the Conclusion part of the thesis. In the Appendix, additional results (A) and the programming code (B can be found.

# 2 Data and Methods

## 2.1 Data

The data is obtained from the Open Data Portal from the House of Representatives, where all parliamentary debates between 25 June 2013 until 7 March 2024 are transcribed (Tweede Kamer der Staten-Generaal, n.d.-b). By using the OData API (Tweede Kamer der Staten-Generaal, n.d.-a), all reports were downloaded as XML file and transformed to a Python Pandas DataFrame with the following attributes (see Table 1). All rows with no text or title are deleted.

Table 1

*Attributes extracted from the reports.*

| **Attribute** | **Content** |
|---|---|
| *Text* | Text said by speaker in one turn |
| *Speaker* | Speaker |
| *Date* | Date of debate |
| *Activity* | Debate part |
| *Title* | Title of the debate part (subject of debate) |
| *Speaking turn* | Index of speaking turn for the speaker |
| *Party* | Party of the speaker |
| *Cabinet* | Boolean to mark when speaker is (prime) minister or Secretary of State |

Several key moments occurred throughout the data collection period in both Ukraine and the United Kingdom. The analysis results will be compared between these moments for both subjects (see Table 2).

Table 2

*Defining moments happening in the subjects Ukraine and UK.*

| **Ukraine events** | **Date** | **UK events** | **Date** |
|---|---|---|---|
| Annexation Krim | *18 March 2014* | Referendum | *23 June 2016* |
| MH17 | *17 July 2014* | Brexit | *31 January 2020* |
| Referendum | *6 April 2016* | | |
| Invasion by Russia | *24 February 2022* | | |

## 2.2 Preprocessing

For data preprocessing, the following steps are taken. To optimize runtime efficiency within the limited timeframe of this research, an initial filtration is performed to reduce the dataset size. The Ukraine dataset retains the part (based on *Title*) of the debate (based on *Date*) where the text contains the string "*Oekra*" (not "*Oekraïne*" due to the diaeresis). Similarly, the United Kingdom (UK) dataset retains the debate parts containing the strings "*Brexit*" or "*Verenigd Koninkrijk*" (both to ensure that all discussions about Brexit that do not use the word *Brexit* are kept). Additionally, only debates where *Oekraïne* or *Brexit/Verenigd Koninkrijk* appear more than three times are retained. This is done to eliminate the possibility of the term being used randomly rather than being the actual subject of discussion. Next, to clean the data, all irrelevant political technicalities without substantive discussion are filtered out from the Ukraine and UK datasets: "*Opening*", "*Mededelingen*", "*Stemmingen*", "*Afscheid*", "*Hamerstukken*", "*Regeling van werkzaamheden*". At last, all three datasets (all reports, Ukraine reports, and UK reports) are saved as an csv file. The Ukraine dataset consists of 251 debate parts, and the UK dataset of 151 debate parts (refer to Table 3 for distribution across the years).

Table 3

*Number of debate parts in both Ukraine and UK dataset.*

| **Year** | *2013* | *2014* | *2015* | *2016* | *2017* | *2018* | *2019* | *2020* | *2021* | *2022* | *2023* | *2024* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ukraine** | 4 | 17 | 20 | 21 | 12 | 11 | 14 | 4 | 10 | 92 | 38 | 13 |
| **UK** | 1 | 11 | 11 | 16 | 17 | 21 | 21 | 16 | 20 | 12 | 3 | 2 |

To prepare the data for topic modelling and sentiment analysis, the following steps have been taken. First, all text is tokenized. Specifically for the UK dataset, the strings *Verenigd Koninkrijk* are replaced with *VERENIGDKONINKRIJK* to prevent them for vanishing during tokenization. Second, all punctuation is deleted, except for hyphens, *ã*, and *ï* due to its occurrence in relevant words (for example, Ukraine is written as *OekraÃïne*). Third, all stop words are deleted, and to complete the stop words list, an additional 191 words have been incorporated (Countwordsfree, n.d.). The following words are also included to the stop words list due to their frequent use in political politeness and terminology: *heer*, *mevrouw*, *staatssecretaris*, *minister*, *voorzitter*, *kamer*, *kabinet*, *motie*, *vraag*. Fourth, with the use of part of speech tagging from `spacy` (specifically the Dutch trained pipeline `nl_core_news_sm`), all auxiliaries are removed due to their frequent use but meaninglessness. Fifth, all empty documents are deleted. Sixth, all reports are grouped by *Title* and *Date* such that each row is one debate part (i.e. document).

## 2.3 Methods

### 2.3.1 Topic modelling

For topic modelling in this type of research, dynamic topic modelling (DTM) is recommended (Blei and Lafferty, 2006, Müller-Hansen et al., 2021, Silva et al., 2021). Unlike Latent Dirichlet Allocation (LDA), which assumes that topics are independent (i.e. uncorrelated), DTM is better suited for comparing topics over time. This DTM is a dynamic extension to LDA (Blei and Lafferty, 2006) and addresses the issue of topic evolution by allowing topics to change and evolve over time. This research uses BERTopic model for the DTM , which has proven to create coherent topics (Hadiat, 2022). Specifically, BERTopic uses pre-trained transformer-based language models to create document embeddings, cluster them, and build topic representations using the class-based TF-IDF technique (Grootendorst, 2022).

To implement BERTopic, the model `bertopic` is imported and the language is set on Dutch. The parameter fine-tuning and specifications used are as follows (Grootendorst, 2024):

- `verbose` is set on `True` to be able to track the stages of the model.
- `global_tuning` is turned off, because this will create a global representation of the topics which is not needed for this research. Only `evolutionary_tuning` is turned on since this research aims to gain insight in how the topics evolve over time.
- *Date* is reformed to a *Year* attribute such that no timestamp bins are needed. These bins will distribute all documents into equal-sized segments. However, because the data is not evenly balanced across time, using bins may lead to timestamps that are difficult to interpret. Therefore, *Year* is used instead of *Date*.
- `min_topic_size` is set on 5 because the input is a relatively small dataset.
- `nr_topics` is set to ”auto” because the model already reduces the calculated topics using a Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) and determining the number of topics beforehand is tricky.
- To diversify the topic representations, `MaximalMarginalRelevance` (MMR) from `bertopic.representations` is used, which is useful for the interpretation of the topics. However, using MMR does slightly reduce the coherence score. Nevertheless, this difference is not significant, and it enhances the interpretability of topics, thereby facilitating analysis. Therefore, the diversity parameter is set to 0.1.

To be able to answer the research question, the topics will be plotted against the years along with their frequency across those years. To evaluate the performance of the model, the *U_mass* coherence score (i.e. how often two words occur together in the corpus) is calculated, which is best when closest to zero (Mimno et al., 2011).

#### 2.3.2 Sentiment Analysis

For the sentiment analysis, the robBERT v2 model (Delobelle et al., 2020) is used. The robBERT v2 model is one of the best performing pre-trained language models in Dutch and outperforms other versions of BERT models, especially on sentiment analysis (Delobelle et al., 2020). The pre-trained version `DTAI-KULeuven/robbert-v2-dutch-sentiment` on Huggingface (DTAI-KULeuven, 2022) is trained on book reviews, to classify them positive or negative with a sentiment score. So, for this research, this model is utilized to classify the debate parts as positive of negative and compute the sentiment scores. To evaluate the performance of the analysis, a random debate segment is selected from each year, and three random strings within that segment are assigned sentiment labels and scores. This approach allows for checking the sentiment evaluation against intuitive feelings associated with each string.

# 3 Results

## 3.1 Topic modelling

### 3.1.1 Ukraine

The BERTopic model generated 25 topics for the Ukraine dataset, see Table 9 in Appendix to view all topics and its coherence scores. The overall coherence score ($U_{mass} = -0.791$) is quite good, since it is smaller than $-1$ and therefore close to zero.

To analyse the generated topics, all topics are grouped on similarity. For example, topics 17 and 18 are grouped under *Condemning Russian Actions* since the following words are used for these topics respectively: *Russia*, *Ukraine*, *sanctions*, *Netherlands*, *criminal*, and *sanctions*, *Netherlands*, *European*. Likewise, some topics are grouped together in *Irrelevant* because they do not seem to relate to any defining moments for Ukraine. For instance, topic 1 contains the following words: *farmers*, *nature*, *animals*, and *agriculture*. All topic groups can be seen in Table 4.

To lay these groups next to the timeline of defining moments in Ukraine, the words per year linked to the topic are checked. First, *Condemning Russian actions*, in which topic 17 is the most interesting because in 2015 and 2022 the words *gas* and *Ukraine* are linked. However, it does not have the best coherence score ($U_{mass} = -2.091$). Second, in *Safety* no linked words appear to have anything to do with the defining moments (*ATMs* and *virus*). Third, for *Refugees*, topic 9 ($U_{mass} = -0.209$) contains different countries from which refugees cause problems in the Netherlands before 2022 and in 2022 *Ukraine* appears, which could be due to the Russian invasion in 2022 (see Table 5). Fourth, in group *Energy crisis* in 2017 the words *sanction* and *list of sanctions* appear which could be related to the sanctions for Russia. Additionally,

Table 4

*Topic groups for Ukraine dataset.*

| **Group** | **Topics** |
|---|---|
| *Condemning Russian actions* | 17, 18 |
| *Safety* | 6, 20, 21, 22 |
| *Refugees* | 9, 11 |
| *Energy crisis* | 19 |
| *Trade* | 15, 23 |
| *Military* | 8, 13 |
| *Referendum* | 5 |
| *Irrelevant* | 1, 2, 3, 4, 7, 10, 12, 14, 16, 24, 25 |

there is a gap from 2018 to 2022 (see Figure 1) where the topic is not discussed whereas in 2022 and 2023 words related to the energy crisis caused by the lack of trading with Russia appear (*fuel*, *price gap*). Fifth, in *Trade* there are no words that can be linked to the defining moments, except for one report in 2015 mentioning *Ukraine*, *Russia*, and *Europe*. Sixth, in *Military* in 2022 and 2024 *Ukraine* is linked to the topics as expected. However, in 2023 only words related to Israel are linked to the topics which is highly unexpected. Seventh, for *Referendum* it is quite unexpected that there are respectively 13 and 9 debates about this topic (see Figure 1). Lastly, the words linked in the *Irrelevant* group are clearly irrelevant to the defining moments (*ISIS*, *media law*, *biodiversity*). However, these irrelevant topics do have relatively high coherence scores, ranging from $-0.083$ (Topic 3) to $-1.164$ (Topic 16).

Table 5

*Words linked to topic 9 (Refugees) per year.*

| Year | Frequency | Words |
|---|---|---|
| 2014 | 1 | kinderpardon, kinderen, criterium, gemeentetoezicht, regering |
| 2015 | 1 | turkije, europese, vluchtelingen, nederland, libanon |
| 2016 | 2 | griekenland, vluchtelingen, nederland, maatregelen, relocatie |
| 2017 | 1 | asielzoekers, afgewezen, landen, marokko, illegaliteit |
| 2018 | 2 | azmani, politie, migratie, pact, nederland |
| 2021 | 1 | school, gemeentes, inspectie, ouders, nederland |
| 2022 | 6 | oekraïne, opvang, mensen, vluchtelingen, gemeenten |
| 2023 | 4 | zorg, dementie, nederland, aantal, zorgkosten |

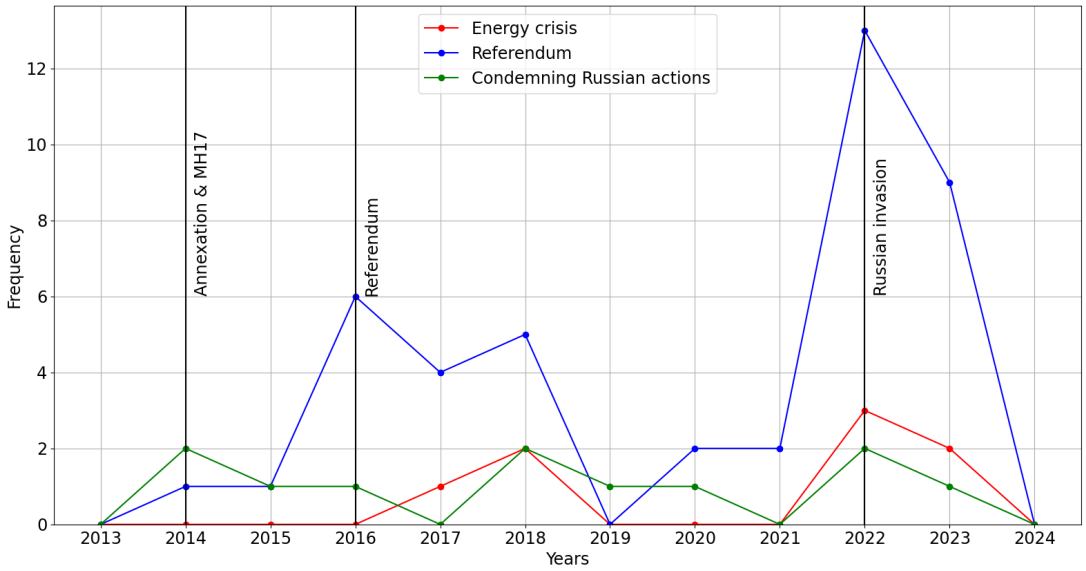


*Figure 1.* Occurrences of topics 19 (Energy crisis, $U_{mass} = -0.457$), 5 (Referendum, $U_{mass} = -0.280$), and 17 (Condemning Russian actions, $U_{mass} = -2.091$) per year

### 3.1.2 United Kingdom

The BERTopic model generated 23 topics for the UK dataset, see Table 10 in Appendix to view all topics and its coherence scores. The overall coherence score ($U_{mass} = -0.910$) is quite good, since it is smaller than $-1$ and therefore close to zero.

Like the Ukraine dataset, group are created on similarity. For instance, topics 2 and 17 are grouped in *International conflicts* due to the words being related to conflicts outside of the Netherlands: *Assad*, *Israel*, *Ukraine*, and *criminal*. On the other hand, topic 4 is grouped as irrelevant since it contains less meaningful words: *pact*, *naturally*, *think*. All topic groups can be seen in Table 6.

To analyse all topics, the linked words are viewed and determined whether they fit into the timeline of defining moments. First, *Health care and COVID*, topic 3 is linked to words such as *Ebola*, *children*, and *parents* before 2020, whereas in 2020 and 2021 it is as expected linked to *virus* and *testing* with way more reports (see Figure 2). After 2021 it is linked to *questions*, *postcovid*, and *report* (see Table 7). Topic 5 is focused on the organisation of health care by being linked to *medicine*, *health insurance*, and *antibiotics*. Second, topic 6 is mostly linked to words like *farmers*, *animals*, and *nature*, with a significantly higher number of reports in 2021 (see Figure 2). Additionally, topic 10 is focused on words related to sustainability: *plastic*, *biomass*, and *waste*. Third, topic 2 is focused on the turmoil in the Middle East (*Israel*, *Syria*, *Iraq*), and *China* and *Turkey*. Whereas topic 17 is focused on the turmoil in *Ukraine* and *Russia*, only appearing

Table 6

*Topic groups for UK dataset.*

| **Group** | **Topics** |
|---|---|
| *Health care and COVID* | 3,5 |
| *Agriculture and climate* | 6, 10 |
| *International conflicts* | 2, 17 |
| *Financial* | 11, 13, 15, 21, 22 |
| *Infrastructure* | 8, 12 |
| *Media and Journalism* | 19, 20 |
| *Energy* | 16 |
| *Irrelevant* | 1, 4, 7, 9, 14, 18, 23 |

between 2015 and 2019 (see Figure 2). Fourth, for *Financial* group, it is most interesting to see that there is a peak in 2017, linked to words as *companies*, *shareholders*, and *sector*. Fifth, for group *Infrastructure*, it is quite uninterpretable due to randomness of linked words to both topics: *spy*, *airport*, *school advise*, and *asphalt*, *nuclear power plant*, *Vietnamese*. Sixth, for group *Media and Journalism*, topic 19 is very random: *education*, *radio*, *sport*, whereas topic 20 focuses on *Turkey* and *freedom of speech* (see Figure 2). Seventh, topic 16 is also linked to quite random words, *mining*, *ProRail*, and *energy*. Lastly, for the irrelevant group, all topics are quite evenly distributed and do not change over time as one would expect along side the defining moments in the UK.

Table 7

*Topic 3 over time.*

| Year | Frequency | Words |
|---|---|---|
| 2014 | 1 | fritsma, overleg, ebola, geboorteakte, kinderen |
| 2015 | 2 | advies, kinderen, inspectie, salarissen, banken |
| 2019 | 1 | vaccinatiegraad, ouders, gezondheidsraad, vaccineren, vaccinaties |
| 2020 | 20 | testen, virus, vragen, aantal, natuurlijk |
| 2021 | 18 | natuurlijk, zorg, vaccineren, testen, vragen |
| 2022 | 2 | dank, besmettingen, vragen, aantal, zorg |
| 2023 | 3 | postcovid, maatregelen, rapport, minister, vragen |

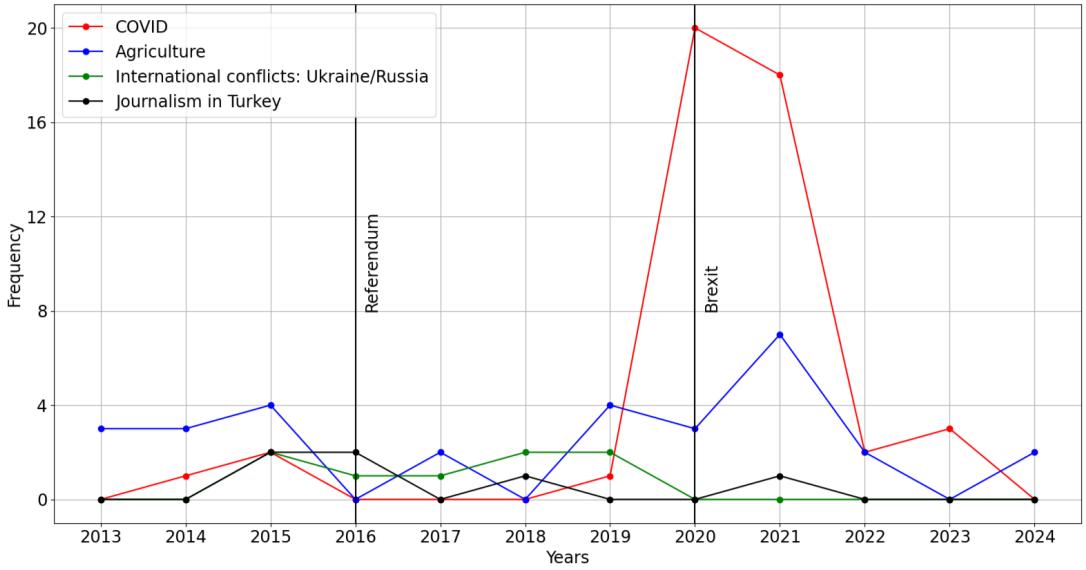


*Figure 2.* Occurrences of topics 3 (COVID, $U_{mass} = -0.630$), 6 (Agriculture, $U_{mass} = -0.586$), 17 (International Conflicts, $U_{mass} = -1.433$), and 20 (Journalism in Turkey, $U_{mass} = -0.319$) per year

## 3.2 Sentiment Analysis

### 3.2.1 Ukraine

The pre-trained version `DTAI-KULeuven/robbert-v2-dutch-sentiment` of the robBERT v2 model calculated the label (*Positive* or *Negative*) and its sentiment score for each debate part of the Ukraine dataset. The positive labelled sentiment score is clearly higher after 2020 with a similar or smaller standard deviation (see Table 8). The negative labelled sentiment score does not show a similar change. In Figure 3, there are two interesting things to see. First, only in 2016, the year of the referendum, there are more debate parts labelled as negative than positive. Second, there is a big growth of number of debate parts as well as number of positive labelled debate parts in 2022, the year of the Russian invasion. To evaluate the performance, the sentiment label and score is calculated for a random sample of words (see Table 11 in Appendix). Some words (*behaalde* and *geschrapt*) are evaluated as expected (respectively *Positive* and *Negative*), whereas some words (*belemmeren* and *verwoest*) are not intuitively evaluated (both *Positive*).

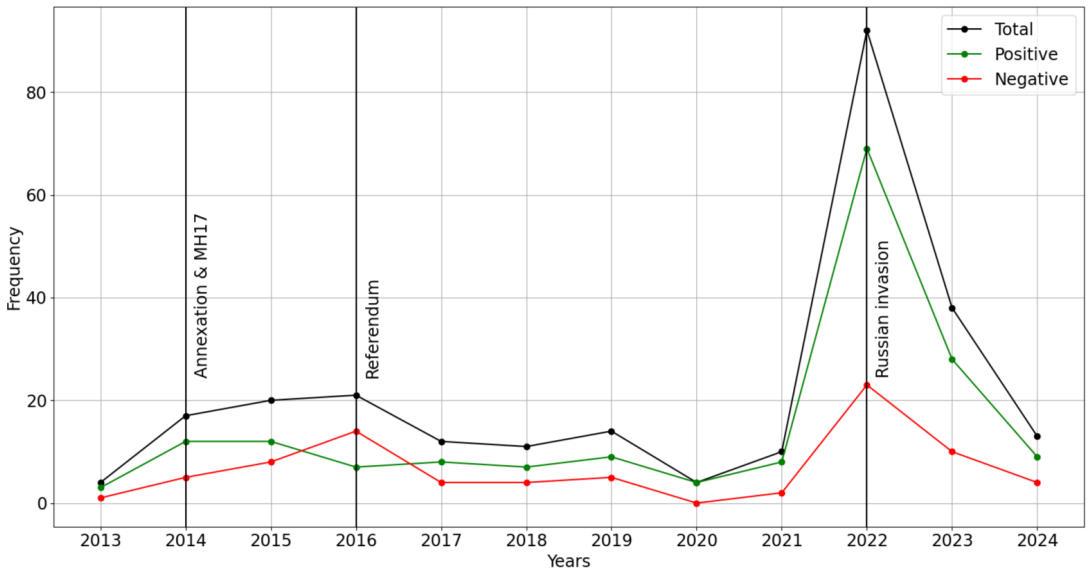


*Figure 3.* Frequency of sentiment labels per year for the Ukraine dataset.

### 3.2.2 United Kingdom

Correspondingly, the robBERT v2 model calculated the label and its sentiment score for each debate part of the UK dataset. The negative labelled sentiment score is the highest in 2015 ($\mu = 0.975$, $\sigma = 0.017$) and 2022 ($\mu = 0.995$, $\sigma = 0.002$) (see Table 8). In Figure 4, it is interesting to see that in 2016 the amount of positive labelled debate parts is the same as negative labelled debated parts, where everywhere else there are more debate parts labelled positive. Especially since the number of total debate parts has quite a lift. To evaluate the performance, the sentiment label and score is calculated for a random sample of words (see Table 12 in Appendix). Some words (*schrok* and *migrantencrisis*) are evaluated as expected (both *Negative*), whereas some words (*laagste* and *optreden*) are not intuitively evaluated (both *Positive*).

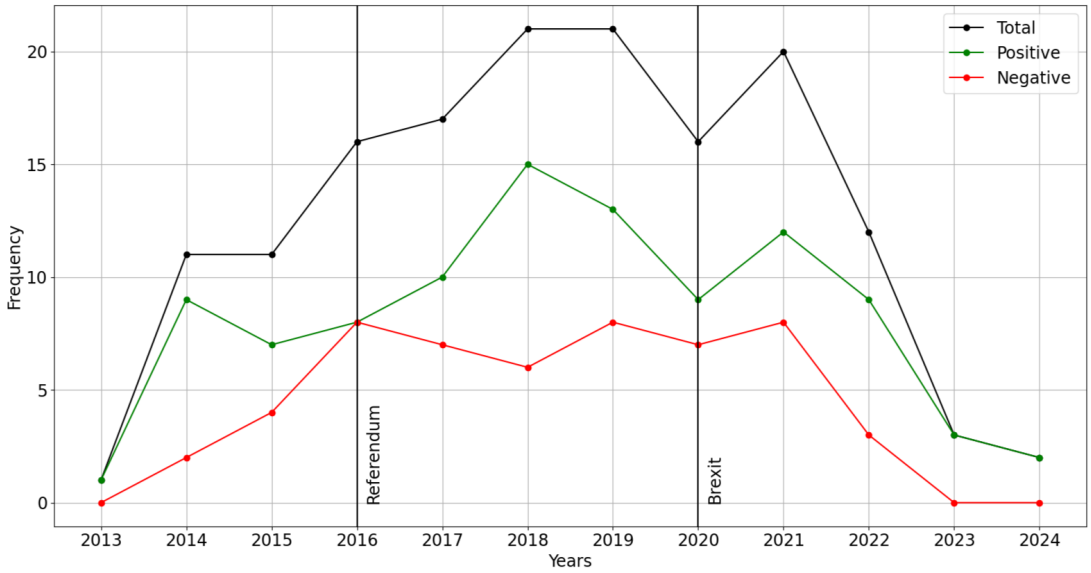


*Figure 4.* Frequency of sentiment labels per year for the UK dataset.

Table 8

*Mean and standard deviation per year per sentiment label for both datasets.*

| | **Ukraine** | | | | **UK** | | | |
|---|---|---|---|---|---|---|---|---|
| | Positive sentiment | | Negative sentiment | | Positive sentiment | | Negative sentiment | |
| *Year* | *Mean* | *Std Dev* | *Mean* | *Std Dev* | *Mean* | *Std Dev* | *Mean* | *Std Dev* |
| 2013 | 0.853 | 0.220 | 0.868 | NA | 0.996 | NA | NA | NA |
| 2014 | 0.905 | 0.149 | 0.836 | 0.210 | 0.892 | 0.164 | 0.787 | 0.278 |
| 2015 | 0.891 | 0.162 | 0.861 | 0.215 | 0.854 | 0.222 | 0.975 | 0.017 |
| 2016 | 0.872 | 0.160 | 0.803 | 0.210 | 0.931 | 0.071 | 0.896 | 0.156 |
| 2017 | 0.953 | 0.075 | 0.921 | 0.034 | 0.897 | 0.123 | 0.810 | 0.168 |
| 2018 | 0.888 | 0.140 | 0.982 | 0.022 | 0.882 | 0.166 | 0.852 | 0.125 |
| 2019 | 0.881 | 0.160 | 0.894 | 0.211 | 0.924 | 0.127 | 0.886 | 0.126 |
| 2020 | 0.883 | 0.184 | NA | NA | 0.912 | 0.085 | 0.855 | 0.204 |
| 2021 | 0.924 | 0.142 | 0.863 | 0.181 | 0.921 | 0.139 | 0.836 | 0.158 |
| 2022 | 0.944 | 0.100 | 0.888 | 0.137 | 0.961 | 0.058 | 0.995 | 0.002 |
| 2023 | 0.928 | 0.125 | 0.890 | 0.180 | 0.857 | 0.103 | NA | NA |
| 2024 | 0.968 | 0.055 | 0.896 | 0.130 | 0.916 | 0.062 | NA | NA |

# 4 Discussion

To recap the results, for the topic modelling most of the coherence scores are quite good (i.e. close to zero). However, the interpretation of the topics does not always make sense in the context of Ukraine or UK. For the sentiment analysis, the single sample words are partly not intuitively evaluated. However, the frequency of the classification in the context of Ukraine or UK is quite as expected.

Despite the good coherence scores, this $U_{mass}$ score is a small representation of a topic model's performance, so to look at the topics and see if they make sense is still the best way to evaluate its performance. However, some topics seem quite random and not all words are coherent in meaning. Therefore, a better, more extensive data cleaning process is needed to be able to check the modelled topics with the timelines of Ukraine and UK. For example, more irrelevant words (*thanks*, *for example*, *as*) could be deleted as stop words. In this research however, there was too little time to extent this process.

Additionally, the use of evaluating a sample of words with a sentiment to check the performance of robBERT v2 is not the best way. The pre-trained version (DTAI-KULeuven, 2022) is trained to classify a complete text, not separate words, which gives the model a very small amount of information to classify its sentiment. A better way to test its performance is to train the version on similar labelled text, other transcripts of parliamentary debates for example, and to test this on a validation set (Delobelle et al., 2020). Unfortunately, in this research there was a lack of labelled transcripts and too little time to label.

Apart from the method, there are also a few limitations in the data. First, this study utilized a small dataset of 256 and 151 debate parts for respectively the Ukraine and UK dataset. These are relatively few documents for especially topic modelling. Second, the imbalance in the Ukraine dataset may have impacted the topic modelling results, suggesting that a more balanced dataset could yield more accurate insights. The UK dataset was more in balance in number of debate parts over time.

Lastly, BERT and similar NLP models operate as "black boxes," meaning their internal workings are not easily interpretable. This lack of transparency makes it difficult to understand how these models generate outputs from given inputs. Additionally, it poses challenges in assessing their performance and understanding their behaviour, which is particularly problematic when traditional metrics are applied without clear insight into the model's decision-making process (in this research the coherence score). Due to this black box characteristic, the performance in this research is difficult to indicate.

All these limitations seem to make this research not of essence. Nevertheless, applying these new models contributes to the evolution of using BERT models in a scientific analysis. Additionally, applying it to a new sort of textual data, parliamentary debates instead of Twitter or review based, creates a new testing field for these models and whether they work as well as other sources. Moreover, the impact on the theoretical base is also of interest since it gains insight in how the perceptions on a region are affected by political upheaval. It is clear to see a difference in the Ukraine analysis between the defining moment, especially the invasion of Russia. For the UK it is more gradual change and might need a more extensive method to be able to see differences between the defining moments.

To conclude, further research on this subject with a revised method is definitely necessary. To be specific, it would be interesting to apply the same analysis on the whole dataset and compare those to the Ukraine and UK filtered datasets. In this analysis, the pre-processing of the data should be more extensive, for example deleting more meaningless words. Also, the sentiment analysis should be done with a trained robBERT v2 on parliamentary debates. After this, the focus could be on different countries with political upheaval to test the generalizability. Lastly, the dataset could be way larger since the parliamentary debate transcripts go further back then 2013. Enlarging the data gives more insight, which was not possible in this research due to limited time.

# 5 Conclusion

To generally conclude, the modelled topics have a good coherence, however, the topics are not as interpretable as hoped. This interpretability is the best way to test the performance of the topic modelling, so it is key to extent the preprocessing steps in further research. Subsequently, the sentiment analysis evaluates a bit as predicted, but there is no valid way to check the performance of the model. Therefore, it is essential to train the model on similar data to be able to check the performance and to improve the analysis.

For the theoretical background, this research is a first step towards implementing a new method to analyse the perspective of an external party on the political upheaval. Specifically, the topic modelling and sentiment analysis on the Ukraine data is mostly as predicted, so it is clear that a big change in a country is reflected on the data analysis. To answer the research question: *What is the effect of political upheaval on the way of speaking about a country by analysing changes in discussed topics and sentiments through political debates?* There is further research needed with a revised method in which time is too limited to be able to fully answer the question. However, since there is more of an effect to see in the Ukraine dataset than the UK dataset, one could say that the more sudden the political upheaval, the clearer a change in the analysis. Additionally, in the Ukraine defining moments it is a matter of life or death whereas the UK defining moments are more financially focused. This might be of effect in the analysis. Therefore, widening the view and including more countries with political upheaval will be of essence to state these results with more certainty.

# References


Abercrombie, G., & Batista-Navarro, R. T. (2020). Parlvote: A corpus for sentiment analysis of political debates. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, *LREC 2020*, 5073–5078.

Agostini, M., & van Zomeren, M. (2021). Toward a comprehensive and potentially cross-cultural model of why people engage in collective action: A quantitative research synthesis of four motivations and structural constraints. *Psychological Bulletin*, *147*(7), 667.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120.

Chaudhry, H. N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z. I., Shoaib, U., & Janjua, S. H. (2021). Sentiment analysis of before and after elections: Twitter data of us election 2020. *Electronics*, *10*(17), 2082. https://doi.org/https://doi.org/10.3390/electronics10172082

Chayinska, M., Minescu, A., & McGarty, C. (2017). Political solidarity through action (and inaction): How international relations changed intracultural perceptions in ukraine. *Group Processes & Intergroup Relations*, *20*(3), 396–408.

Countwordsfree. (n.d.). *The list of stop words*. https://countwordsfree.com/stopwords/dutch Accessed May 2024.

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: A Dutch RoBERTa-based Language Model. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3255–3265. https://doi.org/10.18653/v1/2020.findings-emnlp.292

DTAI-KULeuven. (2022). RobBERT v2: Dutch Sentiment Analysis Model. RobBERT finetuned for sentiment analysis on DBRD [Online; accessed June 12, 2024]. https://huggingface.co/DTAI-KULeuven/robbert-v2-dutch-sentiment

Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Overview of the transformer-based models for nlp tasks. *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 179–183.

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, *25*(1), 77–94.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint*, *arXiv:2203.05794*.

Grootendorst, M. (2024). *BERTopic: Topic modeling with bert* [[Online; accessed June 12, 2024]]. https://maartengr.github.io/BERTopic/index.html

Hadiat, A. R. (2022). Topic modeling evaluations: The relationship between coherency and accuracy. *Doctoral dissertation*.

Ilyas, S., Soomro, Z., 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 (dg.o '20)*, 6. https://doi.org/https://doi.org/10.1145/3396956.3396973

Ivanov, S., Gavrilina, M., Webster, C., & Ralko, V. (2017). Impacts of political instability on the tourism industry in ukraine. *Journal of Policy Research in Tourism, Leisure and Events*, *9*(1), 100–127.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272.

Müller-Hansen, F., Callaghan, M. W., Lee, Y. T., Leipprand, A., Flachsland, C., & Minx, J. C. (2021). Who cares about coal? analyzing 70 years of german parliamentary debates on coal with dynamic topic modeling. *Energy Research & Social Science*, *72*. https://doi.org/https://doi.org/10.1016/j.erss.2020.101869

Peng, J., Strijker, D., & Wu, Q. (2020). Place identity: How far have we come in exploring its meanings? *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00294

Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of twitter data during critical events through bayesian networks classifiers. *Future Generation Computer Systems*, *106*, 92–104. https://doi.org/https://doi.org/10.1016/j.future.2020.01.005

Silva, N. F. D., Silva, M. C. R., Pereira, F. S., Tarrega, J. P. M., Beinotti, J. V. P., Fonseca, M., & de Carvalho, A. C. D. L. (2021). Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, *10*, 104–120.

Tweede Kamer der Staten-Generaal. (n.d.-a). *Open Data Portaal: OData API*. https://opendata.tweedekamer.nl/documentatie/odata-api Accessed April 2024.

Tweede Kamer der Staten-Generaal. (n.d.-b). *Open Data Portaal: Verslag*. https://opendata.tweedekamer.nl/documentatie/verslag Accessed April 2024.

# A Appendix: Extra analysis output

Table 9

*All topics for Ukraine, with coherence score per topic.*

| **Topic 1** *-0.389* | **Topic 2** *-0.688* | **Topic 3** *-0.112* | **Topic 4** *-0.083* | **Topic 5** *-0.280* |
|---|---|---|---|---|
| boeren | europese | nederland | nederland | referendum |
| natuur | nederland | jaar | komen | wetsvoorstel |
| jaar | president | natuurlijk | zeggen | grondwet |
| maken | minister | denk | dank | voorstel |
| komen | europa | landen | gezegd | democratie |
| dieren | natuurlijk | dank | jaar | zeggen |
| nederland | commissie | regering | regering | amendement |
| dank | landen | europese | debat | vragen |
| landbouw | dank | staat | partijen | nederland |
| mee | oekraïne | geld | natuurlijk | jaar |
| **Topic 6** *-0.193* | **Topic 7** *-0.123* | **Topic 8** *-0.649* | **Topic 9** *-0.209* | **Topic 10** *-0.889* |
| politie | nederland | israël | mensen | gemeenten |
| vragen | syrië | nederland | opvang | overheid |
| openbaar | vragen | palestijnse | oekraïne | koning |
| ministerie | zeggen | zeggen | nederland | minister |
| zorgen | nederlandse | staat | vluchtelingen | president |
| nederland | groepen | jaar | asielzoekers | zeggen |
| bericht | informatie | regering | landen | natuurlijk |
| veiligheid | gezegd | kernwapens | dank | echt |
| geval | zoals | nodig | zorgen | democratie |
| zoals | debat | veiligheid | bijvoorbeeld | koninklijk |
| **Topic 11** *-0.386* | **Topic 12** *-0.172* | **Topic 13** *-0.960* | **Topic 14** *-0.230* | **Topic 15** *-5.049* |
| werken | zorg | defensie | dank | verdrag |
| werk | agema | krijgsmacht | constaterende | europese |
| arbeidsmarkt | vragen | militairen | vragen | handel |
| werkgevers | natuurlijk | nederland | verzoekt | nederland |
| denk | tekorten | samenwerking | economie | landen |
| groep | patiënten | nederlandse | overwegende | nederlandse |
| sociale | bijvoorbeeld | dijk | voorgesteld | canadese |
| zorgen | jaar | oekraïne | beraadslaging | boeren |
| werknemers | zorgen | boswijk | regering | europa |
| bijvoorbeeld | zoals | militaire | gesprek | handelsverdrag |

| **Topic 16**<br>*-1.164* | **Topic 17**<br>*-2.091* | **Topic 18**<br>*-0.584* | **Topic 19**<br>*-0.457* | **Topic 20**<br>*-0.251* |
|---|---|---|---|---|
| pensioen | rusland | hof | gas | veiligheid |
| nieuwe | russische | sancties | huishoudens | nationale |
| pensioenen | oekraïne | nederlandse | natuurlijk | nederland |
| pensioenfondsen | sancties | nederland | nederland | onderzoek |
| pensioenstelsel | nederland | verdrag | betalen | bijvoorbeeld |
| gepensioneerden | regering | rechten | energie | verkeer |
| pensioenfonds | landen | rechter | prijzen | zorgen |
| amendement | staatsaansprakelijkheid | mensenrechten | maatregelen | bedrijven |
| pensioenakkoord | nederlandse | europees | geld | zaken |
| fondsen | strafrechtelijk | europese | contracten | geval |

| **Topic 21**<br>*-0.722* | **Topic 22**<br>*-0.925* | **Topic 23**<br>*-2.209* | **Topic 24**<br>*-0.493* | **Topic 25**<br>*-0.468* |
|---|---|---|---|---|
| zorg | kinderen | nederland | president | marokko |
| testen | nederland | landen | minister | vragen |
| maatregelen | veiligheid | handel | vragen | nederland |
| natuurlijk | ouders | ontwikkelingssamenwerking | dividendbelasting | politie |
| virus | politie | wereld | nederland | marokkanen |
| echt | onderwijs | nederlandse | vragenuur | zorg |
| vragen | journaTopicen | vrouwen | debat | huisartsen |
| denk | school | europa | natuurlijk | aantal |
| eigenlijk | justitie | buitenlandse | premier | zoals |
| horeca | scholen | organisaties | zeggen | onderzoek |

Table 10

*All topics for the UK dataset, with coherence score per topic.*

| **Topic 1** | **Topic 2** | **Topic 3** | **Topic 4** | **Topic 5** |
|---|---|---|---|---|
| *-0.044* | *-0.112* | *-0.630* | *-0.023* | *-0.597* |
| nederland | syrië | maatregelen | pact | zorg |
| jaar | nederland | natuurlijk | nederland | patiënten |
| europese | nederlandse | testen | natuurlijk | zorgverzekeraars |
| zeggen | regering | vragen | zeggen | zorgverzekeraar |
| natuurlijk | steun | aantal | komen | kwaliteit |
| dank | israël | virus | denk | nodig |
| komt | assad | denk | gezegd | nederland |
| zoals | informatie | besmettingen | zegt | dank |
| landen | turkije | week | zoals | belangrijk |
| graag | debat | eigenlijk | eigenlijk | verzekeraars |
| **Topic 6** | **Topic 7** | **Topic 8** | **Topic 9** | **Topic 10** |
| *-0.586* | *-0.613* | *-0.487* | *-1.531* | *-0.594* |
| boeren | politie | airport | studenten | plastic |
| dieren | kwetsbaarheden | nr | onderwijs | circulaire |
| natuur | wetsvoorstel | luchtvaart | scholen | stoffen |
| landbouw | seksuele | luchthavens | wetsvoorstel | water |
| nederland | rechter | luchtruim | instellingen | verzoekt |
| maken | staat | verzoekt | student | regering |
| jaar | seksueel | regering | amendement | orde |
| sector | nederland | orde | leraren | voorgesteld |
| vissers | aantal | vluchten | docenten | drinkwater |
| commissie | zoals | dag | leerlingen | blijkt |
| **Topic 11** | **Topic 12** | **Topic 13** | **Topic 14** | **Topic 15** |
| *-1.531* | *-4.629* | *-0.523* | *-1.385* | *-1.125* |
| nederland | apk | bedrijven | wetsvoorstel | banken |
| belasting | rijden | bedenktijd | beveiligers | financiën |
| belastingontwijking | verkeersveiligheid | nederland | nederland | sector |
| country | kentekenplicht | natuurlijk | verdrag | amendement |
| belastingdienst | km | belang | landen | wetsvoorstel |
| belastingverdragen | vvd | publieke | mensensmokkel | bonussen |
| ontwikkelingslanden | voertuigen | zoals | particuliere | beloningsbeleid |
| fiscale | regering | overheid | protocol | bank |
| nederlandse | automobilist | aandeelhouders | europese | toezichthouder |
| vliegbelasting | wegen | dank | rechten | bedrijven |

| **Topic 16**<br>*-1.625* | **Topic 17**<br>*-1.433* | **Topic 18**<br>*-0.154* | **Topic 19**<br>*-1.787* | **Topic 20**<br>*-0.319* |
|---|---|---|---|---|
| amendement | rusland | vluchtelingen | publieke | turkije |
| schade | oekraïne | begroting | regionale | turkse |
| mijnbouwwet | russische | nederland | radio | persvrijheid |
| gas | landen | turkije | concessiebeleidsplan | nederland |
| bewijslast | staatsaansprakelijkheid | geld | commissariaat | nederlandse |
| amendementen | nederland | griekenland | transparantie | journalisten |
| richtlijn | internationale | miljard | toezicht | regering |
| veiligheid | buitenlandse | europese | media | vrijheid |
| banken | strafrechtelijk | regering | omroepverenigingen | europese |
| gaswinning | nederlandse | financiã | wetsvoorstel | buitenlandse |

| **Topic 21**<br>*-0.990* | **Topic 22**<br>*-0.531* | **Topic 23**<br>*-0.206* |
|---|---|---|
| pensioen | amendement | europese |
| pensioenfondsen | belastingplan | commissie |
| pensioenfonds | fiscale | verzoekt |
| pensioenen | belasting | orde |
| pensioenstelsel | belastingdienst | parlement |
| gepensioneerden | miljard | albanië |
| fondsen | betalen | regering |
| risico | voorzitter | voorgesteld |
| nabestaandenpensioen | maatregelen | beraadslaging |
| werkgevers | miljoen | turkije |

Table 11

*Random sample of words and its sentiment evaluation for the Ukraine dataset.*

| Year | Word | Sentiment score | Sentiment label |
|---|---|---|---|
| 2013 | incentivepotje | 0.982 | Positive |
| | behaalde | 0.973 | Positive |
| | uitgebreidere | 0.808 | Positive |
| 2014 | aanscherpen | 0.977 | Positive |
| | uitkomsten | 0.960 | Negative |
| | projectie | 0.890 | Negative |
| 2015 | naadloos | 0.653 | Negative |
| | thematiek | 0.948 | Negative |
| | geschrapt | 0.999 | Negative |
| 2016 | Investigation | 0.863 | Positive |
| | daarbij | 0.882 | Negative |
| | deel | 0.943 | Negative |
| 2017 | herindelingen | 0.971 | Positive |
| | representatieve | 0.996 | Negative |
| | werking | 0.559 | Negative |
| 2018 | drie | 0.856 | Positive |
| | belemmeren | 0.998 | Positive |
| | effecten | 0.868 | Negative |
| 2019 | informatieve | 0.999 | Positive |
| | gekkigheid | 0.966 | Negative |
| | Aanhangers | 0.535 | Positive |
| 2020 | orgaan | 0.905 | Positive |
| | vergaande | 0.994 | Negative |
| | opkomt | 0.705 | Positive |
| 2021 | verwoest | 0.965 | Positive |
| | opdringen | 0.931 | Positive |
| | indringend | 0.989 | Positive |
| 2022 | woorden | 0.562 | Positive |
| | gek | 0.997 | Negative |
| | komen | 0.882 | Positive |
| 2023 | gevallen | 0.944 | Negative |
| | groepsdenken | 0.998 | Positive |
| | achteren | 0.992 | Positive |
| 2024 | verwachting | 0.527 | Negative |
| | Valstar | 0.998 | Positive |
| | vaststelling | 0.930 | Positive |

Table 12

*Random sample of words and its sentiment evaluation for the UK dataset.*

| Year | Word | Sentiment score | Sentiment label |
|---|---|---|---|
| 2013 | doorrekeningen | 0.972 | Negative |
| | vormgeven | 0.910 | Positive |
| | onderkennen | 0.999 | Positive |
| 2014 | voorzorgsprincipe | 0.832 | Positive |
| | realiseer | 0.994 | Positive |
| | hernieuwbaar | 0.998 | Positive |
| 2015 | dienen | 0.985 | Negative |
| | verwachting | 0.527 | Negative |
| | migrantencrisis | 0.587 | Negative |
| 2016 | daarmee | 0.865 | Negative |
| | meenemen | 0.964 | Positive |
| | liggende | 0.984 | Positive |
| 2017 | gebaseerd | 0.774 | Positive |
| | schrok | 0.913 | Negative |
| | recht | 0.997 | Positive |
| 2018 | genuanceerd | 0.993 | Positive |
| | laagste | 0.951 | Positive |
| | 2021-2027 | 0.962 | Positive |
| 2019 | verengelst | 0.576 | Negative |
| | verstrekken | 0.559 | Negative |
| | onthouden | 0.987 | Negative |
| 2020 | dekken | 0.966 | Positive |
| | standaardkarikatuur | 0.951 | Negative |
| | praat | 0.803 | Positive |
| 2021 | afspraken | 0.636 | Positive |
| | schrift | 0.881 | Negative |
| | eenheid | 0.913 | Positive |
| 2022 | intens | 0.998 | Positive |
| | Zij | 0.969 | Positive |
| | landbouwgrond | 0.613 | Positive |
| 2023 | diensten | 0.781 | Positive |
| | democratie | 0.938 | Positive |
| | verzetten | 0.952 | Negative |
| 2024 | oordeelt | 0.870 | Negative |
| | deelname | 0.931 | Positive |
| | optreden | 0.998 | Positive |

# B Appendix: Programming code

## B.1 XML to DataFrame code

```
import os
from bs4 import BeautifulSoup
import pandas as pd

# Initialize lists to store data
texts = []
speakers = []
dates = []
indices = []
activities = []
titles = []

file_path = "C:/Users/annet/Documents/Applied Data Science/Thesis/data/verslagen"
file_names = os.listdir(file_path)

for file_name in file_names:
    with open(os.path.join(file_path, file_name), 'r', encoding='utf-8') as text:
        data = text.read()

    # Parse the XML
    soup = BeautifulSoup(data, 'xml')
    elements = soup.find_all('tekst')
    date = file_name.split('_')[0]

    for index, element in enumerate(elements):
        alinea_items = element.find_all("alineaitem")

        if alinea_items:
            speaker = alinea_items[0].get_text().strip()
            text = element.get_text()
            text_clean = text.replace(speaker, "")

            # Find subject within debate
            parent_activiteit = None
            for parent in element.parents:
                if parent.name == 'activiteit':
                    parent_activiteit = parent
                    break

            if parent_activiteit:
                parent = parent_activiteit['soort']
                title = parent_activiteit.find('titel').text
```

```
            else:
                parent = None
                title = None

            # Append data to lists
            texts.append(text_clean)
            speakers.append(speaker)
            dates.append(date)
            indices.append(index)
            activities.append(parent)
            titles.append(title)
        else:
            print("No alinea items found in file:", file_name)
    print("File " + file_name + " done!")

# Create DataFrame from lists
df = pd.DataFrame({
    'text': texts,
    'speaker': speakers,
    'date': dates,
    'activity': activities,
    'title': titles,
    'speaking turn': indices,
    'party': None,
    'cabinet': None
})

# Cleaning
# Extract party of speaker
df["party"] = df["speaker"].str.extract(r'\(([^)]+)\):')
# Extract whether the cabinet is speaking ((prime)minister/secretary of state)
df["cabinet"] = df["speaker"].str.match(r'^(Minister|Staatssecretaris)(\s+(\w+))*:$').
    astype(bool)

# Delete all empty text and title rows
df.loc[:, ['text', 'title']] = df.loc[:, ['text', 'title']].fillna("")
df_filtered = df[df['text'].str.contains(r'[a-zA-Z0-9]') & df['title'].str.contains(r'[a-
    zA-Z0-9]')]

# Save DataFrame to CSV
df_filtered.to_csv("C:/Users/annet/Documents/Applied Data Science/Thesis/data/verslagen.
    csv", sep=';', encoding='utf-8', index=False)

# Oekraine filtering
# Filter rows where the text contains "Oekra"
df_oekra = df_filtered[df_filtered['text'].str.contains("Oekra")]
# Extract unique dates and titles from filtered rows
```

```
dates_with_oekra = df_oekra['date'].unique()
titles_with_oekra = df_oekra['title'].unique()
# Filter the original DataFrame based on both dates and titles
df_oekra = df_filtered[(df_filtered['date'].isin(dates_with_oekra)) & (df_filtered['title
    '].isin(titles_with_oekra))]

# Irrelevant activities removed
activities_to_remove = ["Opening", "Mededelingen", "Stemmingen", "Afscheid", "
    Hamerstukken", "Regeling van werkzaamheden"]
df_oekra = df_oekra[~df_oekra['activity'].isin(activities_to_remove)]

# Save DataFrame to CSV
df_oekra.to_csv("C:/Users/annet/Documents/Applied Data Science/Thesis/data/
    verslagen_oekra.csv", sep=';', encoding='utf-8', index=False)

# VK+Brexit filtering
df_vk = df_filtered[df_filtered['text'].str.contains(pat = '(Brexit|Verenigd\sKoninkrijk)
    ', regex = True)]
# Extract unique dates from filtered rows
dates_with_vkbrexit = df_vk['date'].unique()
titles_with_vkbrexit = df_vk['title'].unique()
# Filter the original DataFrame based on both dates and titles
df_vk = df_filtered[(df_filtered['date'].isin(dates_with_vkbrexit)) & (df_filtered['title
    '].isin(titles_with_vkbrexit))]

# Irrelevant activities removed
df_vk = df_vk[~df_vk['activity'].isin(activities_to_remove)]

# Save DataFrame to CSV
df_vk.to_csv("C:/Users/annet/Documents/Applied Data Science/Thesis/data/
    verslagen_vkbrexit.csv", sep=';', encoding='utf-8', index=False)
```

## B.2 Preprocessing code

```
import os
import pandas as pd
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk
nltk.download('stopwords')
import plotly.io as pio
pio.renderers.default='browser'
import spacy

```

```
path = "C:/Users/annet/Documents/Applied Data Science/Thesis/data/"
os.chdir(path)

df_oek = pd.read_csv('verslagen_vkbrexit.csv', sep = ";")
df_vk = pd.read_csv('verslagen_vkbrexit.csv', sep = ";")

# Create stop words list
stop_words = set(stopwords.words('dutch'))
unique_items = set()
with open('C:/Users/annet/Documents/Applied Data Science/Thesis/stop_words_dutch.txt', 'r
    ') as file:
   for line in file:
      unique_items.add(line.strip()) # Strip any whitespace characters and add the line
          to the set
extra_stop_words = unique_items.union(stop_words)
extra_stop_words.update(["heer", "mevrouw", "staatssecretaris", "minister", "voorzitter",
     "kamer", "kabinet", "motie", "vraag"])

# Preprocessing function
def preprocess(text):
   def is_valid_word(word):
      return any(char.isalpha() for char in word) or '\tilde{a}' in word or '\ddot{i}' in
           word or '-' in word

   # Replace "Verenigd Koninkrijk" with a unique placeholder
   placeholder = "VERENIGDKONINKRIJK"
   text = text.replace("Verenigd Koninkrijk", placeholder)
   # Split text into words while preserving special characters
   tokens = word_tokenize(text)
   # Filter out words that do not meet the criteria
   tokens = [word for word in tokens if is_valid_word(word)]
   # Remove stopwords in a case-insensitive manner
   tokens = [word for word in tokens if word.lower() not in extra_stop_words]
   return tokens

nlp = spacy.load("nl_core_news_sm")

# Function to process and tag text using SpaCy
def tag_text_filt(text_list):
   tokens = []
   for text in text_list:
      doc = nlp(text)
      tokens.extend([token.text for token in doc if token.pos_ != 'AUX'])
   return tokens

# Function to create a string by joining elements of list
def to_tokenize(text):
```

```
    return text.split()

def filter_and_group_documents(df, name):
    # Apply preprocessing and tagging
    df["pros_token"] = df["text"].apply(preprocess)
    df["pros_token_filt"] = df["pros_token"].apply(tag_text_filt)

    # Filter out empty documents
    df = df[df['pros_token_filt'].map(len) > 0]

    # Group the tokens per debate part (title - date combination)
    df_grouped = df.groupby(['date', 'title']).agg(
        {'pros_token_filt': lambda x: ' '.join([item for sublist in x for item in sublist])
            }
    ).reset_index()

    # Make string/text column
    df_grouped['pros_string_filt'] = df_grouped['pros_token_filt'].apply(to_tokenize)

    if name == "UK":
        # Count occurrences of "VERENIGDKONINKRIJK" in strings (> 3)
        df_grouped['vk_count'] = df_grouped['pros_token_filt'].str.count("
            VERENIGDKONINKRIJK")
        df_filt = df_grouped[df_grouped['vk_count'] > 3]
    else:
        # Count occurrences of "Oekra" in strings (> 3)
        df_grouped['oek_count'] = df_grouped['pros_token_filt'].str.count("Oekra")
        df_filt = df_grouped[df_grouped['oek_count'] > 3]

    # Convert 'date' column to datetime format
    df_filt['date'] = pd.to_datetime(df_filt['date'], format='%Y%m%d')
    # Extract year from 'date' column
    df_filt['year'] = df_filt['date'].dt.year

    return df_filt

df_oek_filt = df_oek.apply(filter_and_group_documents(df_oek, "Ukraine"))
df_oek_filt.to_csv("C:/Users/annet/Documents/Applied Data Science/Thesis/data/
    processed_oek.csv", sep=';', encoding='utf-8', index=False)

df_vk_filt = df_oek.apply(filter_and_group_documents(df_oek, "UK"))
df_vk_filt.to_csv("C:/Users/annet/Documents/Applied Data Science/Thesis/data/processed_vk
    .csv", sep=';', encoding='utf-8', index=False)
```

## B.3 BERTopic (topic modelling) implementation code

```
import os
import pandas as pd
import bertopic
import plotly.io as pio
pio.renderers.default='browser'
import gensim.corpora as corpora
from gensim.models.coherencemodel import CoherenceModel
from bertopic.representation import MaximalMarginalRelevance

path = "C:/Users/annet/Documents/Applied Data Science/Thesis/data/"
os.chdir(path)

df_oek = pd.read_csv('processed_oek.csv', sep = ";")
df_vk = pd.read_csv('processed_vk.csv', sep = ";")

def topic_modelling(df):
    representation_model = MaximalMarginalRelevance(diversity=0.1)

    # BERTopic model
    model = bertopic.BERTopic(language="Dutch", min_topic_size=5, verbose=True, nr_topics=
        'auto', representation_model=representation_model)
    topics, probs = model.fit_transform(df['pros_token_filt'])
    topics_over_time = model.topics_over_time(df['pros_token_filt'], df['year'],
        global_tuning=False)

    model.visualize_topics_over_time(topics_over_time)

    # Save topics_over_time plot and model
    fig = model.visualize_topics_over_time(topics_over_time)
    fig.write_html("topics_VK_plot.html")

    model.save("BERTopic_vk", serialization="pickle")

    # Coherence score
    # Initialize model and docs
    docs = df['pros_token_filt']

    # Extract vectorizer and tokenizer from BERTopic
    vectorizer = model.vectorizer_model
    tokenizer = vectorizer.build_tokenizer()

    # Extract features for Topic Coherence evaluation
    words = vectorizer.get_feature_names_out()
    tokens = [tokenizer(doc) for doc in docs]
```

```
    dictionary = corpora.Dictionary(tokens)
    corpus = [dictionary.doc2bow(token) for token in tokens]
    topic_words = [[words for words, _ in model.get_topic(topic)]
                for topic in range(len(set(topics))-1)]

    # Evaluate coherence scores
    coherence_model = CoherenceModel(topics=topic_words,
                                texts=tokens,
                                corpus=corpus,
                                dictionary=dictionary,
                                coherence='u_mass')
    coherence_score = coherence_model.get_coherence()

    cs_per_topic = coherence_model.get_coherence_per_topic()

    return model, topics_over_time, coherence_score, cs_per_topic

df_oek_tm = df_oek.apply(topic_modelling)
df_vk_tm = df_vk.apply(topic_modelling)
```

## B.4 robBERT (sentiment analysis) implementation code

```
import os
import pandas as pd
from transformers import AutoTokenizer, pipeline, AutoModelForSequenceClassification
import random
import ast

tokenizer = AutoTokenizer.from_pretrained("DTAI-KULeuven/robbert-v2-dutch-sentiment")
model = AutoModelForSequenceClassification.from_pretrained("DTAI-KULeuven/robbert-v2-dutch-sentiment")
sentiment_pipeline = pipeline("sentiment-analysis", model=model, tokenizer=tokenizer, truncation=True)

path = "C:/Users/annet/Documents/Applied Data Science/Thesis/data/"
os.chdir(path)

# Function to calculate sentiment per debate part
def sentiment_calculation(input_df):
    senties = []
    for index, row in input_df.iterrows():
        sentiment = sentiment_pipeline(row['pros_token_filt'])
        senties.append({
            'title': row['title'],
            'date': row['date'],
            'year': row['year'],
            'sentiment_score': sentiment[0].get('score'),
            'sentiment_label': sentiment[0].get('label')
        })
    return pd.DataFrame(senties)

# Function to analyse sentiment per year
def analyse_sentiment(df):
    result = df.groupby('year').agg(
        avg_pos_senti_score=('sentiment_score', lambda x: x[df['sentiment_label'] == 'Positive'].mean()),
        std_pos_senti_score=('sentiment_score', lambda x: x[df['sentiment_label'] == 'Positive'].std()),
        avg_neg_senti_score=('sentiment_score', lambda x: x[df['sentiment_label'] == 'Negative'].mean()),
        std_neg_senti_score=('sentiment_score', lambda x: x[df['sentiment_label'] == 'Negative'].std()),
        total_count=('sentiment_score', 'count'),
        count_neg=('sentiment_label', lambda x: (x == 'Negative').sum()),
        count_pos=('sentiment_label', lambda x: (x == 'Positive').sum())
    ).reset_index()
```

```
    return result

df_oek = pd.read_csv('processed_oek.csv', sep = ";")
df_oek_filt = df_oek[df_oek['oekra_count'] > 3]
df_oek_result = sentiment_calculation(df_oek_filt)
analyse_oek_result = analyse_sentiment(df_oek_result)

df_vk = pd.read_csv('processed_vk.csv', sep = ";")
df_vk_filt = df_vk[df_vk['vk_count'] > 3]
df_vk_result = sentiment_calculation(df_vk_filt)
analyse_vk_result = analyse_sentiment(df_vk_result)

# Performance test for the sentiment analysis
def performance_test(df):
    # Create a subset dataframe and extract the words
    subset = df[['pros_string_filt', 'year']].rename(columns={"pros_string_filt": "string"
        })
    subset["words"] = subset["string"].apply(ast.literal_eval)

    # Sample one row per year
    years = range(2013, 2025)
    sampled_subset = pd.concat([subset[subset['year'] == year].sample(n=1) for year in
        years])

    # Sample three unique words per row if possible
    sampled_subset['sampled_words'] = sampled_subset['words'].apply(lambda words: random.
        sample(list(set(words)), min(3, len(set(words)))))

    # Collect sentiment analysis results
    senties = [
        {
            'year': row['year'],
            'word': word,
            'sentiment_score': sentiment[0]['score'],
            'sentiment_label': sentiment[0]['label']
        }
        for _, row in sampled_subset.iterrows()
        for word in row['sampled_words']
        for sentiment in [sentiment_pipeline(word)]
    ]
    return pd.DataFrame(senties)

perform_oek = performance_test(df_oek_filt)
perform_vk = performance_test(df_vk_filt)
```