Some Are Never Right, Others Have Nothing Left: Unveiling Ideological Extremes in Parliamentary Debates Using BERT

Mislav Đomlija, Barbara Kos, Jerko Šegvić

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
mislav.domlija@fer.hr, barbara.kos@fer.hr, jerko.segvic@fer.hr

Abstract

The speeches in parliamentary debates can often outline the overall political situation in the countries. In this paper, we fine-tune BERT on a downstream task of classifying political speeches by their political orientation. Afterward, we use representations of speeches to compare political stances between the countries, as well as the difference between right and left policies within the specific country. Because of the fine-tuning task, we can peek at the attention scores of the last layer to improve the understanding of what words are associated with different ideologies in a political setting.

1. Introduction

Political speeches in parliamentary debates are a vital component of democratic governance, offering insight into the ideological landscape and policy priorities of different countries. Understanding the political ideology conveyed through these speeches is crucial for researchers, policymakers, and the public, as it shapes national opinion and influences legislative outcomes.

This paper presents a comparative analysis of parliamentary speeches from 28 countries, utilizing a dataset from the CLEF 2024 Touché Lab [1], with all speeches translated into English. Our goal is to predict and compare the political ideologies expressed within these speeches. By leveraging advanced computational models, we seek to quantify how speeches from various countries align with left-wing or right-wing ideologies, and to examine how extreme these positions are in a global context.

The motivation for our study stems from the need to understand political speech across different political systems and cultural contexts. Analyzing parliamentary speeches can reveal underlying ideological shifts and highlight the diversity of political thought across nations. The main goal is to compare the extremity of ideological positions in parliamentary speeches from different countries.

In the following sections, we describe the methodology, data processing, model development, and results of our analysis, highlighting key findings and their implications for understanding political ideologies in parliamentary discussions.

2. Related work

Political ideology detection from text has been a growing area of interest in Natural Language Processing (NLP). Previous work in this field has primarily focused on using machine learning models to identify the ideological bias of political texts.

One notable study by Iyyer et al. (2014) utilized recursive neural networks (RNNs) to detect political ideology at the sentence level, focusing on the importance of subsentential elements. Their model, which incorporates phrase-level annotations, outperformed traditional bag-of-words methods by capturing the compositional aspects of language. In our approach, we leveraged BERT to capture contextual embeddings and handle longer dependencies within text, offering a more comprehensive analysis of complete speeches rather than just sentences.

Another significant contribution was by Baly et al. (2020), who developed a framework to predict political bias in news articles by creating a dataset annotated for political ideology at the article level. They utilized adversarial media adaptation and a triplet loss to prevent the model from learning the source rather than the bias, enhancing the performance of pre-trained Transformers by incorporating background information about the news source. In contrast, our focus is on parliamentary speeches from 28 European countries, utilizing a balanced dataset of left-wing and right-wing speeches. We analyze these speeches to compare the extremity of ideological positions across different countries, providing a direct comparison of political ideologies in various political systems.

Our approach differs from mentioned work by providing a comparison of how extreme the ideological positions are between countries, while offering insights into the diversity of political thought across different political systems and cultural contexts.

3. Dataset

The dataset used in this study comes from the CLEF 2024 Touché Lab and includes parliamentary speeches from 28 European countries and regions. These countries are diverse, spanning from Austria and Bosnia and Herzegovina to the United Kingdom and Ukraine. The country with the most examples in the dataset is Great Britain. The dataset provides a comprehensive view of political discourse across a wide range of political systems and cultural contexts.

In total, the dataset consists of around 40% left-wing speeches and around 60% right-wing speeches. To ensure the quality and consistency of our analysis, we filtered the dataset to include only speeches that can be tokenized into

[1] https://touche.webis.de/clef24/touche24-web/

512 tokens or fewer. This was necessary because we used a smaller version of the BERT model, which has a maximum token limit of 512 tokens. This step was crucial to maintain the integrity of the text inputs for our computational models.

To ensure balanced representation, we split the speeches from each country into training, testing, and validation sets with an $80/10/10$ ratio. Specifically, we allocated 31.411 speeches to the training set, 3.914 speeches to the testing set, and 3.922 speeches to the validation set. This approach ensured that the diversity and representativeness of speeches from each country were maintained across all sets, allowing for a balanced and fair comparison.

4. Experimental setup

In this section, we outline the experimental setup used to fine-tune BERT for political ideology classification. We discuss the tokenizer, model architecture, and the fine-tuning process in detail.

4.1. Tokenizer

We employed the AutoTokenizer from the HuggingFace Transformers [2] library to preprocess the text. This tokenizer ensures that each input sequence is formatted with the [CLS] token at the beginning and the [SEP] token at the end, with padding and truncation to a maximum length of 512 tokens. The truncation was necessary to ensure that all input sequences fit within the BERT model's maximum token limit.

4.2. Model architecture

To learn the intricacies of political rhetoric and ideologies we utilised BERT from Devlin et al. (2019) on downstream task of speech classification. We used **BERT$_{\text{BASE}}$** (L=12, H=768, A=12, Total Parameters ~110M) with classification head in line with (Sun et al. (2020). By using the [CLS] token for classification, the model learns to capture ideological cues and features in the embedding of [CLS] token. This embedding along with classification probability can be used to discern the degree of extremism in a speech. The architecture of the model is illustrated in Figure 1, providing a detailed view of its structure and components.

4.3. Fine-tuning process

For fine-tuning, we created two experimental setups. In the first one, we used speeches in native languages for creating token embeddings, while in the second we used English translations. We chose to use the English **BERT$_{\text{BASE}}$** model for all languages to investigate its cross-lingual capabilities and to see how well this widely recognized and robust model performs on multilingual datasets. For both experiments, we used the Adam optimizer and Exponential learning rate scheduler. We experimented with different training hyperparameters and got the best results with a learning rate of 1×10^{-5} and scheduler learning rate decay of 0.75 while training the model for five epochs. For each of those setups, we trained the model five times and took the ones with the best results on the validation dataset for further evaluation.

[2] https://huggingface.co/

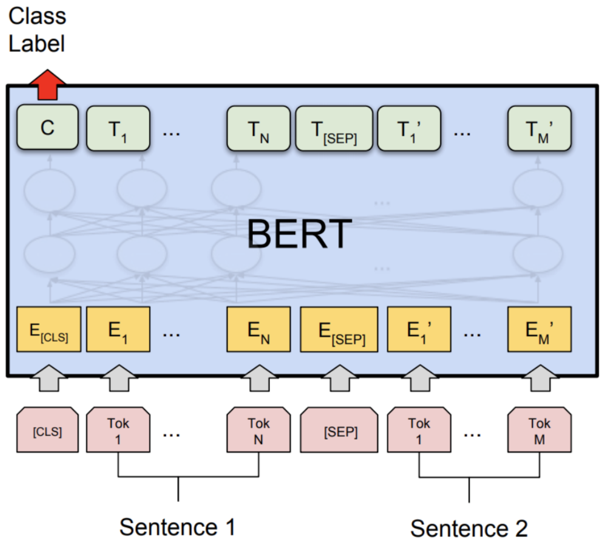

Figure 1: BERT architecture

5. Results

5.1. Results on original classification task

The results on an (unseen) test set of our best runs are shown in Table 1. The bar chart showing weighted F1 for every country of native-language setup is shown in Figure 2 and for English setup in Figure 3. We used the weighted F1 as a main metric because we found it most informative for our dataset. Since many countries have very imbalanced datasets, the ordinary F1 largely depends on which of the labels are considered positive.

Native	**PR**	**R**	**F1**	**Acc.**
Left	0.68	0.64	0.66	0.73
Right	0.76	0.79	0.78	
Weighted average	0.73	0.73	0.73	0.73
English	**PR**	**R**	**F1**	**Acc.**
Left	0.70	0.64	0.67	0.74
Right	0.77	0.81	0.79	
Weighted average	0.74	0.74	0.74	0.74

Table 1: Binary Classification Report

The results are especially interesting considering the fact that the same model (that was pre-trained primarily on English corpora) got similar results on English speeches and native language speeches. We argue that non-English speeches are often badly translated and the results in English setup could be much better.

For further analysis, we calculated the mean probabilities of the model's outputs for left and right speeches for every country in the training dataset. We took all left speeches in the country calculated the model's mean probability for them and did the same for the right speeches. In other

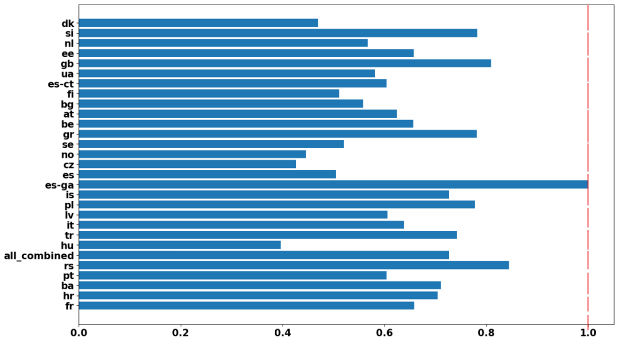

Figure 2: Weighted F1 scores by country in native language setup

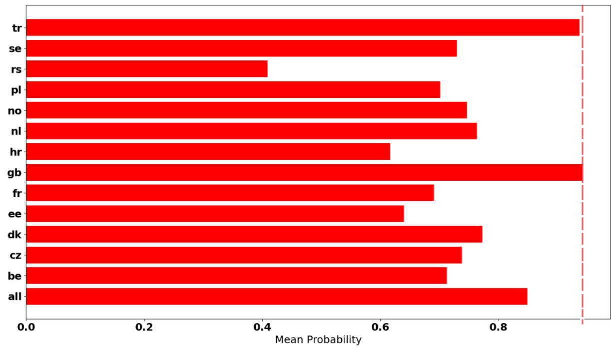

Figure 4: Mean probability of left speeches by country

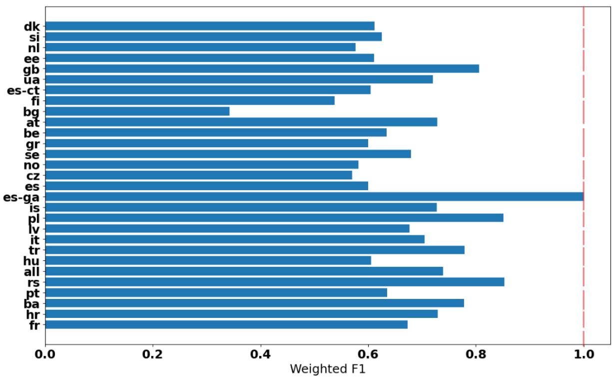

Figure 3: Weighted F1 scores by country in English setup

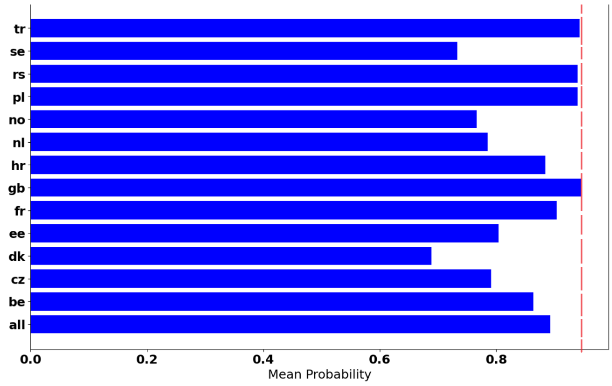

Figure 5: Mean probability of right speeches by country

words, for each country we calculated

$$\overline{P}_{left} = \frac{1}{|L|} \sum_{x \in L} P_{model}(left|x)$$

$$\overline{P}_{right} = \frac{1}{|R|} \sum_{x \in R} P_{model}(right|x)$$

where L is a set of all left speeches in the country and R is a set of all right speeches in the country. The results for left speeches are plotted in Figure 4 and for right speeches in Figure 5. We showed only countries that had more than 500 examples in the training dataset. Those with high mean probabilities have speeches that were easier for the model to label as right or left. This can be because the speeches are more extreme to the left or to the right or because they are just easy to classify. We argue that it is a mix of both, because countries like Serbia(rs) and Poland(pl) have a huge mean probability of right speeches but much lower for the left speeches, and we know from the domain knowledge that those countries tend to have more right politics. On the other hand, countries like Denmark(dk) and Norway(no) have lower mean probabilities for right speeches and higher for left.

5.2. Analyzing the hidden state of classification token

Since some of the speeches are not indicative of political opinions (e.g. talking about certain procedures or general inquiry of government policy) we consider that the model has successfully captured the intricacies of political speeches and that it can provide valuable insight. The dataset consists of 28 different countries and regions, so we computed the mean [CLS] token embedding for right and left speeches in the training dataset.

By using PCA for dimensionality reduction we can effectively visualize extremism of political opinions between countries. By doing this we can quantify how the political labels (left and right) change between the countries and general country position regarding politics. Furthermore, we can analyze how different political currents are in each country. The results of this experiment are shown in Figure 6 (again we included only those countries with more than 500 examples on the training dataset). We can see that the model learned some kind of pattern in the hidden state of the [CLS] token. We can interpret that the horizontal axis represents the country's left or right orientation. Also, we can see that those countries that had high mean probabilities of left and right speeches in the last experiment have more separated mean right and mean left representations (like Great Britain(gb) and Turkey(tr)). Also, those with lower scores (like Sweden(se)) have less separated means.

5.3. Finding most decisive words

By inspecting the attention scores of final layer before the classification head we can gain great insight in how different words in the speech influence its political orientation and the degree of extremism. As the classification of speeches is based on the [CLS] token, the model is greatly

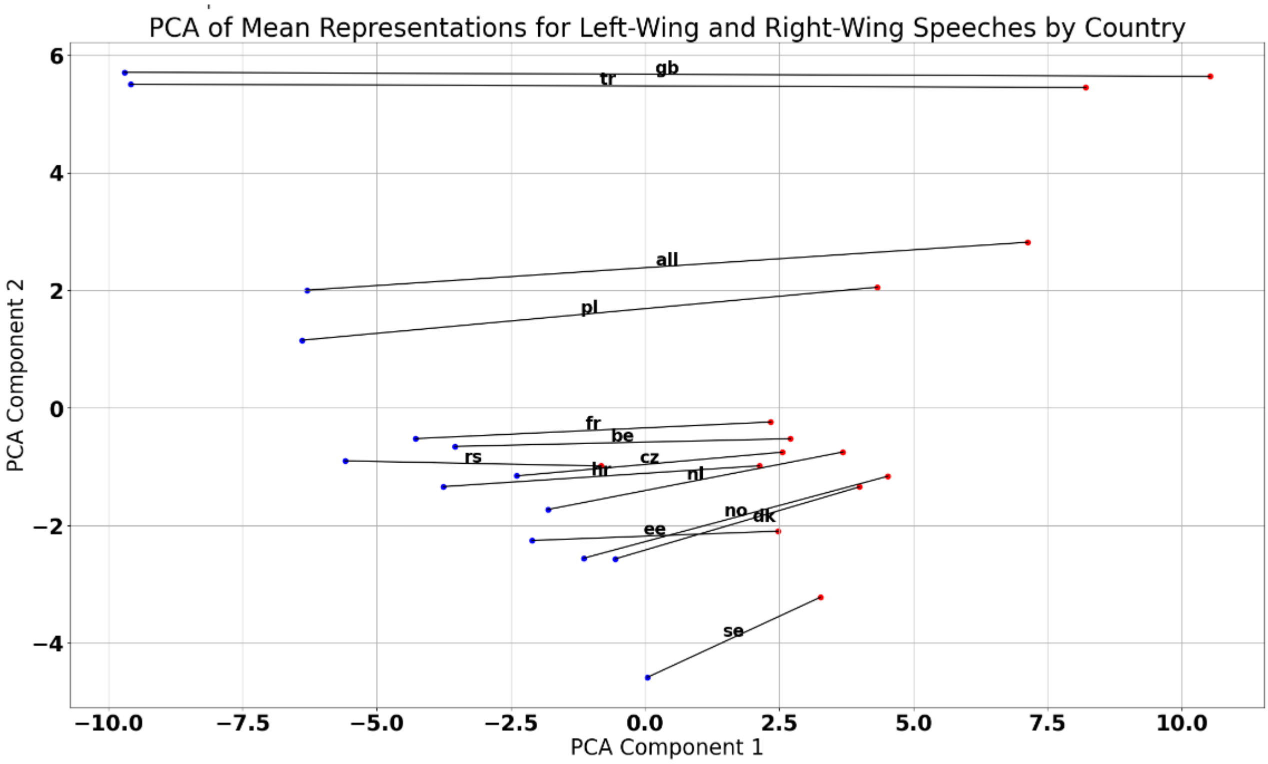

Figure 6: PCA analysis of mean representations. Blue dots represent mean right, while red dots represent mean left.

Madam President, I would like Recently, we have been able to read many stories about **labour immigrants** who are to be expelled because of small mistakes, which have often been **made** by previous employers, and things that have happened several years ago. This is basically a new practice from the **Migration** Court, which now means that thousands of people will have their cases tried in a completely new way. It is about people who behave themselves, pay taxes and contribute to society, people who just want to live their lives here. Abuse of the system should, of course, be prosecuted and countered, but this is not about abuse. These are small mistakes, and there is no way whatsoever of correcting them afterwards. This affects the **worker** very hard.

Figure 7: Example of speech from left political spectrum

incentivized to capture the connection between tokens and political orientation in attention scores computed for the [CLS] token. Inspired by (Abnar and Zuidema (2020) we use this connection to highlight key words and phrases that indicate political orientation. We analysed 5000 most extreme speeches for both labels. We then took top 5 tokens sorted by attention scores from each speech and tracked the number of times the tokens were found decisive. Some interesting tokens are presented in table 2. Furthermore we can use this mechanism to highlight important words for orientation identification. An example of speech from the left political spectrum is displayed in Figure 7 with words with high attention scores emphasised in bold.

Table 2: Number of times the token was important for the most extreme speeches.

Token	Left	Right
green	22	8
crisis	16	22
steel	0	27
workers	69	30
health	35	12
children	41	9
tax	13	4
treasury	1	19

6. Conclusion

Politics is often called *the art of possible*, and *the possible* can oftentimes differ between the countries. By training our model on multiple countries we are able to capture how politics vary between the nations. The difference between the left and right political currents within the country also unveils valuable information about the political stances of people.

The analysis of last-layer attention scores is a powerful tool for the vocabulary and topic analysis of different political groups both within and between countries.

The accuracy of ideology prediction can be improved with a better translation of speeches to English, and the accuracy per country can be improved by training the model for each country separately. We leave these prospects for future work.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov, 2020. *We Can Detect Your Bias: Predicting the Political Ideology of News Articles*, pages 4982—-4991.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik, 2014. *Political Ideology Detection Using Recursive Neural Networks*, pages 1113—1122.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?