Modeling Political Distances

Mathieu Breier Guillem Mirabent Rubinat

Barcelona School of Economics

I. Introduction

The spanish political landscape, once dominated by the Spanish Socialist Workers' Party (PSOE) and the Popular Party (PP), fragmented since the mid-2010s. This shift began around 2014-2015 with the emergence of new parties. Podemos, founded in 2014, positioned itself as a radical leftist alternative amid discontent from the financial crisis. Ciudadanos, a liberal centrist party, also rose around the same time, advocating for national unity and economic liberalism. Vox, a far-right party, entered the national parliament in April 2019 and focuses on national unity, opposing Catalan independence, criticizing immigration, and rejecting political correctness.

These changes have created a more polarized and complex political environment in Spain. Coalition governments are now more common, and traditional parties must negotiate compromises.

In this context, we aim to explore party lines and ideologies using a transformer neural networks architecture, seeking insights into how Spanish politicians express their beliefs. We will investigate whether a deep learning model can accurately distinguish speeches from different parties. This study builds upon the work of A. Bennett, M. Handt, and G. Mirabent on speech clustering in Spanish politics using Natural Language Processing techniques. The scope of this study could extend to identifying political ideologies in various speeches, thereby enhancing support for various economic analyses.

II. Related Work

This study contributes to the expanding body of literature focused on detecting political traits and ideologies in speeches through machine learning. Several studies have utilized deep learning models to investigate political ideologies and extract features from text. For instance, Pan et al. [2023] fine-tuned a pre-trained Italian Transformer model to detect the gender and political ideology of users based on their texts written in Italian.

Similarly, Iyyer et al. [2014] applied a recursive neural network (RNN) framework to identify the political position conveyed by a sentence. Their study aimed to determine whether a sentence exhibits ideological bias, specifically a political orientation that is either conservative or liberal.

Furthermore, our study draws from the work of Öztürk and Özcan [2022], who fine-tuned transformer-based models to detect different ideologies from English sentences. Their transformer-based neural network achieved an accuracy of 69% in classifying political sentences as liberal, conservative, or neutral.

III. Data

The data comes from an open source project called ParlaMint. The project team at ParlaMint has worked to create extensive metadata for each of the speeches, including speaker names, party affiliations, gender, age and position. We selected the Spanish corpus for our analysis.

The Spanish corpus consists of 32,551 speeches from the Spanish parliament. The speeches span 8 years from 2015 to 2023, and includes 5 legislative terms. In total, there were over 50 parties in the data including subgroups of the two major parties we are interested in, PP and PSOE.

In order to evaluate our model on unseen data, we split our data into 2 datasets:

1. **PSOE - PP** (main center-left, center-right national parties): Our **training** and **validation** set
2. **PODEMOS - VOX** (more clearly left and right national parties): Our **testing** set

Our objective is to train the model in such a way that it can classify speeches from different parties with respect to the training, but being trained on just the two main parties in Spain (PSOE-PP). Then we use this same trained model on speeches from different political parties.

IV. Preprocessing

The preprocessing involved several steps to prepare the parliamentary speeches dataset for our analysis. Initially, the data was divided into multiple .tsv and .txt files. The first step of the preprocessing was to consolidate this data into a single dataframe. We then filtered out non-congressional speakers, focusing only on interventions by "diputado/a" (congressman/woman).

Next, we performed data cleaning and transformation. We coded the Party affiliations into binary labels, with "PSOE" as 0 and "PP" as 1. We corrected typos in speaker names, labeled leftist parties as 0 and rightist parties as 1, and created a dummy variable to indicate whether the speaker's party was in power. We converted the birth years into ages, and typos in speaker IDs were fixed. We also encoded the legislative term, and ensured that all rows had associated speeches by removing any rows with missing speech data.

Once the dataset ready, we followed with tokenization and filtering. We tokenized the speeches, applied lemmatization, and removed stopwords. Speeches shorter than 300 tokens were filtered out, while those longer than 500 tokens were split into smaller chunks of consistent length. This splitting ensured that lengthy speeches did not introduce variability that could affect model performance.

The fine-tuning involved removing speeches that were too short after splitting and experimenting different minimum and maximum text lengths to optimize model performance. We also created a custom stopword list to exclude names and proper nouns, which were identified as significant in politicians' debate strategies. This step improved the model's accuracy by more than 5%.

Transformer Based Neural Networks

The transformer-based neural network is a novel architecture designed to solve sequence-to-sequence tasks and handle long-range dependencies with ease (figure 1). It uses an attention mechanism to weigh the importance of different sections of the input data, and has been widely adopted in natural language processing (NLP). This mechanism provides context for each position in the input stream independently.

Our transformer model classifies political speeches using self-attention to capture long-range dependencies. It features a token and positional embedding layer with an embedding dimension of 32. The core transformer block includes multi-head attention and feed-forward networks with dropout. The model also uses global average pooling and dense layers with RELU activation, and ends with a softmax layer for classification. To improve the model's performances, we tried to implement a Word2Vec embedding but it found no improvement.

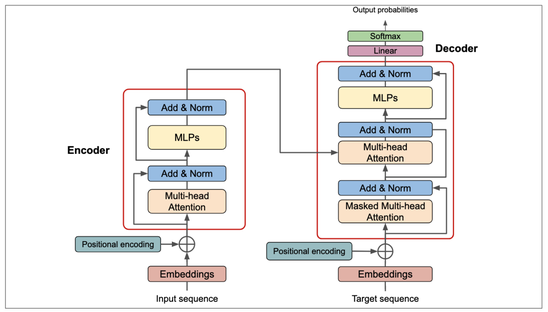

Figure 1. Transformer Blueprint

Baseline Models

For our evaluation, we consider several baseline models to compare their different performances with the transformer neural network:

- **Random:** This baseline randomly selects a label.
- **Random Forest (Rf):** A Random Forest model, which uses an ensemble of decision trees.
- **eXtreme Gradient Boosting (XGB):** An eXtreme Gradient Boosting model, that iteratively improves model accuracy by focusing on and correcting the errors of previous models.

VI. Results

The quantitative results of each model are shown in Table 1

Models	Accuracy	Precision	Recall	F1-Score	AUC
Random	0.49	0.49	0.49	0.49	0.50
Rf	0.82	0.82	0.82	0.82	0.90
XGB	0.81	0.81	0.81	0.81	0.91
Transformer	0.85	0.85	0.84	0.85	0.92

Table 1. Results of Transformer Trained on PSOE-PP Datasets and Evaluated on VOX-PODEMOS Speeches

Our results indicate that the Transformer-based neural network outperforms the other baseline models, achieving an area under the curve (AUC) of 0.92 on the VOX-PODEMOS test dataset. This slight improvement by our simple transformer architecture suggests that employing a more complex model, such as a large language model (LLM), could further enhance our results.

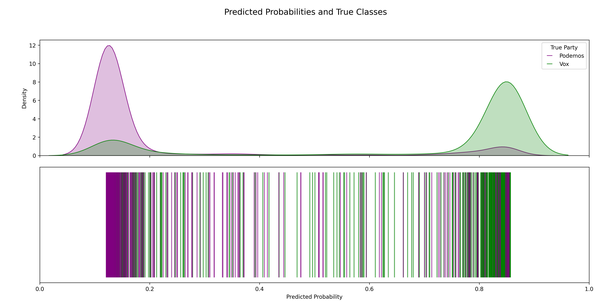

Figure 2. Podems/Vox Classification Predictions

Finally, we plot the predicted labels for the speeches in our test dataset (figure 2). Speeches labeled as from PODEMOS are shown in purple, while those from VOX are in green. We observe that most PODEMOS speeches are detected as far left, whereas most VOX speeches are recognized as far right.

References

Mirabent Rubinat Guillem, Bennett Tharaeparambil Andrew, and Handt Fueyo Miguel. Party lines: Speech clustering in spanish politics: An analysis of the differences in the speeches of members from the main parties in spain. 2024.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 1113–1122, 2014.

Oktay Öztürk and Alper Özcan. Ideology detection using transformer-based machine learning models, 2022.

Ronghao Pan, Ángela Almela, and Francisco García-Sánchez. Umuteam at politicit-evalita2023: Evaluating transformer model for detecting political ideology in italian texts. 2023.

mathieu.breier@bse.eu guillem.mirabent@bse.eu