Topic Analysis in the Basqueparl Dataset

Jon Irastorza Ancín

1 Introduction

Political speeches are a valuable source of data that reflect the concerns, approaches, and arguments of political leaders and their parties. However, manually analyzing these speeches can be a tedious task that consumes a considerable amount of time and resources. Moreover, large-scale analysis of political speeches can reveal changes in the topics addressed over time or identify emerging or declining issues.

This research project focuses on analyzing the speeches from the Basque parliament from various perspectives to examine the topics discussed. For this purpose, we have the Basqueparl dataset, which will be described in section 2.1.

2 Methodology

This section describes the technical aspects followed to carry out this research project, including important details about experimentation and analysis.

2.1 Data Description

The analysis was conducted on BasqueParl, a bilingual corpus for political discourse analysis. It encompasses transcriptions from the Basque Autonomous Community Parliament over eight years. The considered transcriptions include parliamentary sessions from December 3, 2012, to February 7, 2020, covering two legislatures, namely 2012-2016 and 2016-2020.

2.1.1 Data Structure

The BasqueParl corpus is formatted as a tab-separated values (TSV) file. [Table 1] illustrates the information included for each paragraph in BasqueParl.

2.1.2 Data Preprocessing

As shown in [Table 1], the lemmas_stw column contains speech fragments that are lemmatized and without stopwords. On one hand, lemmatization reduces words in the speech to their root form, which reduces the dimensionality of the feature space as it helps group words with similar meanings. On the other hand, removing stopwords involves eliminating common words that do not add meaning to the text analysis, i.e., these words are usually not related to topics and introduce noise into the analysis. Combining these two techniques helps reduce noise and focus on the most important words, improving model performance.

Additionally, a search for collocations was performed to capture more complex lexical relationships between words. This helps detect specific meanings that would not be obtained by analyzing individual words, providing thematic consistency to the identified topics. In total, 2000 collocations were replaced; [Figure 1] shows some of the collocations found and replaced.

	Colocación resultante
Eh Bildu	Eh_Bildu
Comunidad Autónoma	Comunidad_Autónoma
Unión Europea	Unión_Europea
Elkarrekin Podemos	Elkarrekin_Podemos

Figure 1: Example of found collocations

Finally, the speeches were grouped by speech_id and date. Combining the speech segments preserves the coherence and structure of the entire speech. Additionally, it reduces noise and ambiguity that may arise when analyzing individual frag-

Field	Description
date	Date of the speech.
speech_id	Identifier number for the speech within its date.
text_id	Identifier number for the paragraph within its speech.
speaker	Surnames of the speaker, including their title if they have one.
birth	Year of birth of the speaker.
gender	Gender of the speaker.
party	Political party of the speaker.
language	Language assigned to a paragraph.
text	Paragraph of the speech text.
lemmas	Lemmatized paragraph.
lemmas_stw	Lemmatized paragraph without stopwords.
entities	Named entities extracted from the paragraph.
entities_stw	Named entities extracted from the paragraph without keywords.

Table 1: Fields of the Basqueparl dataset

ments. Having complete speeches increases the presence of keywords, improving the quality of topic analysis. On the other hand, mixing speeches in different languages may introduce noise and dilute the signal of specific topics we aim to capture, so we focus on Spanish speeches for the analysis. Finally, we have 18016 speeches in Spanish and discard 23401 speeches in Basque.

2.1.3 Data Distribution Analysis

This section focuses on observing the data distribution after preprocessing.

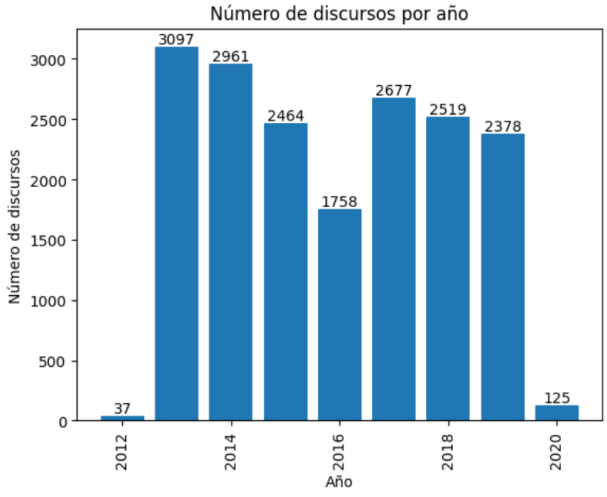

Figure 2: Distribution of speeches per year

As seen in [Figure 2], the years 2012, 2016, and 2020 have the fewest speeches, as they are years with legislative changes. Additionally, the first and last years are the beginning and end of the corpus.

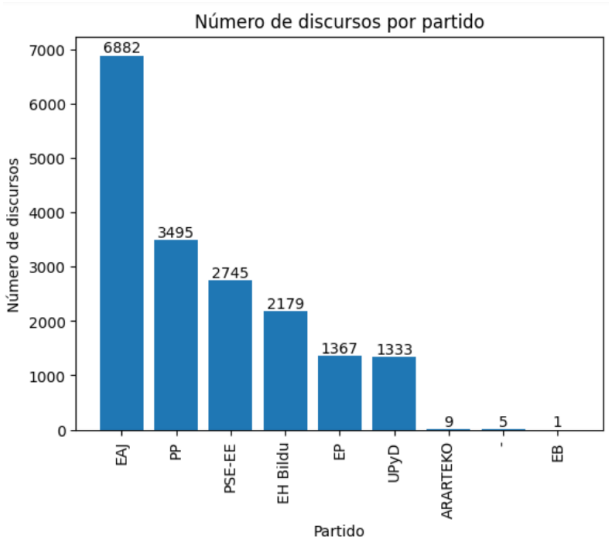

Figure 3: Distribution of speeches per party

As seen in [Figure 3], over these years, several political parties participated in the debates: Euzko Alderdi Jeltzalea-Partido Nacionalista Vasco (EAJ-PNV), Euskal Herria Bildu (EH Bildu), Partido Socialista de Euskadi-Euskadiko Ezkerra (PSE-EE), Partido Popular (PP), Elkarrekin Podemos (EP), and Unión, Progreso y Democracia (UPyD). Other parties and institutions also participated, such as the Basque Ombudsman (Ararteko) or Ezker Batua (EB) through a regional parliamentary spokesperson. EAJ-PNV is the main ruling party and the president's party throughout the years considered, and therefore the author of the most speeches. UPyD had representation until 2016, and that year EP entered the parliament. The rest of the groups main-

tain their presence in parliament throughout the time considered.

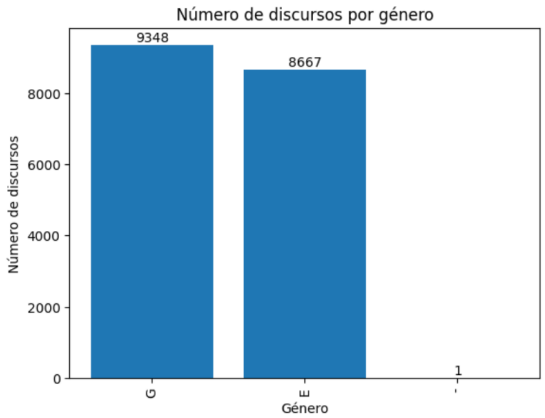

Figure 4: Distribution of speeches by gender

In [Figure 4], it can be seen that more speeches were given by men (G) than by women (E); however, this difference is not significant.

2.2 Models Used

For the analysis of the preprocessed speeches, the LDA (Latent Dirichlet Allocation) algorithm was used. This is a statistical model and unsupervised machine learning algorithm used to discover hidden or latent topics in a set of documents. It was first proposed by David Blei, Andrew Ng, and Michael Jordan in 2003. LDA is based on the assumption that each document is composed of multiple topics and that each topic is composed of words with a certain probability distribution. The goal of the LDA algorithm is to find these latent topics and the word distribution in each topic.

To perform a deeper analysis and compare results, three different models were trained. The parameter configurations of the LDA models can significantly affect the results. Therefore, it is important to conduct a comparative analysis of different configurations to determine their impact on identifying latent topics.

Model 1 aims to find 30 topics in the set of documents and focuses on words that appear in less than 60Model 2 attempts to have greater granularity in identifying topics, so the number of topics was increased to 40. Similarly to the previous model, it focuses on words that appear in less than 60Finally, Model 3 also has 40 topics, like the second model, but the minimum number of times a word must appear was adjusted back to 300.

2.3 Evaluation Methodology

The Basqueparl dataset does not have speeches labeled according to the topic(s) they address, so there is no metric to evaluate the quality of the learned models. However, reviewing keywords and comparing the obtained results with reality can help assess their quality. If the topics are interpretable and make intuitive sense, it is a positive indicator of the model's performance.

3 Presentation of Results and Analysis

This section presents the results obtained with the different models. The results will be defined by their respective graphs and tables, which will be analyzed to verify that the obtained results are satisfactory. The number of total features obtained with each of the configurations described above will also be considered.

For each model, each of the learned topics is characterized by a series of words, which have a weight assigned according to their importance within the topic. To clarify the analysis, some of the topics have been selected based on an analysis of their component words, to finally assign a name that characterizes each topic.

For each model, all documents were classified with the topic of highest probability.

3.1 Model 1

The following shows the results obtained with Model 1.

Using the parameters explained in section 2.2, a total of 1497 features, characteristics, or words were obtained, which were used to construct the 30 topics.

As seen in [Figure 5], the topics are completely separated, implying that the learned topics are distinct and have little or no similarity between them.

Ten topics were selected from a total of 30, and a name was assigned to each.

In [Figure 7], the evolution of topics addressed in parliament over the years is shown. It is normal

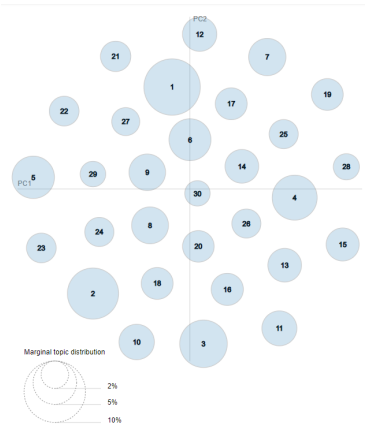

Figure 5: Topic distribution of Model 1

Figure 6: Selected topics Model 1

for the years 2012 and 2020 to have fewer topics, as they are the boundary years of the dataset and have fewer speeches. However, it is interesting that in 2020 the topic of health became important, as it was a year marked by the COVID-19 pandemic, which was already being heard in the media in January and February.

In [Figure 8], the thematic areas of different parties can be identified and compared. In this case, Ezker Batua (EB) does not address any of the selected topics, while parties like EH Bildu and UPyD talk a lot about terrorism. On the other hand, EP and Ararteko discuss gender equality quite a bit. It is normal for Ararteko to have few topics addressed as they have only nine interventions, and EB only has one.

In [Figure 9], the topics discussed by women (E) and men (G) can be seen. Apparently, women address gender equality much more frequently than men, while men focus more on business and industry.

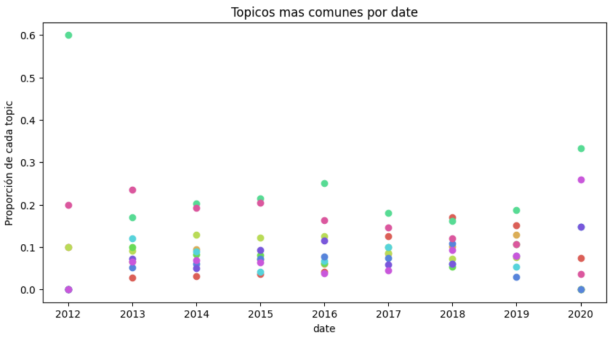

Figure 7: Topics by year Model 1

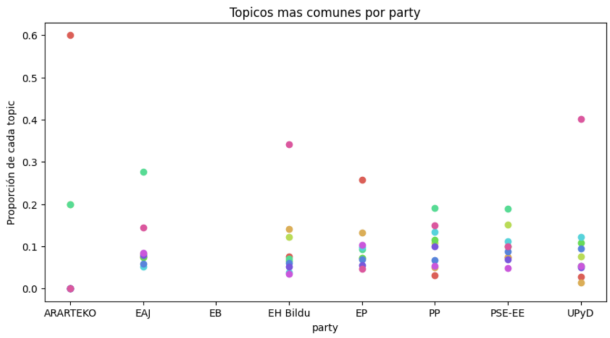

Figure 8: Topics by political party Model 1

3.2 Model 2

The following shows the results obtained with Model 2.

Using the parameters explained in section 2.2, a total of 2610 features, characteristics, or words were obtained, which were used to construct the 40 topics.

As seen in [Figure 10], some of the learned topics are very close and overlap. This indicates that they have similarities and share terms, i.e., there are common thematic areas among them.

Sixteen topics were selected from a total of 40, and a name was assigned to each.

In [Figure 12], the evolution of topics addressed in parliament over the years is shown. Compared to the results obtained with Model 1, 2020 has two more topics, as the number of selected topics increased from 10 to 16.

In [Figure 13], the thematic areas of different parties can be identified and compared. In this case, Ararteko does not address any of the selected topics. The figure does not include the Ezker Batua (EB) party to spread the points more, but in

Figure 9: Topics by gender Model 1

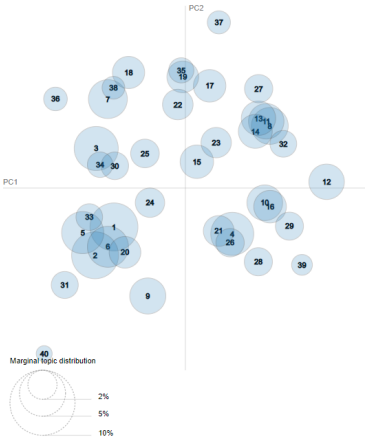

Figure 10: Topic distribution of Model 2

Figure 11: Selected topics Model 2

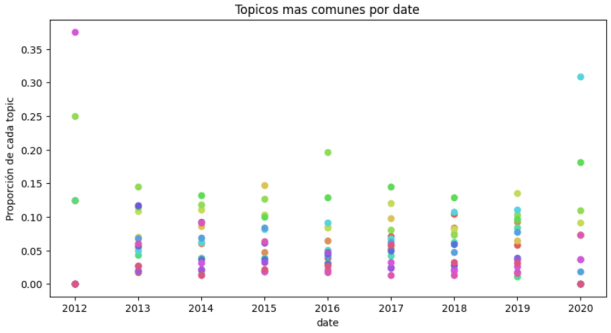

Figure 12: Topics by year Model 2

this case, unlike Model 1, their only speech was classified as legislation.

In [Figure 14], the topics discussed by women (E) and men (G) can be seen. In this case, gender equality is still more frequently addressed by women; however, the difference is not as large. On the other hand, women seem to discuss budgets and education more frequently than men.

3.3 Model 3

The following shows the results obtained with Model 3.

Using the parameters explained in section 2.2, as with Model 1, a total of 1497 features, characteristics, or words were obtained. However, in this case, the 40 topics were constructed instead of 30.

As seen in [Figure 15], like with Model 2, some of the learned topics are very close and overlap.

Fifteen topics were selected from a total of 40, and a name was assigned to each.

In [Figure 17], the evolution of topics addressed in parliament over the years is shown. In this case, 2020 also has more topics than in the case of Model 1. However, it appears that in 2020, the topic of corruption became even more important than health.

In [Figure 18], the thematic areas of different parties can be identified and compared. In this case, Ezker Batua (EB) does not address any of the selected topics. The figure does not include the Ararteko party, which only addresses one topic, gender equality.

In [Figure 19], the topics discussed by women (E) and men (G) can be seen. In this case, women again discuss gender equality more than men. On the other hand, men talk more about corruption than women.

4 Conclusions

The analysis of political speeches using different LDA models revealed different but complementary results. These results highlight the complexity of text analysis and the importance of considering multiple approaches in academic research. By combining different models and carefully

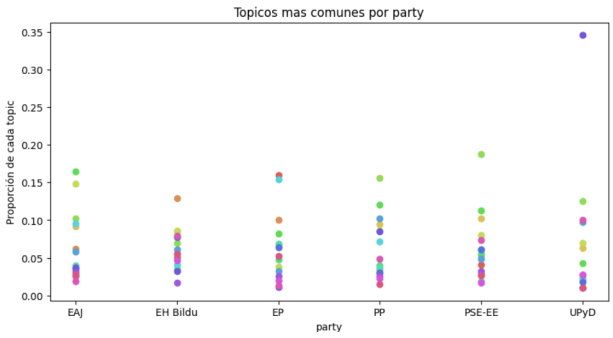

Figure 13: Topics by political party Model 2

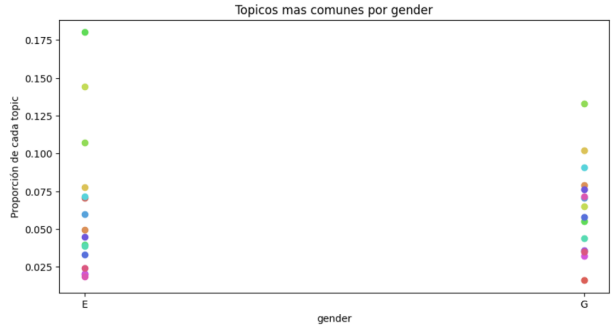

Figure 14: Topics by gender Model 2

interpreting the results, we can obtain a more comprehensive and accurate view of political speeches and underlying trends.

5 Colab

If you want a more precise view of the results or want to search for specific examples for a particular topic, it is recommended to visit the Jupyter Notebook where the analysis was conducted.

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[2] BasqueParl: A Bilingual Corpus of Basque Parliamentary Transcriptions (Escribano et al., LREC 2022)

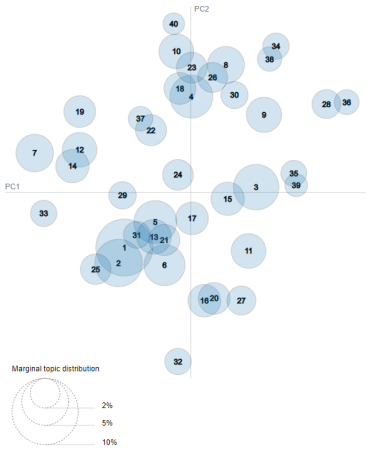

Figure 15: Topic distribution of Model 3

Figure 16: Selected topics Model 3

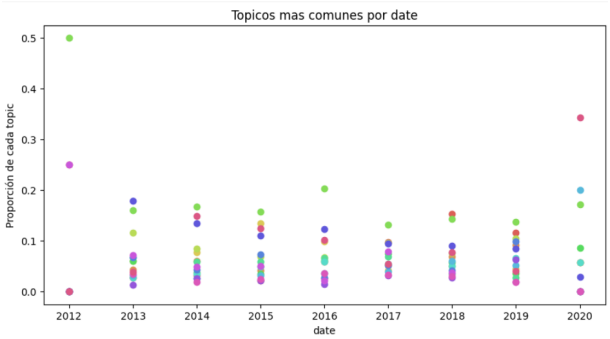

Figure 17: Topics by year Model 3

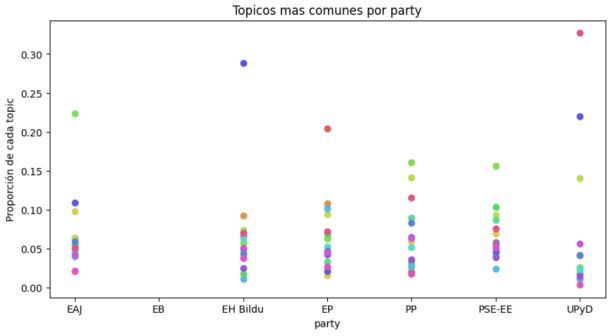

Figure 18: Topics by political party Model 3

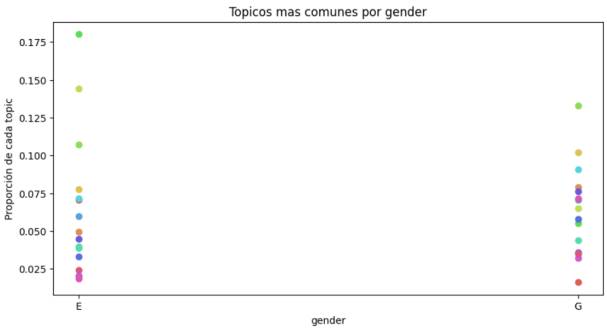

Figure 19: Topics by gender Model 3