A quantitative analysis on the Italian Senate

Bruno Giacomo
bruno.giacomo@studbocconi.it
and Leonardo Bandera
leonardo.bandera@studbocconi.it

Abstract

This study analyzes ideological shifts in the Italian Senate from 1948 to 2024, focusing on the "Seconda Repubblica" era. We utilized doc2vec and Bert-based models to examine political alignment evolution and the blurring of traditional left-right distinctions. Our findings reveal increased ideological heterogeneity and prediction errors, particularly post-Tangentopoli, indicating a dynamic political landscape where new parties display fluid ideological stances.

1 Introduction

The Italian political landscape, spanning from 1948 to 2024, has witnessed significant transformations, particularly emphasized during the "Seconda Repubblica". This era, marked by the rise of populism and major political crisis such as the Tangentopoli scandal, has introduced a period of remarkable fluidity in political alignments, with traditional parties dissolving and new movements taking the stage.

In this study, we constructed a comprehensive dataset of the Italian Senate's transcripts, meticulously organized by legislature and further categorized by individual senators and their parliamentary speeches. Our objective was to quantitatively analyze the ideological shifts and polarization within the Italian Senate, with a particular focus on the last 20 years.

To capture the nuanced spectrum of Italian politics, we labeled political parties with values ranging from -1 (extreme-left) to 1 (extreme-right) and framed this as a regression problem. We utilized doc2vec to embed senatorial speeches from four distinct periods of Italian politics and build a classical ML regression model on the embeddings. In addition we finetuned two versions of "Bertino", an italian distilBert model, one trained on data from 1948 to 1998 and another on the most recent 15 years.

These models aimed to discern if political parties are consistently aligning with their historical ideological positions or exhibiting unpredicted heterogeneity in their discourses, indicative of the ongoing realignments and the blurring of traditional left-right distinctions. This approach allowed us to explore whether there is a clear division and polarization within parties or a complex mix-up, reflecting the evolving and dynamic nature of Italian political alignments.

Our findings indicate that the most recent legislature has been particularly challenging in defining clear ideological boundaries. Since the Tangentopoli scandal, there has been a noticeable increase in prediction error, which supports the notion that traditional left-right distinctions are increasingly blurred. This shift is exemplified by the behaviors of newer parties such as Movimento 5 Stelle (M5S) and Lega, which not only lack strictly defined ideological values but also spatially move across the entire political spectrum. Unlike the more ideologically consistent parties of the 20th century, these newer entities demonstrate a broad range of positions, reflecting their adaptability and fluidity.

2 The Data

A brief introduction to Italian politics

The Italian Senate, is one branch of the Italian Parliament. Senators, like their counterparts in the Camera, are elected for five-year terms, which may be cut short by early elections if called by the President of the Republic. Due to shifting political alliances and changes in leadership, it is common for multiple governments to serve within a single legislative period. Senators are initially grouped by political party affiliations, but they have the freedom to change their party alignment during their tenure. Sessions of the Senate, known as "sedute," focus on specific agenda items determined by the President of the Senate in cooperation with the government. These sessions are critical for discussing national issues and advancing legislative proposals. Additionally, government members regularly address the Senate to update on current issues or respond during question times, clarifying government actions and policies.

The landscape of Italian politics underwent a seismic shift following the Tangentopoli scandal in the early 1990s. This corruption scandal, which exposed widespread graft within the political ecosystem and business sectors, marked a definitive end to what is commonly referred to as the "First Republic". The aftermath of Tangentopoli led to a complete overhaul of the political system, ushering in what is now known as the "Second Republic" This new era is characterized by increased fragmentation, a reconfiguration of party systems, and shifts in political ideology. Since its first legislature in 1948, Italy has progressed through 19 legislatures, with the latest commencing in 2022.

How We Got the Data

Obtaining the data was one of the most demanding parts of our project. We faced several issues, from acquiring the pure text to transforming it into usable data. Our initial challenge was the costly task of obtaining text files of Senate discussion transcripts. We developed a script to scrape the Senato website[1] and download PDF transcripts of each session for every year, totaling 11,688 PDF files.

For the first 11 legislatures, these PDFs, especially from earlier years, required OCR technology; we used Tesseract OCR to convert them into text files. Each OCR scan per legislature took about one day. Although it was expensive, it was necessary to process the text and obtain all the data. For the remaining legislatures, the selectable text in the PDFs was of high quality, which simplified the conversion process.

Once we obtained the text of the Senate sessions, we removed all non-interesting parts such as introductions, discussions by the president of the Senate, and other formal text segments. Additionally, we collected the list of all Senators and government officials for each legislature to identify who spoke at each session and grouped them in two tables, you can see the examples here above.

These two tables were fundamental. Using regex, we extracted the names and corresponding texts from each session. This process allowed us to identify who spoke and what was said. We then created a CSV file for each session. In these CSVs, each row represented a speaker, with columns for the speaker's name, party affiliation, and their speech during that session.

Table 1: Example of one row in `senators.csv`. Total of 8096 senators

Member	Legislatura	Group Name
SPANO Velio	1	Comunista

Table 2: Example of one row in `government.csv`. Total lines = 5089

Name	Government	Ministry	Name
1	De Gasperi-IV	Premier	DE GASPERI A.

How We Used This Data

Given all the CSVs for each session, we had several ways in which we could proceed. We consolidated the data from each Senate session into a single CSV per legislative period. Each row in this merged dataset represented a speaker, with their speeches from the entire period concatenated in the text column. To facilitate text analysis with spacy, we initially cleaned the data by using string manipulation techniques to remove custom stop words, reducing the text volume by about 30%, and stripping all punctuation then with multiprocessing we lemmatised the cleaned text with spacy. Finally, we categorized all political parties and government positions into six groups: far right, right, center-right, center, center-left, and extreme left. We then assigned values ranging from -1 for extreme left to 1 for extreme right to each category and labeled each speaker. This approach helped us create a final table that organized the data clearly in a final table that looked like this:

Name	Party	Legislature
Salvini	Lega	18

Political Category	Political Spectrum	Text	Lemma clean
Right	0.5	text ...	cleaned text ...

Disclaimer In utilizing the data, we have classified the political parties from 1948 to 2022 into a spectrum that ranges from extreme right to extreme left. This classification is a significant disclaimer because it introduces a substantial inductive bias. The labels were assigned based on our interpretation and understanding of Italian politics, which

[1] http://senato.it/

is inherently subjective and can vary significantly from one individual to another. While we are prepared to publish the dataset without these labels, allowing for analysis from alternative perspectives, we must acknowledge that our categorization may influence the interpretation of the data. Furthermore, it is important to recognize that the textual and CSV files for each parliamentary session are of great value and can be used to construct analyses in myriad ways. For example, researchers could choose to filter and analyze only the speeches of specific ministries such as Defense or Finance to track their evolution over time, or they might focus on the textual content to study the evolution of language used in politics. Analyses could also be directed at specific senators or governments. Our focus, however, has been on studying the polarization and the evolution of the political landscape, which necessitated the categorization of political parties.

3 Analysys and Results

To tackle the task of measuring the political polarization, we employed two distinct methodologies: Doc2Vec and BERT-based models. Initially, we utilized Doc2Vec to build embeddings for groups of legislative periods, aiming to discern the coherence and consistency of party ideologies over time. On top of these embeddings, we employed classical machine learning models for regression, such as the random forest regressor and Support Vector Regressors, to predict and evaluate the political alignment based on speech content. Subsequently, to leverage the expressive power of BERT, we transitioned to a more robust approach using 'Bertino', a specialized Italian version of DistilBERT adapted for regression tasks. By tracking losses and predictions at both the individual and party levels, we gained a deeper understanding of the political landscape, which will be presented in this section.

3.1 Doc2vec

We structured our exploration of Italian political history into four distinct epochs, each spanning approximately 20 years, and trained Doc2Vec to create embedding spaces for each. Our aim was to construct a classical machine learning regressor for each period and assess both the internal coherence and the temporal continuity of political ideologies.

Initially, we evaluated the internal coherence within each epoch by testing the embeddings on a test set from the same period. This approach helped us to understand the clarity and consistency of party guidelines and political ideas within each epoch. The results showed mean squared errors ranging from 0.1 to 0.15 for the first three chunks, indicating distinct ideological lines. However, the last period (2000-present) exhibited a much higher MSE of nearly 0.3, suggesting a blurring of political boundaries in recent times.

We then examined the continuity of political values over time by embedding speeches from subsequent legislatures into the embeddings of earlier epochs. The MSE for these tests was around 0.2, slightly higher than within-chunk errors but still indicative of well-defined and consistent political values until 1996, in which tangentopoli caused the error to increase slightly for the conitnuity from the 3 to the 4. Further exploring the adaptability of contemporary political discourse, we conducted a "present into past" experiment by embedding the 18th legislature into the embedding spaces of the first three epochs. This resulted in a significant error of approximately 0.5, underlining the difficulty of interpreting recent political changes through the traditional left-right spectrum and pointing to a major shift in political discourse.

We employed various regressors and utilized cross-validation for hyperparameter tuning, with the Support Vector Regressor yielding the most robust and reliable predictions. Despite exceptions like the MSI (Italian Social Movement), which showed high error rates during the second period due to labeling biases and its ambiguous far-right stance, the overall trend indicated clear ideological distinctions in earlier periods.

3.2 Bertino I

Transitioning from Doc2Vec to a more expressive model, we adapted Bertino for the regression task, so with one output neuron. We then applied tanh activation function to its final layer that constrains output values between -1 and 1 and utilized Mean Squared Error as the loss function. The decision to fine-tune Bertino was driven by the clear division and coherence among political parties observed in the First Republic with doc2vec (covering 12 legislatures) until 1996. Given the Bert architecture's token limit, we segmented speeches into chunks of no more than 510 tokens, assembling a substantial training dataset of 461,680 chunks.

After five epochs of training Bertino, we achieved a training error of 0.1 and a validation error of 0.2, with the second epoch yielding the best results. For testing, we analyzed data from the Second Republic across three periods: 1996-2006, 2006-2013, and 2013-2022. The testing errors increased over time, culminating in the highest mean squared error of 0.5 in the latest period, while the first two periods registered errors between 0.3 and 0.4, largely due to the variable discourse of Lega Nord. 10

Lega Nord, advocating for the autonomy of Northern Italy, notably diverged from traditional right-wing values of the First Republic. This ideological shift resulted in substantial losses in model performance, with Lega senators' speeches covering a broad ideological range, this is confirmed by violin plots showing their wide spectrum of political positions. 9

In contrast, the new party Forza Italia, led by the new politician Berlusconi, remained close to center-right, reflecting traditional liberal-right values. However, L'Ulivo, a coalition of left-wing parties formed to counter Berlusconi, generated high prediction errors due to its diverse ideological composition. 8

The 2014-2022 period saw the rise of Movimento 5 Stelle, a populist and ideologically nonconformist party, leading to the highest losses in our model with an average error rate of 1.2. Their speeches demonstrated a lack of clear ideological positioning, spanning the full spectrum from -1 to 1. Despite this, government officials' discourse remained consistently moderate, clustering around the political center, except during the populist Conte I government, born by the coalition of M5S and Lega, which exhibited a high error rate as illustrated in our plots. 7

3.3 Bertino II

Given the significant errors in recent periods, especially during the 18th legislature (2013-2018), we retrained Bertino on the more recent legislative data from the 15th to the 17th legislatures. This retraining aimed to provide nuanced insights into the 18th legislature, focusing on understanding the complex political dynamics during significant events like the Conte I government and the COVID-19 pandemic.

Our analysis with the updated Bertino showed improved understanding of Movimento 5 Stelle (M5S), positioning it close to the extreme left with a value of -0.75, instead before the bert placed it at extreme right indicating a more defined ideological stance than previously captured so still a bit of shadows on m5s remains. However, Lega continued to exhibit unpredictability in its shift from advocating northern autonomy to embracing national interests, with key figures like Salvini and Bagnai showing a wide range of ideological positioning from -0.75 to 0.75. attach plot 5

This legislative period was characterized by frequent changes in party allegiance and government coalitions, leading to high losses in model prediction for senators who switched parties, like Francesco Mollame who changed parties three times in one tenure. This instability highlights the fluid nature of Italian politics, where traditional alignments are increasingly blurred.

Bertino showed that stable parties like the Democratic Party (PD) and Forza Italia maintained consistent ideological positions with low prediction errors, reflecting a clear political identity.4 In contrast, the government of Conte I and new populist movements demonstrated a mix of political ideas, resulting in a wide spread across the political spectrum in our analyses. 6

3.4 Conclusion

Our analysis using Doc2Vec and Bertino offers significant insights into the evolution of Italian political discourse. The Draghi government, exemplifying classical centrist values, was accurately classified by all three methods from the outset. In contrast, Movimento 5 Stelle (M5S) showcased high variability in its classification, initially categorized as far-right and later as left-leaning, highlighting its ambiguous ideological position.

The emergence of Lega Nord and Fratelli d'Italia has introduced unpredictability into the right-wing sector, significantly shifting its ideological boundaries. Meanwhile, parties like the PD and Forza Italia, displayed stable identities with low prediction errors. This indicates a continuity in their political ideologies despite the broader shifts.

The study underlines the complex and often opaque structure of Italy's newer political entities. To enhance our understanding, further analysis on senatorial discourse could help quantify this "incoherence".

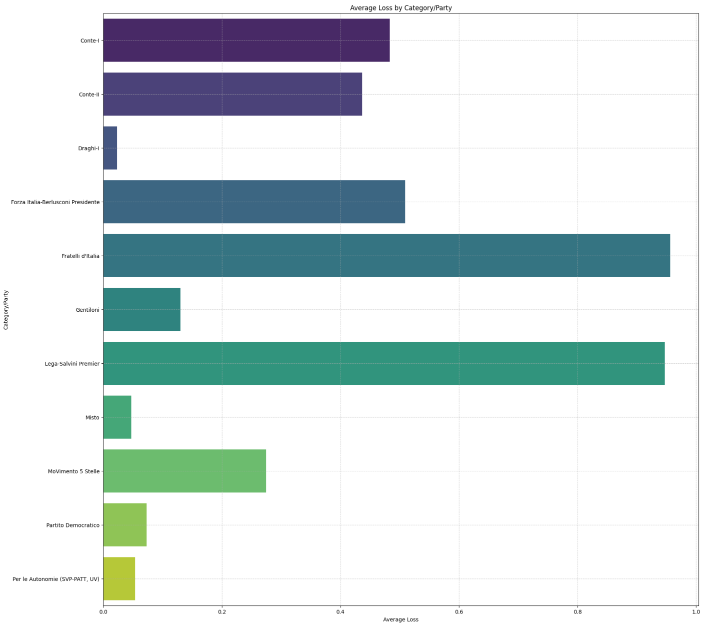

Figure 1: XVIII legislature error w.r.t. first epoch Doc2Vec embedding

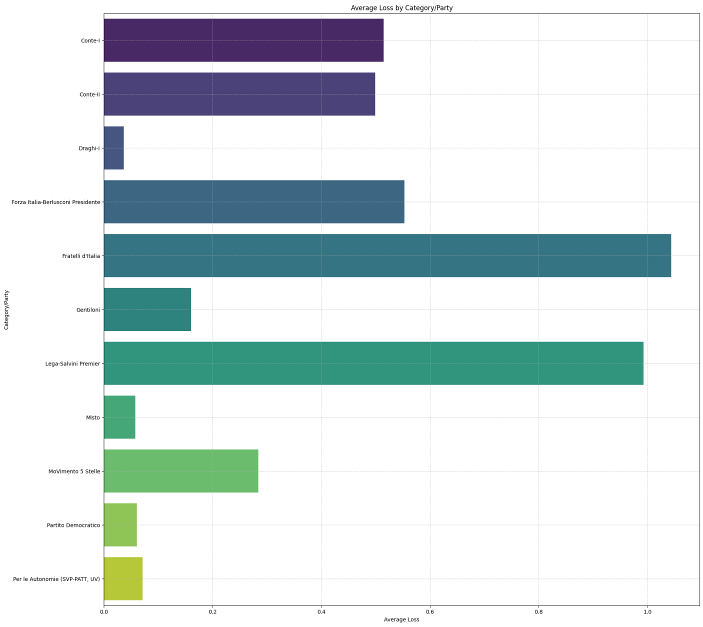

Figure 2: XVIII legislature error w.r.t. second epoch Doc2Vec embedding

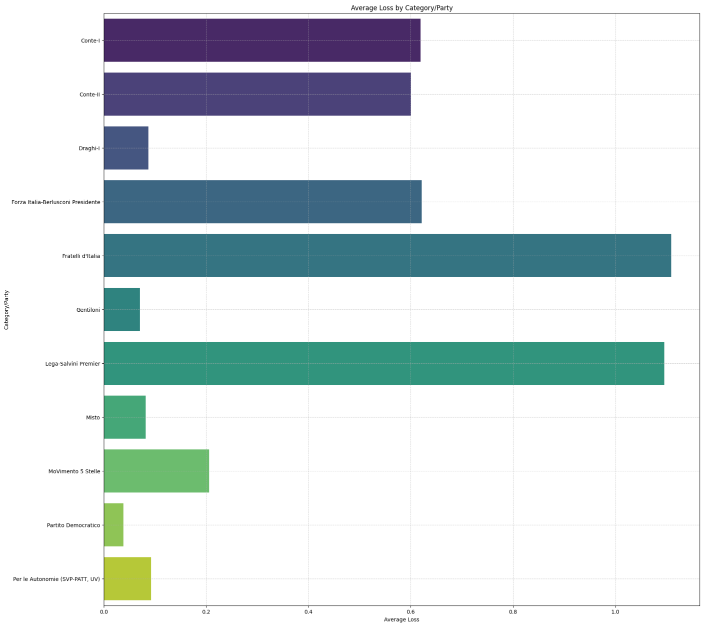

Figure 3: XVIII legislature error w.r.t. third epoch Doc2Vec embedding

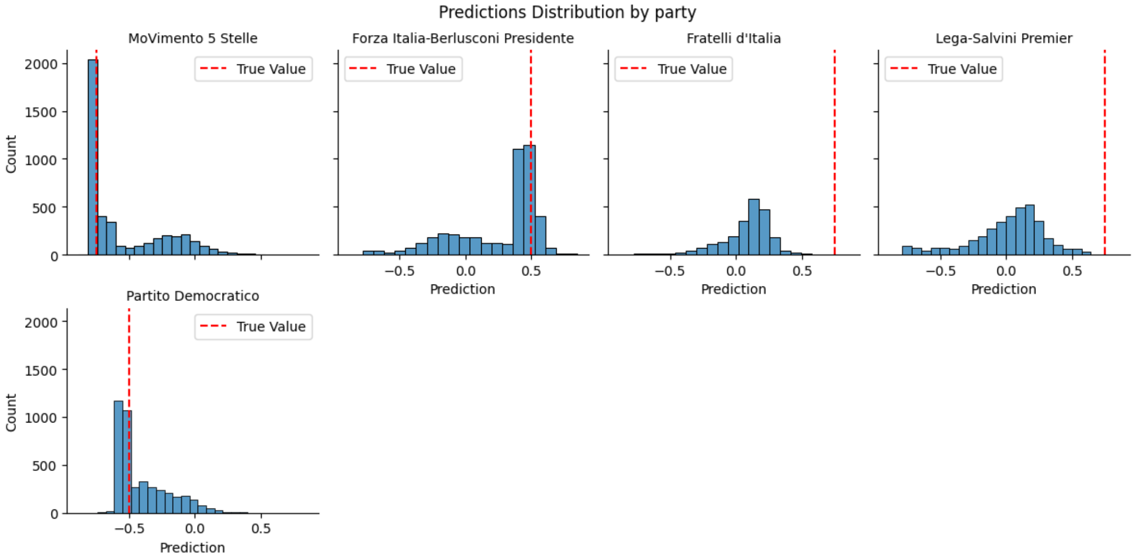

Figure 4: prediction of partiti in the 18: model bertino II

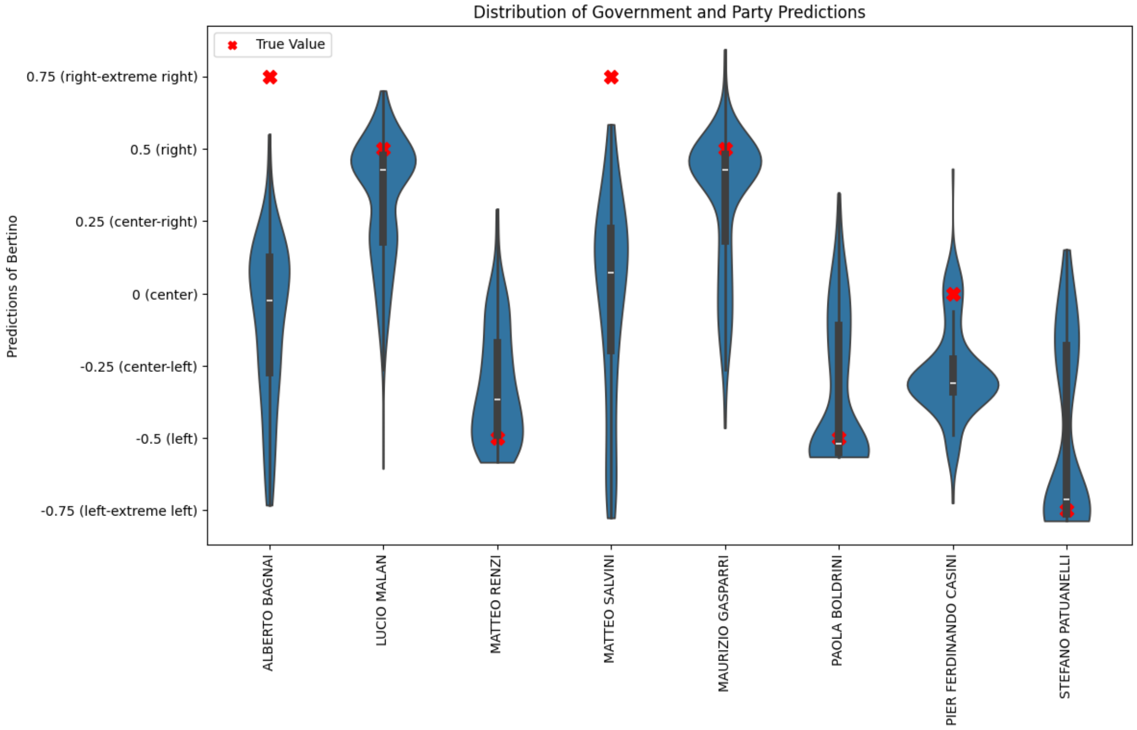

Figure 5: Error of selected senators in the 18 legislatura, model: bertinoII

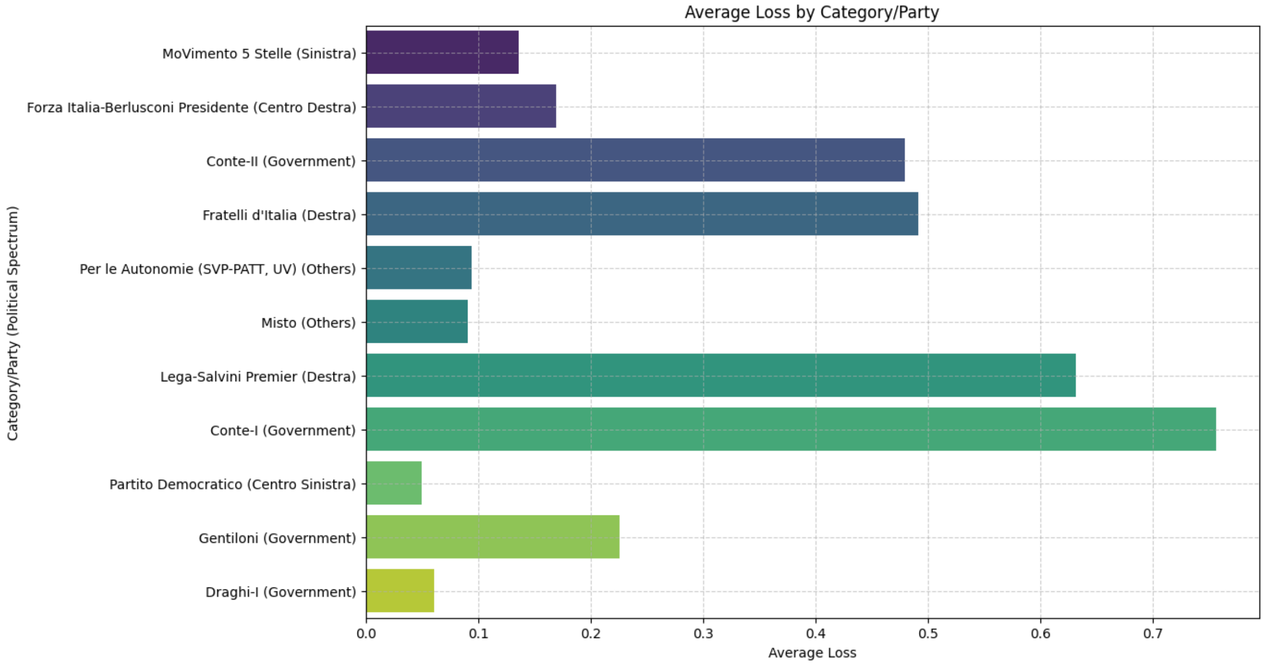

Figure 6: Average loss in the 18 legislatura of parties, model: bertinoII

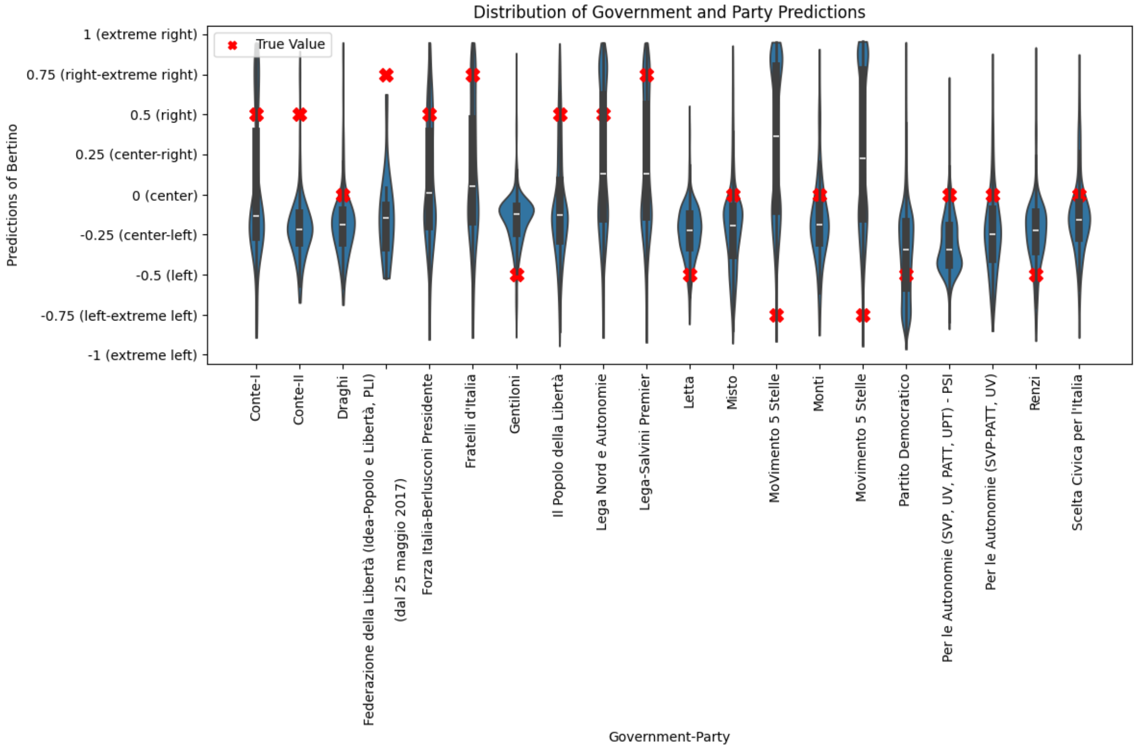

Figure 7: Average positioning-violinplot in the 18 legislatura of parties, model: bertinoI

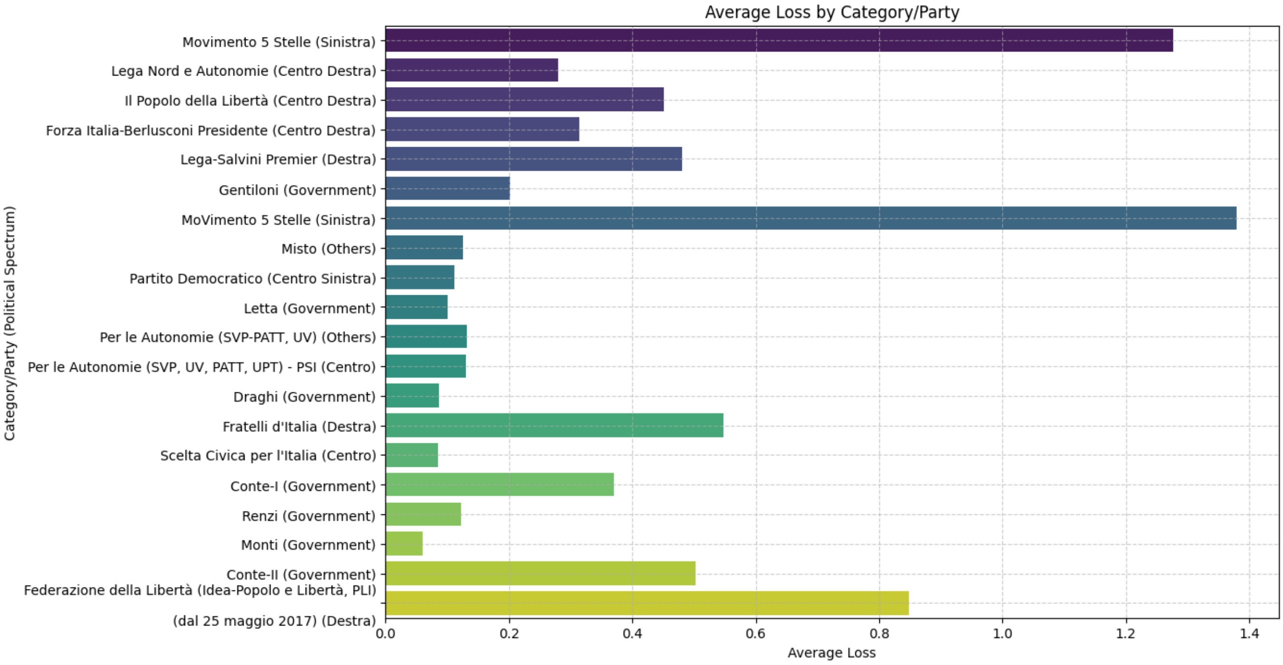

Figure 8: Average loss in the 17-18 legislatura of parties, model: bertinoI

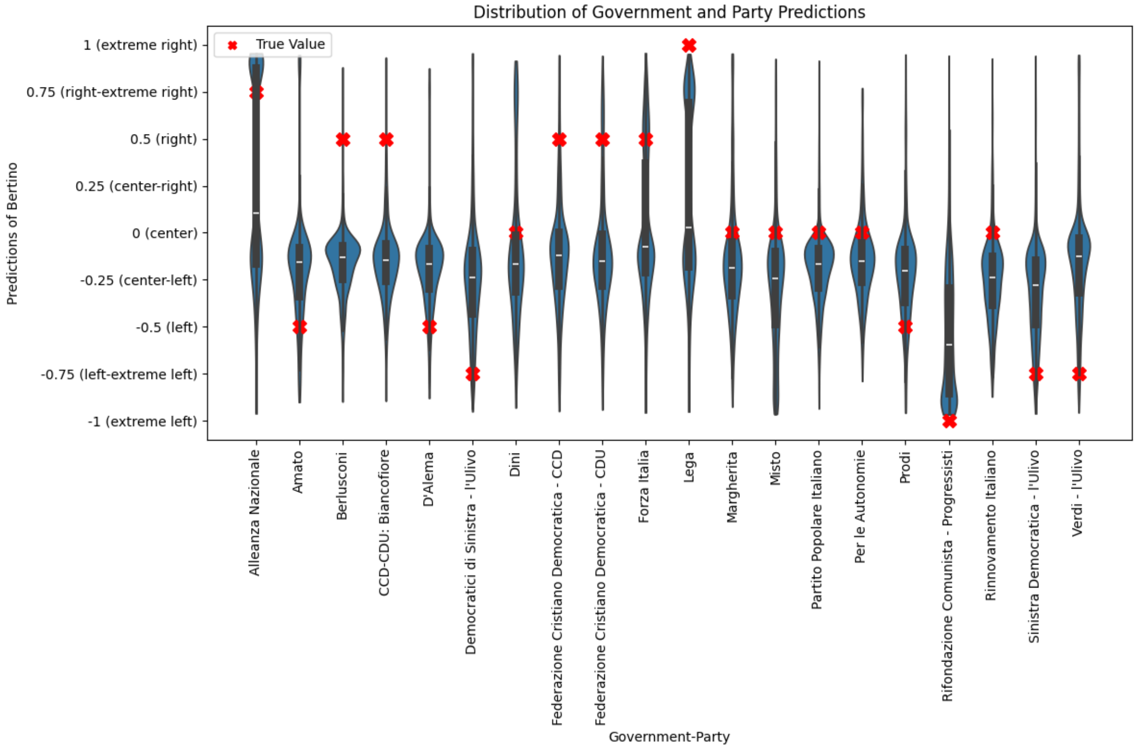

Figure 9: Average loss in the 15-16 legislatura of parties, model: bertinoI

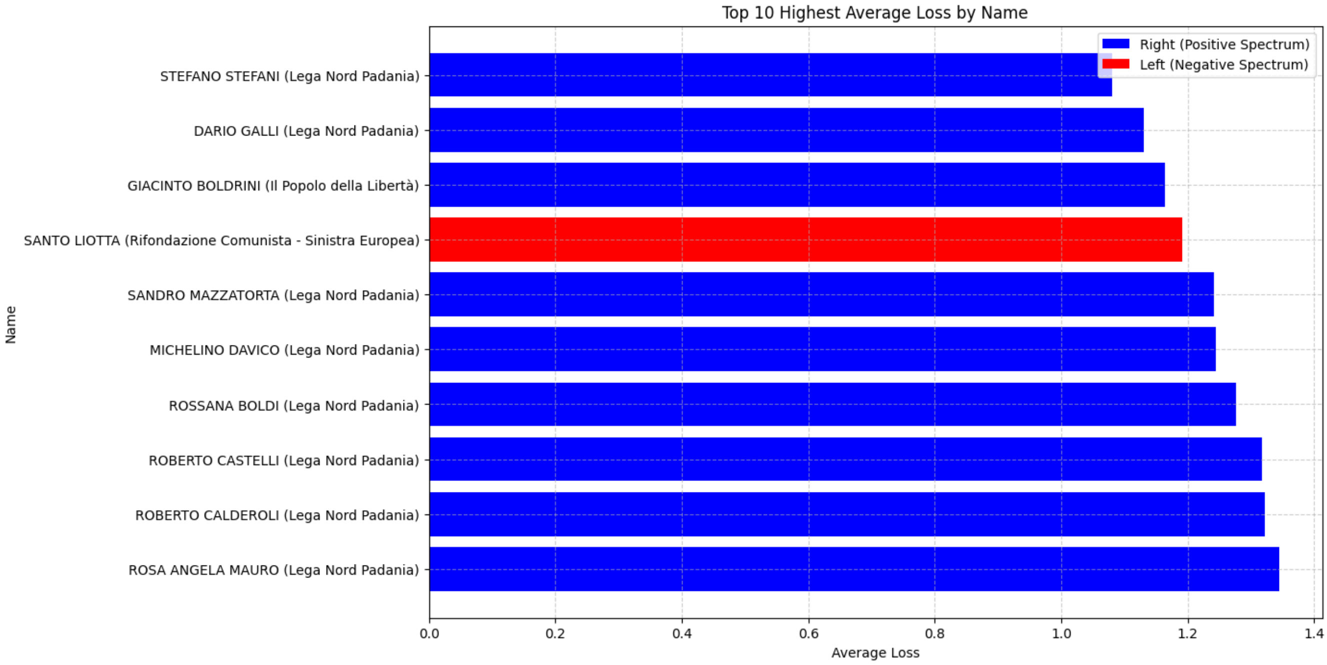

Figure 10: Senator with the highest error, model: bertinoI