DTU Management
Department of Technology, Management and Economics

Analysing Political Polarisation Using Natural Language Processing

Master Thesis

Analysing Political Polarisation Using Natural Language Processing

Master Thesis
May, 2023

By
Daniel Ansia Dibuja

Cover photo: Fernando Diaz Villanueva, CC BY-SA 4.0, via Wikimedia Commons.
Published by: DTU, Department of Technology, Management and Economics, Akademivej, Bygning 358, 2800 Kgs. Lyngby Denmark
www.man.dtu.dk

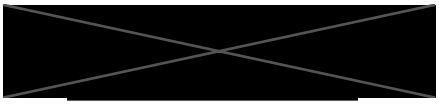
Approval

This Master Thesis report is submitted as a partial fulfillment of the requirements for acquiring the degree Master of Science in Engineering, MSc Eng. Business Analytics at the Technical University of Denmark (DTU).

The project was carried out at the Department of Technology, Management and Economics under the supervision of Francisco Câmara Pereira (DTU) and, as external supervisor, Verónica Sanz (UV).

It is assumed that the reader has a basic knowledge in the areas of natural language processing and politics.

Daniel Ansia Dibuja - S[redacted]



Signature

30 May 2023

Date

Abstract

This thesis analyses the levels of political polarisation in the Spanish parliament for years 2000 until 2023 using text data obtained from parliamentary speeches and natural language processing techniques. Specifically, the project tackles both the affective and ideological aspects of political elite polarisation. For that purpose, a literature review was performed, identifying the current state of the art in terms of polarisation analysis in Spain and internationally, and a dataset was harvested from the parliamentary records and pre-processed utilising several techniques.

The dataset was then processed in three different ways: (1) checking the occurrence of topics and propagation of terminology in the parliament to analyse the discourse of MPs, (2) analysing the levels of ideological polarisation by ideologically placing each political group of each term in a 2-dimensional matrix using a document embedding model and dimensionality reduction, observing their evolution and calculating an ideological polarisation index, and (3) evaluating the sentiment in the speech using a lexicon-based and a transformer-based sentiment classifiers.

Results show an underlying increasing trend in elite ideological polarisation throughout the last seven terms and levels of elite affective polarisation at an all-time high for the time period analysed. The levels of ideological placement and polarisation obtained are comparable to those in available research.

Keywords: *political polarisation, parliamentary corpus, elite polarisation, ideological placement, affective polarisation, ideological polarisation, sentiment analysis, NLP, document embeddings, Spain, Congreso de los Diputados, data science.*

Glossary

MP: Member of Parliament. Each of the elected representatives that compose the parliament.

EP: European Parliament.

NLP: Natural Language Processing.

BoW: Bag-of-Words model.

NN: Artificial Neural Network.

Corpus: language resource composed by a large and structure group of texts. Used for statistical analysis, checking occurrences, etc.

Intervention: every speech given by an MP. In this thesis, the group of speeches given by an MP relative to a topic in the agenda in a specific day.

Tokens: single units that combined constitute an intervention. In this case, a single word.

Political group: combination of all parliamentary groups and political parties that have historically represent the same ideological spectrum. For example, the parliamentary group of *Alianza Popular* and *Partido Popular* are categorised as the political group «PP».

Elite polarisation: Political polarisation between different political representatives — i.e. political elites [1].

Sentiment analysis: classification of the emotional elements of a text. Usually —and in this thesis— scaled between «negative», «neutral» and «positive» text.

Embedding space: in NLP, a mathematical space where words or longer pieces of text are represented as vectors.

Acknowledgements

This project could not have been conducted without the support from supervising professors Francisco Câmara Pereira (Department of Technology, Management and Economics, DTU) and Verónica Sanz González from Universitat de València (UV). Thank you for supervising this project and providing such an extensive an valuable input.

Additionally, I would like to also thank Miguel Folgado García from Universitat Politècnica de València (UPV) for his help and assistance in this thesis.

Lastly, I am grateful to my family and friends for their support and motivation.

Contents

1 Introduction

Political polarisation has been increasing all over the world. Many countries such as the United States, Germany, Brazil, Benin and Spain have increased their levels of polarisation in recent years [2]. Examples of this political polarisation are seen through the increase in toxic and offensive speech between political actors, the increasingly negative view of citizens towards those who support different political alternatives, or the growth of ideological silos [3].

This phenomenon has been analysed in a variety of ways: elite polarisation by analysing the votes in parliamentary chambers —roll-cast votes— [4], mass polarisation by analysing surveys containing ideological views, voter behaviours, opinions on other citizens with diverging political views [3], aggregation of expert opinions [2], etc. In the case of Spain, there was relatively not much research done on the topic. This thesis aims to contribute to that knowledge gap by focusing on a data-driven approach to analyse elite polarisation in Spain from ideological and affective perspectives.

This project will take advantage of a relatively new discipline of data science: Natural Language Processing (NLP). NLP focuses on processing and analysing human language as data. This discipline has been experiencing considerable improvements with the usage of artificial neural networks such as word embeddings and transformer architectures. This thesis takes into account these new methodologies and architectures to process data from political elites –congressional records– to assess their level of affective and ideological polarisation.

Both in Spain, as well as in other countries, polarisation has been an ongoing topic of conversation for political analysts, politicians and journalists. However, there is little research on the subject, and almost no research performed on assessing polarisation using NLP [5] [6]. The third innovative aspect of this thesis is the data source used: parliamentary records.

Parliamentary corpora have been in recent years gaining attention as an interesting data source to research different political aspects by analysing what parliaments talk about and how they do it [6]. According to Abercrombie and Batista-Navarro [6], only one article analysed the Spanish parliamentary corpus as of 2018. In this thesis, we will harvest the Spanish parliamentary corpus and use it as the main data source of this research.

The objective of this Master Thesis is, therefore, to analyse the elite political polarisation in the Spanish parliament, *Congreso de los Diputados*, using Natural Language Processing techniques. Within this thesis, polarisation is analysed from two different angles: (1) ideological polarisation, i.e. the divergence in the discourse of different political groups in the parliament, and (2) affective polarisation, i.e. the measurement of uncivil discourse in the parliament between the different MPs (e.g. insults, strong words, the negativity of the message, etc.) for the period 2000 – 2023.

1.1 Motivation

Political polarisation can have a highly negative effect on societies and democracy by isolating different groups of the population and creating animosity between them, eroding the basis of the democratic system and damaging the possibility of a fruitful integration and coexistence as society [7] [8].

Hence, it appears to be relevant to have tools and mechanisms to measure the level of polarisation so that the magnitude of the phenomenon can be understood by relying on data. With this thesis, I aim at improving this situation by contributing a set of metrics and analyses that can help put into numbers the current state of political polarisation in Spain.

Another reason for this project to exist is the possibility of assessing the influence of the parliament in the polarisation of the mass population. By quantifying the levels of polarisation in the parliament and comparing them to those by the general public, we can assess how these two correlate and obtain further conclusions.

Additionally, it is possible to assess the specific responsibility of each political group in terms of polarisation: which political group is more negative in their discourse? what is the ideological placement of each political group and how does that relate to the others? etc.

This project also allows us to observe the difference in the discourse of the government vs. opposition groups in the parliament, as explained in future sections.

The lack of research on elite polarisation in Spain, data-driven polarisation analysis and parliament corpus analysis make for this project a valuable experience to further the knowledge on these areas.

The results of this project can help other researchers to analyse specific usage of terminology in parliament, the importance of different issues throughout time, etc.

Finally, on a more personal note, I was able to combine my two biggest passions: social and political issues along with data science and machine learning.

1.2 Learning Objectives

The following learning objectives have been defined for this thesis:

- Understand the links between the components of a technical problem.
- Demonstrate the capacity to analyse and define a complex and open problem, put it into its broader context and make a plan for its solution.
- Using a background in a specialised discipline and current international research, to develop new ideas and solve new problems.
- Be able to work, communicate and report research results in written form in Danish or English.
- Be able to use ethical and sustainability principles to evaluate technical solutions and be able to understand their business and societal contexts.
- Be able to find, analyse and critically evaluate information and use it to identify opportunities for novel work.
- Learn and explore NLP techniques in a political and social context.
- Demonstrate a critical and objective perspective regarding ML tools and their application in political texts.
- Understand the full process of unstructured data analysis, from data collection to insights.
- Critically analyse the results, and compare them with literature, particularly considering sentiment analysis, polarisation, topics, and ideological placement.

1.3 Research Questions

To achieve the objectives laid out for this thesis, we aim at solving the following research questions:

- Are politics more polarised nowadays, both ideologically and from an affective perspective?

- How do political groups contribute to polarisation?
- How did the ideological polarisation in the parliament evolve over the years?
- How do specific events affect the level of negativity in the debate?

1.4 Structure of the Thesis

The current chapter is followed by (1) a chapter containing the literature review performed for the thesis, (2) the data gathering process followed to obtain all the necessary data sets, (3) a discourse analysis chapter observing the occurrence of specific issues and the propagation of new terminology across political groups, (4) analysis on the ideological polarisation of the parliament, (5) an analysis on the affective polarisation, by measuring the level of negativity in the speech of MPs, and (6) the conclusions, discussion and results of this thesis as well as future work.

2 Literature Review

For this chapter, I researched two main fields: natural language processing, which involves the set of tools and methodologies that belong to the field of Data Science used in this thesis, political polarisation, and parliamentary systems.

2.1 Spanish Political System

The political system of the Kingdom of Spain is a representative, bicameral and relatively decentralised democracy. At the state level, the legislative branch of government is composed of two chambers: the Congress, *Congreso de los Diputados* and the Senate, *Senado*. The Senate acts as a territorial chamber representing all provinces, communities and autonomous cities. Due to its dynamics and the subjected nature of the Senate with respect to the Congress, its usefulness is often put in question [9].

The *Congreso de los Diputados* is in charge of creating, modifying and approving laws, controlling the work of the executive branch, electing a president, approving the yearly budgets, etc. [10]. It can also dismiss the president —and therefore the entire executive— and modify the constitution, among other attributions.

The most important political debates tend to happen in the Congress, leaving the Senate for a secondary, less important, role. It is for that reason that the corpus of the Senate interventions has not been taken into account for this project.

At the regional level, there are parliaments in each autonomous community of Spain that act as the legislative branch of the regional government and pass laws that correspond to the competencies that each of the regions has. These debates were also excluded from the project.

The Spanish political system is also a plural one. The seats of any parliament are chosen based on a relatively representative formula. Yet, at a state level, government coalitions have only been happening for the last four years in the entire current democratic period (45 years). For most of the previous legislative terms, two main political parties dominated the vast majority of the Congress, PSOE and PP, and have been switching power for decades. PP is a right-wing political party and belongs to the European People's Party group in the European Parliament. PSOE is part of the S&D European Parliament group and occupies the spectrum in the centre-left, according to [11]. Nowadays, the irruption of new political parties and coalitions made it essential to reach agreements between several parties to be able to form a government.

The political groups studied along with PP and PSOE are VOX, *Unidas Podemos* and *Ciudadanos*. VOX is an extreme right-wing political party that emerged in 2019. *Unidas Podemos* is a coalition of newly created *Podemos* (anti-elite and left-wing party arising in 2015 from the 15-M movement and part of GUE/NGL in the EP) and *Izquierda Unida*, the historical left-wing group for the last decades. *Ciudadanos* was a right-wing political party that emerged in 2016. They tried to play a more centrist role than PP and, according to [11], were positioned in the political spectrum slightly on the left of PP.

The terms in Spain last 4 years by default. Every 4 years a new Congress and most of the Senate are elected. The executive is elected by the Congress at the beginning of each term. Until now, Spain has had 14 terms in the current democratic period. Terms 11 and 13 were shorter than a year, characterised by a period of political instability where no executive was appointed and therefore the Congress was dissolved and general elections were held again.

Several types of activities are carried out in the Congress, the most common ones being debates about specific laws —either presented by the executive, members of parliament (MPs), or others[1]—, appointing or removing the president, controlling the activities of the executive —in the form of questions and within the commissions—, political motions, constitutional reforms, commissions, and initiatives connected to the competencies and power of the regions.

Debates happen both in the main chamber of Congress and in commissions. Commissions are working groups formed by some of the MPs that work on specific topics and draft laws connected to their area (e.g. healthcare, education, equality).

Both in commissions and in the main chamber, the rules of debate are regulated. Most interventions have a time limit, they need to adjust to the topic being discussed, the order of debate for each speaker is specified, no speaker can be interrupted, and decorum needs to be respected.

MPs organise in the form of parliamentary groups. In order to form a parliamentary group there is a minimum requirement of MPs needed. MPs that cannot meet these criteria become part of the *Grupo Mixto*, or «mixed group». All the MPs that belong to that group have been excluded from the analysis, as well as the parliamentary groups that have existed for a limited amount of time (i.e. for only 1-2 terms).

2.2 Recent Events in Spanish Politics

The current section explains, chronologically, a few recent events in Spanish politics between the years 2000 and 2022 to assist in understanding the different outcomes of this thesis:

2000 March: PP wins again the general elections by absolute majority.

2004 March: 11M, multiple terrorist attacks by an Islamic group committed in Madrid, 191 were killed.

2004 March: PSOE wins the general elections.

2005 June: Parliament legalises gay marriage.

2006 June: A referendum, seeking more autonomy and recognition as a nation, is approved in Catalonia.

2008 March: PSOE wins the general elections again.

2009 January: Economic recession starts in Spain.

2010 February: The government approves government spending cuts and plans to raise the retirement age to 67.

2010 May: Unemployment rate surpasses 20%.

2011 May: A massive wave of demonstrations, known as 15M, start all over the country demanding more democracy and social justice.

2011 November: PP wins the general elections. More spending cuts (8% GDP) are approved by the new government.

2012 November: ETA, the Basque armed group, announces their disarmament.

2013 April: Unemployment rate peaks at 27.2% and +6 million unemployed workers.

2015 December: New political parties, Podemos and Ciudadanos, enter the parliament.

[1] Citizens and regions can also propose laws to be debated by the Congress.

2016 June: Elections are repeated, due to not being able to form a government.

2016 October: PP forms a minority government thanks to the abstention of PSOE.

2017 August: Terrorist attack in Barcelona kills 16 people.

2017 October: Illegal independence referendum is held in Catalonia. The Spanish Government imposes direct rule in the region. Promoters of the referendum are arrested.

2018 May: ETA officially dissolves.

2018 June: Motion of no-confidence ousts PP from the government in favour of PSOE.

2019 April: General elections. PSOE improves their results and far right-wing party VOX enters the parliament.

2019 November: Repeated general elections. PSOE improves their results and far right-wing party VOX becomes the third biggest party in the parliament.

2020 January: PSOE and UP form a coalition government.

2020 March: COVID-19 Pandemic starts.

2.3 Political Polarisation

Political polarisation is defined as the divergence of political attitudes and behaviours towards extreme positions — i.e. as a «bimodal distribution of observations», as defined in [1]. A further distinction of two types of polarisation is generally used by scholars: affective and ideological polarisation. These two categories are used for this thesis, to distinguish between the emotional aspects of polarisation —the negativity of the speech— and the ideological aspects of polarisation, defined by differences in ideology and policy positions.

We will also take into account a different dimension to categorise polarisation: where it is located. Elite polarisation focuses on the divide between (political) elites, such as party leaders, while mass polarisation focuses on the divergence of political attitudes and behaviours of the general public.

Different methodologies have been used to measure and analyse both elite/mass and ideological/affective polarisation. The following sections focus on analysing these methodologies, the results of those studies, and the Spanish case.

In the case of Spain, the research on both affective polarisation and ideological polarisation from a data-science perspective is very limited [5] [6]. We hope with this thesis to contribute widening the knowledge in this field.

In regards to the different methodologies analysed in this thesis to measure polarisation, I define four main categories: consulting experts, asking citizens, observing parliament behaviours, and —the focus of this thesis— conducting data science studies.

Consulting Experts

This category is defined by the usage of expert opinions to establish measures for the level of mainly ideological polarisation, generally focused on analysing political elite behaviours and ideologies. An example of this methodology is the Chapel Hill Expert Survey (CHES), which is an expert survey for the estimation of national party positions across Europe on several issues such as European integration, ideology, environmental issues, immigration, anti-elite rhetoric, and policy issues. Between 116 and 421 experts have evaluated the state of the political parties in some European countries for the years 1999, 2002, 2006, 2010, 2014 and 2019 [11]. CHES also conducts the same type of research for Latin America since 2020 [12].

Another example of this category is the work performed by the Varieties of Democracy Institute (V-DEM), which conducts yearly reports on the state of democracy across many countries around the world. Their methodology consists of asking experts —5 for every country-year from a pool of 3700— about several aspects of each country, such as the level of polarisation [2].

Asking Citizens
This methodology consists of conducting public opinion surveys on specific policy and ideology issues and processing the results by defining specific metrics and data that can indicate polarisation. Often asked in relation to the voter profile of the respondents.

Several organisations across the world conduct these kinds of studies, such as the Pew Research Center, mainly focused on the US, such as [3]. The research centre established, for that technical report, a measure of ideological polarisation by establishing 10 different ideological dimensions and observing the cohesiveness of opinion intra-block —within democrats or republicans— as well as the level of divergence inter-block for different periods of time.

In Spain, the government institution *Centro de Investigaciones Sociológicas* (CIS) performs public-opinion surveys that are often used to analyse mass ideological and affective polarisation, such as using the level of rejection/affection of citizens to political leaders given their voter behaviour. Orriols and León explore this approach in [13].

Another explored approach consists of analysing the citizens' responses on policy opinions to observe the level of divergence. Miller does that in an article to manifest that citizens' polarisation is higher for ideology and identity than for specific public policies [14].

Other initiatives, such as the work of Torcal et al. [15], are focused on creating a dataset composed of answers to ad-hoc surveys measuring ideological and affective polarisation, called E-DEM.

To define a metric for ideological polarisation by using survey responses, Simón uses surveys on voter ideological self-scaling on a left-right axis, as well as party scaling [16] using the ideological polarisation index developed in [17]. He concludes that between the years 2000 and 2019, there has been a continuous increase in ideological polarisation (Left-Right) for every general election, going from 3.75 to 5.10 points on a 0 - 10 scale.

Finally, the last example corresponds to the work of Miller and Torcal, where they aggregated data from the CSES, CNEP and E-DEM to establish an affective polarisation metric on the sentiment of voters toward the leaders of each political party [18]. The metric defined also has increased, going from 3.39 in the year 2000 to 4.28 for the year 2019.

Observing Parliament Behaviours
The main example of this category is the NOMINATE scaling application [4]. NOMINATE (Nominal Three-Step Estimation) is an estimation done by performing multidimensional scaling to roll-call voting behaviour in the parliament of the US and managing to ideologically place parliamentarians across the different terms on a 2-dimensional Euclidean space. It is considered one of the most important scores for analysing ideology in the US Parliament.

Hohmann et al. use network analysis and Euclidean distance applied to parliamentary voting records to measure the ideological polarisation of MPs [19].

Data Science Studies
I placed inside this category all the studies that involve data mining and analysis, often of texts with NLP techniques. It is possible to further categorise this section by the data sources used:

Social Media: An abundant body of research exists for methodologies that use social media data, usually Twitter data, to analyse and measure affective and ideological polarisation. Examples of

these methods include the work done by Conover et al. analysing the ideological polarisation in Twitter before the 2010 U.S. elections using network analysis [20] and the project of Hohmann et al. where they quantify ideological polarisation using Euclidean distances in network analysis for Twitter debates [19].

For the case of Spain, examples of utilising social media data are, to the best of our knowledge, limited to the analysis of affective polarisation, such as the work conducted by Guerrero-Soler and Philippe on the toxicity of Spanish politics during the COVID-19 pandemic [21] or the work of Folgado and Sanz, where they apply sentiment analysis techniques on politicians' tweets to infer the affective polarisation of the political elites [22].

Party Manifestos: Examples of this approach are the Comparative Manifesto Project (CMP) and the European Manifesto Project (EMP), two projects aimed at systematically coding political party manifestos (electoral programmes) to study parties' policy preferences. These systems rely on the occurrence of specific words or coded data [23].

Parliamentary Corpora: Lately, more research was aimed at utilising as a data source the records of parliamentary sessions. Abercrombie and Batista-Navarro recently performed a systematic literature review about sentiment and position-taking analysis of parliamentary debates. They found a steep increase in published papers tackling this topic in recent years. Their work also shows the lack of research on the Spanish parliament, with only one paper that partially focuses on Spain [6].

One of the most representative examples of using parliamentary corpora to assess ideological polarisation is the work conducted by Rheault and Cochrane [24]. Part of the methodology used in this thesis is based on their work. Rheault and Cochrane use paragraph embeddings to ideologically place the political parties in the Canadian, U.S. and U.K. parliaments. This project replicates part of their methodology to assess the ideological polarisation of the Spanish parliament throughout the last 7 terms.

Political Polarisation in Spain

The results in [18] suggest that mass affective polarisation increased, in their scale, from 3.39 to 4.28 points between the years 2000 and 2019. Simón in [16] suggests the same for an ideological left-right polarisation index, which advanced from 3.75 to 5.10 for the same period of time. Similar results were concluded by the V-DEM, which in [2] show an increase in political polarisation from 1.3 for the year 2011 to 2.4 in 2021.

Finally, CHES data shows a change in elite ideological polarisation from 3.88 to 5.50 on economic aspects and an evolution from 4.60 to 5.05 on cultural aspects for political parties between 2006 and 2019, as elaborated by [25].

2.4 Natural Language Processing

Natural Language Processing (NLP) refers to the subfield of Data Science that focuses on processing, modelling, and analysing texts that are written in plain language. In the next sections, I refer to several methodologies and algorithms used to process texts as data in a computerised way.

In recent years, NLP has gained a lot of popularity and has advanced tremendously as a field of research, thanks to the emergence of text embedding and transformer models, which are analysed in the next sections along with other more rudimentary methods and models.

Bag-of-Words Model

The simplest of all methods, the Bag-of-Words (BoW) model represents texts as the combination of the words that form it, regardless of the order of the words [26]. The next example helps

illustrate this.

```
sentence = 'I like to play board games. Do you like to play board games?'
BoW = {'i': 1, 'like': 2, 'to': 2, 'play': 2, 'board': 2, 'games': 2, 'do': 1, 'you
    ': 1}
```

Listing 2.1: Bag-of-Words Model

With this method, it is only possible to interpret the presence of specific words in the text. Therefore, since the context and word order are not taken into account, it is difficult to perform a higher-level analysis with this method. Another limitation is that it does not care about words that appear several times in all possible texts, thus adding noise to the processing.

Once a group of texts is turned into a bag of words, the BoW can be used as a feature for downstream machine learning models, such as for clustering, classification, etc.

TF-IDF

TF-IDF (term-frequency times inverse document-frequency) is a statistical measure that assigns a value to the presence of a word in a document with respect to a corpus [27]. It takes into account the occurrence of common words in the corpus by dividing the frequency of a word in the document by the frequency of that word in the entire corpus. Doing this, the measure highlights infrequent words that convey more information about the specific text being processed. The formula that defines TF-IDF is the following.

$$TF * IDF(t,d) = TF(t,d) \times log\left(\frac{n}{1+df(d,t)}\right) \tag{2.1}$$

Being $TF(t,d)$ the number of times the term t appears in the document d and $df(d,t)$ the amount of documents in the total number of documents, n, where t is present.

In the same way as for BoW, the outcomes of TF-IDF are generally used as features for downstream tasks.

Word Embeddings

Word Embeddings are models that transform unstructured text data into vector representations of N dimensions. Each word is transformed into a vector, which can be represented in an «embedding space» near other words that co-occur with it. The basic principle of text embeddings is that words that appear together in a text convey a similar semantic meaning, and therefore this model focuses on the context of a word to depict its numerical representation. Words that co-occur together will appear closely in the embedding space and vice-versa [28].

These models are often neural networks that contain a projection layer. The weights of that layer become the vector for that word. The neural networks tend to follow two main alternative architectures represented in figure 2.1:

Continuous Bag of Words (CBOW): This architecture consists of training the NN to infer a missing word within a sentence. Here the order of words does not influence the projection (i.e. the resulting vector).

Continuous Skip-gram: The model is trained to predict similar words before and after the input word within a range of size R. The order of words does influence the projection.

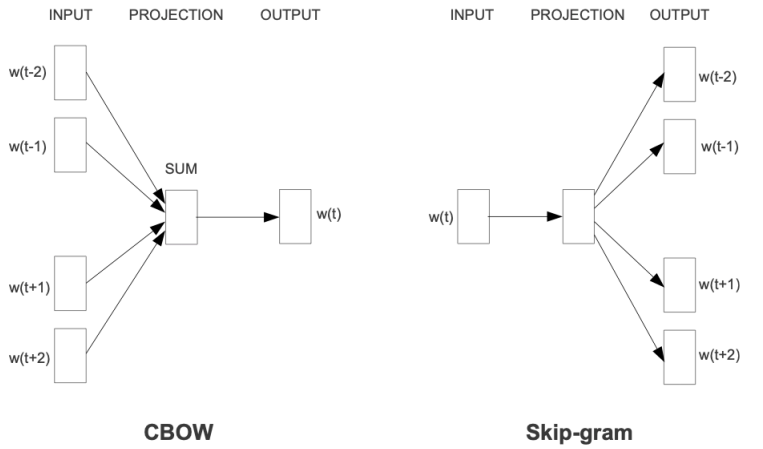


Figure 2.1: Model architectures for word embeddings. Elaborated by Mikolov et al. [28].

Other implementations of word embeddings focus on calculating word-to-word co-occurrence instead of the co-occurrence of a word within its context, such as GloVe [29].

Paragraph Embeddings

Paragraph embeddings, document embeddings or paragraph vectors follow the same idea of word embeddings but are applied to sequences of words [30]. Larger sentences, paragraphs or entire documents are vectorized, obtaining a fixed-length vector for all documents regardless of their length. Two of their main advantages compared to other models are that paragraph vectors take into account the word order as we as semantics. The paragraphs —i.e. pieces of text— are assigned an id which gets associated with a vector generated through training a neural network. In order to train the paragraph vectors, two main approaches exist:

Distributed Memory (PV-DM): The distributed memory approach follows the same idea as for training word embeddings. The difference is that in this case, the label of the paragraph is also a feature when predicting the next word in the sequence. Both the words and the paragraph ids are vectorized, obtaining embeddings for both words and paragraphs/documents. The average or concatenation of the paragraph and word vectors is used to predict the next word. Word vectors are the same across the entire corpus. Paragraph vectors are unique for each tag. In a later implementation of this model, which is used for this thesis, several paragraphs or documents can share the same tag and more than one tag can be assigned to each text. In this way, each tag will have a vector associated with it based on the words that each of the documents with that tag contain as well as their context.

Distributed Bag of Words (PV-DBOW): This approach, which is similar to the continuous skip-gram model of word embeddings, trains paragraph vectors individually to predict words of a small window of context. PV-DBOW does not take into account the word order, which was identified as a negative aspect of this project's application. Both PV-DBOW and PV-DM are represented in Fig. 2.2.

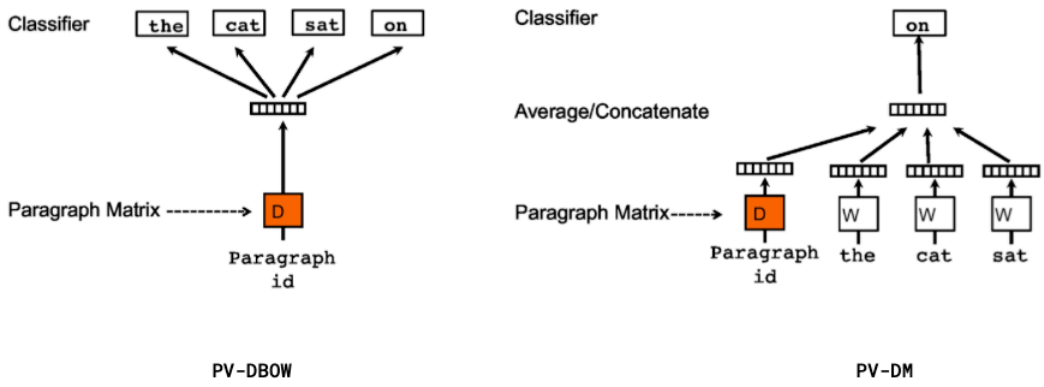


Figure 2.2: Model architectures for document embeddings. Obtained from [30].

For this project's ideological polarisation chapter, the Gensim Doc2Vec[2] implementation of the paragraph embeddings model is used [31].

Transformers

Transformers are deep learning models mainly used in NLP and computer vision. Designed to process sequential data, they are mainly based on the concept of attention to overcome one of the biggest limitations of recurrent neural network models (RNNs). Transformers assign attention coefficients to each item in its context, as well as taking into account their position in the text. These models input all tokens at once, unlike with RNNs, allowing for parallelisation of input and bigger scalability. This has allowed researchers to obtain transformer models with billions of parameters and increasingly larger sizes [32].

Transformers have become some of the most-used models for NLP tasks, as they can perform several tasks such as information extraction, sentence completion, translation, or sentiment analysis, among others [33].

When it comes to their limitations, the most important one is that there is a maximum number of tokens that the models can process, usually between 100 and 500 tokens at a time. This limit is being improved, with models such as GPT-4 allowing inputs of 8.000 to 32.000 tokens. Unfortunately, these models were not accessible while conducting this project.

A pre-trained transformer will be used in chapter 6 for conducting sentiment analysis tasks on the parliament interventions.

[2]See https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html.

3 Data

The project required metadata on parliamentary sessions from the Spanish Parliament (e.g. who was speaking and, their political group, about which topic, etc.) as well as the transcript from each MP's intervention.

Surprisingly, this data was not available online, as discussed in section 3.2. It was therefore essential to harvest and pre-process the data manually. This took a considerable effort, as described below.

Several methods were applied for the harvesting and pre-processing of the data, such as web crawling, PDF parsing, detection of regular expressions, tokenisation, punctuation removal, specific stopwords removal, lemmatisation, stemming, translation to English, etc. All these techniques are explained in detail in this chapter.

The dataset obtained as well as the code used will be publicly available in the GitHub repository of this project after publication.[1]

The collection of the data consisted of three main steps: (1) trying to find the data as open-sourced, (2) obtaining the metadata of the interventions —i.e. date, term, speaker, topic, link to the PDF with the transcript, author of the initiative, etc.— and (3), the harvesting of the texts corresponding to each row of the metadata and its posterior pre-processing.

A diagram of the list of steps followed to obtain the data used for the next three chapters is presented in figure 3.1.

[1] Code, models and data will be made available at https://github.com/dibuja/polarisation-nlp.

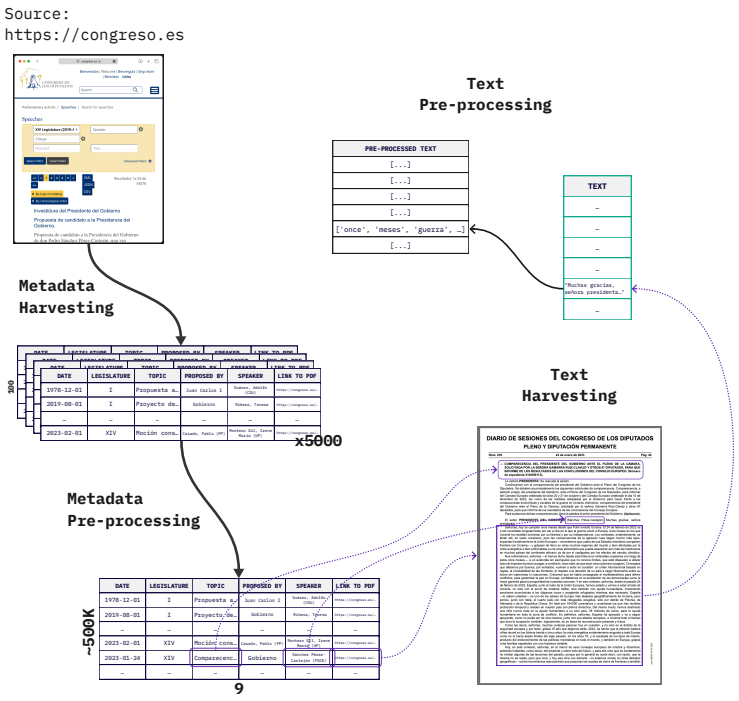


Figure 3.1: *Diagram of the data harvesting and pre-processing.*

3.1 Spanish Parliamentary Debates

As explained in section 2.1, the parliamentary debates in the Congress are strictly regulated regarding order, time and process. The President of Congress —i.e. the Speaker of the Chamber in English-speaking countries— moderates the debate, gives the word to MPs and conducts the voting. No one can interrupt the MP speaking besides the president and the president speaks in between everyone else in order to moderate the session.

Although called «debates», in the parliamentary sessions —and especially in the main chamber— MPs tend to not debate their arguments with respect to a specific proposal, but instead they use their interventions to make their point, to explain their position towards a proposal. In the commissions, parliamentary groups do discuss specific amendments for the debated proposals, however, negotiations and agreements tend to happen in private meetings outside of parliamentary sessions.

The text being processed has specific peculiarities. Because the language employed is very formal and follows decorum, MPs tend to always thank the president of the parliament and refer to each other in a formal way. In addition, the language employed includes considerable amounts of technical words and parliamentary and political jargon. The speeches are also more nuanced than other forms of text.

Due to all these reasons, it is more difficult to infer aspects of each intervention such as their position or the sentiment of the intervention. Since most sentiment analysis algorithms are trained on tweets, which are short texts with a clear meaning and concise, informal language, they seem to not work as well with the texts used in this thesis.

3.2 Available Datasets

Parliamentary Corpora are useful for natural language processing and constitute a valuable resource for performing data science tasks oriented towards political science, such as estimating political attention or even analysing the frequency of specific terminology in a given period of time [34] [35].

It is common to find open source datasets containing parliament corpora, such as the UK Hansard records[2], the Canadian parliamentary records [3] or the European Parliament Records as well as other corpora from European countries [36].

For the case of Spain, after a detailed search, I found 3 datasets:

1. The first dataset consisted of a compilation of several parliament corpora produced by the CLARIN European Research Infrastructure Consortium (ParlaMint) for several countries in Europe [37]. The data only spanned between 2015 and mid-2020, which was too short for the purpose of this thesis.

2. The second dataset, called «ParlSpeech V2», contained data from 9 parliaments across Europe —Spain included— for around 21 to 32 years [38]. For the *Congreso de los Diputados*, the data was missing several fields that were deemed essential, such as the speaker, their political group and the topic being discussed. Therefore, this dataset was also not useful for the thesis.

3. Finally, the third dataset[4], created for [39] did not include any feature besides the text itself, hence, it was also discarded for the project.

Once confirmed that no dataset was useful for this project, we decided to harvest and pre-process the data manually. The data will be made publicly available. Hopefully, it will be a useful resource for researchers.

3.3 Harvesting of Metadata

The first steps consisted of accessing the website of the Spanish Congress[5] and the parliamentary minutes. The website has an open data platform that allows downloading intervention metadata —i.e. term, date, body, time, initiative discussed, author, speaker and political group, link to minutes in PDF, link to a video of the intervention— for only 100 data points at a time.

Since I needed to obtain all the interventions, I created a web crawler that would execute the JavaScript functions of the website and fill out the form to download all the data points day by day for the current democratic period in Spain — i.e. from 1979 to 2023.

I reviewed the `robots.txt` file, the legal notice and the terms and conditions of the Congress website to make sure it was allowed to harvest the data using a crawler. However, the server had in place some measures to prevent DDS attacks which consisted of blocking the IP addresses of devices sending too many requests to the website in a limited amount of time [40].

Using Selenium [41], I coded a script that would download all metadata of the interventions day-by-day while also waiting in between to not overload the servers and prevent me from being blocked. The data search had yet another limitation: only searches that would yield less than 1000 data points would be valid. Since this was not the case for some of the days, I implemented a conditional case to download the data with a lower resolution. Later on, it was found that

[2] See https://search.politicalmashup.nl/
[3] See https://lipad.ca/
[4] See https://github.com/prodriguezsosa/EmbeddingsPaperReplication
[5] See https://congreso.es

the 1000 data-point limitation only works in the user interface but the script was still managing to download those data points. The script, `/src/data/harvest.py`, will be available in the GitHub repository of the thesis.

The data harvested belongs to the period between 23/03/1979 until 05/02/2023, corresponding to a total of 454.533 single interventions excluding the ones from the President of Congress.

3.4 Pre-processing of Metadata

Once harvested, two scripts were developed in order to pre-process the metadata and combine all the files into one.

`/src/data/concatenate.py` concatenated the hundreds of files downloaded by the harvesting algorithm and combines them into one file for each term.

`/src/data/preprocess_metadata.py` is the script created to pre-process the metadata. The script:

1. concatenates all term files into one;
2. removes irrelevant columns, keeping only the features: term in office, date, initiative, initiative code, authors, name of the session, speaker and PDF link;
3. removes a few rows containing misplaced data from terms 3, 4 and 6;
4. turns the dates into datetime format;
5. removes from the PDF link the reference to each page, in order to speed up the parsing process and because it is irrelevant data for the scripts;
6. removes duplicates and empty fields and sorts by date and URL; and
7. replaces Roman numbers for Arabic numerals for the terms in office.

The data presented several fields that were incomplete for each term, as well as some data points with errors that needed to be manually fixed or removed.

Only the latest terms contained links to video recordings or the exact time at which interventions happened. Those fields were removed as they were also not relevant.

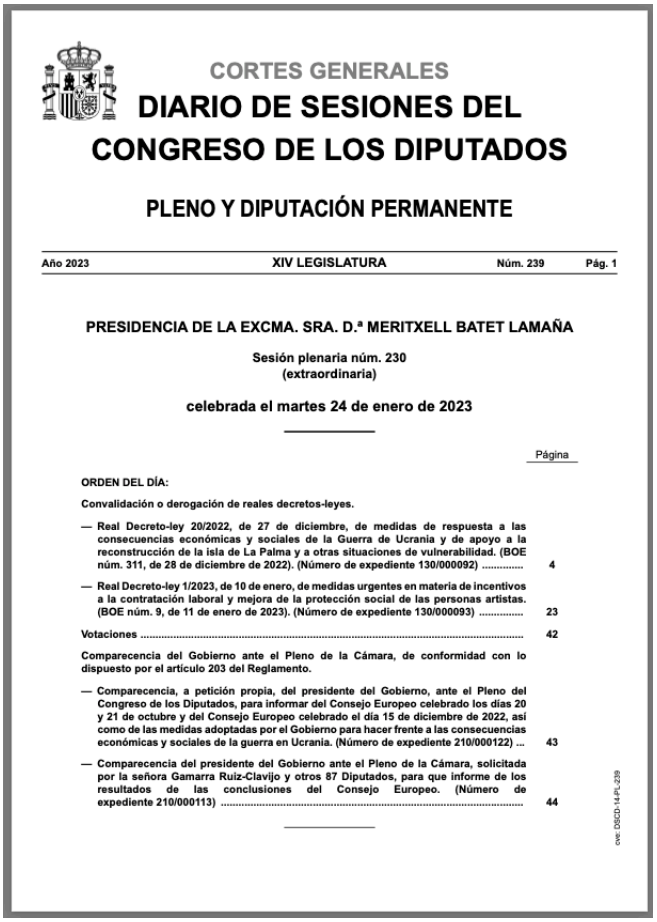
Since MPs speak several times for one topic, there were many duplicated rows in the dataset and, at the same time, actual duplicates due to the harvesting process. In order to solve this issue, I decided to remove all duplicates and, when obtaining the texts, concatenate all the times an MP speaks for the same topic on the same day into one single text (referred to as «**intervention**»).

Additionally, a new feature was created, called «political group» which includes the parliamentary group of the person speaking. The parliamentary group was later on turned into a specific political party or coalition, which I called a «political group», unifying all the political groups under consistent labels. The parliamentary groups were extracted from the name of the person speaking which were included in the harvested metadata.

Besides the straightforward conversions for political groups that change name across terms, the political groups of *Izquierda Unida (IU), Izquierda Plural (IP), Iniciativa per Catalunya Verds (ICV)* and *Unidas Podemos (UP-EC-EM)* were all merged together under the label of «UP» because all these political groups have been allies, have been defending the same ideas and in recent years became a coalition of parties. The political group created by *Izquierda Unida* and *Esquerra Republicana de Catalunya* during the ninth term was labelled as «ERC» since the majority of its members belonged to this party.

3.5 Harvesting of Texts

The next step in the process was to extract the content of each intervention. The link to the pdf document present in the metadata was used for this. I created a script that (1) downloads each pdf document containing the minutes, or *Diario de Sesiones*, (2) parses all its content, (3) using regular expressions (RegEx) identifies sections that correspond to the topic in the metadata, (4) finds the speaker for each intervention, (5) concatenates the text corresponding to the same topic, day and speaker, and (6) populates a new feature in the dataset called «text» with the content. The script is available in `/src/data/obtain_texts.py`. An excerpt of the *Diario de Sesiones* can be seen in figure 3.2.



CORTES GENERALES

DIARIO DE SESIONES DEL CONGRESO DE LOS DIPUTADOS

PLENO Y DIPUTACIÓN PERMANENTE

Año 2023 | XIV LEGISLATURA | Núm. 239 | Pág. 1

PRESIDENCIA DE LA EXCMA. SRA. D.ª MERITXELL BATET LAMAÑA

Sesión plenaria núm. 230
(extraordinaria)

celebrada el martes 24 de enero de 2023

Página

ORDEN DEL DÍA:

Convalidación o derogación de reales decretos-leyes.

— Real Decreto-ley 20/2022, de 27 de diciembre, de medidas de respuesta a las consecuencias económicas y sociales de la Guerra de Ucrania y de apoyo a la reconstrucción de la isla de La Palma y a otras situaciones de vulnerabilidad. (BOE núm. 311, de 28 de diciembre de 2022). (Número de expediente 130/000092) 4

— Real Decreto-ley 1/2023, de 10 de enero, de medidas urgentes en materia de incentivos a la contratación laboral y mejora de la protección social de las personas artistas. (BOE núm. 9, de 11 de enero de 2023). (Número de expediente 130/000093) 23

Votaciones 42

Comparecencia del Gobierno ante el Pleno de la Cámara, de conformidad con lo dispuesto por el artículo 203 del Reglamento.

— Comparecencia, a petición propia, del presidente del Gobierno, ante el Pleno del Congreso de los Diputados, para informar del Consejo Europeo celebrado los días 20 y 21 de octubre y del Consejo Europeo celebrado el día 15 de diciembre de 2022, así como de las medidas adoptadas por el Gobierno para hacer frente a las consecuencias económicas y sociales de la guerra en Ucrania. (Número de expediente 210/000122) ... 43

— Comparecencia del presidente del Gobierno ante el Pleno de la Cámara, solicitada por la señora Gamarra Ruiz-Clavijo y otros 87 Diputados, para que informe de los resultados de las conclusiones del Consejo Europeo. (Número de expediente 210/000113) 44

cve: DSCD-14-PL-239

Figure 3.2: *Diario de Sesiones, Congreso de los Diputados.*

The *Diario de Sesiones* has changed format three times throughout the 45 years of the current democratic period. The first format consisted of documents written with a typewriter and later on, scanned into PDF. The format is consistent throughout the first 6 terms. Due to the very low quality of the documents as well as the lack of alignment of the pages, it was not possible to parse those documents. Therefore, the period of time analysed spans between term 7, starting on 05/04/2000 until the time the data was gathered, during term 14 on 05/02/2023.

The format of term 7 until the middle of term 10 was already made with a computer, and therefore it was possible to parse. The text was distributed in two columns per page, unlike in terms 10 to 14, where the format changed to only include 1 column per page and the *Diario de Sesiones* stopped being printed physically to become only available online.

It was noticed that all the time a topic is about to be discussed, a convention is followed. That

convention, consisting of capital letters followed by the topic code, was used, with RegEx, to find all topics in a document [42]. Using also RegEx it was possible to identify the speaker of each text and assign all those texts to each of the speakers.

Several error-handling elements needed to be included to prevent the Congress servers from blocking the connection —although, as explained in section 3.3, there is no legal notice forbidding such activity— as well as other errors that the parsing functions needed to deal with.

Due to the computational effort required to parse PDF documents, I tried to run the scripts from one of DTU's High-Performance Computers (HPC). This attempt resulted unsuccessful due to software compatibility issues so finally, I opted for using the script, term by term, from my own machine. The total amount of time for obtaining all texts took around 17 hours.

After analysing the results of the script, I noticed that not all the interventions were able to be captured. Out of the initial 235.071 interventions for terms 7 to 14 after removing duplicates, only interventions by MPs were kept, accounting for a total of 199.842 interventions. The «Grupo Mixto» political groups were removed, as they are composed of very small political parties and they do not share a common ideology, which resulted in 163.525 interventions corresponding to the *Partido Popular (PP), Partido Socialista Obrero Español (PSOE), Unidas Podemos (UP), Vox, Ciudadanos (CS), Partido Nacionalista Vasco (PNV), Convergencia i Unió (CIU), Esquerra Republicana de Catalunya (ERC)* and *EH Bildu (EHB)*. Out of the 163.525 interventions, the texts were captured for 75.618 of them, which represents around 46% of the total corpus available. The amount of texts captured oscillates between 40% and 55% for all the political groups.

The distribution of the length of each intervention is shown in figure 3.3.

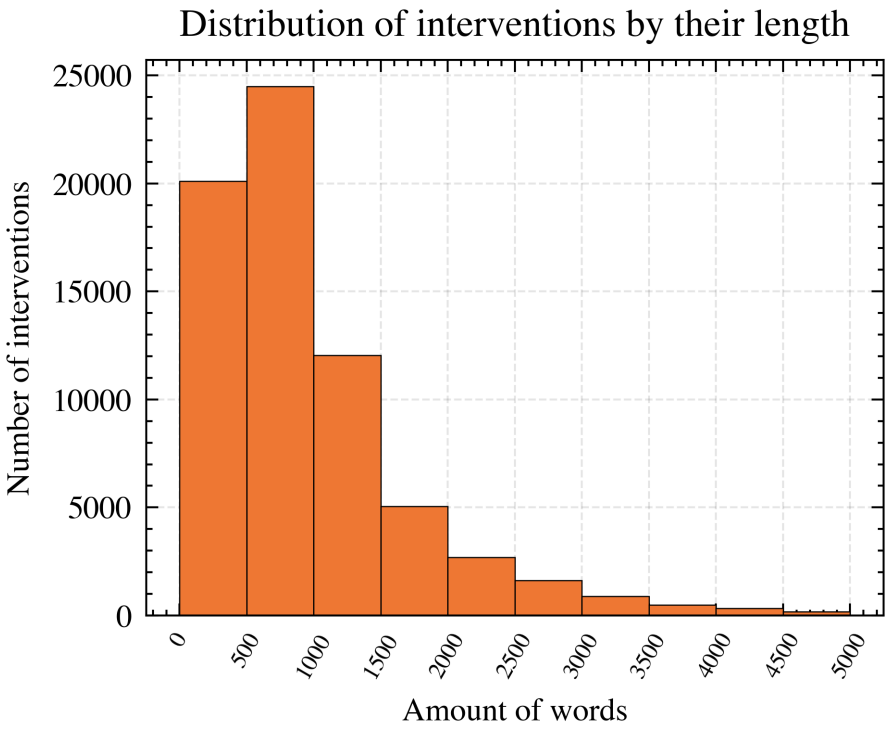


Figure 3.3: Distribution of interventions by their length.

We can observe a skewed distribution, with almost 33% of all interventions ranging between 500 and 1000 words. Regarding time, this would correspond to 3 – 7 min, which is also the most

common time length for interventions in the parliament. The vast majority of interventions range from 0 to 1500 words.

3.6 Pre-processing of Texts

The texts were pre-processed using the script `/src/features/preprocess_texts.py`. The following processes were applied:

Combining broken words: Many words were split in two due to the width of the documents. I re-join those words by using RegEx to identify their patterns.

Removal of punctuation, odd characters and digits. Commas, em-dashes, dots, newlines, tabular spaces, double spaces and other similar characters were removed from the text, leaving only 1 space as a separation between words.

Lemmatisation: Once the odd characters and punctuation were removed, I turn all words into lemmas. To do that, I used the SpaCy package, which was the only lemmatisation package I could find for Spanish [43]. The results of the lemmatisation algorithm were deficient since it lemmatised different words as the same lemma and invented words that do not exist in Spanish. Therefore, we decided not to use it.

Tokenisation: This process consisted of turning sentences into tokens, i.e. into a list of items composed of individual words.

Stemming: Since lemmatisation yielded very bad results, stemming was tried out. Stemming consists of removing the suffix from the root of nouns and verbs to keep only the root of the word, combining all words that belong to the same root in one token, e.g. «procedure», «procedures» and «procedural» would be combined as one. This process was one of the options used before training the document embeddings model of chapter 5. The results were also unsatisfactory, so no stemming was applied to the definitive model.

Creation of bigrams and trigrams: To capture concepts that involve two or three words —e.g. human rights, national healthcare system, tax increase, social rights, the feminist movement, social security, etc.— I programmed the script, using Gensim, to turn co-occurring groups of two and three words into one token. This approach yielded great results, being able to identify the most used concepts in parliament, providing a huge improvement to the models trained later on.

Short word removal. I removed words containing two characters or less since their relevance is very limited.

Stopwords removal. Since the texts contained a lot of irrelevant and misleading words, those words were eliminated manually from the list of tokens. Two default lists were used, one of them belonging to the Natural Language Toolkit package and another one from an individual researcher [44]. Later on, I created 2 ad-hoc lists of stopwords: procedural stopwords, containing all the words that belong to parliamentary decorum and procedural language — e.g. «president–e/–a», «señoría/s», «moción/es», «proyecto de ley», «proposición/es», «interpelación/es», «comisión/es», «decreto/s»— and a list of stopwords that contain other irrelevant words that were previously not in the other lists, mentions to political parties in order not to bias the models, and words in Euskera, Catalan and Galician, since they are used by very specific political parties and would be a confounding factor for the model.

English translation. Due to sentiment analysis models not working for Spanish, we translated all the original texts into English using the deep translator package[6] and did no further pre-

[6] Available at https://pypi.org/project/deep-translator/.

processing, as explained in chapter 6.

By the end of this process, the corpus was composed of 75.618 interventions and a total of 26.267.985 tokens. The distribution of interventions per political party is irregular, as shown in Fig. 3.4 since their participation in the parliament has varied vastly based on the percentage of representation in each term for each political group.

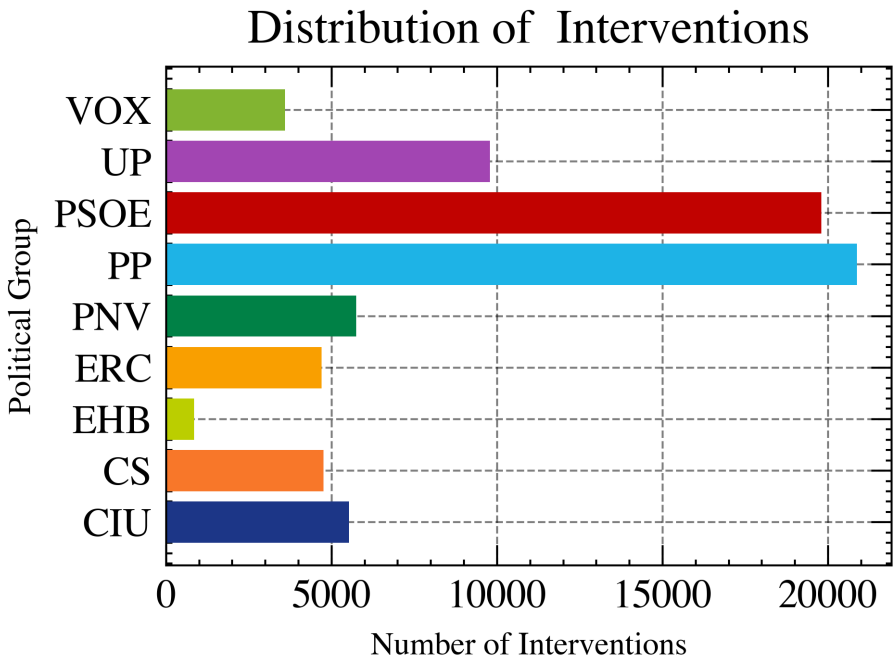


Figure 3.4: Distribution of interventions captured. More than half of the interventions belong to the two main political groups: PP and PSOE. UP is the third one with around 10.000.

The final dataset contained the following fields:

- **Term in office**. The term that each intervention belongs to, ranging from 7 to 14.
- **Initiative's purpose**. The purpose of why the session is being held, e.g. to elect a president, vote on the amendments of a law, etc.
- **File number**. A unique number that corresponds to the topic being discussed. There is a file number for each unique initiative.
- **Authors**. The individuals authoring the initiative. It can be MPs, the government, the king, etc.
- **Name of Session**. Type of session. For example, the main chamber, the labour commissions, the budgetary commission, etc.
- **Speaker**. The MP speaking for each intervention.
- **PDF link**. The link to the PDF file where the minutes are in the *Diario de Sesiones*.
- **Political group**. Feature created to have a unified label for the political group across terms.
- **Text**. The harvested text for each of the obtained interventions.
- **Pre-processed Text**. The text after being pre-processed.

- **Translated text**. The natural text after being translated to English for chapter 6.

4 Discourse Analysis

In this chapter, I performed an analysis of the discourse observed in the parliament. For such analysis, the occurrence of specific topics and the proximity of embeddings between them are used. The analysis spans the period between the years 2000 and 2022.

This chapter used two main methodologies, explained in each section. The analysis was done in an iterative way, trying out different topics and leaving the ones that present more fluctuation throughout the years or a more interesting evolution.

4.1 Topic Occurrence

For the occurrence count of specific topics, I performed a selection of topics based on current relevant themes in politics. Afterwards, I used the embedding model developed in chapter 5 to find terms that compose that topic. For that, I obtained the N topics most similar to a specific term, e.g. the topics most similar to «equality» and «feminism». After gathering all the terms, I counted the interventions in parliament that contained any of the words that compose each topic. Each of the sections below contains a figure for each topic and annotations on relevant events connected to it. The list of terms included in each topic is presented in the appendix A.

The main topics analysed were: Feminism, LGBTQ+ rights, labour rights, environment, Catalan territorial conflict and terrorism.

Although this methodology is simple, the lack of literature —and publicly available data— on this shows the relevance of this chapter.

To normalise these numbers, the percentage of interventions where a specific term is mentioned is used as a metric. Therefore, this section tackles the percentage of times specific topics are mentioned every year in the parliament between 2000 and 2022.

This section shows a small number of examples of how the corpus can be used to analyse the discourse of politicians. However, the code and terms used, along with the data, will be available in the GitHub repository of the thesis, at `/notebooks/term_occurrence.ipynb`, so that it is possible to use these tools to explore other topics.

Feminism

Figure 4.1 shows the occurrence, in percentage, of topics related to feminism in the parliament corpus.

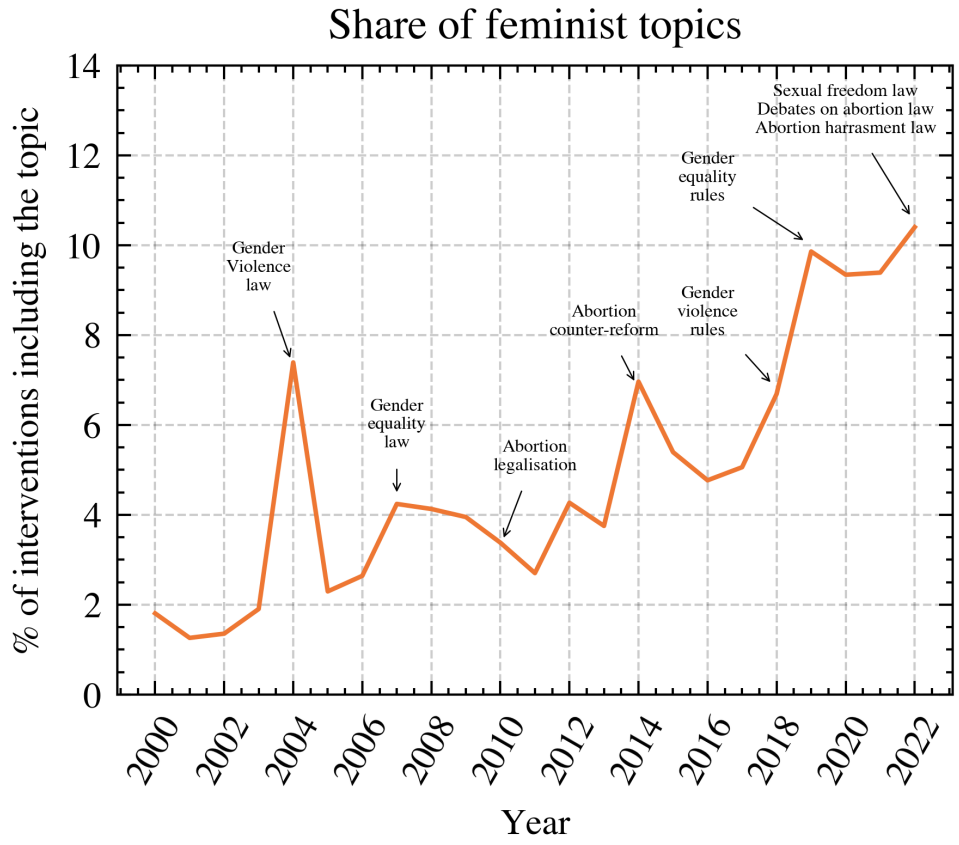


Figure 4.1: Share of feminism-related issues in the parliament corpus between 2000 - 2022.

The topics comprise anything connected with gender equality and feminist rights: abortion rights, the pay gap, sexual freedom rights, equality laws, etc. Along with the figure, annotation is provided for relevant events in the parliament related to feminism. These events were:

1. the promulgation of the first gender-based violence law;
2. the promulgation of the first gender-equality law;
3. the legalisation of abortion in 2010 within a 14-week period;
4. the abortion counter-reform conducted by PP limiting some of the aspects of the previous law;
5. the update of gender-based violence rules;
6. the promulgation of updated gender-equality rules; and
7. the debates and approvals of a sexual freedom law, a new abortion law and a law to prevent patients' harassment outside of abortion clinics.

As can be seen in the graph, there has been a pronounced increase in the topics connected to feminism in the parliament. During 2004 the promulgation of the first law against gender-based violence increased the share of equality-related topics to around 7.5% of all interventions for that year. From 2018 onward, the topic of feminist rights has been above 7%. The share of these topics has increased from a 2% in the year 2000 to an all-time high of 10.5% in 2022, which represents a more than 5x increase.

The increase in mentions of feminist issues could partially be explained by the increase in visibility and relevance of the feminist movements in Spain, especially since the 2018's mass strike that, according to [45], «*amazed both international and national public opinion*».

LGBTQ+ Rights

This section captured the share of mentions in the parliament to the topic of LGBTQ+ rights: same-sex marriage, same-sex adoption rights, anti-discrimination laws, gender self-determination, etc. The occurrence of this topic is presented in figure 4.2.

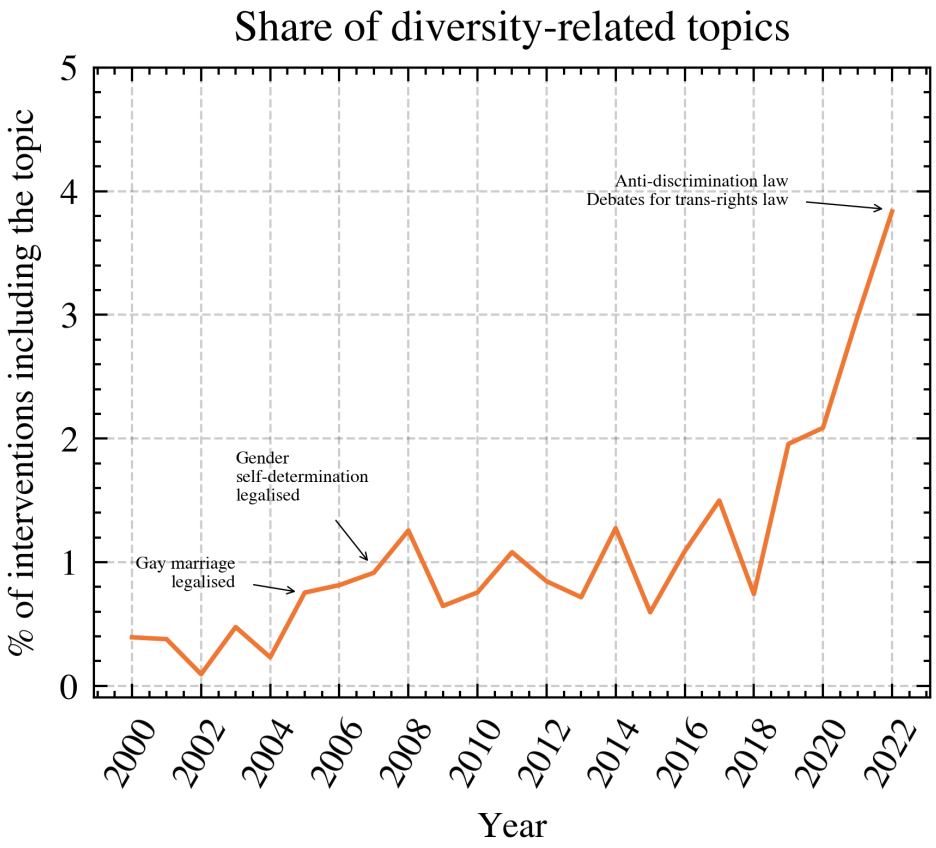


Figure 4.2: Share of mentions to LGBTQ+ rights in the parliament corpus between 2000 - 2022.

In 2005, same-sex marriage was legalised in Spain, and in 2007 the right to gender self-determination was granted under specific conditions. The chart shows an increase in mentions of diversity-related topics between 2005 and 2008. A decrease in mentions followed by a share fluctuating around 1% of all interventions in the parliament. Between 2018 and 2022, the usage of the term has increased from ≈1% to ≈4%. That is partially explained by the debates on a new LGBTQ+ rights law including more-relaxed gender self-determination rules and another law with anti-discrimination provisions for the LGBTQ+ population.

Labour Rights

The topic of labour rights included mentions of any term related to unemployment, precariousness, badly-paid jobs, time-bound contracts, collective bargaining, labour reforms, etc.

The evolution in the occurrence, observed in Fig. 4.3, has been an important topic of discussion in the parliament for the entire time series. I have included as well the yearly unemployment rate of Spain for the same period of time.

Figure 4.3: Share of mentions to labour rights in the parliament corpus between 2000 - 2022.

The minimum share that labour-rights topics received was between 2005 and 2007 when between 12% and 15% of the interventions mention it. This coincides with the lowest level of unemployment rate. Generally, the correlation between the yearly unemployment rate and the % of share that labour-related topics have obtained in the parliament is very high, as can be seen in the graph. The year with the highest share corresponds to 2012, with more than a third of all interventions in the corpus mentioning labour-related topics.

As expected, labour reforms and other legal changes to the framework of labour rules overlap with increases in mentions of the topic.

Environment

For the topic of environment, I included all mentions of any of the current ecological crises and their consequences: droughts, climate change, the biodiversity crisis, sea-level rise, desertification, plastic and air pollution, ecosystem restoration, etc. The figure below shows the share that this topic has had in the parliament for the last 22 years.

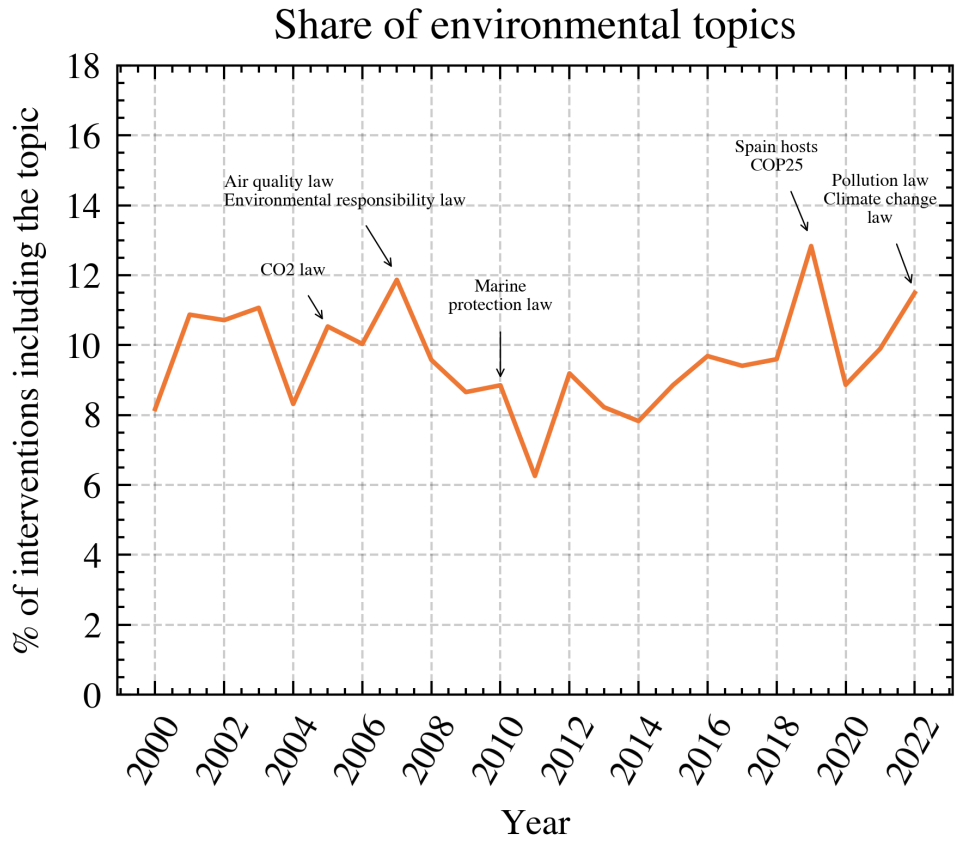


Figure 4.4: Share of mentions to environmental issues in the parliament corpus between 2000 - 2022.

The mentions in the parliament of environmental issues have remained stable oscillating between 8% and 12% with a minimum of 6% in 2011. The maximum value, 13% in 2019, coincides with the hosting of COP25 by Spain and the beginning of the debate on the Climate Change law. In 2022 two new environment-related laws were approved, likely being the reason for the increase in share in that year.

Catalan Territorial Conflict

For the topic of Catalan territorial conflict, I included mentions of the approval of the statute of autonomy of Catalonia in 2006 and the independence movement and Catalan nationalism. Figure 4.5 shows the share of the topic throughout time.

Figure 4.5: Share of mentions to the Catalan territorial conflict in the parliament corpus between 2000 - 2022.

Two main spikes occur in the time series, in 2006 and 2019. The high share in 2006 is likely caused by the approval of a new statute of autonomy in Catalonia that declared Catalonia a nation and increased independence from the state government. In 2006 a referendum was held where the citizens of Catalonia approved the new rules. In 2017, an illegal independence referendum was held followed by a unilateral declaration of independence on October 27. The Spanish government took control of the Catalan government, dissolved the parliament and called for elections in December of that year. The political leaders that called for the referendum were imprisoned and sentenced to jail in 2019. This led to a series of mass protests and civil blockades in Catalonia. The protests were a big topic of the electoral campaign happening in Spain at the end of 2019 and generally debated in the parliament, which explains the 9% of share during that year.

Terrorism

This topic mainly relates to terrorism, including domestic terrorism by *ETA*[1]. Figure 4.6 shows the occurrence of this topic in the parliament between 2000 and 2022.

[1] *Euskadi Ta Askatasuna (ETA)* was an armed group operating in the Basque Country that committed several terrorist attacks between 1968 and 2010 [46].

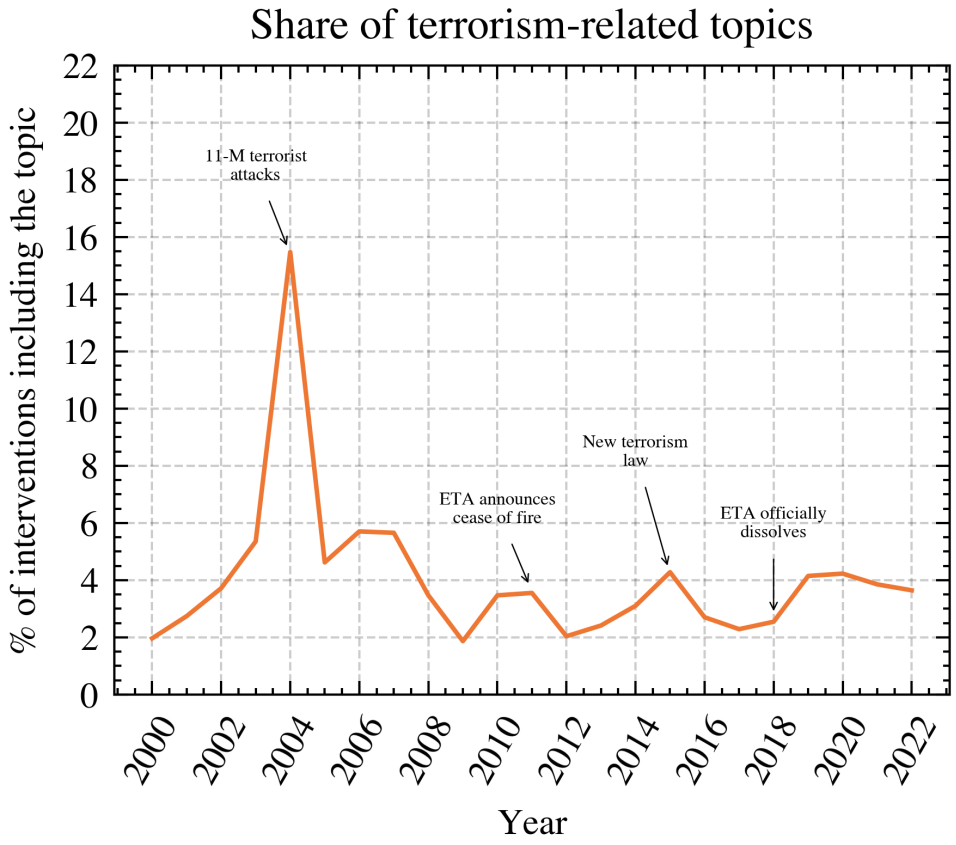


Figure 4.6: Share of mentions to terrorism in the parliament corpus between 2000 - 2022.

The graph presents a big increase in the occurrence of the topic in 2004. During that year, ≈16% of all the obtained parliamentary interventions include terms related to terrorism. This is due to the terrorist attacks that occurred in the country in 2004. We can also observe an increase in share in 2010, 2011 —when the group announced a cease-fire—, 2015 —when the parliament passes a new anti-terrorist law and there was an attack in Paris— and in 2019 onward.

Between 2019 and 2022, the mentions of terrorism have increased considerably to around 4% of all interventions, although ETA announced its dissolution in 2018 [47].

The occurrence of these terms not only shows the presence of terrorist attacks in Spain but also serves, to some extent, as a thermometer of polarisation: some MPs and political parties from the right-wing spectrum tend to refer to some of their political opponents as «friends of ETA» or «(friends of) terrorists», especially since 2019, due to the proximity of the current coalition government to a left-wing Basque political party.

4.2 Term Propagation

This section focuses on analysing novel terminology from the parliament and observing which political groups started using them and how those terms were later on adopted by other political groups. Here, we zoom in to observe the transmission of terms across political groups with very specific tokens not mentioned before in the parliament.

For this section a search of novel terms was conducted, identifying as such: «filoetarra/s», «bilduetarra/s», «socialcomunista/s», «emergencia/crisis climática», «constitucionalista/s». The term «populista/s» is also included in the analysis due to its increased usage in recent years.

Although other words were analysed —e.g. «fascista/s», «golpista/s»—, none of them showed results that would showcase what this section aims to show. We focused on terminology that tends to be used as an offence between MPs, rather than terminology connected to public policies or common terms, since it is easier to interpret without context and contains a higher ideological and polarising baggage.

Fig. 4.7 shows the occurrence of the term «filoetarra/s» in the obtained corpus of the parliament for each of the 5 main political groups. Both the «filoetarra» and «bilduetarra» terms refer to an individual that supports the dissolved separatist and terrorist group ETA. «Etarra» refers to the name of a member of this separatist group. «EH Bildu» is a basque self-denominated left-wing political party.

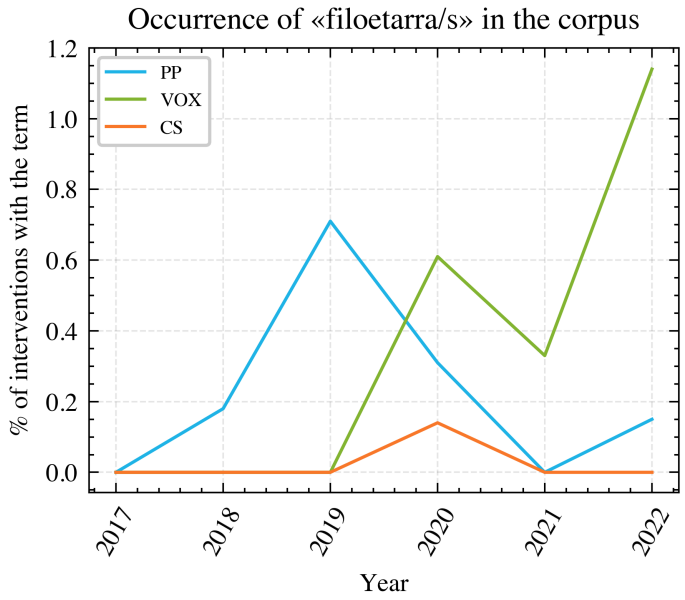


Figure 4.7: Occurrence of «filoetarra/s» in the corpus for each political group.

The occurrence of the term starts in 2018 in the parliament, used by PP. In 2019 it is used again by the same political group. From 2020 onward we see the other two political groups that are closer to PP adopt the term and use it several times more than themselves. We can find in this corpus 15 interventions in 2022 alone where the words «filoetarra/s» were mentioned.

Figure 4.8 shows the occurrence of the term «bilduetarra/s».

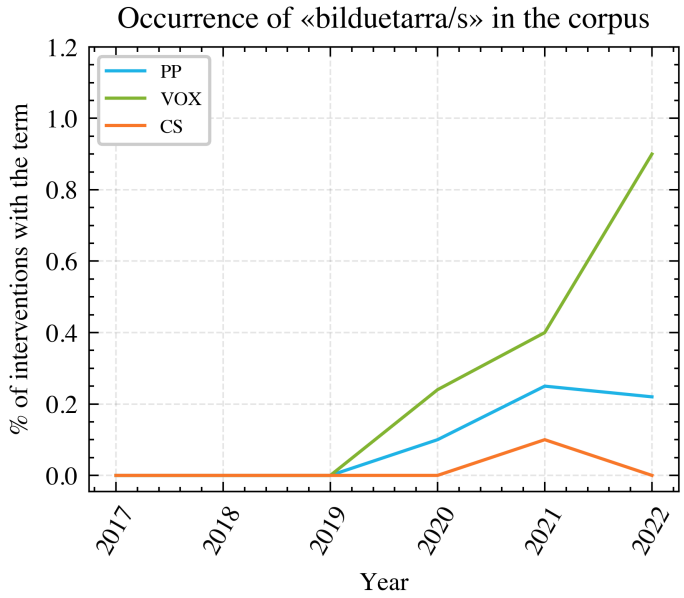


Figure 4.8: Occurrence of «bilduetarra/s» in the corpus for each political group.

The term starts existing in 2020 —the first political group to use it is PP in February of that year— by PP and VOX. Afterwards, its usage increases in 2021 and CS uses it once. In 2022, VOX is the main party using the term, with 11 interventions in the corpus across 2022 including it. Although, according to the data gathered, PP is the first party to use the concept, VOX ended up using it more.

The term «socialcomunista/s» has been used in the parliament to refer, mainly, to the coalition government formed during the 14th term. As we can see in figure 4.9, the term was not used before in the parliament, according to our data. The first time the word appeared in our data was in February 2020 by PP and VOX. Right after that, the word also started being used —with irony— by the two groups that are part of the coalition government, PSOE and UP. The usage of the term has increased afterwards, reaching a peak in 2021, with 94 interventions in our dataset containing it.

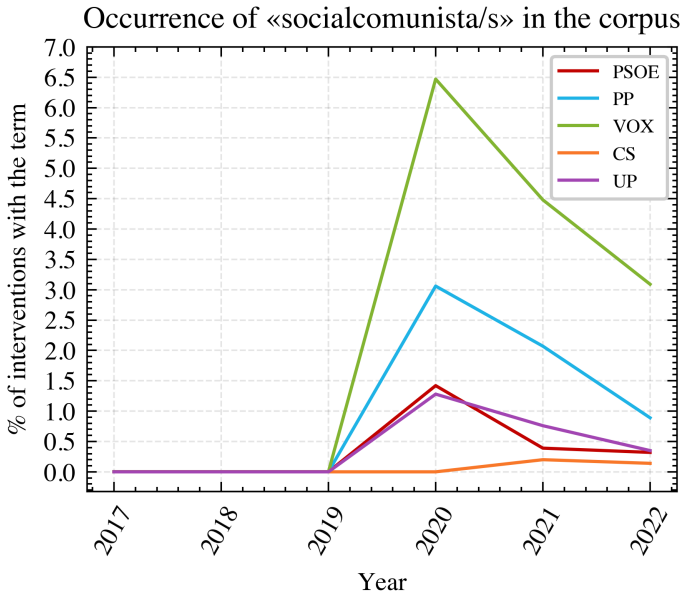


Figure 4.9: Occurrence of «socialcomunista/s» in the corpus for each political group.

The term «populista/s» refers, according to Collins Dictionary, to politicians that conduct political activities or ideas that claim to promote the interests and opinions of ordinary people.

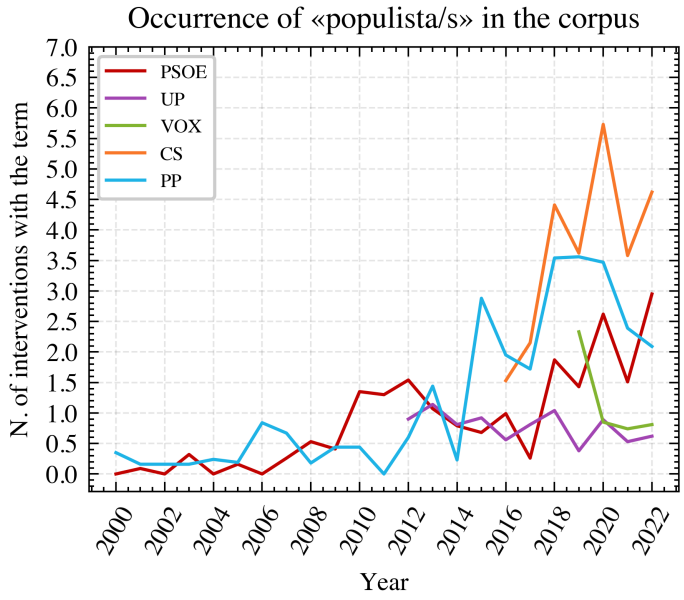


Figure 4.10: Occurrence of «populista/s» in the corpus for each political group.

The term was used during the whole time period of the dataset, however, we can observe in figure 4.10 a notable increase in recent years. Reasons for this include the usage of the term to refer to UP and VOX by PSOE, PP and CS.

The occurrence of the terms «climate crisis» and «climate emergency» was also analysed in the following figure.

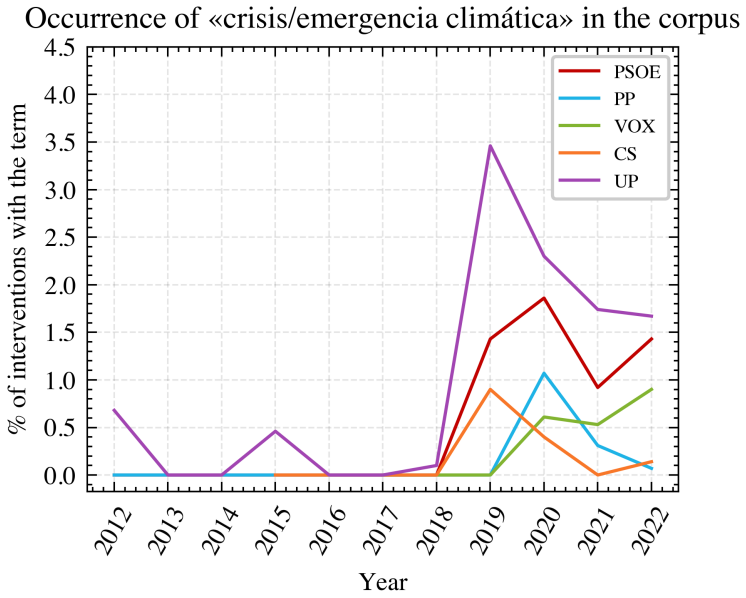


Figure 4.11: Occurrence of «crisis/emergencia climática» in the corpus for each political group.

The term is mentioned the most by UP and CS in 2019 and PSOE and PP in 2020. During this period, several policies were debated and enforced in the parliament connected to climate change. The term first appears in the corpus in 2012 used by UP, but it is not until 2018 that it is being used at a wider scale. PP and VOX start using the term only in 2020, in the case of VOX denying climate change or sarcastically referring to the term, as the next excerpt from an intervention by an MP of VOX illustrates:

«No hay ninguna emergencia climática en España, y, si la hay, ahora mismo le aseguro que les importa un rábano a los españoles, que solo ven la verdadera emergencia social y económica a la que su desastrosa incompetencia los ha llevado.» (Translated: "There is no climate emergency in Spain and, if there is, I assure you that spaniards don't care a single bit, since they see the true social and economic emergency that your disastrous incompetence has lead them to.").

Finally, we analyse the occurrence of the term «constitucionalista/s». This term refers to individuals or, in this case, political groups that defend the Constitution or constitutionalism. It was recently used by PP, VOX, CS and also PSOE within the context of the independence Catalan movement, but not only. Fig. 4.12 shows the results of the occurrence.

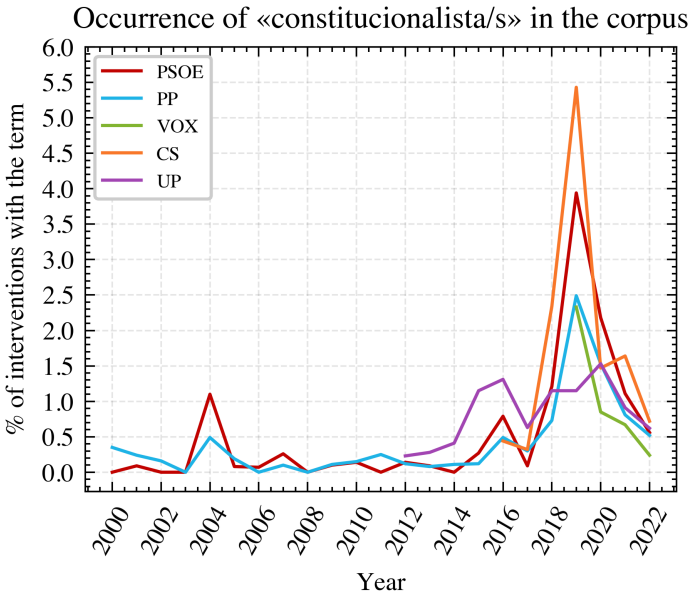


Figure 4.12: Occurrence of «constitucionalista/s» in the corpus for each political group.

During 2019, there were barely any ordinary sessions in the parliament as the elections were repeated twice in 2019. The high occurrence of the term probably has to do with the attempts of political groups to form a government and the usage of the term to refer to themselves in contrast to their political opponents. CS, a group that placed themselves in the «central» political spectrum, was the group that used the term the most, in up to 5.5% of all recorded interventions in 2019.

5 Ideological Polarisation

In this chapter, we analyse the ideological polarisation of the parliament over the last 2 decades. For that purpose, we try to analyse the discourse proximity —and therefore, the ideological placement— between political groups for the last 7 terms and assess if political groups are having a more diverging or a more similar discourse from the point of view of the concepts and ideas they talk about. The methodology of this chapter was performed following the work of [24]. The code for this chapter can be found in the GitHub repository of this thesis, in `src/notebooks/ideological_polarisation.ipynb`.

The chapter explains the processing applied to the data and the state of the training dataset, the model used and the alternatives discarded, the hyper-parameter tuning applied to the model, its evaluation, the reduction of dimensionality for visualisation, and the results obtained.

5.1 Training Dataset

The resulting dataset from the pre-processing applied in chapter 3 needed to be adjusted to comply with the requirements of the Doc2Vec model. A new dataset containing three columns: political group - term; term; and pre-processed texts was built as Tagged Document data types. An example can be seen in table 5.1.

Table 5.1: First 5 elements of the dataset used to train the Doc2Vec embeddings model.

	Political Group	**Term in office**	**Text**
0	UP-7	L07	['candidato', 'primer_lugar', 'lógico', 'reconcimiento'...]
1	PSOE-7	L07	['aznar', 'quiero', 'primeras_palabras', 'reconocimiento'...]
2	PP-7	L07	['cosa', 'acuerdo', 'efectivamente', 'país_vasco'...]
3	UP-7	L07	['izquierda_unida', 'presentado', 'serie', 'catalán'...]
4	PP-7	L07	['comenzar_intervención', 'permítanme', 'felicitar'...]
...	...	...	...

Since the purpose of this chapter is to assess the ideological proximity of the different political groups across terms, each political group for each term constitutes a «tag» or «document» for the model. All the interventions that were performed by members of that political group during that specific term in office form the tag and each of them counts as a different instance of it. Due to this level of granularity, the model provides a vector that corresponds to each political group per term, which allows us to place them in the embedding space and compare their relative distances.

A second tag was added, corresponding only to the term in office for each intervention. This was a crucial step which allowed us to isolate the term-dependent tokens and focus on ideological differences between political groups. Without this two-tag system, the model was just creating embedding spaces that were placing the different political groups nearby due to the term in office they belonged to, since they were using the same tokens due to tackling the same topics. The second tag isolates the effect of each individual term, making it possible to create vectors that focus on the proximity of political groups regardless of the topics in a specific term.

Further processing was applied to the data before using it for training. Small and regional political groups were removed from the training dataset, leaving only state-wide groups, namely PP, PSOE, UP, VOX and CS.

The final training dataset consisted of 58.812 interventions containing 20.937.824 tokens distributed across terms in office as presented in the figure below.

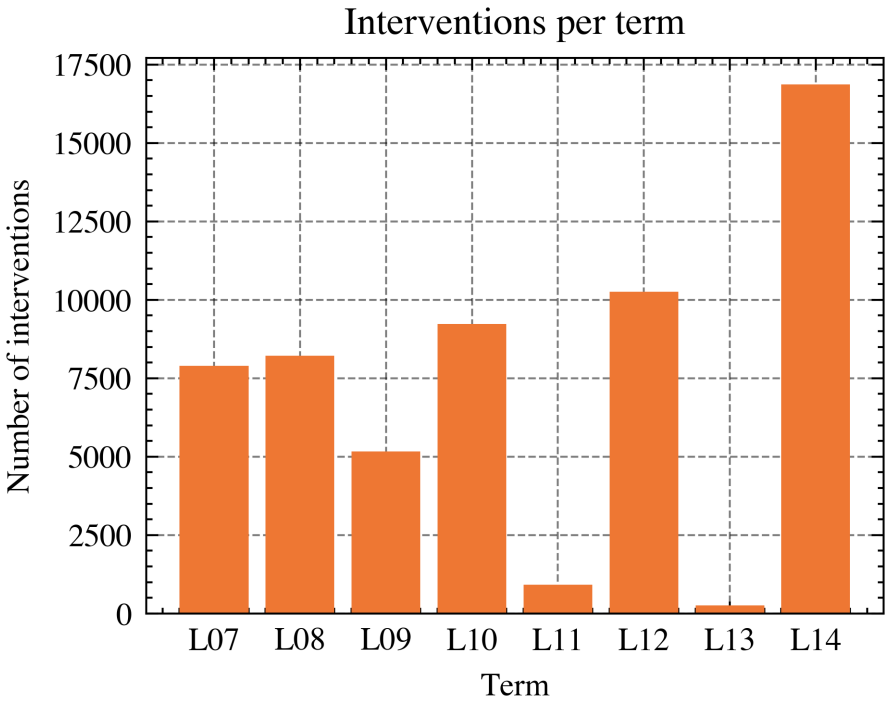


Figure 5.1: Number of interventions for each term included in the training dataset. Terms 11 and 13 contain a very limited amount of interventions due to their limited duration caused by the repetition of general elections due to not forming a government within the legal deadline. Other differences in numbers are mainly explained due to the increased number of political initiatives discussed, more political groups and fragmented speaking time among several spokespeople.

5.2 Models

The models used for this chapter were document embeddings (or «paragraph embeddings»). Specifically, the definitive model used was the Gensim Doc2Vec model [31].

This document embedding model consists of a neural network that learns word associations from a large corpus of text. The words are vectorised, obtaining a vector of length N corresponding to each unique word in the corpus. The weights of the first neural network layer become the vectors that correspond to each word. A more elaborate explanation of how document embedding models work is included in chapter 2.

In the model used in this chapter, the size of N was 50. The embedding model also aggregates the vectors corresponding to each word in a document and creates document vectors, for each of the tags used in the model. In my case, I obtained 29 vectors representing the 5 political groups for the terms where they had political representation and a vector to represent the 159.774 unique tokens in the corpus.

I trained several models to try to obtain document embeddings both for each token in the corpus and for the political groups of each term. The hyper-parameters were tuned and both qualitative and quantitative evaluations were performed.

Once the models were trained, PCA was applied to interpret the results by observing them in a 2-dimensional plot, therefore building a 2-axes semantic-proximity map of political groups

throughout the last 7 terms in office. Along with this map, another one containing some of the most relevant tokens utilised in the parliament was drawn.

The model used was the Document embedding model outlined in [30]. The implementation of the model used was the *doc2vec* model from the Gensim package.

5.3 Discarded Alternatives

Besides generally exploring text-to-vector models, such as Word2Vec and GloVe, other alternatives were further evaluated and discarded.

One of the alternatives consisted to use a pre-trained embedding model. This solution was discarded since (1) we obtained good results with the implemented model, (2) there are very few pre-trained Spanish models of document embeddings and (3) the existing models are mainly trained using tweets or news articles, whose structure, vocabulary and linguistic register differ greatly.

Additionally, a model using as a tag a politician-term was also trained. This model was not very successful since it did not manage to place together politicians belonging to the same political group. The reason behind this rests in the very nature of interventions in the parliament: the variance of speeches per MP is high. Ranging from hundreds of interventions to less than 10 depending on the MP. It is assumed that unless a big corpus of texts is available, it is not possible to distinguish, focusing on a low number of interventions, what is the political ideology of the person speaking. However, having hundreds of interventions to analyse, it becomes easier to draw patterns within them that give us an indication of their ideology. This reason also did not allow me to create an embedding model using tags for the political group-year. Perhaps it would be possible if the entirety of the corpus would have been captured during the harvesting process.

5.4 Hyper-Parameter Tuning

Several hyper-parameters were available for tuning the model. Those were: whether to use the distributed memory (DM) or distributed bag-of-words (DBOW) training algorithms, the size of vector representation, the window or context of words, the minimum amount of times a token needs to appear to be included, the number of epochs and the threshold at which high-frequency words are randomly down-sampled.

Below, I explain the approach followed to tune each of the hyper-parameters.

- **Distributed Memory vs. Distributed Bag-of-Words**: the DM model was chosen. First of all, because it takes into account the order in which words are presented in the corpus, unlike with DBOW; secondly, because after training two different models, each following one of the approaches, the DBOW model completely failed to approximate the different political groups across terms and the representation of embeddings did not take as a reference any relevant terminology.
- **Size of vector representation**: the size chosen was 50 after performing an experiment where models with vector size values from three orders of magnitude from 10 to 1000 were trained. Afterwards, several models were trained with vector sizes 20, 50 and 100. In both experiments the model with vector size 50 qualitatively performed better than others. This was assessed by observing the representation of embeddings, where the proximity within the same political group was notorious, and by also analysing the terminology associated with each PCA-reduced dimension. The terminology used to place the vectors in the embedding space included ideologically-rich concepts.

- **Window**: Due to parliamentary interventions being of high length, it was assessed that a larger-than-usual window was necessary to grasp the context in which a word is being used. After training several models with different window sizes, the window size chosen was 20.

- **Minimum count**: Increased slightly from the default (2) to 5, due to having a large corpus with +20M tokens.

- **Epochs**: The best accuracy was achieved with around 20 epochs, compared to 5, 10, 100, 200 and 500. When having fewer or more epochs, the model performance drifted and took into account tokens that were not related to the political ideology of the political groups.

- **Sample**: following [48], a specific focus was put into the down-sampling threshold. For that, different down-sampling thresholds were analysed (between 10^{-5} and 10^{-1}). We found out that the best value was 10^{-3} since it was the model that contained the best vocabulary for discerning the ideology of the different political groups and took into account the least amount of irrelevant words, such as verbs or adverbs.

5.5 Model Evaluation

The evaluation of the models represented a big challenge for this thesis, since, having a Spanish Doc2Vec model and having such a specific corpus did not allow the use of any of the already-existing annotated evaluation methods, such as word analogies. Additionally, due to document embedding models' interpretability not being easy, dimensionality reduction was implemented right away as an intermediate step to qualitatively interpret the model and observe its results.

We evaluated the performance of the models using both qualitative and quantitative methods.

The model was evaluated quantitatively by computing an accuracy metric. The accuracy metric was calculated by leaving aside 10% of the interventions, computing their embedding vectors, and afterwards the proximity of the non-included interventions to their political group. If the interventions were placed the nearest to other interventions by the same political group, they were considered to be correctly placed. The accuracy metric is the amount of correctly-placed interventions divided by the total amount of analysed interventions. Figure 5.3 shows the results of this metric for the accuracy of an intervention being placed the closest to its political group, and among the top 2, 3 and 4 closest groups, out of a total of 5 groups (PP, PSOE, CS, VOX, UP).

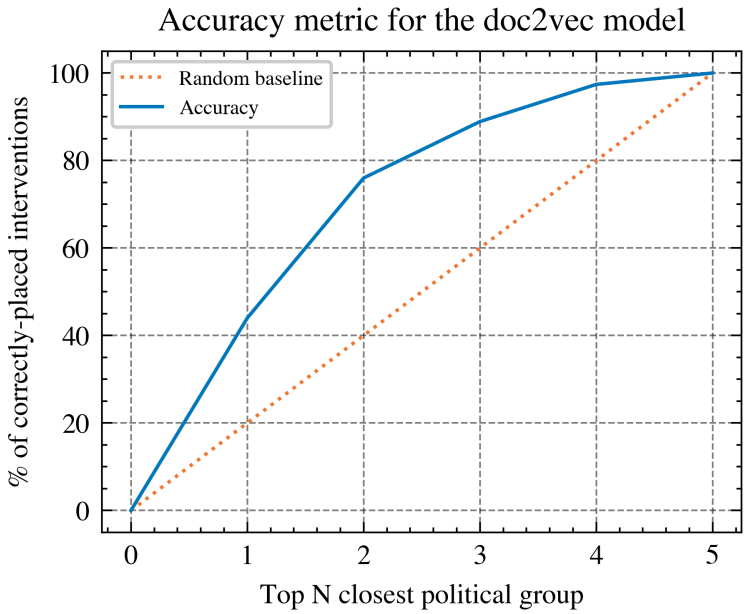


Figure 5.2: Accuracy metric for the document embedding model.

Cross-validation was performed by removing 10 times a random 20% of the interventions and training different models and dimensionality reduction algorithms to obtain a cloud of points corresponding to the location of each tag in the embedding space. The following figure shows the result of the cross-validation. It is the same figure as 5.4 but it presents clouds of points corresponding to the regions that contain each political group-term.

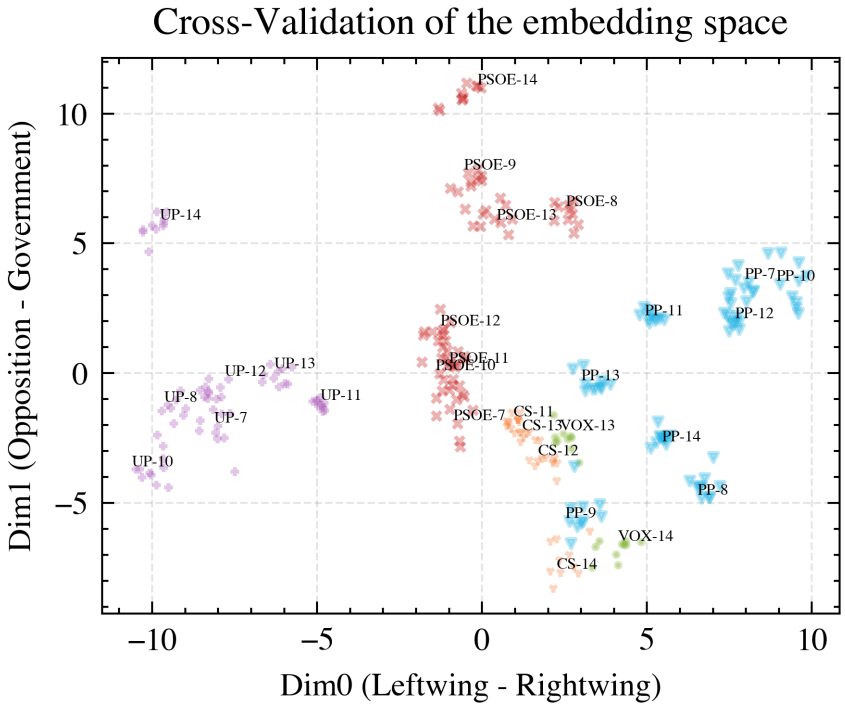


Figure 5.3: Cross-validation of the document embedding model.

The qualitative evaluation consisted of observing the map of concepts and its proximity to each

political group and the reference words associated with each extreme of the PCA two dimensions.

Additionally, I evaluated the model by comparing it with the results of other polarisation metrics for the parliament. These results are found in section 7.1.

5.6 Results

I was able to create a vector space that positions political groups and tokens in proximity to each other. The results suggest that, after performing dimensionality reduction, the model is able to place political groups based on their ideology and based on whether they are in the government or in the opposition.

In the following sections, I present the main results containing the reduced embedding space representations, the evolution of political groups regarding the two dimensions, a map of ideological concepts in that embedding space, proximity representations between political groups and concepts and finally an ideological polarisation index.

Embedding Space

The model was successfully trained, being able to place political groups in a 2-dimensional space based on their parliamentary interventions. The embedding space is presented in Fig. 5.4. The dimensionality reduction was able to capture approximately 70% of the variance in the model for three components. However, we disregard the third component and only focus on the first to (60% variance) for the analysis and interpretation of the model.

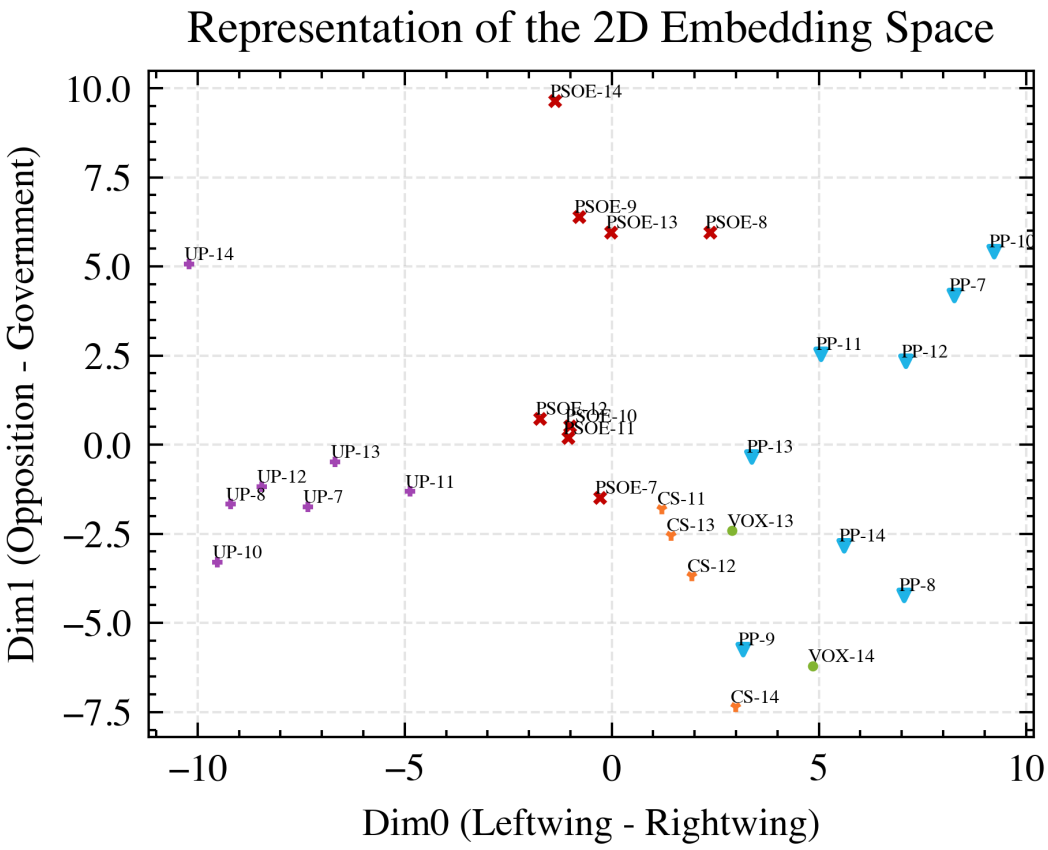


Figure 5.4: Representation of the 2-dimensional embedding space. Ideologically-close political groups appear nearby, while more different political groups show at a higher distance.

The figure shows an embedding space where political groups perceived as right-wing appear in

the right area of the plot, while the groups perceived as left-wing appear on the left. Three main areas are present: the left area, including only UP, the middle and top area, including only PSOE, and the right area, including PP, CS and VOX. This classification also goes in line with the way political groups vote in the parliament and their ideological alignment, however, it was rather surprising to see the distance between PSOE-14 and UP-14 which are two political groups that are part of the same coalition government.

Among all legislative terms, it is observable that the latest, 14, presents the highest level of ideological polarisation since the distance between PP-14, VOX-14, CS-14, PSOE-14 and UP-14 is the greatest in the plot. In order to visualise these distances, two more figures are presented, one for the time-series plot of each dimension.

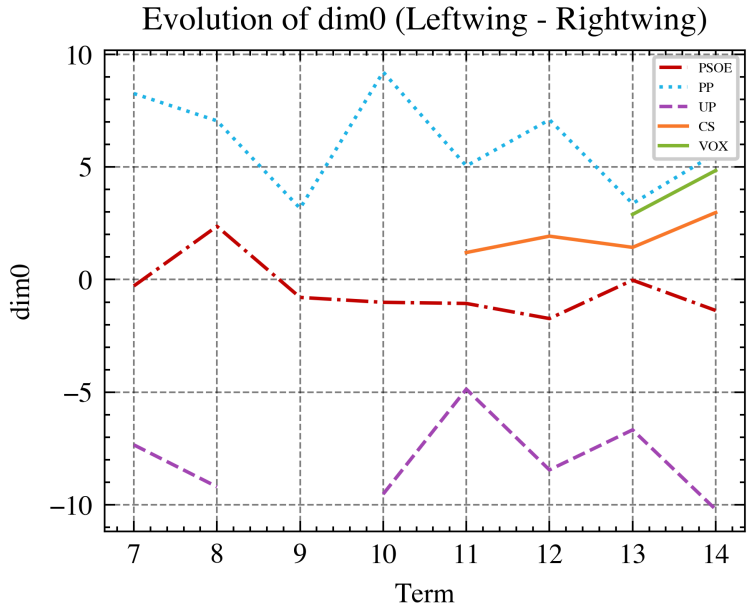


Figure 5.5: Evolution of dimension 0 of the embedding space throughout terms for each political group.

The above plot shows the distances between political groups across terms for dimension 0. Based on the terminology associated with each area of the embedding space, as shown in table 5.2, it would be safe to assume that the current plot shows ideological differences between the groups. UP and PP tend to be the most distant political groups. When CS appeared, it occupied a space between PSOE and PP. VOX and PP are very close to each other in the plot, suggesting that the terminology they use is similar.

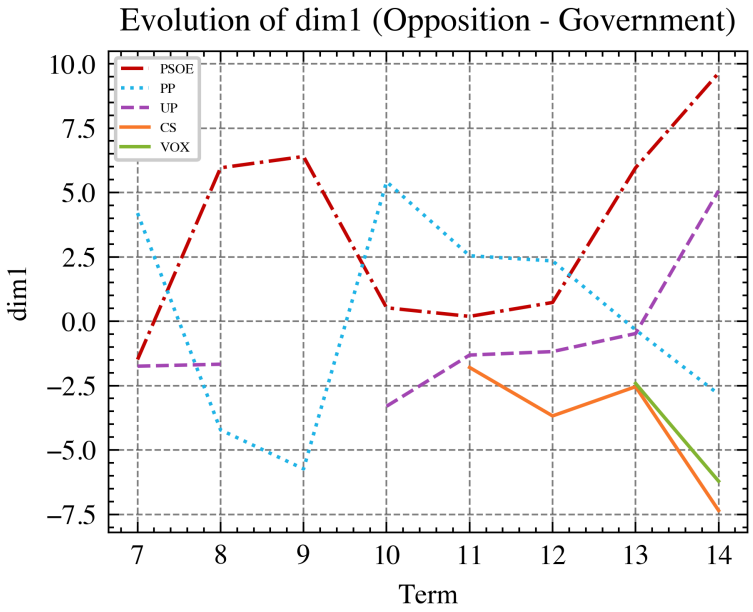


Figure 5.6: Evolution of dimension 1 of the embedding space throughout terms for each political group.

The values in the above plot show a high correlation with a government-opposition axis. During terms 7, 10, 11[1] and partially 12[2] PP was in power. In the plot, the values of PP for those terms are higher than any other political group. During terms 8, 9, partially 12, 13[3] PSOE was in power. In the plot, PSOE has the highest values for those terms. Finally, during term 14, both PSOE + UP are part of a government coalition. It seems that for this dimension, political groups governing obtain high results and political groups in the opposition, low ones. This suggests that being in the government or being in the opposition inherently changes the discourse of a political group, something we can also see analysing the terminology in table 5.2.

Both dimensions show an increase in distance between the different political groups, therefore, based on these results, it is possible to state that the ideological polarisation —seen from the point of view of the terminology employed by political groups— has been growing in the last years. Section 5.6 measures that level of polarisation developing a single index for it.

Table 5.2 presents a list of the 50 tokens that are the closest to each of the edges of the two dimensions.

[1]During the 11th term there was only a caretaker government in office.

[2]During term 12 there was a successful vote of no-confidence, so the government at the time was ousted from power.

[3]During the 13th term there was only a caretaker government in office.

Table 5.2: Top 50 tokens associated with each edge of the two reduced dimensions (in Spanish).

	Edge	Tokens
Dimension 0	Leftwing	Recortes, derechos laborales, privatización, trabajadores y trabajadoras, ciudadanía, recorte, privatizaciones, planteamos, derechos sociales, grandes empresas, cuidados, poderosos, derechas, movimiento feminista, organizaciones sociales, movilizaciones, planteábamos, neoliberal, recortar, salario mínimo, sanidad pública, trabajadoras, Trillo, ambiental, Montoro, desahucios, sindical, ultraderecha, precariedad, servicios públicos, estafa, especulación, Gallardón, empresas privadas, grandes corporaciones, paraísos fiscales, derecha, represión, LGTBI, entendíamos, Troika, grandes fortunas, sindicatos, inversión pública, franquismo, pública, iglesia católica, salarios, eléctricas, franquista.
	Rightwing	Víctimas del terrorismo, ETA, impuesto renta personas físicas, Sánchez, Zapatero, concordia, nación española, autónomos, territorio nacional, impuesto renta, explosivos, dirección general de tráfico, deportistas, Rodríguez Zapatero, Pérez Rubalcaba, nación, Coalición Canaria, PYMES y autónomos, ejército de tierra, País Vasco, subida de impuestos, terroristas, contribuyente, lucha terrorismo, directores generales, unidad de mercado, Otegi, vasco, abro comillas, castellano, inmigración ilegal, sociedad pública de alquiler, miente, proveedores, sector pesquero, Argelia, Cervantes, Gran Canaria, Celaá, consejo fiscal financiera, pescadores, creación de empleo, caldera, Ábalos, dispersión, padres, buques.
Dimension 1	Government	Pronuncia palabras perciben, pide la palabra, Gamarra, gobernamos, ruido, despliegue, concertación, incrementó, Casado, participaron, gobernaban, Olabarría, quiero agradecer, incendios, Abascal, derecha, poner en valor, pensiones mínimas, quiero aprovechar, voy a dar, intenso, duda alguna, primera intervención, permitirá, gestos negativos, disponibles, gobernaron, total millones de euros, comenzó, conocen, interlocutores sociales, emergencia sanitaria, avanzando, resiliencia, primera vez, protestas, reforzando, impulsó, Ayala, mejorando, momentos difíciles, seguir mejorando, amplía, Zaplana, Galicia, reforzado, tranquilidad, protección social, puso en marcha, agradecer.
	Opposition	Vetado, pedíamos, vetan, funcionarios de prisiones, denunciado, proponíamos, dinero público, inaceptable, dimitir, bipartidismo, incomprensible, tememos, arreglar, dinero, ineficiente, promesas, arbitrariedad, pedimos, reprobación, escándalo, AIREF, depender, metido, papel mojado, mecenazgo, acaben, prometió, vetar, exigimos, cese, propusimos, alarmante, indulto, consideramos, meter, servicio exterior, inminente, solicitamos, desaparece, vemos, marca España, suicidio, separatistas, dotación presupuestaria, personal laboral, caro, parche, urgente, colmo, altos cargos.

The table shows that left-wing political groups, placed to the left of the embedding space, tend to be closer to concepts such as «labour rights», «social rights», «neoliberal», «Montoro» (ex-minister of finance of PP.), «unions», «human rights», «public services», «precariousness», «tax

havens», «minimum wage». The political groups on the right of the embedding space are closer to concepts such as «taxes», «terrorism victims», «public deficit», «terrorists», and «freelancers». Specific names of politicians tend to be in the extremes of the plot, as they are usually used by their political opponents when referring to them.

Top and Bottom differences are, in the table, more connected with terminology related to being in the government or in the oppositions, such as «continue improving», «deployment», «increased», «moving forward», «reduce the deficit», «governed», or, on the other hand, «vetoed», «we were demanding», «we were proposing», «unacceptable», «[you] promised», «inefficient», «we demand», «scandal», «censure», «resign», «wet paper» (an expression to refer to something unfruitful and not credible).

Map of Concepts

Since in the Doc2Vec model, both word vectors and document vectors are trained in the same embedding space and the first is calculated through performing mathematical operations with the latter, it is possible to represent a series of relevant tokens into the embedding space. Fig. 5.7 shows a map of the tokens placed in the same embedding space as the political groups.

The scale of numbers in the map of concepts and in the embedding space was differing, but since these numbers only matter for the sake of placing them with relative distances between each other, the values of the embedding space were scaled down to match the values of the map of concepts.

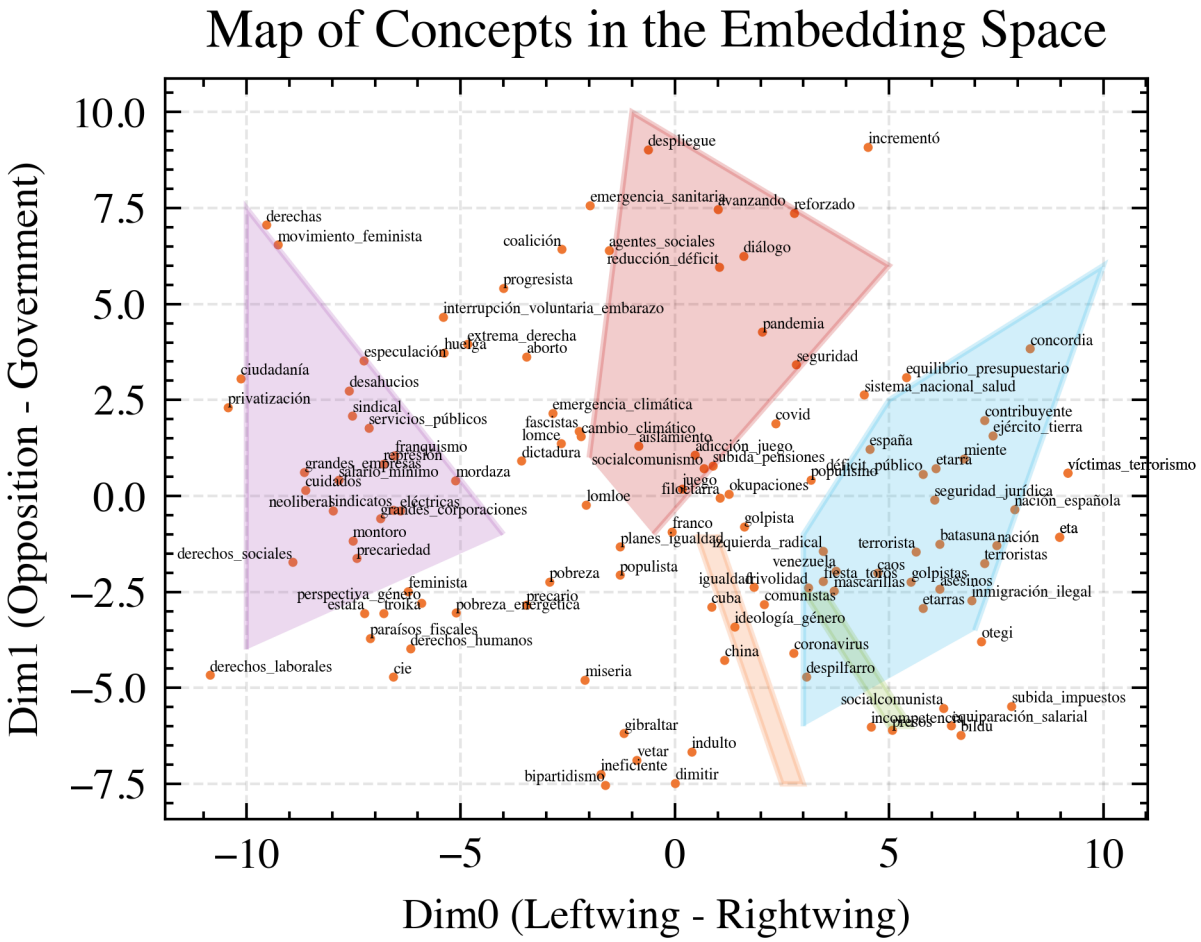


Figure 5.7: Map of concepts in the embedding space. Concepts more closely connected to political groups perceived as left-wing are on the left of the plot and vice-versa. The areas represent the position of the political groups in the embedding space.

Vector Proximity

The model also allows us to compute the proximity among political groups, concepts, and between political groups and concepts. Table 5.3 illustrates this, by showing the top 10 most similar concepts to each political group in the 14th term.

Table 5.3: Top 20 concepts associated with each political group in the 14th term (in Spanish).

Political group	Most similar concepts
PSOE-14	Derecha, Gamarra, Casado, gobernamos, significa, primera vez, ruido, fortaleza, concertación, saben, permite, duda alguna, Zaplana, coalición, derechas, derecha y ultraderecha, liderado Pedro Sánchez, partir, compromiso firme, permiten.
PP-14	Sánchez, Ábalos, socialcomunista, Pedro Sánchez, Zapatero, populistas indepenentistas, españoles, Celaá, doctor Sánchez, sanchismo, preguntamos, incapacidad, soberbia, Moncloa, incompetencia, caos, Marlaska, Ribera, socios separatistas, Rodríguez Zapatero.
UP-14	Derecha, derechas, trabajadores y trabajadoras, gente trabajadora, ciudadanía, ultraderecha, movimiento feminista, allá, gente, trabajadoras, terreno, movilización, dejéis solos, fíjense, unidas mayoría población, fuertes, Asturies, clase trabajadora, públicas.
VOX-14	Sánchez, ruina, socialcomunista, separatistas, españoles, siniestra, populistas independentistas, asesores, haberse, sustitución, moncloa, Iglesias, enemigos, sistema, agenda globalista, bilduetarras, socios, cuyo, nación.
CS-14	Vetan, proponíamos, vetado, pedíamos, propusimos, Esquerra, creemos, propuesto, Sánchez, presentábamos, vemos, flexibilización contratación incentivar talento, desaparecido, pongan sectores estratégicos, sido vetada, veta, muchísimas comenzar repitamos, seguimos pensando, independiente, lamentamos.

A lot of the words above correspond to specific politicians with leadership roles in their political groups (e.g. spokesperson, minister, leader). In the parliament, it is very common to hear interventions where MPs refer to specific politicians, either to reply to them or to criticise/praise their work. Besides that, other concepts such as «arrogance», «incompetence», «chaos», «ruin», «social-communist» and «enemies» are highly associated with PP and VOX, CS also has tokens connected to being in the opposition («[you] vetoed», «we proposed», «we asked for».). Both UP and PSOE refer to the opposition as the right, and while PSOE has concepts associated with specific politicians, UP seems to have more terminology associated with the wider public («workers», «citizens», «the majority of the population», «working class»).

As an example of the computation of proximity between political groups and concepts, Figure 5.8 presents the proximity between specific terminology, in this case, «feminist movement» and the political groups across terms [24].

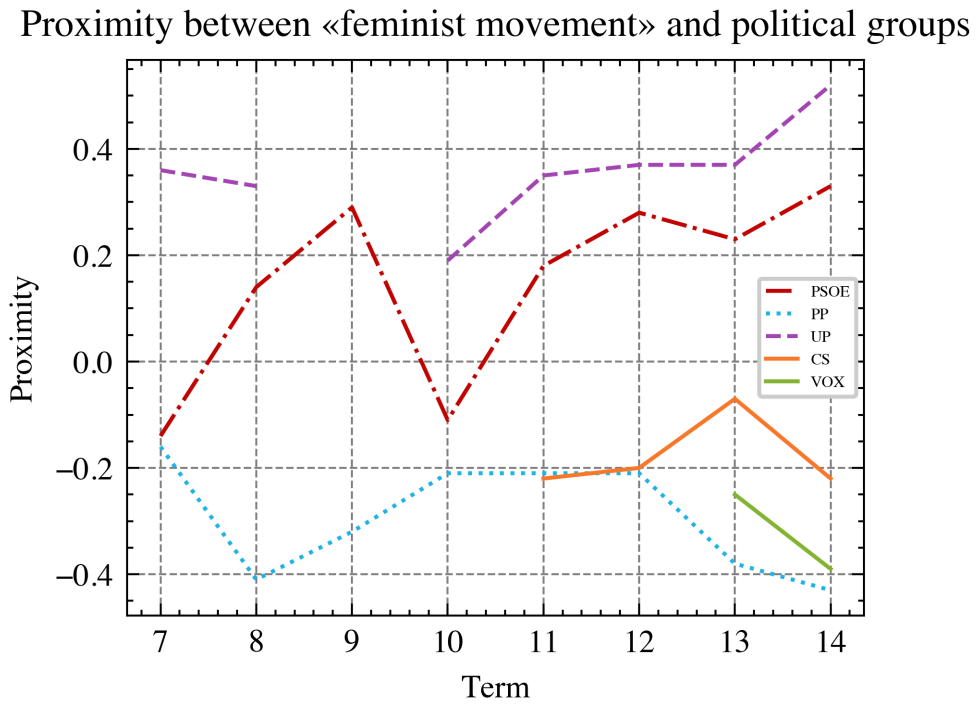


Figure 5.8: Proximity between «feminist movement» and political groups across terms. The higher, the closer.

There is a high relation between the concept «feminist movement» and UP and PSOE, especially in recent terms. CS, VOX and PP are on opposite sides since the proximity between them and the concept presents very low numbers. In recent years, the proximity between the concept and both UP and PSOE has increased, while with PP, CS, and VOX, it has decreased.

The same analysis could be applied to any other token that is present in the vocabulary of our model. This would allow us to analyse up to 159.774 tokens.

Distances between Political Groups

In order to assess the ideological differences between political groups, the Euclidean distance between the two political groups can be calculated. As an example, we show the Euclidean distance between PSOE and PP. Figure 5.9 shows this metric for the 7 terms.

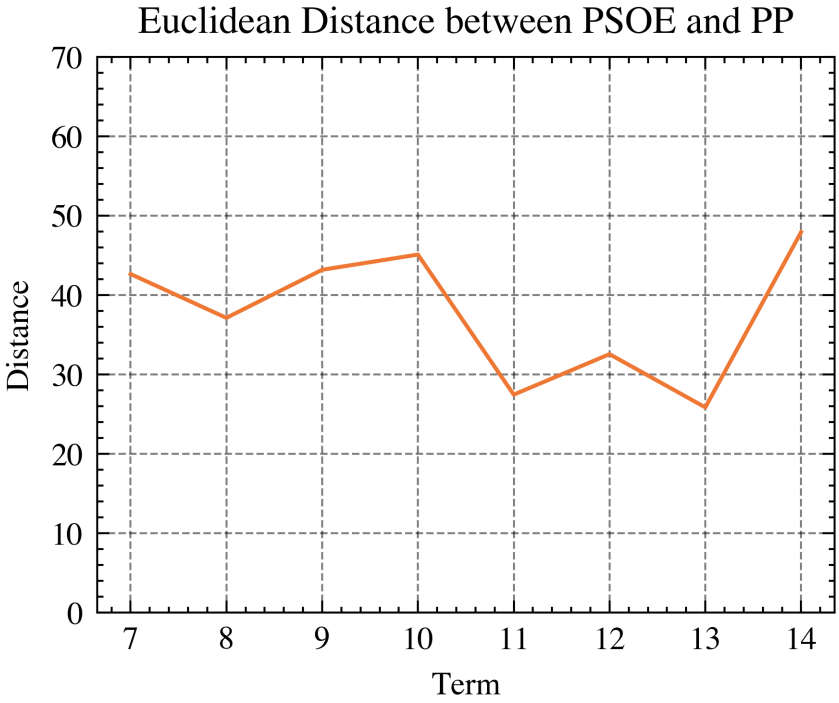


Figure 5.9: Euclidean distance between PSOE and PP.

The Euclidean distance for the vectors corresponding to PSOE and PP are, for the current term in office, at the highest distance from each other in the entire series. This means that their discourse is the most distant for the last 7 legislative terms. We can also observe a steep decline during term 11, 12 and 13. That could be explained due to terms 11 and 13 only lasting a few months without a government formation, which does not allow the model to determine significant differences because of the limited amount of interventions. The 12th term had initially a government from PP and later on, a change of government through a vote of no confidence that gave PSOE the power, since both political groups were in the government for a while and the term was an uncommon one, it is understandable that the distance between the parties was one of the lowest.

Ideological Polarisation Index

I computed a general index for ideological polarisation following the formula developed by Dalton in [17]. This index calculation is commonly accepted as a standard measure of ideological polarisation in a parliament, so computing it will allow for comparison with other research. Computing this index is especially useful to compare our results with the work conducted in [16], since it has been the only other elite ideological polarisation metric found for the case of Spain.

The index takes into account the voter share of the political group as well as their score in the Left-Right scale to obtain a general ideological polarisation index where 0 equals no ideological divergence and 10 represents a system where all parties would be in opposite extremes of the scale.

To calculate the ideological polarisation index, instead of voting share, we took into account the percentage of seats in the parliament. The index is presented in figure 5.10. For obtaining the L-R scaling, we normalised the left-right dimension of the reduced embedding model to the range [0 - 1]. Usually, that scale is obtained by asking citizens or experts through surveys.

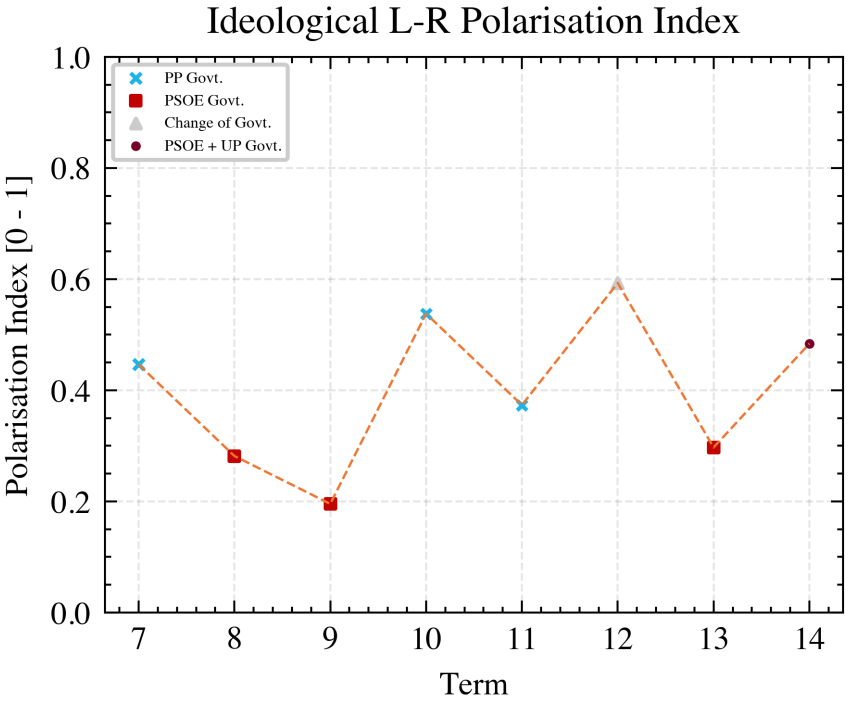


Figure 5.10: Ideological Polarisation Index among groups in the Spanish parliament for terms 7 to 14. Data for terms 14 is incomplete, as it is ongoing.

In our case, the index is partially incomplete because we are not measuring political parties, but political groups. This is due to the difficulties to discern between interventions of political parties that belong to common parliamentary groups because of not being big enough to form one on their own.

The figure shows a high fluctuation in polarisation, ranging from 0.2 to 0.6. The low levels in term 9 are because UP did not form a parliamentary group and therefore only PP and PSOE were used to compute the index. The low levels in terms 11 and 13 are explained by the fact that they only lasted a few months, no government was formed and therefore there was little room in the parliament to exhibit ideological differences. Finally, the data from the 14th term is incomplete, only accounting for $\approx 75\%$ of the data, since it has not finished yet.

The index shows, however, that levels of ideological polarisation are some of the highest for terms 14 and 12, but ideological polarisation seems to be somehow stable for the last 6 years between 0.35 and 0.6 points.

6 Affective Polarisation

This chapter develops the work conducted to assess the state of affective polarisation among the Spanish political elites by analysing their speech in the Congress over the last two decades. We infer the level of affective polarisation by obtaining the level of negativity in the speech by conducting two different sentiment analysis —or «sentiment classification»— techniques.

We assume that, if the negativity perceived in the interventions of the parliament increased over time, we can conclude that the affective polarisation also did and vice-versa. We apply this assumption throughout the entire chapter and use the negativity of parliamentary interventions as a proxy to measure affective polarisation.

The code for this chapter can be found in the GitHub repository of this thesis, in `src/notebooks/affective_polarisation.ipynb`.

6.1 Dataset

The data used in this chapter were the texts corresponding to each intervention translated with the deep translator package, as specified in Chapter 3. No further pre-processing was applied to it besides splitting each intervention into sentences for applying the transformer model sentiment analysis as explained in section 6.2.

Due to both models using all words to infer a positive or negative sentiment, no word was removed from the texts to get an accurate picture of the negativity levels in the parliament. Although words corresponding to decorum do influence the sentiment analysis algorithms since interventions following include sentences such as «Thank you, madam President» or «honourable members», I decided to not remove any of these from the corpus due to all interventions including them and them having a negligible impact in the overall sentiment of an intervention. Therefore, not interfering with the results.

6.2 Models

Two different techniques were used to perform sentiment classification of the parliament corpus: a lexicon-based approach and a transformer model.

Lexicon-based approach: This approach consisted of using a manually annotated dictionary called VADER that contains different words and their scaling between -4 —extremely negative sentiment— and +4 —extremely positive sentiment. Words like «awful» and «harm» would be considered negative, while words such as «gorgeous» and «improvement» would be ranked as positive. The total sum of scores of all the words in a sentence is then normalised to obtain a value between –1 and +1. VADER also takes into account punctuation, word capitalisation, degree modifiers —e.g. «a bit», «slightly»—, the existence of negation in a sentence and shifts in polarity in a sentence —e.g. «but», «however», «nonetheless»— to establish the sentiment of a text [49].

Although VADER tends to work best with shorter, informal texts such as tweets, its performance with longer texts is still acceptable, with an F_1-score of 0.55 for longer texts. Therefore, it was still used for this project [49].

We processed all the interventions using VADER and obtained a classification between «negative» and «positive» interventions. Later on, we counted the percentage of negatively-labelled interventions over the total. The processing of interventions was done in around 2.5h.

Transformer-based approach: The second technique used consisted of a transformer model, as explained in section 2.4. We tried to utilise a transformer model that accepted a high amount of tokens, such as GPT-4, but the model was not available. Finally, the model used was the so-called «XLM RoBERTa base» multilingual model, which was fine-tuned for sentiment analysis trained on 198 million tweets written between 2018 and 2020 and fine-tuned for the purpose of sentiment analysis [50]. The model achieves an F_1-score of 0.70 for classifying the sentiment of tweets as negative, neutral or positive as compared to the dataset annotated by humans. Using the transformer model to classify the sentiment of texts give us a label and the score corresponding to the percentage of the category observed in the sentence, e.g. «positive, 0.66».

The model only accepts a limited amount of tokens to be processed at the same time. That limit corresponds to approximately 500 words. To overcome that limitation, we label interventions sentence by sentence and later on calculate the overall sentiment for the entire intervention, so that the resulting metric is comparable to the results obtained using VADER. The metric is calculated by classifying an intervention as «negative» when it is composed in its majority of negative sentences.

It is expected for the results of VADER and the transformer to differ, due to their approaches being very different — the transformer does not only take words into account, but it also uses the context of each sentence, sarcasm or hidden criticism.

The processing of all the interventions using the transformer model required 93h of computing time, or ≈ 37 times more than with the lexicon-based approach.

6.3 Results

This section includes the results obtained using the two different algorithms for the whole time series, quarterly for the last six years and separated by political group.

In order to obtain more robust results, I present the results obtained with both VADER and the transformer model.

Entire-series Analysis

After processing the corpus of interventions available, I obtained a measure of negatively-labelled interventions for the entire time series, both with VADER and with the transformer. Fig. 6.1 shows the negativity in the Congress yearly for the last 22 years.

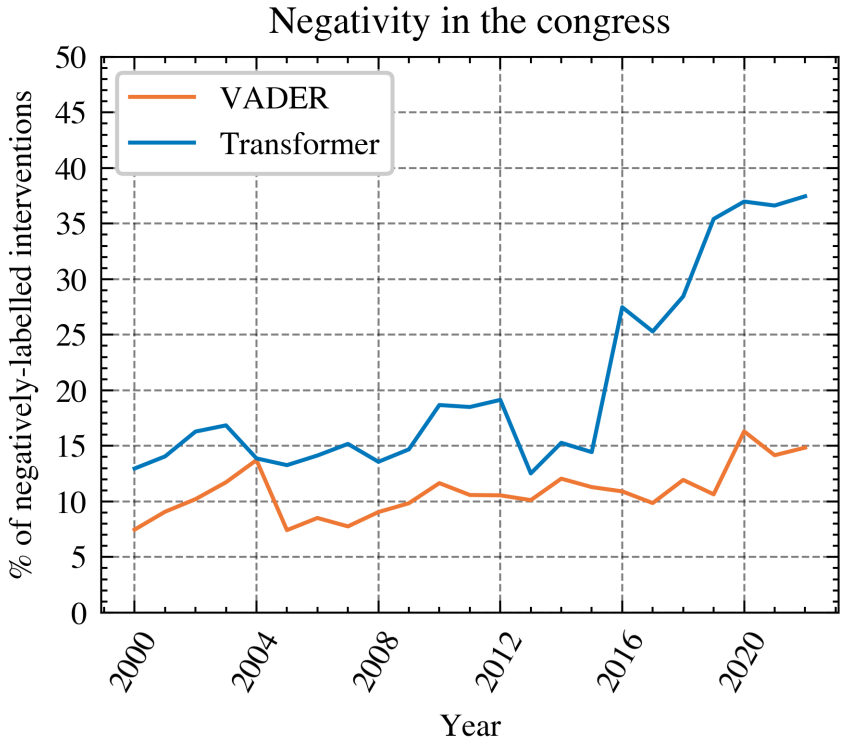


Figure 6.1: Global yearly affective polarisation for years 2000 to 2022.

The figure shows two time series, one for each method used, containing the percentage of negatively-labelled interventions for the years 2020 to 2022.

The VADER time series presents a steep increase in negativity between the years 2000 to 2004, followed by a decrease in 2005 and a moderate trend afterwards. 2020 was the year with the most negatively-labelled interventions in recent history, with 16% of them being considered as such. From 2020 onward there has been a moderation in negativity but values are still at an all-time high.

The transformer time series presents a 3x increase comparing the year 2000 with 2022: in the year 2000, the transformer model identified 13% of the interventions as negative, while for 2022 that number was 38%. This time series also shows a higher variation than VADER, with different fluctuations over the years.

The differences observed in the figure are caused by the different methodologies that the two approaches employ to classify the sentiment of a text. However, both series show an increasing trend —7% to 15% for VADER and 13% to 38% for the transformer— and that for the period 2020 - 2022 the affective polarisation, interpreted taking as a proxy for the percentage of negatively-labelled interventions in the Congress, was at its highest level.

Regarding differences in methodology, aside from the different procedures of each technique, VADER was inputted the entire intervention and resulted in either negatively- or positively-labelled interventions. Since that was not possible with the transformer due to input size limitation, we processed each sentence individually into a negative, neutral or positive label and labelled interventions as negative if the majority of sentences were also considered negative by the transformer model.

As shown in the previous figures, both sentiment-analysis processes show similarities for the

levels of affective polarisation for the last two decades, pointing to the conclusion that affective polarisation has been indeed increasing in the last years and we are currently at all-time highs.

Analysis per Political Group

To analyse how each political group is contributing to the affective polarisation —measured with the percentage of negative speech—, Figure 6.2 shows the percentage of negatively-labelled interventions for each political group according to VADER.

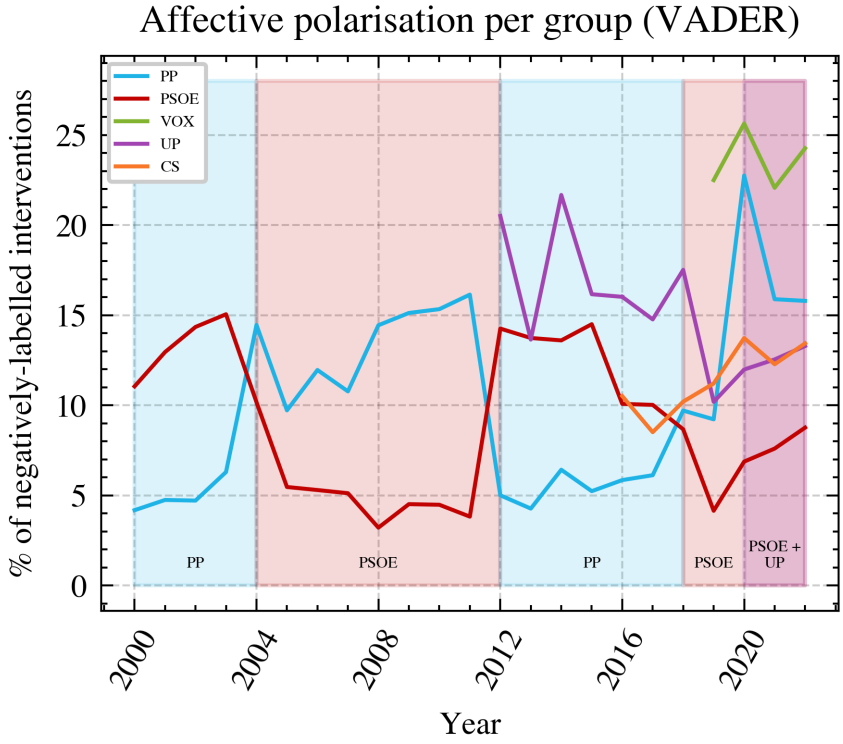


Figure 6.2: Affective polarisation per political group according to VADER. For periods 2000-2004, and 2012-2018 June, PP was in the government. For periods 2004-2012, and 2018 June-2020, PSOE was in the government. During 2020-2022 PSOE + UP governed in a coalition. UP is only covered since 2012 due to a lack of data and incomplete time series.

It is possible to observe a fluctuating pattern for all political groups —especially for PSOE and PP due to having complete time series— where depending on if the political group forms part of the government or the opposition, the percentage of negativity in their speech is very different.

For the case of PSOE, for example, the average percentage of negativity was approximately 5% for all the periods in time where they were part of the government, while for the moments they were in the opposition, the percentage of negative speech oscillated between 10% to 15%, which corresponds with a 2-times or 3-times increase. The same fluctuations apply for PP except for the years 2020 onward, where we can observe an increase in negativity, reaching levels above 15% with a maximum of 23% in 2020, at levels that are closer to the levels of negativity of VOX, which is the highest for the 5 political groups evaluated.

A very interesting situation shows in the previous figure: during 2008 there were two different governments, PP for the first half of the year and PSOE for the second. Their levels of negativity are similar, in the range of 10% since both held the opposition and government condition for the same amount of time.

UP had levels of negativity between 15% and 20% between 2012 and 2018. When the parliament formed a PSOE government, their levels of negativity decreased to 10%. UP became part of the government in 2020, however, their levels of negativity have been increasing since, as has happened for PSOE. These last trends represent a change compared to other periods in time, where the political group in power tends to have lowers levels of negativity. This probably has to do with the fact that PSOE and UP share a coalition government, and therefore have diverging opinions on policy issues and they cannot deploy the entirety of their agenda, as it would be more likely to happen in a single-group government.

The same plot is shown in figure 6.3 but with the classification of negatively-labelled interventions performed by the transformer.

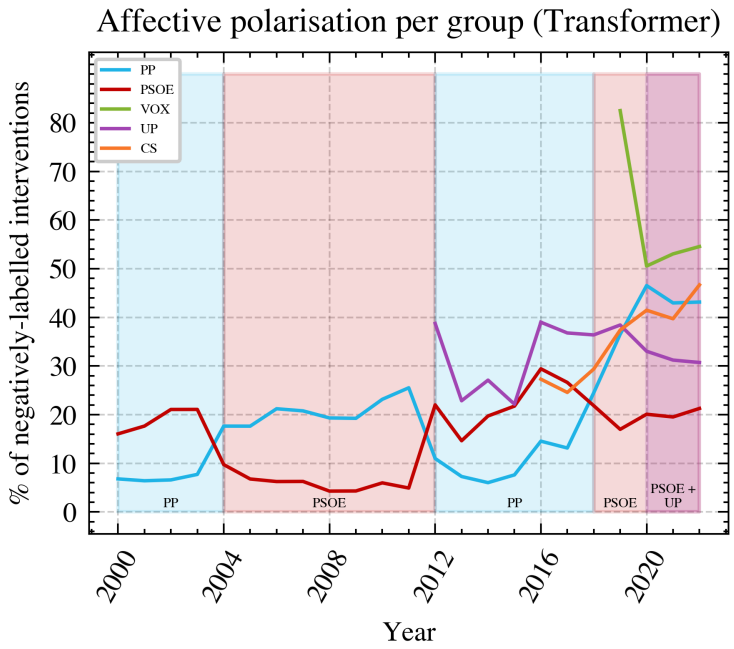


Figure 6.3: Affective polarisation per political group according to the transformer. For periods 2000-2004, and 2012-2018 June, PP was in the government. For periods 2004-2012, and 2018 June-2020, PSOE was in the government. During 2020-2022 PSOE + UP governed in a coalition. UP is only covered since 2012 due to a lack of data and incomplete time series.

Again, although exact numbers differ, the figure shows similar patterns as for the results with VADER. PP and PSOE have been exchanging levels of negativity between 2000 and 2017 depending on their presence or lack thereof in the Executive. Especially between the years 2000 and 2012, we observe very similar levels of polarisation for both parties: 5-7% when in the government and 15-25% when in the opposition. This pattern is interrupted by the appearance of new political parties CS and VOX after 2016 that present, along with PP, a high level of negativity for the period 2016-2022. Also shown by the VADER results, PSOE is the political group with the lowest percentage of negativity in their speeches, followed by UP with a 30% in 2022. PP, CS and VOX appeared to have 45-55% of their interventions labelled as negative by the transformer model.

Quarterly Analysis

The current section aims at correlating fluctuations in the calculated levels of polarisation with recent events in Spanish politics. A zoomed-in visualisation of the last 3 years is shown below.

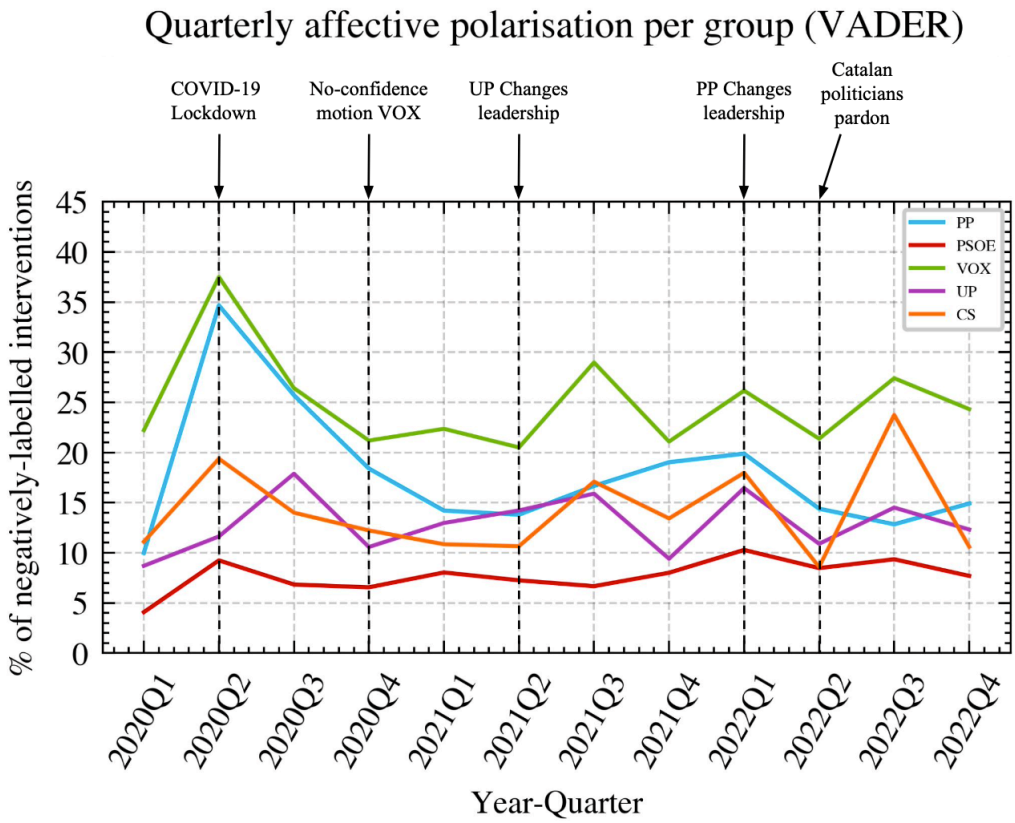


Figure 6.4: Quarterly affective polarisation per political group according to VADER for 2020 - 2022.

When comparing the levels of negativity along with the COVID-19 lockdown start, we can see that the level of polarisation was at its maximum during that time in the parliament, according to VADER. It is not possible to observe a correlation between other events and the levels of polarisation for each party besides the change in leadership of PP, which seems to correlate with a decrease in the levels of affective polarisation for that political group.

To observe how the levels of polarisation detected by the transformer correlate with the same events, fig. 6.5 shows the same plot but with the transformer results.

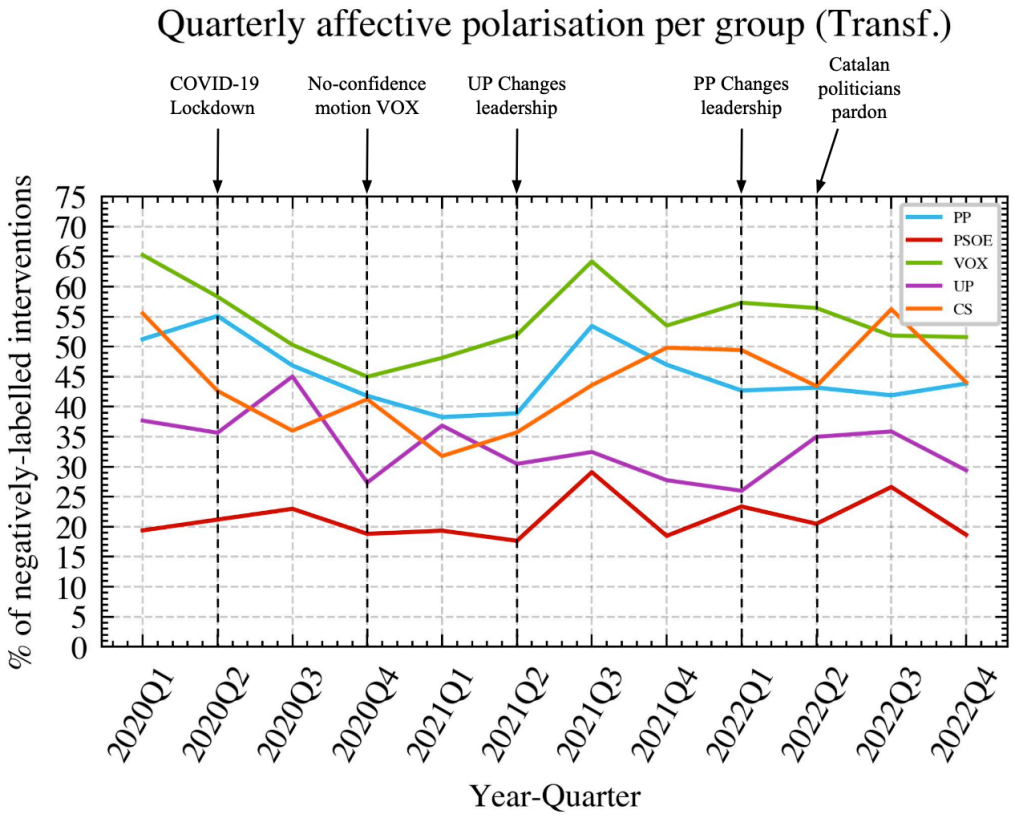


Figure 6.5: Quarterly affective polarisation per political group according to the transformer for 2020 - 2022.

Fig. 6.5 shows fluctuating levels of negativity, with levels peaking for PP, PSOE and VOX during Q3 2021. Affective polarisation levels for CS reach their maximum the quarter after imprisoned Catalan politicians are pardoned by the government in both the VADER and transformer results.

Other events were evaluated both by looking at these figures and examining the available data in the parliamentary corpus. It was not possible to correlate quarterly levels of polarisation with other specific political events such as regional elections, approval of laws or specific policy items. That is due to two main reasons: (1) the debates in the parliament take a long period, sometimes up to two years for a specific law to be approved, and (2) the parliament has rigid rules that difficult the introduction of immediate points to be discussed in the agenda besides for questions to the government, i.e. if a political event not connected to the parliament happens, it is difficult to see its impact in the tone of the MPs speech.

7 Conclusion and Future Work

Throughout this thesis, I identified, harvested and pre-processed a valuable dataset containing parliamentary records with rich metadata for the Spanish parliament. Additionally, that data was processed using several NLP techniques to analyse the discourse and the ideological and affective polarisation in the parliament for the time period 2000 – 2022, between terms 7 and 14.

The main state-wide political groups were ideologically placed throughout the seven terms with the use of document embeddings and their ideological polarisation was therefore analysed in detail, showing a continued increase across the time period analysed and identifying the existence of intrinsic differences both in the semantics and sentiment of speech for groups in the government or in the opposition.

Two approaches were carried out to analyse political polarisation: document embeddings for ideological polarisation and sentiment classification by using VADER and transformer models for affective polarisation. The results of both analyses conclude that the Spanish parliament is currently at their highest levels of both ideological and affective polarisation since the year 2000 and provide a set of tools to analyse ideological differences between political groups —such as a map of concepts that define the discourse of each of the main political groups in the parliament— and metrics for affective polarisation per political group.

Another contribution of this thesis consists of the analysis of the discourse throughout the seven terms in office for topics such as LGBTQ+ rights, feminism, environmental issues, labour rights, terrorism and the Catalan territorial conflict.

To the best of my knowledge, no analysis of the polarisation of Spanish political elites had been carried out before using parliamentary records. This thesis fills that gap by providing a detailed analysis both from an ideological and affective perspective.

Additionally, no dataset containing rich metadata on the Spanish parliament interventions had been published before. This thesis also partially mitigates this problem and lays the foundation to obtain a fully-complete corpus for the time period analysed.

7.1 Discussion

This section compares the conclusions of this thesis with the available metrics on polarisation to analyse possible differences between the methodologies and results. For this purpose, we will be using the literature mentioned in Chapter 2 that refers to the Spanish case.

Affective Polarisation

The comparison of the results from [18] with the obtained metric for affective polarisation in chapter 6 is presented in the following figure.

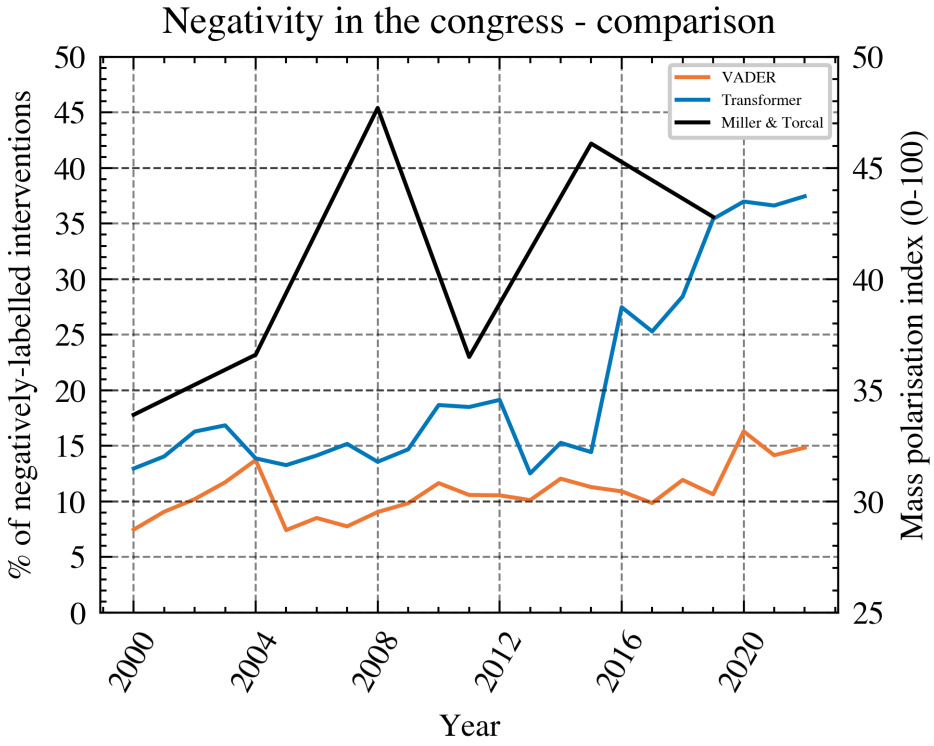


Figure 7.1: Comparison of the levels of mass affective polarisation calculated by [18] and in chapter 6.

The figure shows little to no correlation between the elite affective polarisation index calculated in chapter 6 and the mass polarisation index calculated by [18]. This might is probably due to differences in methodology population: the calculations performed in this thesis refer to political elites in the parliament and are based on data analysis and NLP techniques; the mass polarisation index by Miller & Torcal is based on the affective polarisation that citizens express while rating their supporting vs. their opposing political leaders, which corresponds to mass affective polarisation based on survey responses.

Assuming a scenario where both calculations would accurately represent the levels of polarisation of citizens and of MPs, we could conclude that although there is in both cases an increasing trend, increases in elite affective polarisation in the parliament do not necessarily translate into increases in mass affective polarisation.

Ideological Polarisation

To compare the level of ideological polarisation, I have gathered the data for general Left-Right placement from CHES for the main five political groups [11]. The results of the Left-Right dimension obtained in chapter 5 were normalised by scaling them to a 0-1 range using the following formula:

$$Z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{7.1}$$

The following graph shows the comparison between the two methods.

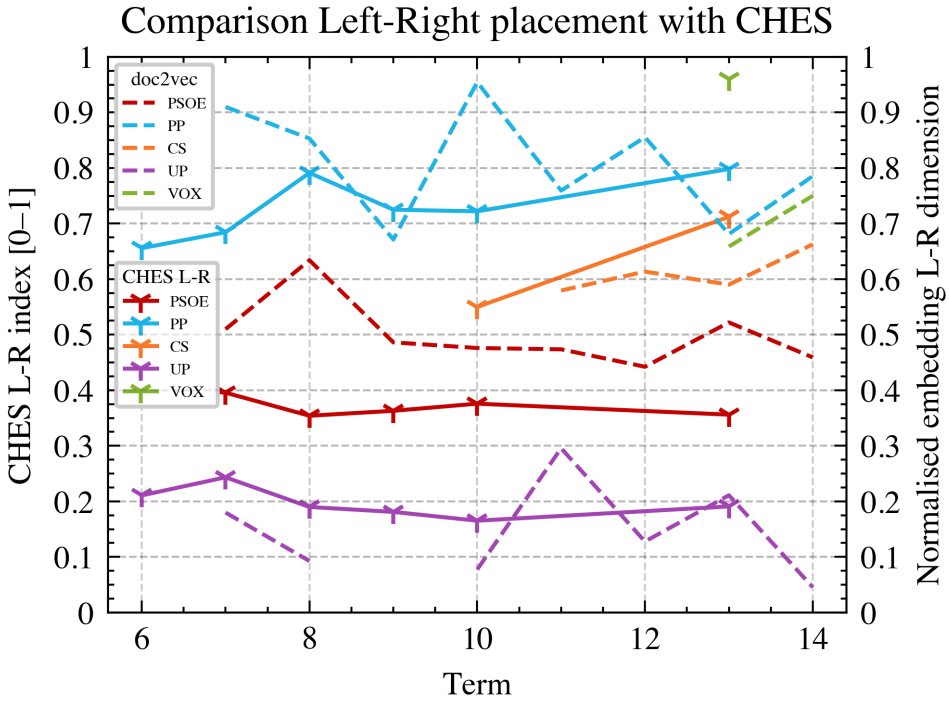


Figure 7.2: Comparison of the levels of ideological placement performed by CHES [11] and those calculated in chapter 5.

The results show a high level of similarity. In both scales, the placement of political groups from left to right is aligned except for the case of VOX in the embedding model. The embedding model does not accurately place VOX in the scale, taking the CHES results as a gold standard. The embedding model shows a higher fluctuation across terms and places PSOE in the range of around 0.5 while the consulted experts by CHES place it in the range of 0.4. It is interesting to see that the document embedding model manages, by analysing the interventions in the parliament, to ideologically place the political groups in a similar way to the panel of experts consulted by CHES.

Reasons for the misplacement of VOX are related to their discourse. This political group sometimes tries to appeal to the «working class» and talk about social issues. That might be the reason why they are placed closer to groups in the left-wing spectrum than PP.

Finally, as the last comparison, we will be observing the relationship between the work conducted by Simón in [16] and the ideological polarisation index calculated in chapter 5. The index calculated by the author of the article uses data obtained from voter surveys on the ideological placement of the different political parties on a Left-Right scale, being 0 extreme left and 10 extreme right. Simón uses the formula developed in [17] to compute the general ideological polarisation index, in the same it is done in this thesis in section 5.6.

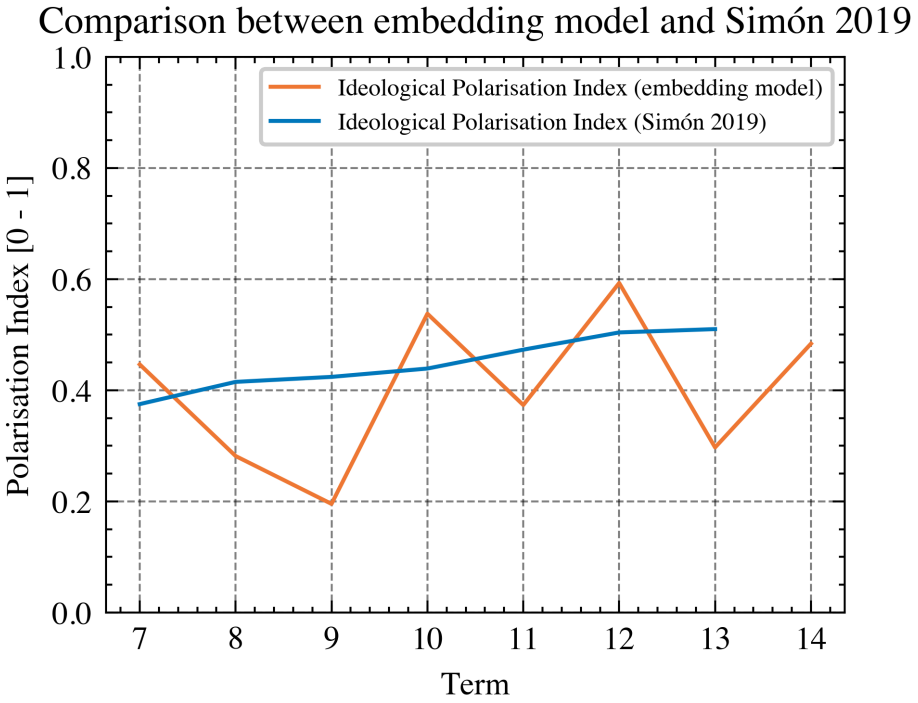


Figure 7.3: Comparison of the ideological polarisation index calculated by Simón [16] and the one calculated in chapter 5.

The figure shows different patterns, with low fluctuations for the index developed by Simón and high ones for the embedding model. These fluctuations are explained in section 5.6. Remarkably, both indexes are in similar ranges, especially for terms 10 onward. They both show a similar range and evolution.

Additionally, it is important to understand that polarisation indexes use radically different data. The index developed in [16] was calculated with electoral surveys to citizens, and the embedding model was computed using exclusively the interventions of MPs in the parliament.

7.2 Future Work

Within the framework of this thesis, a new dataset containing a partial parliamentary corpus for Spain with rich metadata has been harvested. This creates new possibilities for analysing what MPs talk about and how they do it.

Perhaps one of the most impactful steps that could be taken to further this work consists of modifying the logic in the algorithms that harvest the text data from the parliamentary records to obtain all available texts in a machine-readable format. This was not possible in the given time frame for this thesis, however, I will be performing this modification in the future so that a fully-complete parliamentary corpus is available for the 2000–2023 time period. All the analyses and models performed for this thesis can benefit from being trained with all the available data.

I have discussed and compared the results of both the ideological and affective levels of polarisation with other available literature, nevertheless, it would be beneficial to deepen this analysis and try to find other relations between elite and mass polarisation.

This thesis leaves the door open to explore the polarisation around specific policy issues, such as tax reforms and new social rights, by providing data on affective polarisation for each intervention along with information related to the topic that is being discussed in each of those interventions.

By doing topic modelling and analysis, it would be possible to further extract value from the data, such as by analysing the main topics discussed in the parliament, the position of each political actor concerning it, etc.

About the ideological polarisation chapter, the trained embeddings model opens a new window of possibilities when used along with new data. The model can be used to infer the position of any text in the embedding space, which would allow ideologically placing any sort of institution, think tank, or newspaper, in the embedding space and try to infer the proximity in discourse to the political groups in parliament. We will be trying out this approach shortly.

Additionally, it would be possible —provided that the data is accessible— to re-train an embedding model that contains more than just parliamentary records, such as social media data of the political groups, party manifestos, news articles, etc. to try to create a wider ideological map. The model could also be re-trained to show a yearly evolution, which hopefully would help with the outliers that terms 11 and 13 represent.

Finally, the work performed regarding affective polarisation could be automated and regularly updated to create a continuously-updated affective polarisation index for the Spanish parliament.

Bibliography

[1] Morris P. Fiorina and Samuel J. Abrams. "Political Polarization in the American Public". In: *Annual Review of Political Science* 11.1 (2008), pp. 563–588. DOI: 10.1146/annurev.polisci.11.053106.153836. URL: https://doi.org/10.1146/annurev.polisci.11.053106.153836.

[2] Vanessa A. Boese, Nazifa Alizada, Martin Lundstedt, Kelly Morrison, Natalia Natsika, Yuko Sato, Hugo Tai, and Staffan I. Lindberg. *Autocratization Changing Nature? Democracy Report 2022*. Tech. rep. Varieties of Democracy Institute (V-Dem), 2022.

[3] Pew Research Center. *Political Polarization in the American Public*. Tech. rep. Washington, D.C., 2014. URL: https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/.

[4] Keith T Poole and Howard Rosenthal. "A spatial model for legislative roll call analysis". In: *American journal of political science* (1985), pp. 357–384.

[5] Antonio Garrido, M Antonia Martínez Rodríguez, and Alberto Mora Rodríguez. "Polarización afectiva en España". In: *Más Poder Local* 45 (2021), pp. 21–40.

[6] Gavin Abercrombie and Riza Batista-Navarro. "Sentiment and position-taking analysis of parliamentary debates: a systematic literature review". In: *Journal of Computational Social Science* 3.1 (2020), pp. 245–270. URL: https://doi.org/10.1007/s42001-019-00060-w.

[7] Jennifer McCoy, Tahmina Rahman, and Murat Somer. "Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities". In: *American Behavioral Scientist* 62.1 (2018), pp. 16–42. DOI: 10.1177/0002764218759576. URL: https://doi.org/10.1177/0002764218759576.

[8] Jennifer McCoy and Murat Somer. "Toward a Theory of Pernicious Polarization and How It Harms Democracies: Comparative Evidence and Possible Remedies". In: *The ANNALS of the American Academy of Political and Social Science* 681.1 (2019), pp. 234–271. DOI: 10.1177/0002716218818782. URL: https://doi.org/10.1177/0002716218818782.

[9] Diego López Garrido. "El valor del Senado: Funciones patentes y latentes". In: *Revista de las Cortes Generales* (2018), pp. 89–112.

[10] Antonia Martínez. *El Congreso de los Diputados en España: funciones y rendimiento*. Tecnos, 2000.

[11] Seth Jolly, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. "Chapel Hill expert survey trend file, 1999–2019". In: *Electoral studies* 75 (2022), p. 102420.

[12] Cecilia Martínez-Gallardo, Nicolás de la Cerda, Jonathan Hartlyn, Liesbet Hooghe, Gary Marks, and Ryan Bakker. "Revisiting party system structuration in Latin America and Europe: Economic and socio-cultural dimensions". In: *Party Politics* (2022), p. 13540688221090604.

[13] Lluís Orriols and Sandra León. "Looking for affective polarisation in Spain: PSOE and Podemos from conflict to coalition". In: *South European Society and Politics* 25.3-4 (2020), pp. 351–379.

[14] Luis Miller. "Polarización en España: más divididos por ideología e identidad que por políticas públicas". In: *EsadeEcPol Insight* 18 (2020), pp. 1–14.

[15] Mariano Torcal, Andrés Santana, Emily Carty, and Josep Maria Comellas. "Political and affective polarisation in a democracy in crisis: The E-Dem panel survey dataset (Spain, 2018–2019)". In: *Data in brief* 32 (2020), p. 106059.

[16] Pablo Simón. "The multiple Spanish elections of April and May 2019: The impact of territorial and left-right polarisation". In: *South European Society and Politics* 25.3-4 (2020), pp. 441–474.

[17] Russell J Dalton. "The quantity and the quality of party systems: Party system polarization, its measurement, and its consequences". In: *Comparative political studies* 41.7 (2008), pp. 899–920.
[18] Luis Miller and Mariano Torcal. "Veinticinco años de polarización afectiva en España". In: *The Conversation* 31 (2020). URL: https://theconversation.com/veinticinco-anos-de-polarizacion-afectiva-en-espana-149237.
[19] Marilena Hohmann, Karel Devriendt, and Michele Coscia. "Quantifying ideological polarization on a network using generalized Euclidean distance". In: *Science Advances* 9.9 (2023), eabq2044. DOI: 10.1126/sciadv.abq2044. URL: https://www.science.org/doi/abs/10.1126/sciadv.abq2044.
[20] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. "Political polarization on twitter". In: *Proceedings of the international aaai conference on web and social media*. Vol. 5. 1. 2011, pp. 89–96.
[21] Frederic Guerrero-Solé and Olivier Philippe. "La toxicidad de la política española en Twitter durante la pandemia de la COVID-19". In: *Hipertext. net* 21 (2020), pp. 133–139.
[22] Miguel G Folgado and Veronica Sanz. "Exploring the political pulse of a country using data science tools". In: *Journal of Computational Social Science* 5.1 (2022), pp. 987–1000.
[23] Thomas König, Moritz Marbach, and Moritz Osnabrügge. "Estimating party positions across countries and time—a dynamic latent variable model for manifesto data". In: *Political analysis* 21.4 (2013), pp. 468–491.
[24] Ludovic Rheault and Christopher Cochrane. "Word embeddings for the analysis of ideological placement in parliamentary corpora". In: *Political Analysis* 28.1 (2020), pp. 112–133.
[25] Russell J Dalton and Carl C Berning. "Ideological polarization and far-right parties in Europe". In: *Rechtspopulismus in Deutschland: Wahlverhalten in Zeiten politischer Polarisierung*. Springer, 2022, pp. 13–35.
[26] Zellig S Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.
[27] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* 28.1 (1972), pp. 11–21.
[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).
[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.
[30] Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents". In: *International conference on machine learning*. PMLR. 2014, pp. 1188–1196.
[31] Radim Rehurek and Petr Sojka. "Gensim–python framework for vector space modelling". In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011).
[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All You Need". In: 2017. URL: https://arxiv.org/pdf/1706.03762.pdf.
[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 2020, pp. 38–45.
[34] Pedro L Rodriguez and Arthur Spirling. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research". In: *The Journal of Politics* 84.1 (2022), pp. 101–115.

[35] Mykola Andrushchenko, Kirsi Sandberg, Risto Turunen, Jani Marjanen, Mari Hatavara, Jussi Kurunmäki, Timo Nummenmaa, Matti Hyvärinen, Kari Teräs, Jaakko Peltonen, et al. "Using parsed and annotated corpora to analyze parliamentarians' talk in Finland". In: *Journal of the Association for Information Science and Technology* 73.2 (2022), pp. 288–302.

[36] Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, et al. "The ParlaMint corpora of parliamentary proceedings". In: *Language resources and evaluation* (2022), pp. 1–34.

[37] Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, et al. "The ParlaMint corpora of parliamentary proceedings". In: *Language resources and evaluation* (2022), pp. 1–34.

[38] Christian Rauh and Jan Schwalbach. *The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies*. Version V1. 2020. DOI: 10.7910/DVN/L4OAKN. URL: https://doi.org/10.7910/DVN/L4OAKN.

[39] Pedro L Rodriguez and Arthur Spirling. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research". In: *The Journal of Politics* 84.1 (2022), pp. 101–115.

[40] *Web Site of the Congreso de los Diputados*. https://congreso.es. Accessed: 2023-02-05.

[41] *Selenium Web Driver*. https://www.selenium.dev/. Accessed: 2023-02-02.

[42] Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

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

[44] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[45] Inés Campillo. "'If we stop, the world stops': the 2018 feminist strike in Spain". In: *Social Movement Studies* 18.2 (2019), pp. 252–258.

[46] P Janke. "Spanish Separatism: ETA's (Euskadi ta Askatasuna) Threat to Basque Democracy (From Contemporary Terrorism, P 135-166, 1986, William Gutteridge, ed.-See NCJ-107558)". In: (1986).

[47] Ángeles Escrivá. "ETA anuncia el cese definitivo de su 'actividad armada'". In: *El Mundo [Madrid]* 20 (2011).

[48] Jey Han Lau and Timothy Baldwin. *An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation*. 2016. arXiv: 1607.05368 [cs.CL].

[49] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.

[50] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. "Xlm-t: A multilingual language model toolkit for twitter". In: *arXiv e-prints* (2021), arXiv–2104.

A Breakdown of Discourse Analysis Topics

The terms used to define each topic in Chapter 4 were obtained by using the document embedding model to find the most similar terms to words connected to the topic. The lists of words that compose each topic are presented below. Please note that the terms were used following the pre-processing standards of the data, therefore instead of terms such as «interrupción del embarazo», the term used was «interrupción_embarazo». Terms are written in a natural-language format to ease the comprehension by the reader.

Table A.1: Terms that compose each of the topics in Chapter 4.

Topics	**Terms used**
Feminism	movimiento feminista, lucha feminista, violencias sexuales, violencias machistas, feministas, feminismo, feminista, movimientos feministas, abolicionistas, derechos sexuales, derechos feministas, violencia machista, machismo, reproductivos mujeres, patriarcal, patriarcado, violencia de género, 'pacto de Estado contra la violencia de género, libertad sexual, abuso violación, derechos sexuales y reproductivos, aborto, interrupción voluntaria del embarazo, salud sexual y reproductiva e interrupción, abortar, interrupción del embarazo, derecho al aborto, interrumpir el embarazo, aborto legal seguro, catorce semanas, misógina, machista, igualdad de género, mujeres feministas, misoginia, perspectiva feminista, pobreza menstrual, perspectiva de género, sexismo, brecha de género, interrumpir voluntariamente el embarazo, puertas de las clínicas, embarazo deseado, abortistas, provida, abortar clínicas, practicar abortos, aborten, violadas, ACAI, práctica abortos, IVE.
LGBTQ+ Rights	trans, lgtbi, derechos de las personas lgtbi, lgtb, personas lgtbi, personas trans, colectivo lgtbi, intersexuales, gay, comunidad lgtbi, homofobia, transexuales, derechos del colectivo lgtbi, lesbianas gays transexuales y bisexuales, igualdad efectiva de las personas trans, lucha contra la homofobia, felgtb, orgullo lgtbi, derechos lgtbi, motivos de orientación sexual, homófobos, discriminación por la orientación sexual, diversidad sexual, orientación sexual e identidad género, gays y lesbianas, transgénero, derechos homosexuales, igualdad de trato para la no discriminación, mujeres trans, mujeres transexuales, colectivos lgtbi, mujeres lesbianas, identidad de género, intersexual, personas homosexuales, homófoba, homófobas, homosexualidad, orgullo gay, matrimonio gay, bisexual, matrimonio igualitario, matrimonio homosexual, mismo sexo, parejas homosexuales, identidad sexual, independencia de la orientación sexual, lesbianas, colectivo de gays lesbianas.

Topics	Terms used
Labour Rights	precariedad, empleos precarios, empleo precario, bajos salarios, precario, trabajadores pobres, salario mínimo, salario mínimo interprofesional, contrarreforma laboral, precariedad laboral, precariedad y temporalidad, precariedad e inestabilidad, precarización, acabar con la precariedad, jornada completa, horas extra, modelo de relaciones laborales, despidos, despido, salarios de miseria, mercado de trabajo, derechos laborales, subempleo, contratos temporales, sueldo, mileuristas, derogación de la reforma laboral, precariedad y pobreza, bajada salarios, despido, interinos e interinas, trabajo precario, desempleo y precariedad, subir el salario mínimo interprofesional, devaluación salarial, salarios bajos, salarios dignos, trabajos temporales, contrato temporal, salarios, alta tasa de temporalidad, subida del salario mínimo, reforma laboral, facilita el despido, contratación laboral, paro, desempleo, tasa de paro, empleo, destrucción de empleo, millones de parados, hogares con miembros en paro, subida del salario mínimo interprofesional, empleo destruido, tasa de paro juvenil, paro juvenil, desempleo juvenil, crear empleo, millón de parados, tasa de desempleo, abaratar el despido, despidos colectivos, empleo precario, paro de larga duración, creando empleo, condiciones laborales, derechos de los trabajadores, libertad sindical, congelar el salario mínimo interprofesional, salarios, trabajadoras del hogar, trabajadoras y trabajadores, relaciones laborales, estatuto de los trabajadores, negociación colectiva, reformas laborales, trabajadores precarios.
Environmental Issues	planeta, recursos naturales, cambio climático, calentamiento global, consecuencias del cambio climático, océanos, calentamiento del planeta, combustibles fósiles, justicia climática, nivel del mar, climática, lucha contra el cambio climático, preservando el medio ambiente, efectos del cambio climático, crisis climática, deforestación, amenaza del cambio climático, aumento de temperatura, lucha contra la desertificación, pérdida de biodiversidad, contaminación del aire, deterioro del medio ambiente, emergencia climática, reducción de gases de efecto invernadero, frente al cambio climático, cambio climático y transición energética, emisiones de gases de efecto invernadero, adaptación al cambio climático, gases de efecto invernadero, kioto, IPCC, combustibles fósiles, efecto invernadero, luchar contra el cambio climático, reducir emisiones, desertificación, medio ambiente, climático, dióxido de carbono, disminución de emisiones, acuerdo de parís, impactos del cambio climático, protocolo de kioto, biodiversidad, ecosistemas, protección de la biodiversidad, conservación de la biodiversidad, espacios naturales, bosques, parque nacional, red natura, doñana, conservación de especies, especies amenazadas, conservación de la naturaleza, peligro extinción, medioambiente, hábitats, diversidad biológica, medioambiental, organizaciones ecologistas, reforestación, ecologismo, sequía, inundaciones, recursos hídricos, lluvias torrenciales, contaminación atmosférica, mala calidad del aire, contaminación por plásticos, ozono, partículas en suspensión, capa de ozono, dióxido de nitógeno, contaminación de las ciudades.
Catalan Territorial Conflict	independentista, estatuto de autonomía de Cataluña, estatuto de Cataluña, separatista, independentistas, referéndum de autodeterminación, declaración unilateral, referéndum de independencia, autodeterminación, derecho de autodeterminación, secesionistas, secesión, estatut, estatut de Catalunya, sentencia estatut, independentismo, socios separatistas, líderes separatistas, separatistas catalanes, procés, separatismo catalán, referendum independentista, nacionalismo catalán.

Topics	Terms used
Terrorism	terrorism, lucha contra el terrorismo, combatir el terrorismo, terrorista, violencia terrorista, lucha antiterrorista, terrorismo de eta, banda terrorista eta, eta, lucha eta, atentados, banda terrorista, eta eta, delitos de terrorismo, atentados terroristas, víctimas del terrorismo, organización terrorista, presos de eta, eta y batasuna, acercamiento de presos, proetarras, eta mataba, batasuna eta, asesinos etarras, apología terrorismo, kale borroka, bilduetarras, amigos eta, filoetarras, 11m.

Technical
University of
Denmark

Akademivej, Bygning 358
2800 Kgs. Lyngby
Tlf. 4525 1700

www.man.dtu.dk