UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ECONOMIA
PROGRAMA DE PÓS GRADUAÇÃO EM ECONOMIA

Samuel de Mendonça Souza

Measuring Rhetorical Styles to Understand Political Ideology and Economic Discourse in the Brazilian Senate

Juiz de Fora

2026

Samuel de Mendonça Souza

Measuring Rhetorical Styles to Understand Political Ideology and Economic Discourse in the Brazilian Senate

Dissertação apresentada ao Programa de Pós Graduação em Economia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Economia. Área de concentração: Economia

Orientador: Prof. Dr. Douglas Silveira

Coorientador: Prof. Dr. Rafael Morais de Souza

Juiz de Fora

2026

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Souza, Samuel de Mendonça.
Measuring Rhetorical Styles to Understand Political Ideology and Economic Discourse in the Brazilian Senate / Samuel de Mendonça Souza. – 2026.
55 f. : il.

Orientador: Douglas Silveira
Coorientador: Rafael Morais de Souza
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Faculdade de Economia. Programa de Pós Graduação em Economia , 2026.

1. political rhetoric. 2. machine learning. 3. topic modeling. 4. ideology. 5. Federal Senate. I. Silveira, Douglas Sad, Souza. II. Rafael Morais. III. Título.

Samuel de Mendonça Souza

Measuring Rhetorical Styles to Understand Political Ideology and Economic Discourse in the Brazilian Senate

Dissertação apresentada ao Programa de Pós-graduação em Economia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Economia Aplicada. Área de concentração: Economia.

Aprovada em 29 de maio de 2026.

BANCA EXAMINADORA

Dr. Douglas Sad Silveira - Orientador

Universidade Federal de Juiz de Fora

Dr. Rafael Morais de Souza - Coorientador

Universidade Federal de Juiz de Fora

Dr. Jairo Francisco de Souza

Universidade Federal de Juiz de Fora

Dr. Bernardo Pinheiro Machado Mueller

Universidade de Brasília

Juiz de Fora, 29/05/2026.

Documento assinado eletronicamente por **Douglas Sad Silveira**, **Professor(a)**, em 01/06/2026, às 08:31, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rafael Morais de Souza**, **Professor(a)**, em 02/06/2026, às 14:45, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bernardo Mueller**, **Usuário Externo**, em 17/06/2026, às 14:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jairo Francisco de Souza**, **Professor(a)**, em 18/06/2026, às 19:15, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

RESUMO

Esta dissertação utiliza a medida Evidence–Minus–Intuition (EMI) para analisar o estilo retórico em pronunciamentos realizados no Senado Federal brasileiro entre 1995 e 2022. O EMI mede a proximidade relativa dos discursos políticos em relação à linguagem orientada por evidências e à linguagem orientada pela intuição. A análise combina word embeddings, mensuração retórica baseada em dicionários, modelagem de tópicos, representações lexicais por TF–IDF e modelos supervisionados de aprendizado de máquina para examinar como o estilo retórico varia ao longo do tempo, entre classes ideológicas e entre macrotemas substantivos. Os resultados mostram uma queda substancial do EMI ao longo do tempo, sugerindo um deslocamento da linguagem orientada por evidências em direção a uma retórica mais orientada pela intuição. No entanto, as distribuições do EMI se sobrepõem substancialmente entre discursos de esquerda, centro e direita. Os exercícios de aprendizado supervisionado mostram que a orientação binária do EMI e os indicadores de tópicos não melhoram a linha de base lexical de TF–IDF no desenho simples de ablação. Especificações estendidas com covariáveis em nível de senador e covariáveis eleitorais melhoram o desempenho em divisões aleatórias da amostra, mas os testes de robustez indicam menor capacidade de generalização. De modo geral, o EMI é melhor interpretado como uma medida complementar de estilo retórico, e não como um classificador independente da ideologia partidária.

Palavras-chave: retórica política; aprendizado de máquina; modelagem de tópicos; ideologia; Senado Federal.

ABSTRACT

This dissertation uses the Evidence–Minus–Intuition (EMI) measure to analyze rhetorical style in speeches delivered in the Brazilian Federal Senate between 1995 and 2022. EMI measures the relative proximity of political speeches to evidence-oriented and intuition-oriented language. The analysis combines word embeddings, dictionary-based rhetorical measurement, topic modeling, TF–IDF lexical representations, and supervised learning models to examine how rhetorical style varies over time, across ideological classes, and across substantive macro-themes. The results show a substantial decline in EMI over time, suggesting a movement away from evidence-oriented language and toward more intuition-oriented rhetoric. However, EMI distributions overlap substantially across left, center, and right-wing speeches. The supervised learning exercises show that binary EMI orientation and topic indicators do not improve the TF–IDF lexical baseline in the simple ablation design. Extended specifications with senator-level and electoral covariates improve random-split performance, but robustness checks indicate weaker generalization. Overall, EMI is best interpreted as a complementary measure of rhetorical style rather than as an independent classifier of party-based ideology.

Keywords: political rhetoric; machine learning; topic modeling; ideology; Federal Senate.

List of Figures

List of Tables

Contents

1 INTRODUCTION

Rhetoric is a tool through which political actors can shape group behavior and influence public opinion to secure support for specific objectives (Downs 1957). Examining this dimension of political discourse is particularly important in a setting marked by volatility and public debate. Despite its historical relevance for political development, rhetoric remains understudied in economics.

Political discourse can adopt different rhetorical strategies, ranging from highly technical, evidence-based arguments to appeals grounded in emotion and intuition. This choice depends on the speaker's goals and the audience they aim to influence (Gennaro and Ash 2022). Recent empirical evidence shows that rhetoric becomes less rooted in reason during periods of crisis and that opposition parties tend to rely more heavily on emotional or intuitive language than the sitting government (Funtowicz 2024; Aroyehun et al. 2025).

In this study, we examine how members of the Brazilian Federal Senate employ rhetoric, focusing on the distinction between evidence-based and intuition-based discourse. We analyze how rhetorical styles respond to political, economic, and institutional changes, such as economic downturns and political scandals, and whether these shifts follow systematic patterns. In line with recent work, we also link rhetorical style to ideological positioning, exploring a relationship that remains relatively unexplored.

To do so, we use word vectors for word representation, allowing us to quantify rhetorical style. Using this method, we construct a measure of evidence versus intuition in speech and employ it in supervised machine-learning models and diagnostic exercises to evaluate how rhetorical style relates to ideological classification and political context.

We focus on Senate speeches because the chamber concentrates much of the country's public debate and is composed of legislators who are typically older and more highly educated. In this environment, parties may use emotional appeals to mobilize identity-based responses or rely on evidence-based arguments to signal technical competence and institutional stability. This variation makes the Senate an ideal setting to study how rhetorical strategies adapt to political incentives and economic conditions.

The analysis of ideological positioning through text and word patterns has a longer tradition (Laver, Benoit, and Garry 2003; Monroe, Colaresi, and Quinn 2008; Lauderdale and Herzog 2016; Rheault and Cochrane 2020). In Brazil, a recent and relevant contribution is Figueredo, Mueller, and Cajueiro (2022), who build a supervised classifier using word frequencies to estimate ideological orientation in Brazilian Senate speeches. Their model captures lexical variation and documents polarization trends, but it does not focus on rhetorical structure or semantic modes of reasoning. Building on these contributions, this dissertation examines whether reasoning modes, such as evidence-

oriented and intuition-oriented language, provide additional information about ideological classification.

To that end, we propose a supervised model that uses rhetorical indicators to predict ideological orientation. Specifically, we incorporate the Evidence-Minus-Intuition (EMI) score developed by Aroyehun et al. (2025), which quantifies the relative presence of rational and intuitive language. We apply this metric to Brazilian Senate speeches from 1995 to 2022 and examine three main questions. First, how does EMI vary over time and around major economic and institutional events? Second, how does rhetorical style vary across ideological classes and substantive macro-themes? Third, does EMI improve the prediction of party-based ideological class when combined with lexical features, topic indicators, and external covariates?

The empirical strategy links rhetorical dynamics to Brazil's political economy agenda. Following the expectation that political actors may rely more heavily on intuition-oriented rhetoric during periods of uncertainty and may use more evidence-oriented language when discussing technical reforms or policy initiatives (Aroyehun et al. 2025), the dissertation evaluates whether this pattern appears in Brazilian Senate speeches. In this sense, the dissertation contributes to political economy and text-as-data research by treating rhetoric as a complementary dimension of legislative behavior: not as a substitute for lexical measures of ideology, but as a way to examine how senators justify, frame, and communicate their positions under changing political and institutional conditions.

The results show us that EMI declines substantially over time, suggesting a movement away from evidence-oriented language and toward more intuition-oriented rhetoric in Senate speeches. This decline is visible in the aggregate series and also appears across selected macro-themes, although economy, development, and institutional issues follow different trajectories. When the analysis is disaggregated by ideology, however, the EMI distributions of left, center, and right-wing speeches overlap substantially, indicating that rhetorical style does not map mechanically onto party-based ideological classes.

The supervised learning exercises confirm this pattern. Binary EMI orientation and topic indicators contain some ideological signal, but they do not improve the TF–IDF lexical baseline in the simple ablation design. Extended specifications that add senator-level and electoral covariates improve random-split performance, but robustness checks by senator and by time show that generalization is more difficult. Overall, EMI is better interpreted as a complementary measure of rhetorical style than as an independent classifier of party-based ideology.

To structure the analysis, the dissertation unfolds across six chapters. Chapter 2 provides a historical overview of Brazil's political and economic context between 1995 and 2022. Chapter 3 details the construction of the dataset and the main variables used in the analysis. Chapter 4 outlines the empirical strategy, including the construction of the EMI

score, topic modeling, and supervised learning tasks. Chapter 5 presents the descriptive and predictive results. Finally, Chapter 6 concludes by summarizing the main findings, discussing limitations, and outlining directions for future research.

2 HISTORICAL BACKGROUND

This chapter provides essential historical context for understanding the key political and economic developments that shaped Brazil from 1995 to 2022. By examining pivotal events across seven legislative periods, we aim to identify the main themes and debates that influenced shifts in ideological positions, political strategies, and rhetorical patterns. These shifts are particularly relevant for analyzing changes in legislative behavior, discourse, and political alignment over time. During this period, Brazil experienced profound transformations, including price stabilization efforts, pension reforms, economic crises, and large-scale political scandals. These factors not only influenced policy but also catalyzed significant changes in political discourse.

During the 1980s, several emerging economies faced pressures related to public debt management, inflation, and the need to gain credibility (Mendonça and Vivian 2008). Brazil followed a similar trajectory, experiencing price instability and successive attempts at stabilization with limited success. The consolidation of a stable economic policy occurred only in the second half of the 1990s. This new phase began in the 50th legislature (1995–1998) with the successful implementation of the *Real Plan* during Fernando Henrique Cardoso's first administration. As Baer (2001) documents, the plan effectively stabilized prices and laid the groundwork for a new macroeconomic framework. The transition was not complete, however, as Brazil remained vulnerable to global financial instability, including the Asian (1997) and Russian (1998) crises, which put pressure on the balance of payments.

In the 51st legislature (1999–2002), in response to these pressures and external shocks, the government adopted a new macroeconomic framework centered on three pillars: a floating exchange-rate regime, inflation targeting, and primary fiscal surpluses, known as the macroeconomic tripod. These reforms aimed to restore investor confidence and prevent further balance-of-payments crises (Fraga, Goldfajn, and Minella 2003; Mendonça and Silva 2008). Despite the introduction of the tripod, new challenges soon emerged as the Argentine crisis raised fears of contagion, with international investors questioning Brazil's fiscal sustainability and macroeconomic stability. Goretti (2005) empirically investigates these developments and evaluates their impact on Brazil's financial credibility. In 2002, the presidential candidate Luiz Inácio Lula da Silva, from the left, gained electoral momentum, and markets reacted with volatility, driven by concerns that Lula might abandon the macroeconomic framework established under Cardoso (Mendonça and Silva 2008). The end of this period was marked by the publication of the "Letter to the Brazilian People" and an International Monetary Fund (IMF) agreement, which helped reduce volatility and enabled a more stable political transition.

Lula's first term (52nd legislature, 2003–2006) preserved the macroeconomic tripod

and expanded social policies, notably *Bolsa Família*, a conditional cash transfer program, along with real increases in minimum wage and targeted social spending. Weisbrot, Johnston, and Lefebvre (2014) document improvements in socioeconomic indicators during this period. In 2005, Brazil repaid its IMF debt early. With a favorable international environment, Brazil benefited from a global commodities boom, driven in part by China's growing demand for raw materials (Gruss 2014). In 2005, however, the *mensalão* scandal revealed payments to legislators in exchange for political support (Flynn 2005; Silva 2014), exposing the limits of coalition presidentialism but not preventing Lula's reelection in 2006.

In the 53rd legislature (2007–2010), the Workers' Party remained in office. Brazil continued to benefit from favorable external conditions. This trajectory was disrupted by the 2008 global financial crisis, triggered by the collapse of Lehman Brothers (Mantoan et al. 2021). In response, the *Programa de Aceleração do Crescimento* (PAC), launched in 2007, became a central axis of economic policy, consistent with the assessment that fiscal policy took on more countercyclical characteristics from 2007 onward (Mendonça and Pinton 2013). With the economy still growing and strong political support, Dilma Rousseff, Lula's successor, was elected in 2010, ensuring continuity.

During the 54th legislature (2011–2014), the government emphasized *Plano Brasil Maior*, which combined sectoral incentives and investment policies. The external environment soon became less favorable, with falling commodity prices and rising public debt. In 2013–2014, inflationary pressures reemerged, forcing monetary tightening and revealing the limits of interventionist policies (Vartanian and Garbe 2019). At the same time, social tension intensified. In June 2013, mass protests erupted across major cities, revealing dissatisfaction with public services and corruption. In 2014, Operation *Lava Jato* (Car Wash) emerged as an investigation into corruption at *Petrobras* and major construction companies. These dynamics shaped the following years.

Dilma Rousseff's second term (55th legislature, 2015–2018) began in recession, with high inflation, rising unemployment, and GDP contraction IMF (2018). Fueled by revelations from Operation *Lava Jato*, Congress impeached Dilma Rousseff in August 2016. The vice president, Michel Temer, assumed the presidency and adopted a program of fiscal adjustment and structural reforms, including a 20-year federal spending cap and a labor reform aimed at increasing labor-market flexibility. Despite a modest recovery, social discontent persisted, as demonstrated by the 2018 truckers' strike. Supported by a critical discourse on the political system, Jair Messias Bolsonaro was elected in the same year. The 55th legislature marked a turning point in Brazilian politics, as long-standing discontent, economic stagnation, and corruption scandals converged to transform the political landscape.

In the 56th legislature (2019–2022), Jair Messias Bolsonaro assumed the presidency

with an agenda that combined liberalizing reforms and conservative positions on social issues. Congress approved a major pension reform in 2019 (Queiroz and Ferreira 2021), but administrative and tax reforms advanced slowly amid shifting coalitions and recurrent tension between the executive, the Supreme Court, and state governors. The Covid-19 pandemic in 2020 produced a sharp recession and prompted a large emergency cash transfer program that expanded the role of federal social assistance and increased budgetary pressures under the spending cap (SARLET 2021). In 2021, vaccination accelerated and activity recovered, while inflation rose with currency depreciation, supply bottlenecks, and commodity shocks, leading to a rapid monetary tightening cycle. The government replaced *Bolsa Família* with *Auxílio Brasil* and pursued measures to ease fuel prices in 2022, including tax changes and extraordinary spending authorizations. Congress approved the partial privatization of Eletrobras, while debates over fiscal rules, Petrobras pricing, and the scope of social transfers intensified. Environmental policy and rising Amazon deforestation drew domestic and international scrutiny (Barbosa, Alves, and Grelle 2021). The period ended with a highly polarized 2022 election, closing a legislature marked by health and economic shocks, institutional friction, and contested policy priorities.

Against this backdrop, our dataset covers the entire period under review and records major debates, sudden shifts, and episodes of economic and political crisis. This coverage allows us to identify which events most strongly shaped parliamentary rhetoric and to compare the influence of political shocks with that of technical or economic agendas. The next section describes the dataset and summarizes key differences across legislatures.

3 DATA

3.1 Data base

This chapter describes the data sources, the compilation process and the main variables used in the analysis. We study speeches delivered in the Brazilian Federal Senate. We focus on the Senate because senators speak more frequently than members of the Chamber of Deputies and tend to have longer careers and broader legislative responsibilities, which helps identify rhetorical patterns over time. The dataset combines two sources. First, we incorporate the corpus compiled by Figueredo, Mueller, and Cajueiro (2022). Second, we extend coverage by scraping the XML records available on the Federal Senate website, which includes the most recent legislature. The final dataset spans the 50th to the 56th legislatures, covering 1995–2022. This time frame corresponds to the earliest period for which the API provides responses. Moreover, it encompasses years marked by major political decisions and events of significant importance for Brazilian politics and democracy. Table 1 reports the number of speeches by legislature. For each speech, we observe the date, the speaker's name, the state, the speaker's party affiliation on the speech date, and an ideological label (left, center, or right).

Table 1 – Speeches by legislature

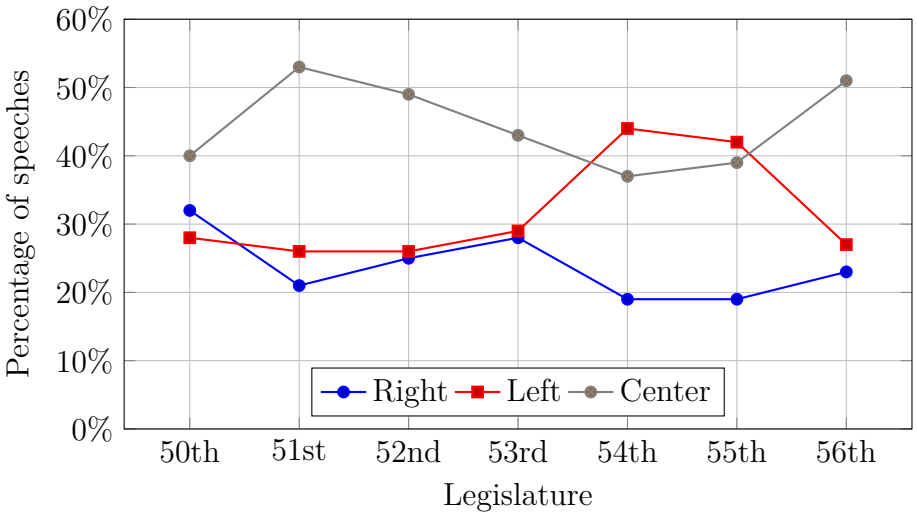
Legislature	Period of the legislature	Nº speeches
50th	1995–1999	10,022
51st	1999–2002	9,578
52nd	2003–2006	18,881
53rd	2007–2010	17,294
54th	2011–2014	17,237
55th	2015–2018	12,328
56th	2019–2022	12,268

Source: authors' calculations.

We follow the ideological classification in Figueredo, Mueller, and Cajueiro (2022), which builds on Power and Zucco Jr (2009) and the literature frequently uses this method in studies of legislative behavior. The approach combines roll-call votes and expert or survey assessments to position Brazilian parties along a left–center–right scale. We assign an ideological label to each speech based on the speaker's party at the time of delivery. Figure 1 summarizes the distribution of speeches by ideology across legislatures. Two patterns stand out. First, the political center accounts for the largest share of speeches in most legislatures. Second, the left's share rises notably in the 54th and 55th legislatures and then falls in the 56th, while the center regains predominance and the right remains below the center in all periods. These are shares of speeches rather than seat shares, and

they should be read as descriptive patterns rather than causal statements about specific events.

Figure 1– Share of speeches by ideology and legislature



Source: authors' calculations.

Table 2 consolidates the party–ideology mapping. After harmonizing renamings, mergers, and splits over time, the sample includes nine parties in the left group (CIDADANIA, PCdoB, PDT, PPS, PSB, PSOL, PT, PV, REDE), nine in the center group (AVANTE, MDB, PMDB, PMN, PODE, Podemos, PROS, PSD, PSDB), and eighteen in the right group (DEM, PDC, PDS, Patriota, PFL, PL, PP, PPB, PR, PRB, PRTB, PSC, PSL, PTB, PTC, REPUBLICANOS, Uniao). Some labels track the same lineage at different points in time (for example, PMDB and MDB; PFL and DEM; PPB and PP; and the formation of Uniao from DEM and PSL).

Table 2 – Party–ideology mapping (Brazil)

Left	**Center**	**Right**
CIDADANIA, PCdoB, PDT, PPS, PSB, PSOL, PT, PV, REDE	AVANTE, MDB, PMDB, PMN, PODE, Podemos, PROS, PSD, PSDB	DEM, PDC, PDS, Patriota, PFL, PL, PP, PPB, PR, PRB, PRTB, PSC, PSL, PTB, PTC, REPUBLICANOS, Uniao

Source: authors' calculations.

In broad terms, left parties emphasize redistribution, social policy, and labor protections; centrist parties occupy a pivotal position with programmatic heterogeneity and a stronger focus on coalition management; and right parties more often advocate market liberalization, fiscal restraint, and greater emphasis on law and order. While these features are commonly associated with ideology in the literature, the concept itself remains contested and difficult to quantify. In this study, we adopt a transparent operational rule:

following Figueredo, Mueller, and Cajueiro (2022), each speech inherits the ideological label of the speaker's party at the time of delivery. This convention provides a consistent baseline for classification while allowing for within-party and temporal variation that the empirical analysis can explore.

In addition to this ideological classification, the database incorporates information obtained from external sources and more metrics derived from the text itself. The empirical analysis considers electoral, biographical, rhetorical, and thematic dimensions.

The first group of additional information comes from the Brazilian Electoral Court (TSE). That includes education, occupation, declared wealth, nominal votes, and the vote share in the state. Since this information is not originally available in the Senate speech records database, we incorporate it through a matching procedure between the Senate data and the TSE candidate data. This linkage uses the election year associated with the mandate, the state, and harmonized name information. After this procedure, we merge the TSE information into the speech-level dataset, so that each speech can be analyzed together with electoral and personal attributes of the legislator.

The analysis uses some measures in raw form, and transforms others to improve comparability and interpretations. Age at the time of speech combines the speaker's date of birth with the speech date. Education can be simplified in some specifications, for example, through an indicator of complete higher education or not. Declared wealth is not used in raw monetary terms, since comparisons over a long period may be affected by inflation, currency changes, and extreme values. For this reason part of the empirical analysis also uses transformed versions of this measure, such as indicators for whether the speaker belongs to the upper part of the wealth distribution in the corresponding electoral year. In some cases, missing-values indicators are also included since the availability of the external covariates is not uniform throughout the sample. Table 3 summarizes the main information blocks used in the analytical dataset.

The database also includes measures constructed with speeches from the Senate API. One of them is the Evidence-Minus-Intuition (EMI) score, which captures whether a speech is closer to evidence-based or intuition-based language. This measure is calculated from the textual content and is explained in more detail in the next section. The data set also includes measures derived from topic modeling. Together with lexical representations such as TF-IDF, they form the textual dimension of the empirical design.

Thus, the final analytical base combines different layers of information. The first is the legislative layer, composed of the metrics derived from speech texts and its metadata. The second is the political legislative layer, composed of the ideological labels and characteristics obtained from TSE records.

Table 3 – Main information blocks in the analytical dataset

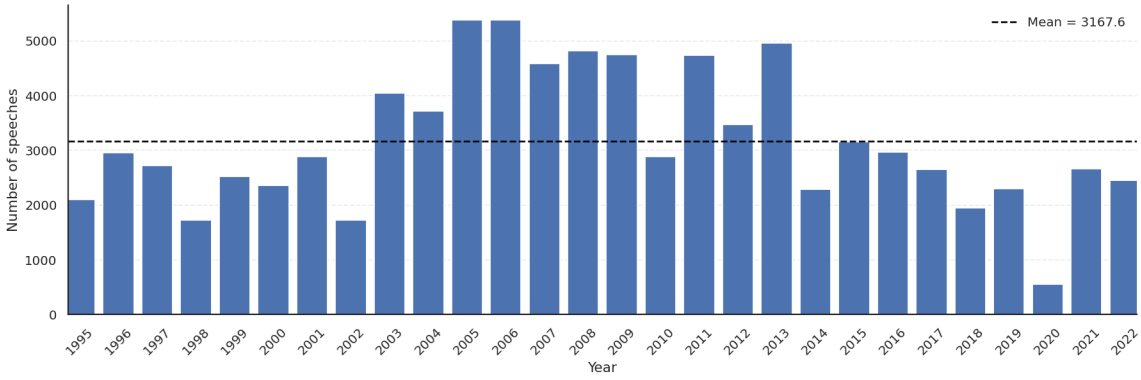
Information block	Main variables	Source and role in the analysis
Speech metadata	Date, legislature, senator, party, state, speech text	Senate speech records. Defines the speech-level unit of analysis and allows aggregation by year, legislature, party, and senator.
Ideological classification	Left, center, right	Party-based label assigned from the senator's party at the time of the speech. Used as the target in the ideological classification task.
Electoral covariates	Nominal votes, state vote share, election cycle	TSE candidate and electoral records. Used to incorporate electoral performance and mandate context.
Biographical covariates	Age, gender, education, occupation, declared wealth, region	Senate and TSE records. Used to incorporate senator-level characteristics into the extended specifications.
Missingness indicators	Missing education, occupation, wealth, and vote-share indicators	Constructed to retain observations with incomplete external covariates and control for data availability.
Rhetorical indicators	Continuous EMI score, binary EMI-orientation indicator	Constructed from the speech text. Used to measure evidence-oriented versus intuition-oriented rhetoric.
Thematic indicators	BERTopic topic labels and macro-theme dummies	Constructed from topic modeling. Used to represent the dominant theme of each speech.
Lexical representation	TF–IDF unigrams and bigrams	Constructed from the cleaned speech text. Used as the main lexical baseline in the supervised models.

Source: Author's elaboration.

3.2 Corpus profile and rhetorical context

We start by examining the number of speeches in the sample by year. The series shows substantial variation over time and suggests three broad periods as shown in Figure 2. The first, from 1995 to 2002, is characterized by modest annual totals, generally below the sample mean. The second, from 2003 to 2013, marks a clear expansion in speech volume. The third begins after 2014, when the total number declines and becomes more uneven, culminating in a sharp collapse in 2020 and only partial recovery in 2021 and 2022.

Figure 2– Number of Senate speeches by year



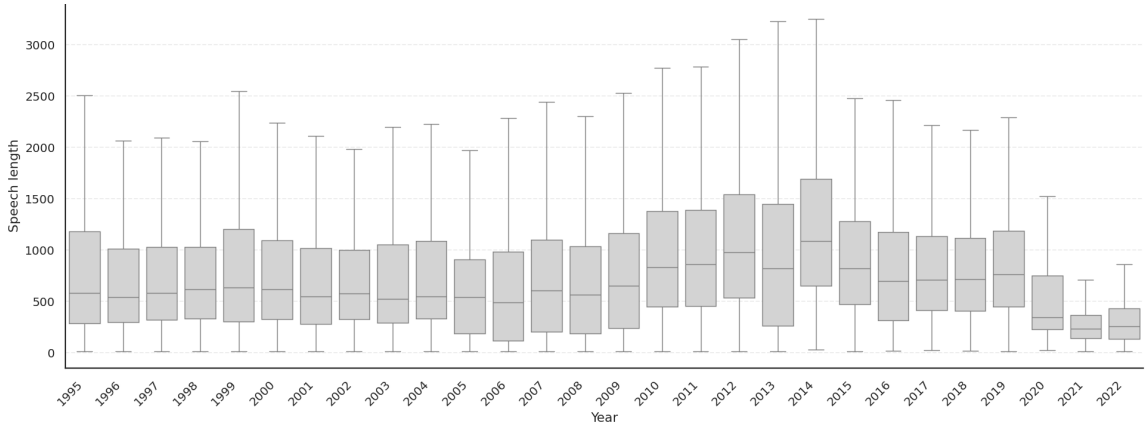
Source: Author's elaboration based on the corpus of Senate speeches.

The most abrupt discontinuity appears in 2020. The number of speeches falls to the lowest point in the entire series, far below both the preceding years and the sample mean. This break is consistent with the disruption associated with the pandemic period and indicates that the temporal distribution of speeches is not a neutral background variation. It is part of the institutional setting in which rhetorical choices are made.

A second relevant dimension is the speech length. Figure 3 therefore reports the distribution of speech length by year, making it possible to examine central tendency and dispersion. The figure indicates that speech length is comparatively stable from the mid-1990s to the late 2000s, with median values concentrated at moderate levels and only limited year-to-year fluctuation. Beginning around 2010, however, the distribution shifts upward. Between 2010 and 2015, speeches become longer on average and also more dispersed, as shown by the higher medians, wider inter-quartile ranges, and longer upper tails. This suggests not only an increase in typical speech length, but also greater heterogeneity in the size of interventions during this period.

After 2015, the distribution gradually moves downward, although speech length remains above the levels observed in much of the earlier period. A second major break occurs from 2020 onward. In 2020, 2021, and 2022, the distribution compresses sharply: medians fall, the interquartile range narrows, and the upper tail becomes much shorter.

Figure 3– Distribution of speech length by year



Source: Author's elaboration based on the corpus of Senate speeches.

Relative to the preceding decade, speeches in these years are substantially shorter and less dispersed.

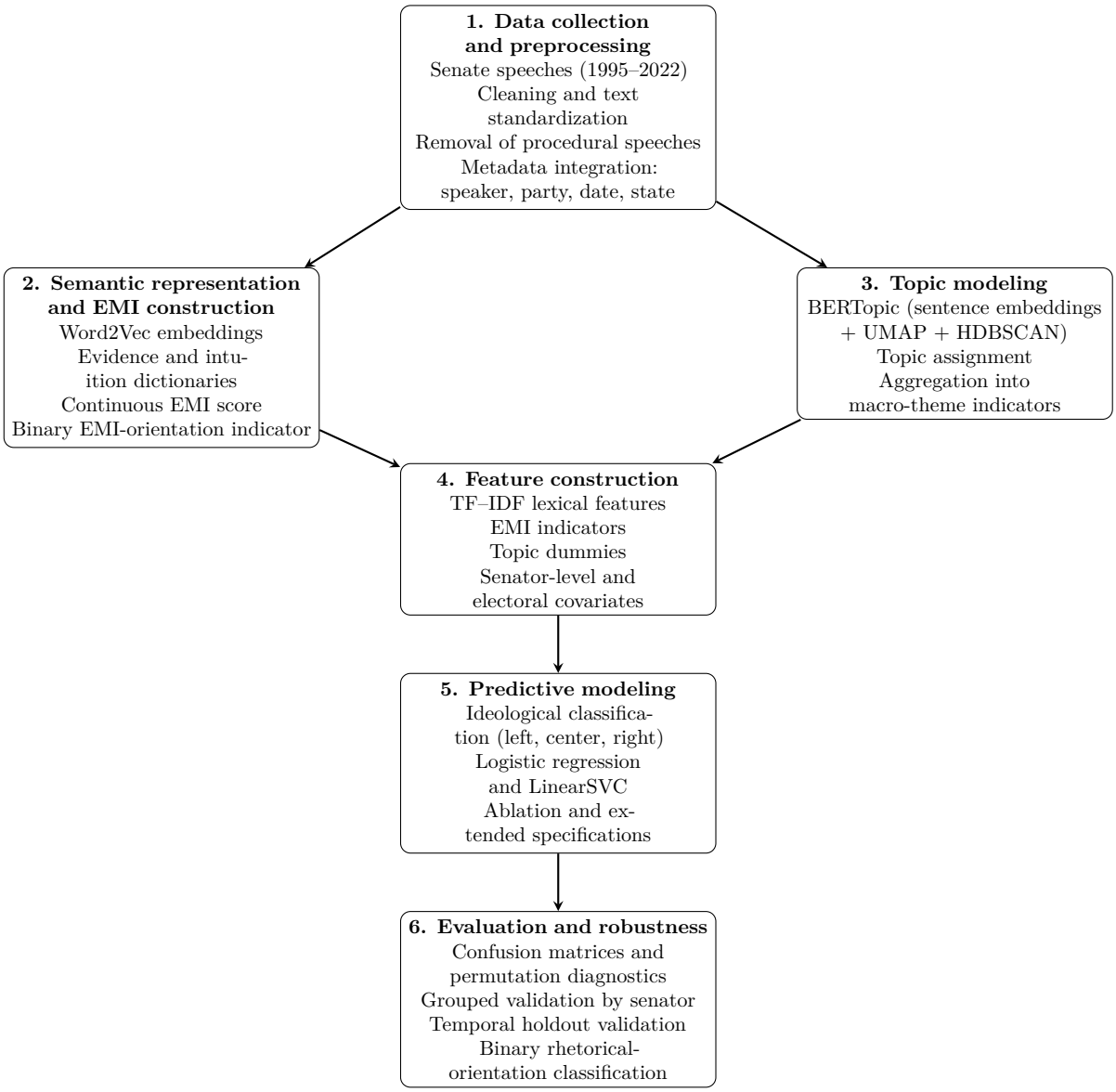
Taken together, Figures 2 and 3 show that the temporal profile of the corpus changes along two dimensions at once: frequency and length. The period of higher speech volume from 2003 to 2013 is followed by a phase of contraction, while the boxplots indicate that speech length rises most clearly in the early-to-mid 2010s before falling sharply after 2020. These shifts are important because they define the empirical context in which the evolution of evidence-based and intuition-based rhetoric requires interpretation. They are also relevant methodologically, since the EMI measure used in the following sections is normalize by speech length to reduce the extent to which longer interventions mechanically produce higher evidence or intuition scores.

4 METHODOLOGY

This chapter presents the empirical pipeline and explains how the data are used to construct the rhetorical measures, topic indicators, and predictive models.

As Figure 4 illustrates, the workflow proceeds from data collection and cleaning through two analytical branches. One branch constructs lexical and rhetorical measures, especially the EMI score. The other recovers thematic structure through topic modeling. These outputs are then merged back into the same analytical base, where they are analyzed together with electoral and biographical information such as education, occupation, wealth, and age.

Figure 4– Methodological workflow



Source: Author's elaboration.

4.1 Analytical strategy and workflow

First, we collect Senate speeches and metadata by web scraping, remove procedural interventions, and standardize the text through tokenization and cleaning (Section 4.2). Pre-processing varies according to the representation required by each method. This is necessary because lexical models, rhetorical measures based on embeddings, and topic models do not require exactly the same textual treatment. The following steps construct the main textual measures, including TF-IDF lexical representation, the Evidence-Minus-Intuition (EMI), and thematic variables derived from topic modeling. Finally, we merge these measures with the external covariates of the database in order to form the final analytical base used in the supervised exercises.

From these features, the empirical analysis follows two main supervised tasks. The first is the prediction of ideological class, based on the left-center-right classification assigned to each speech according to the speaker's party on the speech date. The second is the prediction of rhetorical orientation, distinguishing speeches that are closer to evidence-based language from those that are closer to intuition-based speeches. In both cases, the goal is to compare how different groups of variables contribute to the classification of political and rhetorical patterns.

These structures allow us to compare lexical, rhetorical, thematic, electoral, and biographical information within the same empirical framework. It makes it possible to evaluate whether the inclusion of measures such as the EMI Score, topic indicators, and external covariates improves prediction beyond the information contained in the speech text alone.

4.2 Preprocessing

We begin by removing speeches that are purely procedural. These records announce votes, request order, or describe internal rules, and therefore do not convey substantive positions (Beauchamp 2011). The dataset flags the type of intervention, which allows clean exclusion. We then merge all legislatures into a single corpus for representation learning while preserving document-level metadata for later analysis.

The general standardization steps, such as lowercasing, Unicode normalization, and the removal of duplicated or malformed segments, are common across the textual pipeline. After that, the preprocessing changes according to the objective of each method. In the lexical representation, the main concern is to preserve discriminative terms while removing tokens with little analytical value. In the embedding-based representation, the concern is to preserve semantic context so that the relationship between words is not artificially distorted. In the topic-modeling stage, the emphasis is on reducing elements that may generate clusters based on speaker identity. The following subsections describe

each preprocessing strategy.

4.2.1 Preprocessing for TF-IDF Feature

We filter stopwords using lists from the `NLTK` library (Bird, Klein, and Loper 2009)[1] and extend them with domain-specific tokens that add little semantic value for our purposes. The extended list includes senator names, state names, honorifics (for example, *Sr*, *Sra*), salutations, and common formulaic phrases in floor addresses. We also tokenize the text into words using a Unicode-aware regular expression that captures alphabetic tokens excluding digits and punctuation, following Figueredo, Mueller, and Cajueiro (2022). This is necessary because the objective is to identify which words best distinguish each group and use this information to predict the political group to which a given speech belongs, a common technique in the literature.

The resulting tokens are then used to construct unigram and bigram TF-IDF representations of the speeches. In this context, the lexical preprocessing emphasizes surface differences in word usage, so that the model captures variation in vocabulary associated with ideological and political patterns.

4.2.2 Preprocessing for Word2Vec Feature

To create the data used to calculate the EMI Score, the initial procedure was similar; however, here we did not use the expanded list of stop-words because it is important to preserve the semantic relations between words so that the model can learn how they co-occur in context.

The tokenization process continued to follow the same pattern of the TF-IDF feature. In addition, we segment speeches into sentences before training the Word2Vec model. This ensures that the context window operates within more coherent textual units, reducing the chance that the model learns artificial associations between words that appear far apart in the speech but do not belong to the same semantic context. However, after estimating the embeddings, we calculate the EMI score itself at the speech level. This preserves the speech as the main unit of analysis and allows direct comparison with the other empirical representations used.

4.2.3 Cleaning data to topic modeling

For the topic modeling, the main concern is to improve the coherence of the clusters and reduce the influence of elements that do not reflect substantive content; therefore, it is necessary to remove the senator names and Brazilian states, as some clusters formed around this information rather than the text content itself. Furthermore, the normalization process remained the same as previously presented. This step is important because topic

[1] We use the Portuguese language resources in the libraries.

modeling aims to recover substantive agendas of debate rather than separate speeches according to speaker identity or formulaic conventions of the legislative environment. By reducing these recurrent elements, the clustering procedure becomes more likely to group speeches according to policy themes, institutional issues, or broader political topics, which is consistent with the goal of producing interpretable topics in BERTopic (Grootendorst 2022).

Given the differences among the three databases that compose this work, we can proceed to the presentation of our main models: Word2Vec for the construction of the EMI Score and BERTopic for topic modeling.

4.3 Word Embeddings

We encode lexical meaning with continuous vectors whose geometric proximity reflects similarity in use (Harris 1954). This follows the distributional hypothesis that words appearing in similar contexts tend to share meaning. Let $(w_1, \ldots, w_T)$ denote the token sequence after preprocessing, and let L be the context-window size. For a position t, define the symmetric context set

$$C_t \;=\; \{\, w_{t-L}, \ldots, w_{t-1},\; w_{t+1}, \ldots, w_{t+L} \,\}.$$

The prediction task either infers the center token w_t from its neighbors in C_t or infers each neighbor from w_t.

Based on the logic above, we use the embedding model. However, before formally explaining its construction, it is necessary to provide an intuitive demonstration of how its geometric logic works: In this model, each word appears as a point in a continuous vector space. As a result, words that appear in similar contexts are located near each other, carrying semantic value. This becomes evident in the classic example used by Mikolov, Yih, and Zweig (2013):

$$king - man + woman \approx queen$$

We observe that the dimensions of gender and hierarchy are preserved and aligned in the same vector direction. When subtracting *man* from *king* and adding *woman*, the result is *queen*. This pattern does not arise by imposition but emerges naturally from the co-occurrence of words in a large text corpus. The formalization below explains how this type of relation emerges.

We use Word2Vec, from Mikolov, Yih, and Zweig (2013) and Mikolov, Chen, et al. (2013) to learn target and context embeddings. For each vocabulary term $v \in \mathcal{V}$, the model assigns a target vector $\rho_v \in \mathbb{R}^K$ and a context vector $\alpha_v \in \mathbb{R}^K$, where K is the

embedding dimension. The model learns word–type vectors and applies them to all token occurrences in the corpus. Let the mean context at position (d, n) be [2]

$$\bar{\alpha}_{d,n} \;=\; \frac{1}{|C(w_{d,n})|} \sum_{u \in C(w_{d,n})} \alpha_u.$$

In the CBOW formulation, the probability that the center equals v given its context is

$$\Pr\Big[w_{d,n} = v \mid C(w_{d,n})\Big] \;=\; \frac{\exp\Big(\bar{\alpha}_{d,n}^{\top} \rho_v\Big)}{\sum_{v' \in \mathcal{V}} \exp\Big(\bar{\alpha}_{d,n}^{\top} \rho_{v'}\Big)}.$$

In the Skip-gram formulation, the model predicts each context token $u \in C(w_{d,n})$ from the center $w_{d,n} = v$,

$$\Pr\Big[u \mid w_{d,n} = v\Big] \;=\; \frac{\exp\Big(\alpha_u^{\top} \rho_v\Big)}{\sum_{u' \in \mathcal{V}} \exp\Big(\alpha_{u'}^{\top} \rho_v\Big)}.$$

Training adjusts (ρ, α) to maximize the corresponding log-likelihood over the corpus. Because the exact softmax is expensive for large vocabularies, we use negative sampling, which replaces the softmax with a logistic objective that contrasts observed word–context pairs with noise samples (Mikolov, Sutskever, et al. 2013; Levy and Goldberg 2014).

After estimation, the fitted target vectors $\{\rho_v\}$ serve as word embeddings for each vocabulary term. For each passage, we average the embeddings of its content words to form a document vector that summarizes its local semantics (Arora, Liang, and Ma 2017; Kozlowski, Taddy, and Evans 2019; Aroyehun et al. 2025). We measure semantic proximity using cosine similarity, defined as

$$\cos(\mathbf{p}, \mathbf{c}) \;=\; \frac{\mathbf{p}^{\top} \mathbf{c}}{\|\mathbf{p}\| \, \|\mathbf{c}\|}.$$

In Section 4.4 we compare these passage vectors with concept vectors for evidence and intuition to construct the Evidence–Minus–Intuition score. This representation enables comparisons across time and topics, even when speakers vary their lexical choices, which is central for the rhetorical analysis that follows. That averaging preserves interpretability and aligns directly with the dictionary-based EMI construction, while transformer embeddings are introduced later for unsupervised topic discovery.

4.4 Construction of the EMI Score

We measure rhetorical style with the Evidence Minus Intuition (EMI) score. The construction follows four steps. First, we define seed dictionaries that capture linguistic signals of evidence-based and intuition-based language. We draw these seeds from validated lists in Aroyehun et al. (2025). The evidence list includes terms such as "evidence," "data,"

[2] v denotes the target (center) word, whereas u denotes one of its context words.

and "examine," and the intuition list includes "believe," "feel," and "opinion." Because our corpus is in Brazilian Portuguese, we translate the seed terms and consolidate accents and morphological variants[3] to ensure consistent coverage.

Second, we expand the dictionaries using cosine similarity in a FastText (Bojanowski et al. 2017) pre-trained Portuguese embedding space. We form a composite semantic vector for each rhetorical dimension (Evidence and Intuition) by averaging the FastText embeddings of the corresponding seed words. To ensure that the selection of words for each axis is robust enough for the analysis, we use the original dictionary of the paper by Aroyehun et al. (2025) as a comparison benchmark. This composite vector summarizes each concept in the embedding space. We then compute the cosine similarity between every vocabulary term and each composite vector and rank the terms by their proximity to the evidence and intuition concepts. For each concept, we examine the top twenty words with the highest cosine similarity scores and select those that align with the core rhetorical meaning of each vector. The expanded dictionaries are reported in Appendix .

Third, we represent both rhetorical dimensions and speeches in the same embedding space. We obtain the evidence and intuition concept vectors by averaging the embeddings of the words in each dictionary. The representation of each speech is constructed in the same way, by averaging the embeddings of its preprocessed terms. Let v_s denote the vector representation of speech s, and let c_E and c_I denote the concept vectors for evidence and intuition. We first compute the raw evidence and intuition scores as cosine similarities:

$$e(s) \;=\; \cos\Big(v_s, c_E\Big), \qquad i(s) \;=\; \cos\Big(v_s, c_I\Big).$$

Because speech length may mechanically affect the distribution of these scores, especially in the presence of very short or very long interventions, we adjust both components before calculating the final EMI. Speeches are grouped into quantile-based bins according to their length. Let $b(s)$ denote the length bin of speech s, and let $\bar{e}_{b(s)}$ and $\bar{i}_{b(s)}$ denote the mean evidence and intuition scores within that bin. The adjusted components are defined as

$$e_{\text{adj}}(s) \;=\; e(s) - \bar{e}_{b(s)}, \qquad i_{\text{adj}}(s) \;=\; i(s) - \bar{i}_{b(s)}.$$

We then standardize both adjusted components using a z-score transformation:

$$z_e(s) \;=\; \frac{e_{\text{adj}}(s) - \mu_e}{\sigma_e}, \qquad z_i(s) \;=\; \frac{i_{\text{adj}}(s) - \mu_i}{\sigma_i},$$

where μ_e and σ_e are the mean and standard deviation of the adjusted evidence component in the sample, and μ_i and σ_i are the corresponding quantities for the adjusted intuition component.

The final EMI score is then defined as

$$\text{EMI}(s) \;=\; z_e(s) - z_i(s).$$

[3] For example, *evidência*/*evidências*.

Positive values indicate that the speech is closer to evidence-based language, while negative values indicate greater proximity to intuition-based language.

Finally, we assess the robustness of the measure by comparing alternative dictionary constructions, including the translated version of the original dictionary proposed by Aroyehun et al. (2025). We incorporate the final EMI score into the analytical database as a continuous and binary measure of rhetorical style. This allows the dissertation to use it both as a descriptive indicator of parliamentary rhetoric and as a variable in the supervised exercises presented below.

4.5 BERTopic

Research on political and parliamentary rhetoric shows that rhetorical style varies depending on the issue under debate (Gennaro and Ash 2022; Funtowicz 2024). To study this variation, we need topic modeling methods that recover coherent themes and organize speeches accordingly. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) offers a conventional baseline by treating each document as a mixture of topics and each topic as a distribution over words. Two features limit its usefulness in our setting. The analyst must choose the number of topics in advance even though the thematic space changes across legislatures, and the model assumes independence across topics even though parliamentary interventions often blend multiple agendas within the same speech. To address these issues, we use BERTopic (Grootendorst 2022), which couples transformer-based embeddings with class-based TF–IDF (cTF–IDF).

Conceptually, BERTopic extends the logic of word embeddings to larger textual units. While word embeddings map single words into a semantic space, transformers generate embedding for entire sentences. They capture how each word's meaning shifts with context, allowing the model to recognize that "raising import taxes" and "Increasing tariffs" describe similar policy actions. Rather than counting words, the method represents each passage as a dense vector summarizing its contextual meaning and groups passages that occupy nearby regions in the semantic space. This enables the discovery of coherent themes even when speaker use different vocabularies to discuss related agendas.

Transformers produce contextualized representations that encode lexical content together with local context; Lucchetti and Cajueiro (2025) provides an accessible overview of how these models capture semantic information. The procedure unfolds in four steps.

First, we compute sentence embeddings with a Portuguese model using batched inference.[4] These vectors summarize each passage in a dense representation that reflects nearby words and their relationships. Second, we reduce dimensionality with UMAP so that the representation preserves neighborhood structure while remaining suitable for clustering. Third, we cluster the reduced embeddings with HDBSCAN. This algorithm

[4] Additional details available at: `serafim-335m-portuguese-pt-sentence-encoder`.

does not require a prechosen number of clusters and assigns outliers to a control cluster when they lack a dense neighborhood. Fourth, we apply cTF-IDF within each cluster using a Portuguese vectorizer and then assign human-readable labels from the top terms and prototypical documents. This combination leverages rich semantic representations and a weighting scheme that highlights distinctive vocabulary, which helps identify policy agendas without fixing the number of topics in advance.

Applying BERTopic typically yields many granular clusters. For this reason, the resulting topics are inspected and then organized into broader substantive groupings. This improves interpretability and facilitates their incorporation into the empirical analysis. After this stage, the thematic information is merged back into the analytical database at the speech level. In the final dataset, it is represented by the main topic assigned to each speech, its substantive label, broader macro-theme classifications, and dummy variables derived from these macro-themes.

BERTopic also brings practical considerations. Transformer embeddings increase computational cost in large corpora, and clustering quality depends on sensible choices for dimensionality reduction and density parameters. We mitigate these issues by running embeddings in batches and validating clusters through inspection of top terms, representative documents, and temporal coherence. In the empirical implementation adopted here, topic modeling is applied at the level of the full speech rather than at the level of smaller textual segments. As a result, the thematic output is designed to recover the dominant topic of each speech. This choice improves interpretability and facilitates the incorporation of thematic variables into the empirical analysis. At the same time, it imposes an important limitation. Since topics are assigned from the speech as a whole, the model may not fully capture within-speech thematic variation, especially in long speeches that move across different issues. In such cases, the assigned topic should be interpreted as the dominant topic of the document rather than as an exhaustive representation of all themes contained in it. This also helps explain why some speeches remain in outlier or residual categories when the full document is not concentrated strongly enough around a single substantive topic. A possible improvement for future work would be to estimate topics from smaller textual segments and then aggregate them back to the speech level.

4.6 Supervised learning tasks

This section evaluates whether lexical, rhetorical, thematic, electoral, and biographical information improve the prediction of political and rhetorical patterns in Senate speeches. The empirical strategy starts with transparent supervised benchmarks and then compares how classification changes as additional layers of information are introduced. Table 4 summarizes the supervised stage built on the consolidated analytical database and combines the main textual representations with the external covariates described in

Chapter 3.

The empirical analysis follows two main supervised tasks. The first is the prediction of ideological class. In this case, the target variable takes three categories — left, center, and right — following the party-based classification assigned to each speech according to the speaker's party on the speech date. The second is the prediction of rhetorical orientation. In this case, the target variable is a binary indicator derived from the sign of the EMI score. Let $R(s)$ denote the rhetorical-orientation label of speech

$$R(s) = \begin{cases} 1, & \text{if EMI}(s) > 0, \\ 0, & \text{if EMI}(s) \leq 0. \end{cases}$$

so that speeches with positive EMI receive value 1 and the indicator classifies them as closer to evidence-based language, whereas speeches with $\text{EMI}(s) \leq 0$ receive value 0 and are classified as closer to intuition-based language.

Feature construction blends complementary signals, and the main groups of predictors are summarized in Table 4. First, a lexical baseline uses TF–IDF to represent each passage as a sparse vector of unigrams and bigrams, following Figueredo, Mueller, and Cajueiro (2022). TF–IDF increases the weight of terms that occur frequently within a passage but remain relatively uncommon in the corpus (Gentzkow, Kelly, and Taddy 2019). This representation anchors the lexical benchmark for all subsequent comparisons. Second, the Evidence–Minus–Intuition (EMI) score adds a compact measure of rhetorical style. EMI summarizes the extent to which a passage aligns with evidence-oriented or intuition-oriented language in the embedding space. Because EMI captures stylistic choices that cut across topics and vocabulary, it can add signal where purely lexical features struggle to separate ideologically distinct but lexically similar texts. Third, the thematic structure enters through topic indicators from BERTopic.

In addition, the enriched analytical base allows the models to incorporate external speaker-level and electoral covariates. These include variables such as education, occupation, wealth, age, votes, and vote share. The inclusion of education, occupation, and wealth is also consistent with the literature on the economic backgrounds of politicians, which argues that legislators from different economic backgrounds may think and behave differently in office and therefore may carry politically relevant information beyond the text of the speech itself (Carnes and Lupu 2023).

The supervised models integrate these representations in a straightforward way. TF–IDF provide the lexical baseline, BERTopic contributes thematic variables, EMI enters as a predictor of rhetorical tone when ideology is the target, and external covariates extend the analysis beyond the text itself. Comparing models that include or exclude these groups of variables reveals whether rhetorical, thematic, and speaker-level information add predictive value beyond lexical frequencies alone.

Table 4 – Text Representations and Covariates Used in Supervised Models

Representation	Description	Usage in Prediction
TF–IDF	Lexical weighting of unigrams and bigrams; reflects surface word use and stylistic variation.	Baseline lexical features in both supervised tasks.
EMI	Rhetorical measure that summarizes the relative alignment of a speech with evidence-based versus intuition-based language.	Predictor in ideological classification; used to define the binary outcome in rhetorical-orientation prediction.
BERTopic	Topic modeling output represented through the main topic of the speech, macro-theme classifications, and macro-theme dummies.	Thematic predictors in both supervised tasks.
External covariates	Electoral and biographical variables such as education, occupation, wealth, age, votes, and vote share.	Additional speaker-level and electoral predictors in both supervised tasks.
Ideological label	Party-based left–center–right classification assigned to each speech according to the speaker's party on the speech date.	Target in ideological classification; optional predictor in rhetorical-orientation prediction.

Source: authors' calculations.

4.6.1 Logistic regression

The main benchmark is logistic regression. In the ideological task, the specification is multinomial, since the target has three categories — left, center, and right. In the rhetorical task, the corresponding specification is binary logistic regression. Logistic regression provides a transparent benchmark and allows direct comparison across progressively richer sets of predictors (Izenman 2008; Hosmer Jr, Lemeshow, and Sturdivant 2013; James et al. 2013). This is particularly useful in the present context, since the empirical objective is not only to classify speeches, but also to compare the contribution of lexical, rhetorical, thematic, electoral, and biographical information.

In the ideological task, let $y \in \{1, \ldots, J\}$ denote the class label, with $J = 3$, and let $x \in \mathbb{R}^p$ collect the predictors included in a given specification. For each class j, the linear predictor is

$$\eta_j = \beta_{j0} + x^\top \beta_j,$$

and the corresponding class probability is given by the softmax function:

$$\Pr(y = j \mid x) = \frac{\exp(\eta_j)}{\sum_{k=1}^{J} \exp(\eta_k)}.$$

This specification makes it possible to estimate the probability that a given speech belongs to each ideological category conditional on its lexical, thematic, rhetorical, and speaker-level features.

In the rhetorical-orientation task, the outcome is binary and follows the sign of the EMI score, as described above. The corresponding logistic specification is

$$\Pr(R = 1 \mid x) = \frac{\exp(\alpha + x^\top \beta)}{1 + \exp(\alpha + x^\top \beta)}.$$

This formulation estimates the probability that a speech is closer to evidence-based language rather than intuition-based language, conditional on the predictors included in the model.

High-dimensional textual features motivate the use of regularization, since TF–IDF yields a sparse and potentially very large representation of the speeches. For this reason, the logistic benchmark serves as a stable and interpretable baseline for nested comparisons across alternative feature sets. In practical terms, the main empirical contrasts compare simpler specifications based only on lexical information with richer specifications that add thematic variables, the EMI score, and the external covariates of the database.

In the ideological task, these comparisons make it possible to assess whether rhetorical style, thematic structure, and speaker-level characteristics improve prediction beyond the lexical baseline. In the rhetorical task, the same logic evaluates whether lexical, thematic, ideological, electoral, and biographical information help predict rhetorical orientation once the rhetorical task excludes the EMI score itself from the predictor set.

The analysis excludes variables that trivially reveal the target whenever necessary in order to prevent label leakage. In ideological tasks, we exclude party identifiers. In the rhetorical task, as in the previous model, the rhetorical task excludes the EMI score itself from the predictor set because it is used to define the binary outcome. This keeps the comparison between feature groups substantively meaningful and methodologically coherent.

Together, these logistic specifications provide an interpretable benchmark for both supervised tasks. They allow us to compare predictive performance across alternative feature sets while preserving a direct link between model structure and substantive interpretation.

4.6.2 Linear support vector classification

The supervised exercises also use linear support vector classification (LinearSVC). Support vector methods are widely used in high-dimensional classification settings because they perform well when the predictor matrix is sparse (James et al. 2013). This is especially relevant when speeches are represented through TF–IDF features.

The intuition of the method is to find a separating hyperplane that maximizes the margin between classes. In the binary case, let $x_i \in \mathbb{R}^p$ denote the predictor vector for observation i, and let $y_i \in \{-1, +1\}$ denote its class label. The linear support vector classifier is based on the decision function

$$f(x_i) = \beta_0 + x_i^\top \beta,$$

and assigns the predicted class according to the sign of $f(x_i)$. Estimation seeks a hyperplane that separates the classes while maximizing the margin, allowing for classification errors through a regularization parameter. In its soft-margin form, the optimization problem can be written as

$$\min_{\beta_0, \beta, \xi_i} \quad \frac{1}{2}\|\beta\|^2 + C \sum_{i=1}^{N} \xi_i$$

subject to

$$y_i(\beta_0 + x_i^\top \beta) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad i = 1, \ldots, N,$$

where the slack variables ξ_i allow some observations to fall inside the margin or to be misclassified, and $C > 0$ governs the trade-off between maximizing the margin and penalizing classification errors.

In the ideological task, the outcome has three categories rather than two. In this case, LinearSVC is implemented through a standard multiclass extension, so that the classifier learns linear decision boundaries for the left, center, and right categories. In the rhetorical-orientation task, the outcome is binary, and the same linear decision rule is applied to distinguish speeches that are closer to evidence-based language from those that are closer to intuition-based language.

The role of LinearSVC is not to replace the logistic benchmark, but to provide a second classification rule under the same feature sets. Comparing logistic regression and LinearSVC under common training and evaluation splits makes it possible to assess whether the predictive contribution of lexical, rhetorical, thematic, electoral, and biographical variables is robust across alternative linear classifiers.

As in the logistic specifications, the analysis excludes variables that trivially reveal the target whenever necessary to prevent label leakage. In the ideological task, we exclude party identifiers. In the rhetorical task, the EMI score itself is excluded from the predictor set because it is used to define the binary outcome.

Logistic regression and LinearSVC provide two complementary linear models for supervised analysis. While logistic regression is directly probabilistic and more easily interpretable in terms of predicted probabilities, LinearSVC offers a strong large-margin alternative that is particularly well suited to sparse and high-dimensional text representations. In this way, the supervised analysis does more than identify a single best model. It also compares how different layers of the analytical database contribute to the prediction of ideological and rhetorical outcomes.

4.7 Evaluation and Model Comparison

The analysis relies on a set of complementary evaluation measures because no single statistic fully captures classification performance (Izenman 2008; Hastie, Tibshirani, and Friedman 2009; James et al. 2013). Accuracy reports the share of correct predictions and provides a simple overall benchmark. Although intuitive, accuracy alone may be misleading when performance differs across classes, especially in settings where some categories are easier to recover than others. For this reason, the evaluation also emphasizes F1-based measures and class-specific diagnostics.

To preserve the empirical class distribution, the supervised exercises use a stratified random split. The analytical sample is divided into training and test sets, with twenty percent of the observations reserved for out-of-sample evaluation. The analysis applies this procedure separately in each supervised task and ensures that the relative frequency of the target classes is maintained across the split.

To diagnose classification errors, the analysis reports precision, recall, and the F1 score, in addition to overall accuracy. For a given class in a one-versus-rest formulation, precision $= TP/(TP + FP)$ measures how often positive predictions are correct, whereas recall $= TP/(TP+FN)$ measures how many true instances the model successfully recovers. Because precision and recall may move in opposite directions, the F1 score is a harmonic mean and provides a useful summary that penalizes models that perform well on one dimension but poorly in the other.

In the multiclass setting, macro averages compute each metric by class and then average across classes, thereby assigning equal weight to each category and highlighting minority performance. Weighted averages, in turn, account for the empirical class distribution and therefore summarize performance at the sample level. For this reason, the evaluation reports accuracy, macro F1, weighted F1, and class-specific precision, recall, and F1 scores through the classification report. These measures make it possible to identify heterogeneous patterns of error, such as systematic confusion between neighboring ideological categories.

To further diagnose misclassification, the analysis also examines confusion matrices. These matrices summarize how often speeches from one class are assigned to another and provide a direct view of the structure of prediction errors. In the ideological task, this is useful for distinguishing whether the models mainly confuse left and center, center and right, or all three groups simultaneously. In the rhetorical-orientation task, the confusion matrix indicates whether the models are better at detecting evidence-based or intuition-based language.

Model comparison follows the same logic across the supervised exercises. First, the analysis compares simpler lexical specifications with richer specifications that add

thematic variables, rhetorical measures, and external covariates. This makes it possible to evaluate whether predictive performance improves when the analysis moves beyond the surface vocabulary of the speech. Second, we estimate the same feature sets with logistic regression and LinearSVC, allowing the comparison to assess whether the contribution of each block of variables is robust across alternative linear classifiers.

To isolate the marginal contribution of specific predictors or feature groups, the evaluation also uses ablation and permutation-based comparisons when appropriate. Ablation exercises compare models with and without specific feature blocks while holding the remaining structure fixed. Permutation importance quantifies how much out-of-sample performance deteriorates when a predictor, or a block of predictors, is shuffled in the test data (Altmann et al. 2010; Silveira et al. 2022; Silveira et al. 2023; Brown et al. 2023). The procedure records the drop in a target metric such as macro F1. Larger average losses indicate greater predictive contribution. When appropriate, the analysis also considers grouped permutations for correlated feature families, such as topic variables or external covariates.

Taken together, these metrics and comparisons clarify why accuracy alone is not sufficient, make class-specific errors explicit, and allow us to assess whether rhetorical, thematic, electoral, and biographical information add predictive value beyond the lexical baseline.

5 RESULTS

This chapter presents the descriptive results. It begins by establishing the temporal profile of the corpus and then turns to the rhetorical indicators constructed in Chapter 4. This sequence is substantively important because variation in rhetorical measures may reflect not only changes in rhetorical style, but also broader shifts in the volume and structure of parliamentary speech.

Building on the corpus profile presented in Chapter 3, the analysis proceeds in two steps. First, it examines how the Evidence–Minus–Intuition (EMI) index varies over time. Second, it analyzes whether rhetorical patterns differ across ideological groups and substantive macro-themes.

5.1 Rhetorical analysis

This section examines the behavior of the main rhetorical indicator using the Evidence–Minus-Intuition (EMI) score, which measures the relative proximity of each speech to evidence oriented and intuition-oriented semantic fields. Positive values indicate a greater relative proximity to evidence-based language, while negative values indicate a greater relative proximity to intuition-oriented language.

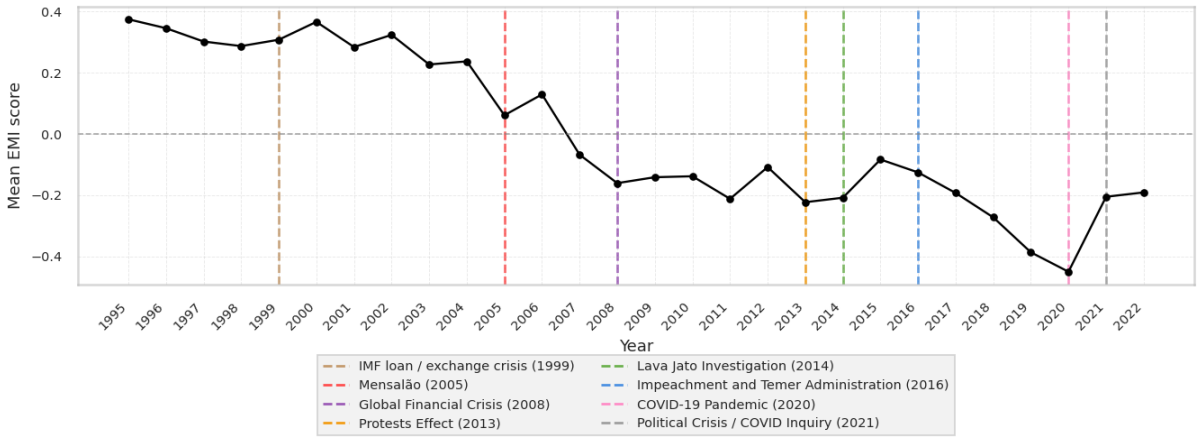
Because the EMI is sensitive to dictionary specification and aggregation choices, the interpretation below focuses primarily on temporal patterns rather than absolute levels. The vertical event lines provide historical reference points. They do not identify causal effects but help situate rhetorical variation within Brazil's political and economic chronology.

5.1.1 Aggregate EMI over time

To start the analysis, we examine the EMI score, as shown in Figure 5. We can observe that in the first years of the analysis, from 1995 to the early 2000s, EMI values are positive and reach the highest levels in the series. This period corresponds to the highest level of evidence-oriented rhetoric was most prominent in Senate speeches. Historically, this pattern is consistent with a period marked by debates mainly concerned with macroeconomic stabilization, institutional credibility, fiscal adjustment, and the consolidation of the policy framework associated with the Real Plan. Even within this short period, however, there is already a tendency toward a reduction in evidence-oriented language in the speeches, which may be related to external shocks, such as global financial crises. Even so, the EMI remains at its highest point in the entire series.

In the early 2000s, the index begins to decline, and this decline has two main moments. The first occurs with the election of President Luiz Inácio Lula da Silva, who was still perceived as an uncertain figure by many political and economic actors. In the

Figure 5– EMI Score



post-election year, 2003, the indicator falls below its previous levels, although this did not yet represent a major structural change. The sharper change occurs in 2005, precisely when the Mensalão scandal took place. Another important point is 2008, with the international subprime crisis. At this moment, the rhetoric appears to reach a new equilibrium point, where it remains in the following years. This new level appears consolidated after the 2010 electoral cycle, when the Senate underwent significant renewal.

Between 2011 and 2014, the index remains close to the same range, indicating a rhetoric more strongly based on intuition. This was a difficult period for Brazilian politics, as it coincided with the beginning of Dilma Rousseff's administration, increasing economic uncertainty, the 2013 protests, and the emergence of Operation Lava Jato in 2014, all of which marked the Brazilian political landscape and became central elements of Brazil's recent history. The persistence of negative values during this period suggests that the rhetorical equilibrium of the Senate shifted in a context of growing institutional and political tension.

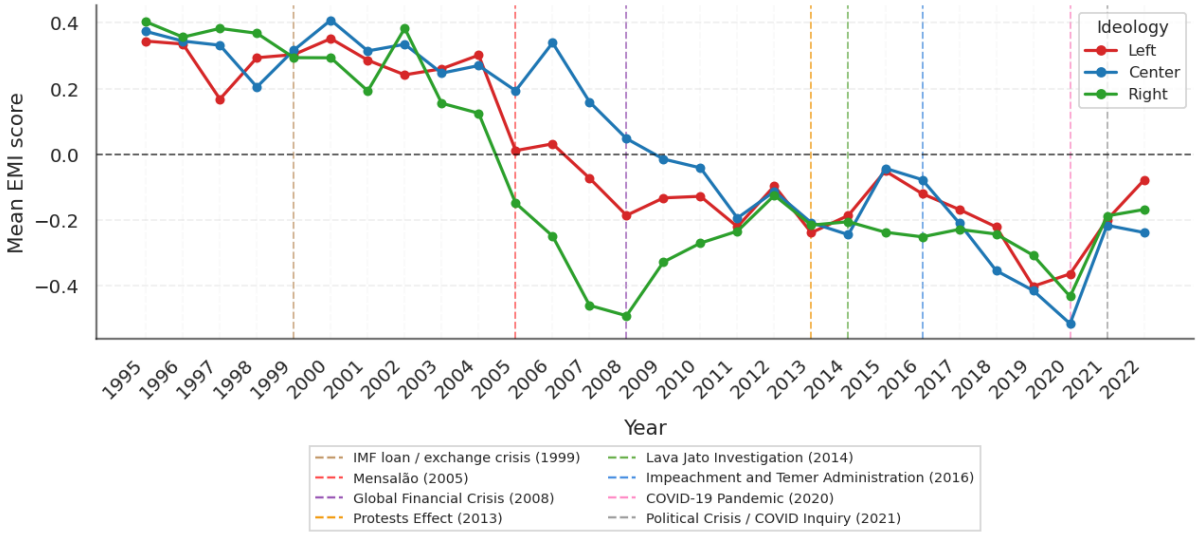
The negative trend continues in the second half of the sample. Although there are short-run recoveries, especially around 2015 and 2021, the EMI remains below the levels observed in the first decade of the series. The lowest values appear around the pandemic period and its aftermath, when the Senate operated under institutional, health, and economic stress.

The Figure 5 suggests that rhetorical style is not constant over time: it varies according to the broader political and institutional context. The partial recovery observed at the end of the series requires caution, as it may reflect both the reduced number of speeches in recent years and the specific rhetorical dynamics associated with the COVID-19 Parliamentary Inquiry Commission, which may have temporarily increased the use of evidence-oriented language. Another important factor to consider is the renewal of the Senate, since new officeholders may use different communication strategies.

5.1.2 Analysis by ideological class

Since our final objective is to identify ideological types and examine how they behave, we now analyze how this indicator varies across political classes: left, center, and right. More specifically, we assess whether these groups use different rhetorical strategies with respect to what we define here as evidence and intuition. We begin with Figure 6, which disaggregates the average EMI by ideology. Overall, the three groups follow broadly similar dynamics, although they differ at a few specific moments. In general, the shift toward more intuition-oriented rhetoric does not seem to be restricted to one side of the ideological spectrum.

Figure 6– EMI score by ideological class



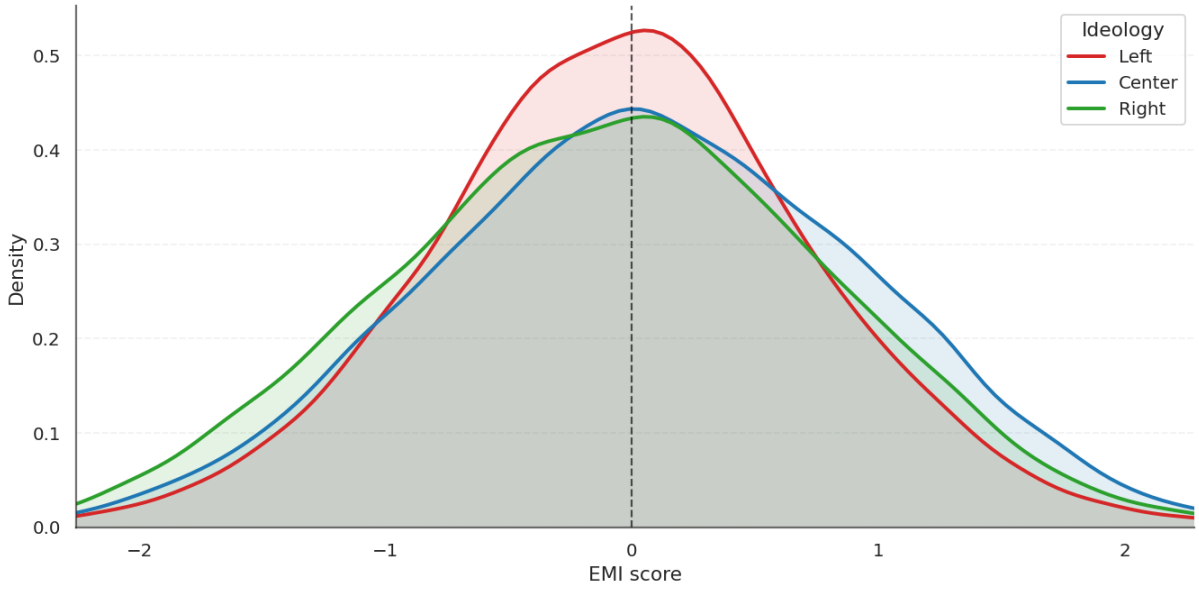
Looking at the series in more detail, in the early years, right-wing and centrist senators tend to present higher EMI values, while left-wing senators remain slightly lower. This pattern is consistent with the political configuration of the period, in which the center and the right were closer to the governing coalition responsible for the stabilization agenda. After the 2000s, however, the differences across groups narrow. The left, center, and right all move toward lower EMI values, although the timing and intensity of this movement vary across groups.

The period in which the groups differ most clearly begins in 2003 and extends until around 2010. During this interval, the right adopts a discourse much more strongly associated with intuition than the other groups, which may reflect its position as opposition to the government at the time. The center, by contrast, appears as the group most consistently associated with evidence-based rhetoric throughout the series. Notably, the center is also the group with the highest number of speeches. A relevant pattern is that the blue line appears closer to the left when the left holds the presidency, but closer to the rhetorical strategy of the right when the right is in government. This dynamic may

reflect the logic of coalition presidentialism, although this claim requires further analysis to support this claim more firmly.

To visualize these differences more clearly, we can examine the distribution of EMI for each political ideology. As shown in Figure 7, the three curves overlap substantially. Therefore, we cannot simply state that left, center, and right-wing speeches are clearly separated along the evidence–intuition dimension. Although the distributions are not identical, they are concentrated around a similar mean. It is also important to note that the global EMI is normalized by the mean; therefore, the pattern observed in the full dataset persists even when speeches are separated by political class.

Figure 7– EMI score density



The vertical dashed line marks the point at which the EMI equals zero. Around this threshold, the left presents a slightly more concentrated distribution, with a higher peak close to zero. The center appears somewhat more shifted toward positive EMI values, while the right displays a slightly wider distribution and somewhat greater mass on the negative side. These differences, however, are modest when compared to the degree of overlap among the three ideological groups.

This result has two important implications. First, it reinforces the interpretation that the EMI captures a rhetorical dimension that cuts across ideological boundaries rather than aligning mechanically with them. Second, it suggests that the relationship between rhetorical style and ideology is not well approximated by simple univariate comparisons. In other words, while the EMI contains meaningful variation, this variation alone is insufficient to clearly distinguish ideological classes.

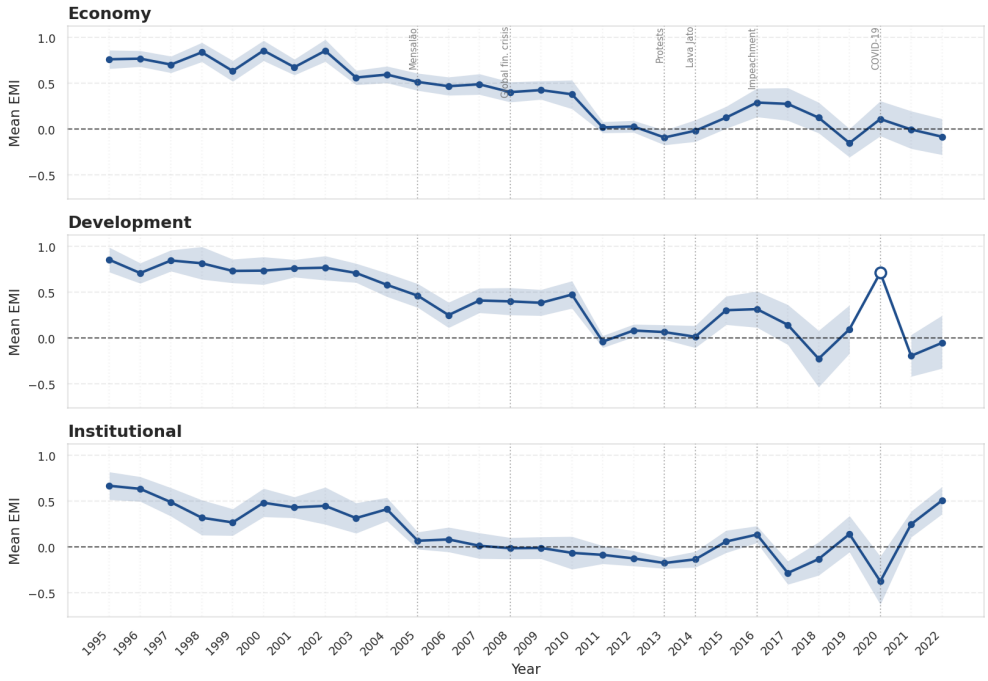
This motivates the next step of the analysis. Rather than relying on a single indicator, the dissertation turns to supervised machine learning models that combine

lexical features, thematic information, rhetorical measures, and speaker-level covariates to recover ideological structure in a high-dimensional setting. If rhetorical style contains latent ideological information, it is more likely to emerge in a multivariate framework than in the univariate density distributions shown in Figure 7.

5.1.3 Dynamics of rhetorical variation by macro-themes

A central question is whether the aggregate decline in EMI is systemic across macro-rhetorical dimensions or reflects specific thematic agendas. To examine this aspect, the analysis examines separately each pillar of the period's political economy: economy, development, and institutions. Figure 8 reports the dynamics of those annual EMI means together with 95% bootstrap confidence intervals.

Figure 8– EMI over time by selected macro-theme



Source: Author's elaboration.

The economy macro-theme presents one of the clearest examples of this movement. At the beginning of the series, EMI values are relatively high, which is consistent with a period in which the economic agenda centers on stabilization, fiscal adjustment, inflation control, and macroeconomic credibility. These debates tend to require a more evidence-oriented rhetorical profile, since they are closely connected to technical diagnoses and institutional policy frameworks. Over time, however, this distinction begins to weaken. The economy series moves toward values closer to zero and becomes more volatile, suggesting that economic discourse gradually loses part of its initially more technical profile and becomes more open to intuitive and political forms of framing.

A similar pattern appears in the development macro-theme, although its post-2010 volatility is particularly visible. In the first part of the series, the emphasis on state capacity, public investment, infrastructure, industrial policy, and growth strategies helps keep EMI at relatively high levels. After the early 2010s, however, we do not observe a simple linear decline. What appears instead is a more irregular trajectory, with periods closer to evidence-oriented rhetoric followed by moments in which the series moves toward a more intuition-oriented profile. This volatility is consistent with the conflictual nature of the Brazilian development debate, in which technical arguments about growth, planning, and public investment often coexist with broader political disputes over the role of the state.

Institutional issues follow a different logic. Unlike economy and development, the institutional macro-theme starts from a lower EMI level and appears more sensitive to moments of political stress. Its trajectory shows reversals around corruption scandals, the impeachment process, the COVID-19 period, and the subsequent parliamentary inquiry. In these contexts, technical and legal arguments tend to become entangled with moral, emotional, and conflict-oriented forms of rhetoric. This helps explain why the institutional series presents a more irregular pattern and does not follow the same path observed in the economy and development macro-themes.

Taken together, our analysis of macro-thematic categories reinforces the aggregate result found previously. The decline in EMI is not restricted to the general series, since it also appears within different substantive agendas. Economy and development begin with a clearer evidence-oriented profile and then lose part of this distinction over time, while institutional issues follow a more irregular trajectory, more directly associated with political stress. Finally, it is important to note that the estimates for the later years require caution, since the smaller number of classified speeches makes theme-specific averages more sensitive to the composition of the sample.

5.2 Predictive exercise

Given that the analysis of the EMI score as a single indicator is insufficient to define the senator's political class, the analysis turns to supervised learning models. The objective is to examine whether rhetorical information, thematic information, and senator-level characteristics add predictive value beyond the lexical content of the speeches. The empirical strategy follows an ablation design. The first specification uses only the binary EMI-orientation indicator, providing a minimal rhetorical benchmark. The second adds topic or thematic indicators. The third uses TF–IDF features as the lexical baseline. Additional specifications then add EMI, thematic variables, and external covariates to the textual representation. This nested structure evaluates the marginal contribution of each feature block. The predictive exercise uses two linear classifiers as the main

benchmarks: multinomial logistic regression and Linear Support Vector Classification. Logistic regression provides a transparent probabilistic benchmark, while LinearSVC offers a robust alternative for high-dimensional sparse text data. The specifications used are presented in the following table.

Table 5 – Feature sets used in the predictive exercise

Feature set	Empirical role
Binary EMI-orientation indicator	Rhetorical benchmark based on evidence–intuition orientation
Binary EMI-orientation indicator + topic indicators	Rhetorical and thematic benchmark
TF–IDF	Lexical baseline based on unigram and bigram frequencies
TF–IDF + binary EMI-orientation indicator	Tests the incremental contribution of rhetorical style beyond lexical content
TF–IDF + binary EMI-orientation indicator + topic indicators	Combines lexical, rhetorical, and thematic information
Non-lexical covariates	Rhetorical, thematic, senator-level, and electoral information
TF–IDF + non-lexical covariates	Extended model combining lexical and contextual information

Source: Author's elaboration.

5.2.1 Ablation results for ideological classification

Our first exercise, and one of the main objectives of this study, is to examine whether rhetorical style helps predict the party-based ideological class of Senate speeches. To do so, we begin by presenting the ablation results for ideological classification, reported in Table 6. The results corroborate the pattern observed in the descriptive analysis. In isolation, the binary EMI orientation indicator has very limited predictive power, ranking among the weakest specifications.

Adding topic indicators improves performance, although it remains substantially below the models that use lexical information. This suggests that rhetorical and thematic characteristics contain some ideological signal, but not enough to distinguish left, center, and right-wing speeches without the richer vocabulary representation that TF–IDF provides. The TF–IDF lexical baseline performs substantially better than the specifications that do not include textual features. This result is expected, since ideology in legislative speech is strongly associated with differences in vocabulary, policy emphasis, and issue framing. Once TF–IDF is included, however, the addition of the binary EMI-orientation indicator and topic indicators does not improve performance. In both model families, the text-only baseline performs slightly better than the specifications that add rhetorical and

thematic indicators. This indicates that, in the ablation exercise, the additional rhetorical and thematic variables do not provide incremental predictive gains beyond the lexical representation.

Overall, these results indicate that most of the predictive signal comes from lexical content. The EMI score, therefore, does not function as an independent ideological classifier. Instead, it functions better as a complementary rhetorical measure: it captures a dimension of discursive style that is substantively meaningful, but only weakly aligned with ideological categories when considered at the speech level. Thus, the evidence suggests that rhetoric more closely associated with evidence or intuition does not clearly distinguish ideological classes on its own. Rather, it appears to interact with broader political, institutional, and speaker-level dynamics. In this exercise, EMI enters as a binary rhetorical-orientation indicator.

Table 6 – Ablation results for ideological classification

Model	Feature set	Accuracy	Macro F1
Logit	Binary EMI orientation	0.4456	0.2055
Logit	Binary EMI orientation + topic indicators	0.4477	0.2257
Logit	TF–IDF	0.7900	0.7767
Logit	TF–IDF + binary EMI orientation	0.7892	0.7758
Logit	TF–IDF + binary EMI orientation + topic indicators	0.7884	0.7751
LinearSVC	Binary EMI orientation	0.4456	0.2055
LinearSVC	Binary EMI orientation + topic indicators	0.4477	0.2257
LinearSVC	TF–IDF	0.7938	0.7833
LinearSVC	TF–IDF + binary EMI orientation	0.7937	0.7831
LinearSVC	TF–IDF + binary EMI orientation + topic indicators	0.7930	0.7823

Source: Author's elaboration.

The previous exercise shows that lexical information provides the strongest predictive baseline for this task. We now examine how an extended specification, combining lexical information with a broader set of non-lexical information, affects prediction. This block includes rhetorical, thematic, and external senator-level and electoral information, such as EMI, topic indicators, age, region, education, occupation, declared wealth, vote share, and mandate-related indicators. Although LinearSVC obtains the highest macro F1 in the text-only ablation, the extended specifications reported in Table 7 are estimated using logistic regression. This choice keeps the classifier fixed across nested feature sets and therefore makes the comparison focus on the marginal contribution of the non-lexical block, rather than on differences between classification algorithms. Logistic regression also provides a transparent probabilistic benchmark, which is useful for comparing progressively

richer specifications.

The table shows that the non-lexical block, in isolation, performs substantially worse than the TF–IDF baseline. However, its performance remains above a naive benchmark, indicating that rhetorical, thematic, and external covariates contain some ideological signal. Still, without lexical information, these variables are not sufficient to recover the party-based ideological classification with the same precision provided by the vocabulary representation.

Combining this non-lexical block with TF–IDF increases predictive performance relative to the text-only baseline. Macro F1 rises from 0.7767 to 0.8063, while accuracy increases from 0.7900 to 0.8159. This gain contrasts with the previous ablation exercise, where adding only the binary EMI-orientation indicator and topic indicators to TF–IDF did not improve the text-only baseline. The improvement in the extended model therefore appears to come primarily from the broader non-lexical block, especially the external covariates, rather than from the rhetorical and thematic indicators alone. The result suggests that information outside the lexical representation contains relevant ideological structure.

This result, however, should be interpreted with caution. Since the train-test split occurs at the speech level, the same senator may appear in both the training and test sets. Stable senator-level and mandate-level covariates may therefore capture persistent political regularities, rather than only generalizable relationships between personal characteristics and ideology. For this reason, the extended specification functions as a predictive benchmark that combines lexical, rhetorical, thematic, and contextual information, not as a causal estimate of the effect of senator-level characteristics on ideological orientation.

Table 7 – Extended specifications with non-lexical covariates

Feature set	Accuracy	Macro F1	Weighted F1
TF–IDF	0.7900	0.7767	0.7873
Binary EMI + thematic indicators + external covariates	0.6173	0.5741	0.6023
TF–IDF + binary EMI + thematic indicators + external covariates	0.8159	0.8063	0.8145

Source: Author's elaboration. Results are estimated using logistic regression.

To complement the aggregate performance metrics, Figure 9 reports row-normalized confusion matrices for the main lexical-rhetorical specification and the extended specification. This diagnostic shows how classification errors are distributed across ideological classes, rather than only reporting average predictive performance.

The confusion matrices also reveal that errors are not symmetrically distributed across classes. In the main lexical-rhetorical specification, the largest classification difficulty

occurs between the right and the center: a substantial share of speeches whose true label is right-wing is predicted as center. This suggests that the boundary between these two categories is lexically and rhetorically less distinct than the separation between left and the other groups. The extended specification reduces this confusion, but the pattern remains substantively informative: ideological classification is especially difficult around the centrist/right-wing boundary.

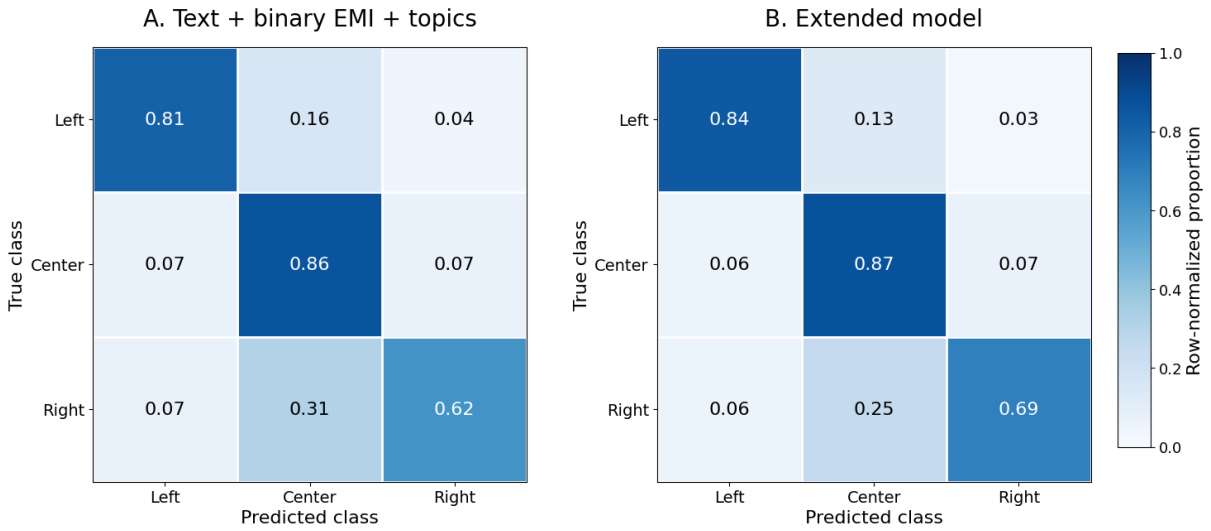


Figure 9– Normalized confusion matrices for ideological classification

Source: Author's elaboration.

The analysis computes the grouped permutation importance for the logistic specification that combines text, binary EMI, controls, and topic indicators. The results in Table 8 confirm the previous evidence. Permuting the textual block produces the largest deterioration in predictive performance, reducing macro F1 by approximately 46.7%. Permuting the block that combines binary EMI and external covariates produces a smaller but still relevant reduction of 17.1%, while permuting topic indicators reduces performance by only 0.1%. This reinforces the interpretation that lexical content is the main source of predictive performance, that the non-lexical block adds secondary information, and that topic indicators contribute little once text and controls are included.

Table 8 – Grouped permutation diagnostic

Feature block	Relative drop in Macro F1
Text	46.7%
Binary EMI + external covariates	17.1%
Topic indicators	0.1%

Source: Author's elaboration. Relative drop is computed as the proportional reduction in macro F1 after jointly permuting each feature block in the test set.

5.2.2 Robustness to senator-level and temporal splits

The previous results rely on a random speech-level split. Although this design is useful for evaluating predictive performance within the observed corpus, it may overstate generalization if the same senators, mandates, or political contexts appear in both the training and test sets. To address this limitation, we conduct two additional robustness exercises. The first uses stratified grouped cross-validation by senator, using *CodigoParlamentar* as the grouping variable, so that speeches from the same senator are not split between training and test folds. The second uses a temporal holdout design, training models on earlier years and evaluating them on the 2018 electoral cycle.

Table 9 reports the grouped cross-validation results. As expected, performance is lower than under the random speech-level split, but it remains informative for a three-class classification task. The best specification in logistic regression is the text alone, with an average accuracy of 0.5465 and a macro F1 of 0.5044. Adding binary EMI, topic indicators, and controls does not improve performance under this grouped design. This suggests that non-lexical covariates are more useful when the same senators can appear across training and test sets than when the model is required to generalize to unseen senators.

Table 9 – Grouped cross-validation by senator

Model	Feature set	Accuracy	Macro F1
Logit	Text only	0.5465	0.5044
Logit	Text + binary EMI + topics + controls	0.5325	0.4970
Logit	Binary EMI + topics + controls	0.4511	0.3938
LinearSVC	Text only	0.5363	0.5018
LinearSVC	Text + binary EMI + topics + controls	0.5317	0.5006
LinearSVC	Binary EMI + topics + controls	0.4629	0.3943

Source: Author's elaboration. Results are based on stratified grouped cross-validation by senator, using *CodigoParlamentar* as the grouping variable.

The temporal holdout exercise produces a similar pattern. When models are trained on earlier years and tested on the 2018 cycle, the text-only LinearSVC specification obtains the best macro F1 of 0.5266. The text-only logistic model reaches the highest accuracy, 0.5708, with a macro F1 of 0.5257. Models that include binary EMI, topic indicators, and tabular controls perform worse in this temporal setting. This suggests that contextual covariates can improve performance within the observed corpus, but they do not necessarily improve generalization to later political cycles.

Taken together, these robustness checks qualify the interpretation of the random-split results. The models capture meaningful ideological structure in the corpus, but part of their performance reflects stable speaker-level, mandate-level, and temporal regularities. When the classifier must generalize to unseen senators or to a later electoral cycle, the task becomes substantially harder. At the same time, the performance remains above a naive

three-class reference point, indicating that the text still contains non-trivial ideological signal under more demanding validation designs.

5.2.3 Rhetorical-orientation classification

As a complementary exercise, we examine whether the type of rhetoric can be predicted from textual and contextual features in the dataset. As in the previous exercise, the analysis uses EMI as a binary variable, where 1 represents evidence-oriented rhetoric and 0 represents intuition-oriented rhetoric. Since the target variable derives from the EMI itself, the model removes the EMI from the set of predictors.

To begin the analysis, Table 10 reports the results for the binary prediction task. The models that include textual features perform substantially better than those based only on profile variables and topic indicators. The best-performing model is the logistic regression model. The best-performing specification is the logistic regression model with text and profile variables, reaching a macro F1 of 0.9035. Adding topic indicators produces virtually the same performance, with a macro F1 of 0.9033. By contrast, the specification without textual information reaches a macro F1 of 0.6279.

Table 10 – Prediction of binary rhetorical orientation

Model	Feature set	Accuracy	Macro F1	Weighted F1
Logit	Text + profile + topics	0.9033	0.9033	0.9033
Logit	Text + profile	0.9035	0.9035	0.9035
LinearSVC	Text + profile	0.8990	0.8990	0.8990
LinearSVC	Text + profile + topics	0.8986	0.8986	0.8986
Logit	Profile + topics	0.6283	0.6279	0.6279
LinearSVC	Profile + topics	0.6281	0.6275	0.6276

Source: Author's elaboration. The target is the binary EMI-orientation indicator.

This pattern is consistent with a rhetorical orientation strongly related to the lexical content of speeches. The EMI measure itself derives from the text, based on the semantic proximity of speeches to evidence-oriented and intuition-oriented language. Therefore, it is natural that models using textual features are better able to recover this classification.

At the same time, adding topic indicators to the text-based model does not improve performance. This suggests that the lexical patterns and profile information primarily capture the distinction between evidence-oriented and intuition-oriented rhetoric. In other words, what matters most for predicting rhetorical orientation is not only the topic being discussed, but the specific vocabulary and semantic structure used in the speech.

Taken together, this complementary task reinforces the interpretation of the previous results. The EMI is not a strong stand-alone classifier of ideology, but the rhetorical orientation it captures is systematic and recoverable from speech content. Therefore,

the EMI functions as a substantively meaningful rhetorical dimension that complements ideological classification, rather than replacing lexical information.

The predictive exercises conducted in this work offer a consistent picture. Lexical information is the strongest predictor of party-based ideological class. The binary EMI-orientation indicator and topic indicators contain some ideological signal when used without text, but they do not improve the lexical baseline in the simple ablation design. Non-lexical covariates improve performance in random speech-level splits, although the robustness tests show that generalization to unseen senators and later electoral cycles is substantially more difficult. Finally, the rhetorical-orientation classification task shows that the EMI captures a systematic linguistic dimension. However, the results do not support the claim that rhetorical style, by itself, strongly predicts party-based ideological class. Therefore, there is no clear evidence that one ideological group systematically relies more on evidence-oriented rhetoric than another.

6 CONCLUSION

This dissertation analyzes whether the rhetorical style employed in a speech provides relevant information for classifying ideology. The central contribution is the adaptation of the EMI metric to a corpus of the Brazilian Federal Senate, covering the period from 1995 to 2022. We sought to combine more information, from lexical representations, topic modeling, electoral information, and senator-level covariates. The analysis examines whether evidence-oriented or intuition-oriented rhetoric provides additional information about political discourse beyond traditional approaches.

The descriptive results show that rhetorical style varied substantially over time, following political trends compatible with Brazil's recent history. At the beginning of the series, evidence-oriented rhetoric reaches its highest point. However, over time, intuition-based rhetoric gains prominence. As reported previously, the main inflection points coincide with moments of political stress. This pattern indicates that rhetorical style is not constant over the period and appears to respond to broader historical and institutional conditions. Future work could explore this point by searching for structural breaks in key moments, such as periods of strikes and crises.

In our case, the results separated by ideological class show that the relationship between rhetorical style and political group is not so clear, at least not in the way we sought to capture. Although left-wing, center, and right-wing speeches present some differences in specific moments, their EMI distributions substantially overlap. This means that evidence-oriented rhetoric and intuition-oriented rhetoric do not belong exclusively to any specific ideological group. This is a dimension that goes beyond a simple change in political paradigms.

The predictive results reinforce this reading. When used in isolation, EMI orientation has low capacity to classify the party ideology of the speeches. The main predictive power comes from lexical information, captured by TF–IDF. In the simple ablation exercise, the inclusion of binary EMI and topic indicators did not improve performance relative to the textual baseline, suggesting that these variables do not add enough information beyond what is already present in the vocabulary.

When we combine TF–IDF, EMI, topics, and external covariates, we find an improvement in performance in relation to the text-only logistic model. This gain seems to be associated mainly with characteristics of the senators, the mandates, and the electoral context. Even so, the robustness tests demonstrate that generalization to unseen senators and to later periods is more difficult. For this reason, these estimates should be interpreted with caution.

Therefore, EMI does not function as an independent classifier of ideology. Its contribution lies elsewhere. Our work demonstrates that the metric allows us to observe

a rhetorical dimension of legislative speech that is not reduced to the categories of left, center, and right. While lexical models better capture "what" senators talk about, EMI helps us understand "how" they justify, frame, and communicate their positions.

Our work also presents important limitations. First, the analysis assigns ideological labels at the party level. This choice makes the classification transparent and reproducible, but it does not capture all the heterogeneity within parties nor possible individual ideological changes over time. Second, the topic model assigns a dominant theme to each speech. This facilitates interpretation, but it may reduce internal thematic variation, especially in longer speeches. Third, our estimates show that random splits at the speech level may overestimate predictive performance, especially when the same senators, mandates, or political contexts appear both in the training set and in the test set. Finally, EMI depends on the construction of dictionaries and on semantic proximity based on embeddings, which means that its interpretation should focus on relative patterns, and not on absolute levels.

Based on these limitations, our work also points to paths for future research. One possibility would be to estimate rhetorical and thematic measures in smaller segments of speeches. This would allow us to better capture speeches that move across more than one theme. Another extension would be to use more granular ideological measures, such as ideal points at the legislator level, instead of labels based only on parties. Future work could also evaluate whether changes in rhetorical style precede institutional crises, electoral realignments, or changes in legislative productivity. Finally, EMI could be applied comparatively to other legislative bodies. This would help verify whether the shift from a more evidence-oriented rhetoric to a more intuition-oriented rhetoric is specific to the Brazilian case or part of a broader transformation in political communication.

Overall, our results demonstrate that rhetorical style is a dimension of legislative speech, but that its relationship with ideology is indirect. The Brazilian Senate does not divide clearly between ideological camps oriented toward evidence and camps oriented toward intuition. Instead, we find a more complex pattern: rhetorical style varies according to the political context, institutional conditions, and the semantic structure of the speeches. The main contribution of this dissertation, therefore, is to show that rhetorical indicators can reveal dimensions of legislative behavior that do not appear only in lexical classification. At the same time, our estimates also highlight the limits of treating rhetorical style as a direct predictor of party ideology. In this sense, EMI contributes less as a classification tool and more as an instrument to understand how senators justify, frame, and communicate their positions in legislative debate.

Bibliography

Altmann, André, Laura Tološi, Oliver Sander, and Thomas Lengauer. 2010. "Permutation Importance: A Corrected Feature Importance Measure." *Bioinformatics* 26 (10): 1340–1347.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. "A simple but tough-to-beat baseline for sentence embeddings." In *International conference on learning representations.*

Aroyehun, Segun T, Almog Simchon, Fabio Carrella, Jana Lasser, Stephan Lewandowsky, and David Garcia. 2025. "Computational analysis of US congressional speeches reveals a shift from evidence to intuition." *Nature Human Behaviour,* 1–12.

Baer, Werner. 2001. *The Brazilian economy: Growth and development.* Bloomsbury Publishing USA.

Barbosa, Luciana Gomes, Maria Alice Santos Alves, and Carlos Eduardo Viveiros Grelle. 2021. "Actions against sustainability: Dismantling of the environmental policies in Brazil." *Land use policy* 104:105384.

Beauchamp, Nick. 2011. "Using text to scale legislatures with uninformative voting." *New York University Mimeo.*

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

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3 (Jan): 993–1022.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching word vectors with subword information." *Transactions of the association for computational linguistics* 5:135–146.

Brown, David P, Daniel O Cajueiro, Andrew Eckert, and Douglas Silveira. 2023. "Information and transparency: Using machine learning to detect communication between firms." *Stan. Computational Antitrust* 3:198.

Carnes, Nicholas, and Noam Lupu. 2023. "The economic backgrounds of politicians." *Annual Review of Political Science* 26 (1): 253–270.

Downs, Anthony. 1957. "An economic theory of political action in a democracy." *Journal of political economy* 65 (2): 135–150.

Figueredo, Felipe C de, Bernardo Mueller, and Daniel O Cajueiro. 2022. "A natural language measure of ideology in the Brazilian Senate." *Revista Brasileira de Ciência Política,* no. 37, e246618.

Flynn, Peter. 2005. "Brazil and Lula, 2005: crisis, corruption and change in political perspective." *Third World Quarterly* 26 (8): 1221–1267.

Fraga, Arminio, Ilan Goldfajn, and Andre Minella. 2003. "Inflation targeting in emerging market economies." *NBER macroeconomics annual* 18:365–400.

Funtowicz, Alan. 2024. "Party competition under dictatorships: evidence from congressional speech in Brazil." PhD diss., Universidade de São Paulo.

Gennaro, Gloria, and Elliott Ash. 2022. "Emotion and reason in political language." *The Economic Journal* 132 (643): 1037–1059.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as data." *Journal of Economic Literature* 57 (3): 535–574.

Goretti, Manuela. 2005. "The Brazilian currency turmoil of 2002: a nonlinear analysis." *International Journal of Finance & Economics* 10 (4): 289–306.

Grootendorst, Maarten. 2022. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." *arXiv preprint arXiv:2203.05794.*

Gruss, Bertrand. 2014. *After the boom–commodity prices and economic growth in Latin America and the Caribbean.* International Monetary Fund.

Harris, Zellig S. 1954. "Distributional structure." *Word* 10 (2-3): 146–162.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied Logistic Regression.* Vol. 398. John Wiley & Sons.

IMF, International Monetary Fund. 2018. "2018 ARTICLE IV CONSULTATION—PRESS RELEASE; STAFF REPORT; AND STATEMENT BY THE EXECUTIVE DIRECTOR FOR BRAZIL." *IMF Country Report No. 18/253.*

Izenman, Alan J. 2008. *Modern Multivariate Statistical Techniques: Regression Classification, and Manifold Learning.* New York, Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* Vol. 112. Springer.

Kozlowski, Austin C, Matt Taddy, and James A Evans. 2019. "The geometry of culture: Analyzing the meanings of class through word embeddings." *American Sociological Review* 84 (5): 905–949.

Lauderdale, Benjamin E, and Alexander Herzog. 2016. "Measuring political positions from legislative speech." *Political Analysis* 24 (3): 374–394.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American political science review* 97 (2): 311–331.

Levy, Omer, and Yoav Goldberg. 2014. "Neural word embedding as implicit matrix factorization." *Advances in neural information processing systems* 27.

Lucchetti, Alexandre Henrique, and Daniel Oliveira Cajueiro. 2025. "Language as Data: A Survey of Natural Language Processing for Economics and Finance." *Journal of Economic Surveys.*

Mantoan, Eduardo, Vinícius Centeno, Carmem Feijo, Financialization, and Development Study Group (FINDE/UFF). 2021. "Why has the Brazilian economy stagnated in the 2010s? A Minskyan analysis of the behavior of non-financial companies in a financialized economy." *Review of Evolutionary Political Economy* 2 (3): 529–550. https://doi.org/10.1007/s43253-021-00051-6.

Mendonça, Helder Ferreira de, and Octavio Vargas Freitas Pinton. 2013. "Fiscal Cyclicality in the Brazilian Economy: Something is Changing." *Latin American Business Review* 14 (2): 163–174.

Mendonça, Helder Ferreira de, and Viviane Santos Vivian. 2008. "Public-debt management: the Brazilian experience." *Cepal Review* 2008 (94): 145–162.

Mendonça, Helder Ferreira de, and Rubens Teixeira da Silva. 2008. "Administração da dívida pública sob um regime de metas para inflação: evidências para o caso brasileiro." *Economia aplicada* 12:635–657.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781.*

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26.

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic regularities in continuous space word representations." In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies,* 746–751.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16 (4): 372–403.

Power, Timothy J, and Cesar Zucco Jr. 2009. "Estimating ideology of Brazilian legislative parties, 1990–2005: a research communication." *Latin American Research Review* 44 (1): 218–246.

Queiroz, Bernardo Lanza, and Matheus Lobo Alves Ferreira. 2021. "The evolution of labor force participation and the expected length of retirement in Brazil." *The Journal of the Economics of Ageing* 18:100304.

Rheault, Ludovic, and Christopher Cochrane. 2020. "Word embeddings for the analysis of ideological placement in parliamentary corpora." *Political analysis* 28 (1): 112–133.

SARLET, Ingo. 2021. *Social Security in Brazil: Public Pension Reform and Responses to the COVID-19 Pandemic.* Technical report. Social Law Report.

Silva, Rafael Silveira e. 2014. "Beyond Brazilian coalition presidentialism: the appropriation of the legislative agenda." *Brazilian Political Science Review* 8 (3): 98–135.

Silveira, Douglas, Lucas B de Moraes, Eduardo PS Fiuza, and Daniel O Cajueiro. 2023. "Who Are You? Cartel Detection Using Unlabeled Data." *International Journal of Industrial Organization* 88:102931.

Silveira, Douglas, Silvinha Vasconcelos, Marcelo Resende, and Daniel O. Cajueiro. 2022. "Won't Get Fooled Again: A Supervised Machine Learning Approach for Screening Gasoline Cartels." *Energy Economics* 105:105711.

Vartanian, Pedro Raffy, and H de S Garbe. 2019. "The Brazilian economic crisis during the period 2014-2016: is there precedence of internal or external factors." *Journal of International and Global Economic Studies* 12 (1): 66–86.

Weisbrot, Mark, Jake Johnston, and Stephan Lefebvre. 2014. "The Brazilian economy in transition: Macroeconomic policy, labor and inequality." *Center for Economic and Policy Research,* 1–27.

APPENDIX A – EMI Dictionaries

Table 11 – Dictionary Terms: Evidence vs. Intuition

Evidence-related terms	**Intuition-related terms**
evidence	believe
fact	opinion
proof	consider
reality	feel
evaluate	intuition
truth	think
examine	say
finding	experience
irrefutable	doubt
verify	trust
confirm	conviction
prove	intuit
evident	imagine
check	tend
corroborate	perception
assumption	re-believe
demonstrate	reason
obvious	want

Notes: The table reports the seed dictionary terms used to represent evidence-oriented and intuition-oriented semantic fields.

APPENDIX B – EMI Dictionaries 2 - in Portuguese

Table 12 – Dictionary Terms: Evidence vs. Intuition (Portuguese Version)

Evidence-related terms		**Intuition-related terms**	
inteligência	dados	conselho	sentimento
exata	fato	dúvida	opinião
precisa	aprender	enganar	ponto de vista
busca	ler	sugestão	senso comum
analisar	real	crença	genuíno
exame	dossiê	falso	perspectiva
investigar	encontrar	enganado	errado
procedimento	lógica	suspeita	engano
mostrar	razão	acreditar	palpite
análise	verdadeiro	fake news	instinto
examinar	educação	desconfiança	propaganda
investigação	descobertas	visão	desonestidade
processo	pesquisa		intuição
estatísticas	verdade		senso
correto	evidência		mentira
especialista	informação		sugerir
conhecimento	método		
prova	ciência		
estudo	verídico		
correção	evidente		
explorar	inquérito		
laboratório	localizar		
pergunta	científico		
teste			

APPENDIX C – Temporal Holdout Robustness Check

Table 13 – Temporal holdout results for ideological classification, 2018

Model	Feature set	Accuracy	Macro F1	Weighted F1	Macro Precision	Macro Recall
Logit	Text only	0.5708	0.5257	0.5595	0.5525	0.5259
Logit	Text + binary EMI + topics + controls	0.5221	0.4804	0.5091	0.4741	0.5041
LinearSVC	Text only	0.5593	0.5266	0.5555	0.5326	0.5275
LinearSVC	Text + binary EMI + topics + controls	0.5459	0.5133	0.5380	0.5098	0.5312

Notes: The table reports temporal holdout results for the party-based ideological classification task using 2018 as the test period. The training sample contains 82,808 observations and the test sample contains 5,886 observations. The feature matrix contains 87 variables before text vectorization.