A natural language measure of ideology in the Brazilian Senate

Felipe C. de Figueredo[1,2]
Bernardo Mueller[3]
Daniel O. Cajueiro[4]

DOI: 10.1590/0103-3352.2022.37.246618

Introduction[5]

Political debate is often a battle. Politicians from across the spectrum engage in combat over topics such as abortion, immigration, gun control, gay marriage, and taxation. When these disputes go beyond the realm of civil debate, they can pose a threat to the very system of liberal democracy that normally thrives on the existence of a diversity of views and constructive debate. In a recent book, Levitsky and Ziblatt (2018) argue that politicians increasingly treat their rivals as enemies, intimidate the free press, and threaten to reject the results of elections, weakening many institutional buffers of democracy, such as courts, intelligence services, and ethics offices. The trend towards increased political polarization observed in many countries across the world is evidence of the radicalization of political debate in recent times. For the United States, for example, Hare and Poole (2014) and Gentzkow, Shapiro and Taddy

[1] For Data Replication see: https://doi.org/10.7910/DVN/JSZI4V

[2] Ph.D. Student at University of Washington, Department of Economics, Seattle, WA, USA. E-mail: fcdf@uw.edu.

[3] Full Professor at University of Brasilia, Department of Economics, Brasilia, DF, Brazil. E-mail: bmueller@unb.br.

[4] Associate Professor at University of Brasilia, Department of Economics, Brasilia, DF, Brazil. Member of INCT-SC and LAMFO. Email: danielcajueiro@unb.br.

[5] We thank John Wilkerson, Lucio Rennó, Lucas Naves, Aurélio Carrilho and Pedro Campelo for comments and suggestions. Bernardo Mueller and Daniel O. Cajueiro are also thankful to CNPQ for partial financial support. The editors and two anonymous reviewers provided especially conscientious feedback. All the errors remain ours.

(2019) find a sharp upswing in polarization.[6] In the Brazilian context, former president Dilma Rousseff's impeachment process followed by the presidential victory of far-right politician, Jair Bolsonaro, reveals a similarly divided scenario. Many European countries are following a similar path. Understanding and measuring the influence of ideology in contemporary politics is simultaneously crucial and challenging.

This paper uses a novel approach that measures the content of ideologies using speech as data. The method treats language as the unit of analysis and applies an automatic classification algorithm to represent dictionaries that best explain each political position.[7]

Underlying this approach is the hypothesis that ideologies give coherence to a person's opinions and attitudes, so that once we have properly identified a person's ideology, we may be able to predict his or her opinions on new or modified issues. In contemporary politics the knowledge that a politician opposes the increases of corporate taxes and of the minimal wage, makes her/him also more likely to favor a balanced budget, oppose affirmative action, etc. In other words, people hold their political views – even those that are logically organized – with passion (POOLE, 2007).[8]

The study of ideology is one of the most difficult tasks in political science, mostly because it is not directly observable. In this effort, scholars such as Poole and Rosenthal (1985) and Power and Zucco Jr. (2009) employ different strategies, ranging from survey data to statistical estimates based on voting records. Their results rely upon a scaling method in which different political preferences can be projected into a latent ideological space. However, models based on multi-dimensional frameworks tend to be better at capturing the idiosyncrasies between different ideologies (DIERMEIER et al., 2011). Because they can take into account a large number of features, they display a richer representation of each political position. Unlike scaling methods for parliamentary

[6] They do, however, diverge about the extent of this phenomenon. Hare and Poole (2014) use a standard measure of ideological polarization based on roll-call votes and find that political polarization in the US began its upward trend around 1970 rather than 1990 – as found by Gentzkow, Shapiro, and Taddy (2019) using speeches.

[7] The complexity of language implies that automated content analysis methods will never replace careful and close reading of texts. Instead, these methods are best thought of as amplifying and augmenting careful reading and thoughtful analysis (GRIMMER; STEWART, 2013). It provides a method to study legislative behavior, intra-party politics, and polarization (LAUDERDALE; HERZOG, 2016).

[8] As argued by Converse (1964), ideology is a belief system – namely, that issues are interrelated or bundled, and that ideology is fundamentally the knowledge of "what-goes-with-what". He pointed out that the sources of constraints on idea-elements are much less logical in the classical sense than they are psychologically — and less psychologically than in social contexts.

roll-call data,[9] they do not restrict their analysis to a small set of features. The high dimensionality of language[10] provides a rich source of data to assess ideologies. Through political speech we can analyze the patterns subsumed in ideological labels. For instance, manually classifying phrases into substantive topics shows that the increase in partisanship is due more to changes in the language used to discuss a given topic (e.g., "estate tax" vs. "death tax") than to changes in the topics parties emphasize (GENTZKOW; SHAPIRO; TADDY, 2019). In another approach, Laver, Benoit and Garry (2003) and Lauderdale and Herzog (2016) use political text to estimate parliamentarians' political preferences. These language- based measures suggest that speech and roll-call votes respond to different incentives and constraints.[11] This encourages the use of complementary methods to study ideology, such as those based on content analysis. While Quinn et al. (2010) employ topic modelling, Yu, Kaufmann and Diermeier (2008) and Diermeier et al. (2011) use a supervised learning method to classify political speech and explore its content.

Our paper fits in the context of a long-standing debate in the Brazilian Political Science literature about the nature and role of parties and the extent to which they are institutionalized and ideologically disciplined and consistent. Open-list proportional electoral rules with state-wide districts in the lower house, together with an idiosyncratic political history and culture, have produced a unique and unparalleled party system in Brazil. Its extreme fragmentation – one of the highest in the world – results in an unwieldly number of parties and low levels of party identification in the electorate. Together with massive party switching by politicians, even across ideological divides, ideologically incoherent coalitions, and ill-defined party platforms, this has led many analysts to argue that unlike in most countries, parties in Brazil are not one of the main dimensions through which the political system should be analyzed and understood. Furthermore, in this view, these characteristics of the political

[9] See an application of multidimensional scaling in Mccarty, Poole and Rosenthal (1997).

[10] Human language makes use of a large and varying number of features (herein words) to represent various aspects of communicative information. One could easily represent language as high-dimensional vectors, with each dimension corresponding to a specific feature.

[11] Roll-call votes may be shaped by strategic considerations related to the passage of legislation and may therefore not reflect legislators' sincere policy preferences. Speech may reflect party differences in values, goals, or persuasive tactics that are distinct from positions on specific pieces of legislation (Gentzkow, Shapiro, and Taddy, 2019). Kim, Londregan and Ratkovic (2018) combine a model that estimates political preferences through both parliamentary speeches and roll-call data. They estimate a two-dimensional latent space. The first dimension, running left-right, corresponds to standard methods such as DW-NOMINATE (MCCARTY; POOLE; ROSENTHAL, 1997); while the second dimension, revealed by the text data, places parliamentarians according to their leadership positions.

system pose serious impediments to governability, weaken democracy, and induce economic inefficiencies and budget deficits (AMES, 2002; LAMOUNIER, 1989; MAINWARING; SCULLY, 1995; SHUGART; MAINWARING, 1997; SAMUELS, 2003; SHUGART; CAREY, 1992; among others).

An opposing view, however, holds that despite the apparent lack of coherence and stability, parties are actually significantly institutionalized and disciplined, especially within Congress where legislative institutions provide party leaders with tools that can be used to create incentives for and achieve significant levels of control over their members, as shown by data on roll-call votes and distribution of pork (AMORIM NETO, 2000; AMORIM NETO; SANTOS, 2001; ALSTON; MUELLER, 2006; FIGUEIREDO; LIMONGI, 1999; 2006; 2007; PEREIRA; MUELLER, 2002; 2003). Although the literature that focuses on the pathologies of the political system originated in a period when the Brazilian economy was performing poorly, while the literature that stresses the functionality of the system arose in a period of greater stability and growth, the debate remains current and unresolved. As the country entered a renewed and prolonged cycle of political and economic crises in 2015, arguments attributing these outcomes to the nature of Brazilian political institutions have reemerged. At the same time, the main proponents of the opposing view have doubled down with a paper that "reasserts the more positive view that points to the capacity of the Brazilian system to produce decisions" (LIMONGI; FIGUEIREDO, 2017, p. 79).

We make two contributions in this paper. The first is to provide a measure of the evolution of polarization, which is has been increasing throughout the world and also in Brazil. We do this, however, with a novel text-based approach. Instead of the usual approach of using a predefined dictionary, we build our own political dictionary that evolves over time. This framework is useful to analyze the stability and transition of beliefs, in an approach similar to Alston et al. (2013), who analyze the change in Brazil's social contract – although we rely on an automatic technique. Using political speech data from the 50th legislature to the 55th legislature of the Federal Senate of Brazil (1995-2018), we analyze changes in ideology over time. Polarization is captured by comparing the evolution of our method's classification accuracy, the intuition being that greater accuracy indicates more easily distinguishable speech across ideological dimensions, that is, more polar positions.

Our second contribution is to debate the nature of ideology in the Brazilian political system and the merit of different measures. Although ideology is necessarily a multidimensional concept, for most countries – and especially for bipartite systems – it turns out that a single left-right dimension captures legislators' ideologies remarkably well. In Brazil, however, measuring ideology is made difficult by the idiosyncratic nature of the political system described above. Early attempts at measuring legislator and party ideologies quickly discovered that the estimated rankings did not uncover pure ideology, but rather a measure contaminated by executive-legislative relations in the context of a strong presidential system. This measure aligns along a government-opposition dimension instead of the traditional left-right one that emerges when the same method is used in other countries (FIGUEIREDO; LIMONGI, 1999; LEONI, 2002; MORGENSTERN, 2003).[12] The reason is that presidents have tools to purchase votes and garner support in Congress. Thus, it is common to see anomalies such as clearly left-wing parties showing up on the right and vice versa, due to the intricacies of coalition building and logrolling especially with the Executive.

In an influential paper Power and Zucco Jr. (2009) argued that because roll-call-based measures of ideology were contaminated by pressures external to the legislators' true beliefs, they did not capture pure ideology. They proposed instead to use perceptual data from surveys and self-declarations by legislators to create a more accurate representation of ideology. They find that "the ideological ordering of parties in Brazil has remained considerably stable, despite significant social and economic change during the past two decades, and that a single left-right dimension is an accurate description of the ideological landscape of the county" (POWER; ZUCCO JR., 2009, p. 239).

More recently new approaches to measuring ideology have appeared in the literature using text data, such as speeches and manifestos. These methods hold the promise of overcoming many of the limitations of the roll-call-based measures as well as of the survey and perception-based measures. Some early papers applying text analysis to Brazilian data are Moreira (2020) and Arnold, Doyle and Wiesehomeier (2017). Medeiros and Izumi (2021) used a

[12] Ferreira and Mueller (2014) find that Brazilian Supreme Court Justices also align on a government-opposition dimension, using data from 2002 to 2012.

Bayesian approach based on sentiment analysis classifications (Pang and Lee, 2008) on over 64,000 speeches by senators and found that a dominant government-opposition dimension emerges, similar to that derived from roll-call data. They note that this is a surprising result because unlike votes, speeches do not have direct impact on outcomes and are often just "cheap talk". They suggest that speeches may play a role as signaling loyalty to party leaders and the Executive, who control resources within congress.

In this paper we also use senators' speeches as data, but instead of sentiment analysis we use a pre-existing classification of parties on the left-right dimension as a dictionary to train a classifier to identify words used by left, center and right parties. We use the Power and Zucco Jr. (2009) classification. By using their classification as the 'correct' ideology to train our algorithm, we assume that they effectively uncover a left-right classification. If this assumption is correct, then the set of words we identify can be interpreted as an expression of ideology. There are, however, some limitations to classifications done using surveys and experts, such as lack of political knowledge, low level of survey responses, and strategic considerations among others. If, the Power and Zucco Jr. (2009) measure suffers from any of these limitations or is also contaminated by coalition management and vote-buying by the executive, then our sets of words will be capturing a similarly contaminated measure of ideology. Though we cannot settle this issue in this paper, we hope our analysis will contribute to the growing literature that taps into the vast amount of available text-based data to measure ideology and other political motivations and actions. An obvious direction in which to seek to settle the debate about the left-right versus government-opposition ideological positioning is to move from single-dimension analysis to multiple dimensions.

Finally, our paper contributes to a growing literature on the Brazilian Senate. As noted by Izumi (2016, p. 95) most of the work on legislative studies in Brazil is on the House of Representatives or on the Congress as a whole, whereas only recently has the number of studies specifically about the Senate started to grow. There are many reasons why there should be more interest on the upper house in the Brazilian Congress. Not only is it one of the strongest upper houses in the world, but its members have, on average, a very different profile than those in the lower house, being older, more educated, richer and more experienced politically (IZUMI, 2016). Together with the smaller size of the Senate (81 senators versus 513 representatives) and with the different electoral rules (majoritarian, 8-year terms versus proportional open list 4-year terms) these characteristics make senators markedly more independent from their party leadership. Neiva and Soares (2013, p. 97-98) cite a series of themes that have recently been studied in the specific context of the Brazilian Senate, including female representation, intracameral dynamics, discipline and party cohesion, propositions, and legislative production, among others.

The specificity of the Brazilian Senate has given rise to several questions that have been pursued by this recent literature concerning the senate's role in policymaking, its specific powers relative to the House, among others. Perhaps the most interesting question has been about the dimensions around which senators' interests are organized in Brazil. Are the main cleavages primarily federative, pitting the Union against the interests of the states and municipalities or are they found to be mostly organized across party lines (IZUMI, 2016; NEIVA; SOARES, 2013; NEIVA, 2011)? In other word, can the voting and other behavior of Brazilian Senators be best understood as lying across an ideological dimension, a federative cleavage, or one that pits pro-government legislators versus opposition?

The outline of the paper is as follows. In the first section, we present our model – describing in detail how the data was prepared and what methods were employed. In the second section, we present the general results of our classification method and analyze their dynamics. Also, we compare our results to the extant political science literature. In the final section, we present a brief conclusion of our findings.

The model

We use automated techniques (i.e., Natural Language Processing, or NLP) to analyze the ideological content of the Brazilian Senators' speeches. Diermeier et al. (2011) employed a similar approach.[13] Our model, then, provides an alternative approach using supervised learning to get at issues of interest to political science.

[13] They analysed the speeches of the US Senate and applied a different classification algorithm, Support Vector Machines (SVM). They did, however, employ the same classification setup as a baseline model.

Data set

The data used in this analysis consists of the speech records from the Federal Senate of Brazil from its 50th to its 55th legislature – dating from 1995 to 2018.[14] This was a unique period in Brazil's recent democracy, when important debates – that drove important decisions – were made. Among others, there were debates and discussions about land reform, fiscal reform, pension reform, and an impeachment process.[15]

We make use of all senators' political speech within this interval.[16] In doing so, we divide the entire data set by legislature. We therefore have one set of speeches (i.e., one document set, or corpus) for each legislature. We treat each political speech as a document d_i, such that $d_i \in \mathbb{R}^N \left(i = 1, \ldots, M\right)$, where N is the total number of words (or, dimensions) and M is the total number of documents. For each document d_i, we have a unique category c_i, such that $c_i \in \{0,1,2\}$. Note that we set only three categories: left (i.e, $c_{left}=0$), center ($c_{center} = 1$), and right ($c_{right} =2$).

The categories c_i of the documents d_i must be determined in advance, typically by inspecting them individually and hand-labelling them (BISHOP, 2006). Since to establish such a coding technique is very costly, we decide to follow the categorization employed by Power and Zucco Jr. (2009). Their classification is widely used in the literature to study legislative behavior, because they produced a rich and in-depth estimate based on both roll-call voting data as well as survey responses. They place the positions of the main Brazilian political parties along an ideological line (the standard left-right dimension). They also provide a classification setup defining the parties among left, center or right. For instance, the classification of three of the most important parties are: PT (Workers' Party, or *Partido dos Trabalhadores*) is located on the left (i.e, $c_{PT} = 0$), while PSDB (Brazilian Social Democracy Party, or *Partido da Social Democracia Brasileira*) is located on the center ($c_{PSDB} =1.$), and PP (Progressive Party, or *Partido Progressista*) is located on the right ($c_{PP} = 2$).

[14] The data is available from the Brazilian Senate website: http://www.senado.gov.br/. See Supplementary Material for the data pre-processing description.

[15] The senators' speeches also help to illuminate how political party guidelines arise and evolve, as many of these speeches were given by the elite members of the respective parties.

[16] The data set consists of 85,466 addresses, including speeches and pronunciations. In the Brazilian Senate each legislature has 81 senators (three for each of the 26 states plus the Federal District). Each senator holds a mandate for two four-year legislatures.

Feature selection

We can represent the documents d_i as counting vectors.[17] In doing so, we ignore the order of words such that d_i is a vector whose length is equal to the N number of words in the vocabulary and whose elements w_{ij} are the number of times word j occurs in document d_i.

We remove stop words as manner of reducing the complexity of our data (JURAFSKY; MARTIN, 2020).[18] Stop words are words that do not bear solid meaning (mostly function words, such as *the*, *a*, and *of*).

Suppose that the text of document d_i is:[19] [20]

> And electoral reasons do not move us. Let's face it.
> And I think it's a risk that should be taken by any parliamentarian, any public man. To disagree here or to please there, this is of the public life.

After removing stop words and punctuation, we have: "electoral reasons move let face think risk taken parliamentarian public man disagree here please there public life". The bag-of-words representation would then have $w_{ij} = 2$ for $j \in \{public\}$ $w_{ij} = 1$ for $j \in \{electoral; reasons; move; let; face; think; risk; taken; parliamentarian; man; disgrace; here; please; there; life\}$, and $w_{ij} = 0$ for all other words in the vocabulary.

To further reduce the dimensionality of our data, we set a minimum term frequency of 50, assuming that words with frequencies below that threshold have low coverage and thus are not useful for classification.[21] In addition, we also set a maximum term frequency of 95%, assuming that words with such high frequencies are not relevant for classification, because most of them do not bring meaningful elements in a language. In addition, all words with fewer than three characters are removed.

[17] The simplest and most common way to represent a document is through counting vectors (or, bag-of-words.) There are different ways to represent text as data: for instance, by providing a Boolean representation, which is an algebraic expression consisting of binary values (1 or 0). It assigns a value of 1 for a simple occurrence of feature j in document i, and when absent it assigns the value of zero.

[18] Stemming is another way of reducing the dimensionality of our space. It reduces words to their base form. We do not apply stemming here, because we want to measure the ideologies' content considering the entire vocabulary.

[19] his example is from a specific speech record from our document set (Senator Arthur Virgílio, 53rd legislature, on April 7, 2010.)

[20] See Supplementary Material for the original text in Portuguese.

[21] By doing so, we also avoid words that might be misspelt. See Supplementary Material for alternative thresholds.

To prevent the classifiers from picking up the potential correlations between the names and party affiliations, we also removed senators' names from the documents. Our last resource to simplify the data was to identify specific expressions embedded in all documents that do not carry significance for our method of classification. We note that certain expressions, such as *Exº* (Your Excellency) and *Srs* (gentlemen), are irrelevant and thus removed them. In general, these expressions represent an example of speech standardization.

A useful approach that excludes both common and rare words is filtering by "term-frequency-inverse-document-frequency", or *tf-idf* (GENTZKOW; KELLY; TADDY, 2017). For a word or other feature *j* in document d_i, term frequency (tf_{ij}) is the count w_{ij} of occurrences of *j* in d_i. Inverse document frequency (idf_j) is the log of one over the share of documents containing *j*: $log\ (M/s_j)$ where $s_j = \sum_i 1_{[wij>0]}$ and *M* is the total number of documents. The object of interest *tf-idf* is the product $tf_{ij}\ x\ idf_j$. Thus, very rare words will have low *tf-idf* scores because tf_{ij} will be low. Moreover, very common words that appear in most or all documents will have low *tf-idf* scores because idf_j will be low. In other words, *tf–idf* assigns to term *j* a weight in document *i* that is (MANNING; RAGHAVAN; SCHÜTZE, 2008): i) highest when *j* occurs many times within a small number of documents (thus lending high discriminating power to those documents); ii) lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal); and, iii) lowest when the term occurs in virtually all documents.

Supervised learning

In supervised models, we observe the documents and categories in a training set ($\mathbf{d}_{\mathbf{train}}$, $\mathbf{c}_{\mathbf{train}}$) such that they may be directly harnessed to inform the model of the data generation process. During the training, also known as the learning phase, we determine the precise form of the function $f(d_i)$ based on the training data. Once we train the model, we can then determine the identity of new documents ($\mathbf{d}_{\mathbf{test}}$; $\mathbf{c}_{\mathbf{test}}$) which are said to comprise a test set. The result of running the machine learning algorithm can be expressed as a function $f(d_i)$ which takes a new document d_i as input and that generates an output vector c_i, encoded in the same way as the categorical vectors. The ability to categorize correctly new examples that differ from those used for training is known as generalization (BISHOP, 2006).

The most common supervised generative model is the naive Bayes classifier (JURAFSKY; MARTIN, 2020), which treats counts for each token as independent with class dependent means.[22] In naive Bayes, c_i – our categorical variable – and the token count distribution is factorized as $p\left(d_i \mid c_i\right) = \prod_j p_j\left(w_{ij} \mid c_i\right)$, thus "naively" specifying conditional independence between tokens j.[23] The parameters of each independent token distribution are estimated, yielding $\hat{p}$ for $j = 1, \dots, p$. We invert the model for prediction, with classification probabilities for the possible class labels obtained via Bayes's rule as:

$$p\left(c \mid d_i\right) = \frac{p\left(d_i \mid c\right)\pi_t}{\sum_a p(d_i \mid a)\pi_a} = \frac{\prod_j p_j(w_{ij} \mid c)\pi_t}{\sum_a \prod_j p_j(w_{ij} \mid a)\pi_a}$$

where π_a is the prior probability on class a (usually just one over the number of possible classes). For a full explanation see the Supplementary Material.

To employ a standard supervised learning approach, it is necessary to follow a validation process to ensure our model's ability to make predictions. As the data set is relatively large, we employ a hold-out test validation procedure. This means we withhold some of the sample data from the model identification and estimation process, then use the model to make predictions for the hold-out data to see how accurate they are. Following that, we shuffle our speeches prior to dividing them into training and test samples: setting two--thirds of our data set to form the training sample $\mathbf{d}_{\mathbf{train}}$ and the remaining one--third to be the test sample $\mathbf{d}_{\mathbf{test}}$.

To guarantee the robustness of our analysis we also use a second validation process: in this case, we observe two measures, Precision and Recall. The former is the fraction of retrieved documents that are relevant, while the latter is the fraction of relevant documents that are retrieved. The measures of precision and recall concentrate the evaluation on the return of true positives, asking what percentage of the relevant documents have been found and how many false positives have also been returned (MANNING; RAGHAVAN; SCHÜTZE, 2008).

[22] For alternative supervised learning models, such as Support Vector Machine and Random Forest, see BISHOP (2006).

[23] An alternative approach – without assuming independence among words – is the unsupervised learning algorithm doc2vec (LE; MIKOLOV, 2014). Even though this approach is more reliable, since it keeps the syntax of language, it also increases the computational cost of our model. Therefore, we decided to employ the simplest classification algorithm without any loss of generality.

Results and discussion

Classification results

We start our analysis by representing each word in each document d_i through their *tf-idf* values w_{ij}.[24] Then, we proceed by training the classification algorithm through our labelled data set – in that case, our training set $\mathbf{d}_{\mathbf{train}}$. Following this classification procedure[25] we are able to test the prediction accuracy of each training set over its respective test set – relative to the hold--out validation (Tables 1 and 2). In short, we take a document $d_i \in \mathbf{d}_{\mathbf{test \neq train}}$ and through our learned function $f(w_{ij})$, we can generalize about its content. Then, we classify the speech regarding its *tf-idf* values.[26] This enables us to investigate whether the politicians underlying each ideology employ a similar language pattern during each legislature. In addition, we analyze the inter-temporal dynamic of political language in Brazil.

Table 1. Prediction accuracy on the training set representations

	50th	**51th**	**52th**	**53th**	**54th**	**55th**
tf-idf	77.1	69.1	79.6	80.1	83.2	83.2

Source: Authors' results.

Table 2. Hold-out validation on the test set

	50th	**51th**	**52th**	**53th**	**54th**	**55th**
tf-idf	67.6	65.1	73.6	75.7	79.2	81.5

Source: Authors' results.

The general accuracy across legislatures is relatively high, indicating validation by the hold-out process (Table 2). We find the top average accuracy score across the models for the 55th legislature.

Peterson and Spirling (2018) demonstrate that machine learning accuracy provides an informative measurement instrument for the degree of aggregate polarization over time. According to them, when accuracy is high, and

[24] The value of each dimension j is the word frequency normalized by the inverted document frequency, that is, the word frequency divided by the document frequency (the number of documents that contain this word in the whole collection).

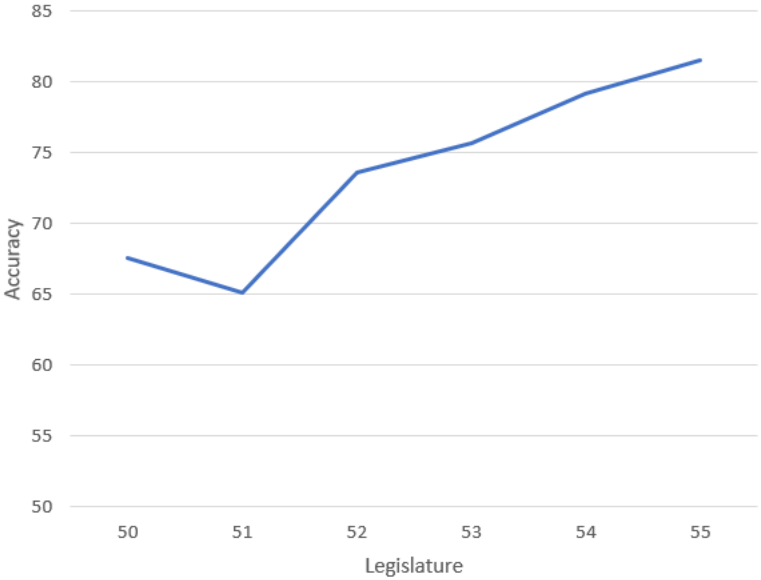
[25] Here, we employ the MultinomialNB algorithm from Pedregosa et al. (2011).

[26] According to our setup, the document di can be classified as ci, s.t. ci ϵ {0,1,2}.

the machine does well at discriminating between partisans based on their utterances – say, with regards to the topics they raise, or the way they express themselves – we are in a more polarized era.

Analyzing Figure 1, we note a decrease in accuracy between the 50th and the 51st legislature. However, starting in the 52nd legislature the trend in accuracy increases. Following our prediction accuracy on the test set we can thus assume that polarization specifically increased in the Brazilian Senate during the last four legislatures. Our results show that most contemporaneous senators employ a truly specific vocabulary, which will be further analyzed in the next sections.

Figure 1. Classification accuracy on the test set across the legislatures



Source: Authors' results.

Our data set – as observed in Tables 3 and 4 – presents a very imbalanced distribution among classes. Inspection of the 50th to the 53rd legislatures indicates that the high frequency of speeches is derived mostly from centrist politicians. Yet, in our two last legislatures, the 54th and the 55th, more than 40% of speeches' are delivered by leftist senators.

Table 3. Speech count

	50th	51th	52th	53th	54th	55th
Left	2,784	2,498	4,833	4,932	7,594	5,175
Center	4,005	5,105	9,288	7,516	6,438	4,864
Right	3,233	1,975	4,760	4,846	3,295	2,324
Total	10,022	9,578	18,881	17,294	17,327	12,364

Source: Authors' results.

Table 4. Speech frequency

	50th	51th	52th	53th	54th	55th
Left	27%	26%	26%	28%	44%	42%
Right	40%	53%	49%	43%	37%	39%
Center	33%	21%	25%	29%	19%	19%

Source: Authors' results.

Imbalanced data sets may not be as accurate. There may be some cases is our data set where a party has few elected senators, which risks conflating the position of the individuals with that of the party. In that case, we need a method to further support the validity of our results. Thus, in order to check the robustness of our prediction accuracy analysis, we present the precision--recall calculations for each class across all legislatures (Table 5).[27]

Table 5. Precision-Recall

	50th		51st		52nd	
	Precision	Recall	Precision	Recall	Precision	Recall
Left	0.80	0.68	0.87	0.46	0.75	0.68
Center	0.59	0.85	0.61	0.97	0.74	0.86
Right	0.79	0.46	0.76	0.07	0.76	0.59
	53rd		54th		55th	
	Precision	Recall	Precision	Recall	Precision	Recall
Left	0.72	0.76	0.73	0.95	0.77	0.91
Center	0.73	0.85	0.83	0.80	0.71	0.85
Right	0.84	0.58	0.98	0.39	0.95	0.20

Source: Authors' results.

[27] Given these values we can calculate the f1-scores, defined as the harmonic mean of precision and recall:
$$F1 = 2\frac{P \times R}{P + R}$$

High precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).[28] The results show high precision for all classes across the different legislatures, whereas the recall is low for the Right in the 50th, 51st, 54th and 55th legislatures (despite being high for the other classes). Since speeches delivered by right-wing senators have on average the smallest distribution, the high rate of false negatives does not impact substantially the performance of our classifier (for example, it would be worse to get a low recall for the Center). This can be measured through the steady increase of our accuracy curve. This stability corroborates our prediction accuracy analysis, supporting its validity. This enables us to apply our classification model for the content analysis of each ideology.

Validation and robustness checks

The results presented above rely on the classification of senator ideology created by Power and Zucco Jr. (2009) based on data from the Brazilian Electoral Survey. In this section we test the validity and the robustness of those results by replicating the analysis with an alternative labelling process that uses a scaling algorithm called Wordfish (SLAPIN; PROKSCH, 2008). This method classifies each political party's ideological position based on their speeches. It generates an estimate for each party on the ideological line.[29] The results are shown in Figure 2 for each legislature.[30] The scaling method shows a well-distributed representation of each political party: left-wing parties are clustered on one side (e.g., PT and PSB) while right-wing parties (e.g., PTB and PP) are clustered on another side.

In Figure 3 we compare the classification using each of the methods. For the 50th and 51st legislatures, the Wordfish classification for each party matches the baseline classification (even after rescaling the results on the interval [-1,1]).

[28] A system with high recall but low precision returns is possible, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels.

[29] This parameter is a unidimensional representation of each party with respect to their speeches. For instance, if the algorithm estimates two close values (for two different parties), it means that these parties are ideologically close to each other.

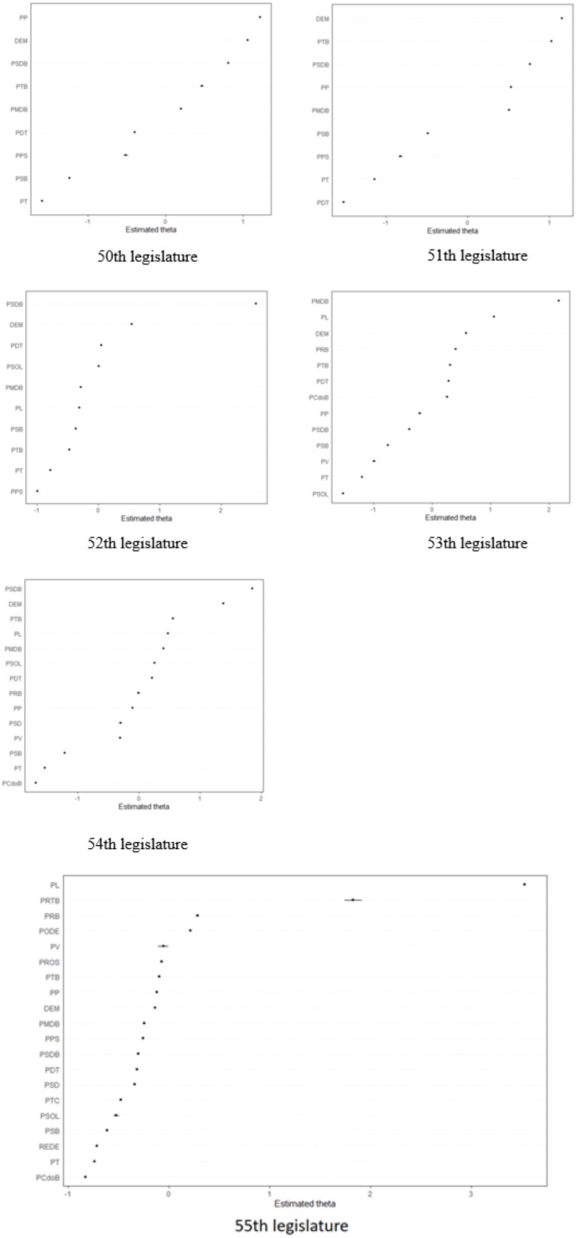
[30] We present the original output before any re-scaling. Parties which had only a small number of speeches (generally either one or two) were removed, these estimates also persistently presented a high standard deviation. In a few exceptions, a party is compounded by just one senator. They are small parties though, and their politicians are the ones that have more prevailing power (e.g. Marcello Crivella and PRB, Roberto Freire and PPS).

Note that despite these similar classifications, the position of the PSDB seems to be underestimated in the survey analysis, especially in the 52nd and 54th legislatures. For the 53rd and 55th legislatures, however, the Wordfish algorithm underestimates the PSDB's position as compared to the survey results. These results show that although the PSDB is perceptually identified on the Left side of our ideological axis (given the survey results) the opposite holds for the results from the speech analysis. In this case, the PSDB seems to be the major representative of Right-wing politics in the Senate. On the other hand, the discrepancy between the results obtained from these two methods (primarily starting from the 52nd legislature) suggests that additional dimensions are relevant for classifying party ideologies in Brazil. That is, one ideological dimension may not be sufficient to translate the differences among all these parties and between the two distinct ideologies. Previous results from Power and Zucco Jr. (2009) that used roll-call voting data show that there may be, in fact, two prevalent dimensions, the ideological (which we are already assuming) and another, representing the support and opposition to the incumbent government.

Another interesting distinction is that, starting from the 52nd legislature, many right-wing parties (such as, PTB and PP) are located near left-wing parties, especially the incumbent PT. This is likely because starting from the 52nd legislature the PT held the presidency and established a government which included some right-wing parties (often referred to as the *Centrão*). Since many of these senators became high-ranking government officials, it is natural to expect some convergence in their discourse, as they now represent the federal government itself. The Brazilian Legislative Survey, however, is also a perceptual measure, and it seems not to reflect the government's coalition dynamic. For this purpose, using speech seems to capture nuances that are not captured by other methods that rely instead on this type of data.

Note that there was a marked change starting with the 54th legislature, when the country went through a severe fiscal and incipient political crisis, weakening the federal government. If this new scenario was to change the political dynamics, this would be primarily identified in the 55th legislature. The positioning of many right-wing parties, such as PL and PRB moved even farther from the left-wing cluster. In Figure 2, moreover, PL is located at the far right of the ideological line – even further from the other right-wing parties.

Figure 2. Ideological positions of parties estimated through the Wordfish algorithm



50th legislature

51th legislature

52th legislature

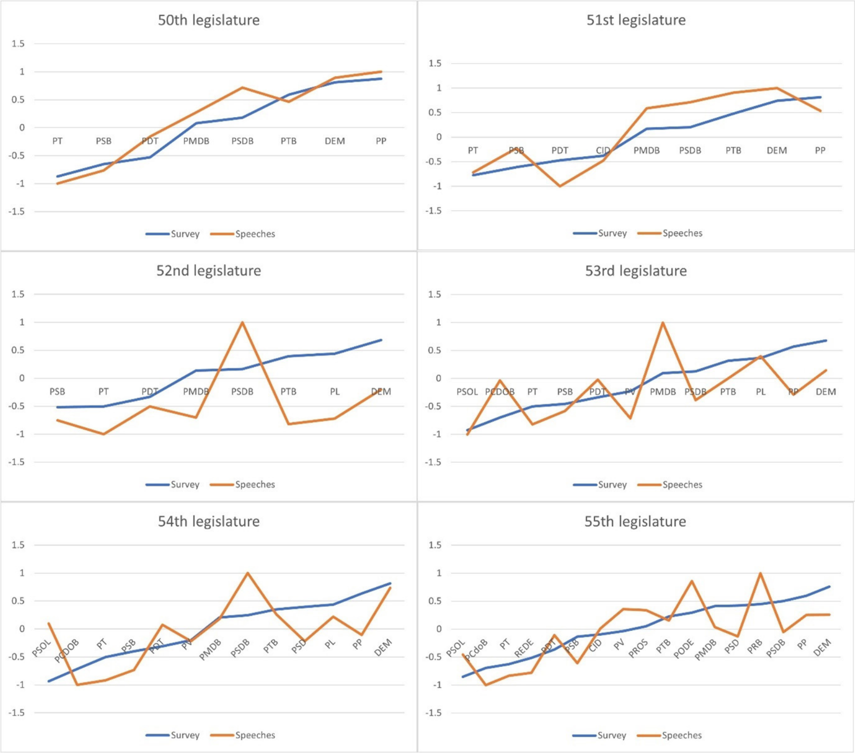
53th legislature

54th legislature

55th legislature

Source: Author's Wordfish estimation.

Figure 3. Comparison of the baseline versus Wordfish party ideology classification



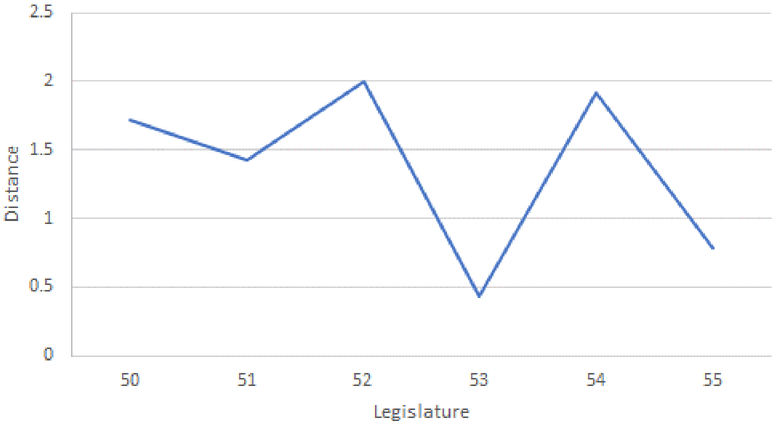
Source: Authors' calculations comparing the classifications by Power and Zucco Jr. (2009) and Wordfish.

As we have shown, the peak of polarization takes place in the 55th legislature. It turns out to be interesting to use the Wordfish estimations to better understand this phenomenon. They allow us to compare the distance between the two main Brazilian parties at this period.

By analyzing the ideological classification of the PT and the PSDB in Figure 4, we can see that the greatest distance is ascertained in the 52nd legislature. For the 53rd legislature the results in Figure 1 suggest a stabilization of polarization. This matches our findings through the Wordfish algorithm, as the distance between these two oligopolistic competitors diminished (see figure 4). Polarization increased relatively more in the 54th legislature compared with the 53rd legislature, which also matches an increase in distance between these two parties. The opposite holds in the comparison between the 54th and 55th legislatures. Despite this relationship, it is not clear whether we can explain the change in polarization only in terms of the government-opposition debate – along with the electoral competition in the background, since it has been translated here into the variation in the distance between PSDB and PT.

As we can see, until the 52nd legislature the behavior of both curves (Figures 1 and 4) are similar. But starting from the 53rd legislature, we observe a monotonic increase in polarization (through the accuracy increase in Figure 1). This suggests, in line with Power and Zucco Jr. (2009), that only part of this polarization was driven by the dichotomy and competition between these two parties. Our content analysis strengthens the hypothesis that polarization is not only a by-product of electoral competition (see next section) and that it goes beyond the rivalry between government and opposition.

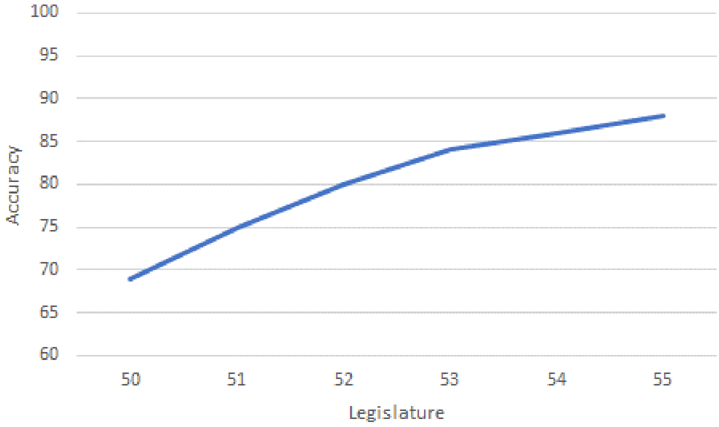
Figure 4. Ideological distance between PT and PSDB



Source: Authors' calculation using the Wordfish algorithm. *Distance is in absolute value.

Finally, we replicate our classification task (see last section) now taking as reference the Wordfish estimation. The first step is to re-scale these estimated parameters within the interval [-1,1]. Given the re-scaled values, we define those parties within the interval [-1,0) as classified on the left, while parties within the interval [0,1] are classified on the right. Note that given the challenge to defining a consistent range for the center across all legislatures, we decided to focus on these two political ideologies: left and right. Despite this difference with respect to the first classification task and although less information is being inputted in the model, we believe that our results are still valid. If the algorithm is capable of classifying speeches in a class out of three, it should be also capable of classifying one out of two classes. Figure 5 shows the results:

Figure 5. Classification accuracy on test set (Wordfish)



Source: Authors' results.

Figure 5 shows the monotonically increasing accuracy of the algorithm across legislatures, translated here as an increase in polarization across our whole period of analysis. Differently than our first classification task, between the 50th and 51st legislatures there is no decrease in accuracy. These new results confirm our previous finding that polarization has been increasing over time.

As a further robustness test, we employ an alternative classification method: Support Vector Machines (PEDREGOSA et al., 2011). In doing so we get different results: polarization achieves its peak in the 52nd legislature, but then decreases until our last period of analysis. Despite this difference, we must notice that the average accuracy is quite higher (over 95%), showing that for this analysis, the SVM is superior to the NB algorithm (though there is a non-negligible increase in the computational costs). It also increases the recall for left-wing speeches. Since the accuracy only varies marginally between legislatures, the informativeness of this measure under the SVM is not that high. This suggests that relying on the results from the NB algorithm is a better choice.

Alternative data pre-processing could be a way to extend and further validate our results (by allowing for a different data representation). For instance, in the pre-processing step we could allow for bigrams (rather than just unigrams representations of our tokens) and stemming.

Feature analysis

In this section we analyze the content of each ideological position starting from *tf-idf* representations to visualize dictionaries that best translate each ideology. We start by visualizing a selection of 15 features from each political spectrum among the top 100 features, respectively.[31][32]

Table 6. ***Tf-idf*** **feature set analysis for the 50th legislature vocabulary**

Left	Center	Right
Land	Health	Development
Reform	Northeast	Labor
People	Cities	Production
Labor	Economy	Agriculture
Agrarian	Men	Northeast
Wealth	Agrarian	Amazon
Cardoso	Land	Communication
Respect	Agriculture	Reform
Rights	Energy	Market
CPI	Plan	Producers
Children	Center	Agricultural
Women	Market	Capital
Reelection	Companies	Rural
Unemployment	Debt	Teaching
Security	Investments	Interest

Source: Authors' results.

We notice a subtle difference in the political spectrum across the first two legislatures (Tables 6 and 7). The left-wing senators speak more about

[31] These features are organized according to their *tf-idf* weights in decreasing order.
[32] See Supplementary Material for the original words in Portuguese.

rights, children and women. While centrist and right-wing senators speak more about topics related with production, using words such as debt, interest and energy. Other interesting features found through these two periods is the usage of different words to talk about similar topics, showing distinct perspectives towards the same issue. Left-wing senators talk about "agrarian" and "land" issues, while right-wing senators employ the term, "agriculture". This reveals idiosyncrasies concerning each political ideology, because, in general, rightist parties are composed of politicians derived from country's elites. On the other hand, leftist senators are more concerned with social issues, such as land reform.

We also notice in the 51st legislature the rightist senators, when addressing topics related to education, use words such as "knowledge", "culture" and "university". This represents a paradox, because, in general in Brazil, the discussion about education is the province of progressive politicians. Possibly this theme became dear to conservative senators due to their preoccupation with issues related to human capital and its impact on economic output, although there could be other interpretations.

The presidential victory of the most important left-wing leader, Lula, marks the 52nd legislature (Table 8). Concomitant with this new government we observe a substantial change in the vocabulary employed by leftist senators: now, they talk more about conservative topics such as tax and social security reforms.[33] Another difference is the appeal concerning the corruption scandal known as the *Mensalão*, that involved several politicians – many of them from PT, such as former cabinet leader, José Dirceu, and the former party president, José Genoíno.[34] Both centrist and right-wing senators have high *tf-idf* weights for CPI (*Comissão Parlamentar de Inquérito*, or Parliamentary Inquiry Commission),[35] reflecting the intent to engage in an institutional battle against the leftist incumbent government.

[33] As noted by Power and Zucco Jr. (2009): the general trend is that both parties (PT and PSDB) moved markedly to the right while in government and tended to move to the left while in opposition. Part of this shift may be due to the difference between government policies and party policies.

[34] Several politicians from other political parties were also involved in this bribery scheme, including elite members from right-wing parties such as PTB and PP.

[35] During the 52nd legislature a joint parliamentary inquiry commission was installed involving both the Chamber of Deputies and the Federal Senate to investigate the corruption scandals.

Table 7. *Tf-idf* feature set analysis for the 51th legislature vocabulary

Left	**Center**	**Right**
Wealth	Energy	Northeast
CPI	Health	Companies
Debt	Democracy	Energy
Wage	Production	Teaching
International	Economy	Agriculture
Economy	Agriculture	Production
Money	Market	Water
Responsibility	CPI	Indians
Rights	Reform	University
Crisis	Northeast	New
Reform	Teaching	Knowledge
Poverty	Violence	Culture
Children	Investment	World
Ethics	Court	Reform
Women	Water	Producers

Source: Authors' results.

Table 8. *Tf-idf* feature set analysis for the 52th legislature vocabulary

Left	**Center**	**Right**
Reform	Newspaper	CPI
Health	Stream	Northeast
Children	CPI	God
Women	Health	Opposition
Security	Folha	Safety
CPI	Growth	Wage
Rights	Bank	Police
Growth	Money	Energy
Economy	Production	Tributary
PEC	Dirceu	Growth
Land	Opposition	Companies
Tributary	Crisis	Interest
Family	Magazine	Economy
Police	Water	Court
Agrarian	Northeast	Corruption

Source: Authors' results.

Table 9. **Tf-idf feature set analysis for the 53th legislature vocabulary**

Left	Center	Right
Health	CPI	Man
Development	Crisis	God
Debate	Court	Life
Rights	Man	Northeast
School	Opposition	CPI
Wealth	Economy	Justice
Fight	History	Police
Women	Servers	Family
Retired	Money	Safety
Children	Magazine	Doctor
Universities	CPMF	Crisis
Wage	Reform	Father
Professor	Safety	*Paranaíba*
Violence	Family	Fight
Oil	Problem	Court

Source: Authors' results.

The topics regarding land conflicts, revealed by words such as agrarian and land, continue to define much of the main characteristics of leftist senators. One interesting point observed through this legislature is the antagonism between left-wing and right-wing senators when they talk about family: as revealed by our *tf-idf* weights, left-wings senators refer to family referring to women and children, while right-wing senators favor references to men. The reference to women by left-wing senators may also be related to women's rights and discrimination. We also notice the emergence of high *tf-idf* weights related to right-wing vocabulary such as "God", "safety" and "police" – reflecting a standard conservative pattern for Brazil.

The 52nd legislature represents an important inflection point in which there was a change in the topics discussed. Power and Zucco Jr. (2009) pointed out that incumbent politicians tend to have a more conservative agenda, which showed up in our analysis as discussions regarding structural reforms. However, the political drift followed by the PSDB as it moved out of office is not easily perceived through this feature analysis. We see that the centrist politicians employ a "blurry" vocabulary, sometimes adopting more conservative words and sometimes adopting more progressive jargon. In general, we find that their focus was on issues related to corruption. The observed polarization is thus primarily over electoral issues and political scandals involving the government than it is about issues concerning public policy.

Analyzing the 53rd legislature (Table 9), we notice a certain lack of variability. The trend measured through the feature analysis indicates that topics of discussion in general are the same, although there are subtle differences. The first is the adoption of a more progressive vocabulary by left-wing senators, as they started to talk about education, school, university, professors and wage. Secondly, the leftist party members seemed to dodge discussions involving the corruption scandals, which the others continued to pursue. Interestingly, we already see early signs of themes that will become central in later legislatures, such as God, family and police.

The 54th and the 55th legislatures are truly distinctive – Tables 10 and 11, respectively. While this first legislature was marked by the continuity of the PT in power – through Dilma Rousseff's victory – with certain economic stability, the last legislature was marked by her impeachment and by a deep economic crisis. These facts are easily perceived through the results of our feature analysis.

While the 54th legislature exhibits a continuity with topics previously discussed by senators, the subsequent legislature exposed the intense conflict that came to prevail. It can be perceived by the use of words such as coup, impeachment and justice. The general motivation of this political battle is expressed through words such as "fiscal" and "socialism", because according to political opinion the principal threats to democracy were the crimes of responsibility (in general, in the fiscal area) and the slant towards a "socialist" agenda. An interesting feature that can be seen here is the proximity between the *tf-idf* weight values for the words Dilma and Lula. That is, when senators initiated a discussion about the Dilma government, they almost always invoked the presence of former president Lula, recognizing his role in putting her in power as well as his influence over her. This pattern is only perceived at this point in the analysis.

Table 10. *Tf-idf* feature set analysis for the 54th legislature vocabulary

Left	Center	Right
Women	Minority	Laborite[36]
Rights	Economy	Strength
Violence	Court	Money
Wealth	Corruption	Doctors
Children	Power	Family
Northeast	Supreme	Safety
Liberty	Document	Energy
Growth	Truth	Industry
Socialist	Growth	Producers
Energy	Cities	Man
Doctors	Companies	God
School	Production	Police
Reform	Opposition	Women
University	Works	Frontier
Plan	CPI	Infrastructure

Source: Authors' results.

Table 11. *Tf-idf* feature set analysis for the 55th legislature vocabulary

Left	Center	Right
Women	Women	Security
Impeachment	Money	Production
Socialism	Socialism	Money
Reform	Energy	Agriculture
Coup	Progressive	Family
Rights	Water	Companies
Violence	Justice	Police
Crime	Security	Rights
Fight	Impeachment	Laborite
PEC	Family	Workers
Fiscal	Workers	Impeachment
Security	Fiscal	Democratic
Michel	Interest	Economic

Source: Authors' results.

[36] Member or supporter of the Worker's Party (*trabalhista*).

After Dilma's impeachment, her vice president assumed the government and tried to implement a reformist agenda. This appears through use of words such as "PEC" (*Proposta de Emenda à Constituição*, or Proposed Amendment to the Constitution) and "security" – where this last expression was employed across the political spectrum. As the Brazilian economy was impacted by a terrible crisis in the 55th legislature, the need to discuss topics related to the economy is revealed by this feature analysis.

There are a variety of approaches and methods that produce multi-dimensional representations of political actors' ideologies and preferences. Although these studies tend to measure several dimensions, the first dimension on its own usually predicts much of the variability in ideology (POOLE; ROSENTHAL, 1985). Typically, that first dimension maps into the right-left divide. Given the strength of the first dimension, many studies simply ignore additional dimensions which add little explanatory power. Another reason the second dimension is often side-lined is due to the difficulty of identifying exactly what it measures. Different studies in different contexts have variously identified the second dimension as referring to moral issues such as race, abortion, gun rights or many others (DIERMEIER et al., 2011; EVANS, 2003). Given their text-based nature, the approach we use in this paper is particularly well-suited to uncovering the nature of these other dimensions.

Conclusion

As quantitative text analysis is increasingly used in political science, many long-standing debates in the Brazilian political science literature have been reenergized. In this paper we use different text-based methods to analyze over 85,000 speeches by Brazilian Senators from 1995 to 2018, to contribute to our understanding of issues such as polarization and the identification of ideology and legislative organization.

We used a supervised learning technique that classified each speech as left, center or right. This was done using the survey-based classification by Power and Zucco Jr. (2009) as the "true" ideological position of legislators. Based on the intuition that greater classification accuracy for the speeches in each legislature is an indicator of greater polarization, we identified a clear increase in polarization starting from the 52nd legislature.

We also addressed the ongoing debate in Brazilian political science on whether the legislature should be understood as self-organized according to a left-right ideological pattern or to government-opposition standings. As noted, our classifier is based on the left-right classification by Power and Zucco Jr. (2009). However, when validating our results with the Wordfish scaling algorithm, we found that our classification is compatible with this alternative ranking of parties for the first two legislatures (50th and 51st), but less so for the subsequent ones, as some traditionally left/right wing parties are placed in unexpected positions. These mixed results may be due to a misplaced reliance on the Power and Zucco Jr. (2009) classification for capturing ideology, but they might similarly or simultaneously be due to the unidimensional limitations of the Wordfish scaling algorithm, as it is clearly possible that more than one dimension determines speech. Text based methods open up countless new sources of data that allow different ways of testing hypotheses, but much work is still needed to establish better how these will affect our understanding of Brazilian political organization.

References

ALSTON, L. J.et al. Changing social contracts: beliefs and dissipative inclusion in Brazil. **Journal of Comparative Economics**, v. 41, n. 1, p. 48-65, 2013.

ALSTON, Lee J.; MUELLER, Bernardo. Pork for policy: executive and legislative exchange in Brazil. **Journal of Law, Economics, and Organization**, v. 22, n. 1, p. 87-114, 2006.

AMES, Barry. **The deadlock of democracy in Brazil**. University of Michigan Press, 2002.

AMORIM NETO, Octavio. Gabinetes presidenciais, ciclos eleitorais e disciplina legislativa no Brasil. **Dados**, Rio de Janeiro, v. 43, n. 3, p. 479-519, 2000.

AMORIM NETO, Octavio; SANTOS, Fabiano. A conexão presidencial: facções pró e antigoverno e disciplina partidária no Brasil. **Dados**, Rio de Janeiro, v. 44, n. 2, p. 291-321, 2001.

ARNOLD, C.; DOYLE, D.; WIESEHOMEIER, N. Presidents, policy compromise, and legislative success. **The Journal of Politics**, v. 79, n. 2, p. 380-395, 2017.

BISHOP, C. **Pattern recognition and machine learning**. Springer, 2006.

CONVERSE, P. The nature of belief systems in mass publics. *In*: APTER, David E. **Ideology and discontent**. New York: Free Press of Glencoe, 1964.

DIERMEIER, D. et al. Language and ideology in Congress. **British Journal of Politics**, v. 42, n. 1, p. 31-55, 2011.

EVANS, J. Have Americans' attitudes become more polarized? – an update. **Social Science Quarterly**, v. 84, n. 1, p. 71-90, 2003.

FERREIRA, Pedro Fernando Almeida Nery; MUELLER, Bernardo. How judges think in the Brazilian Supreme Court: estimating ideal points and identifying dimensions. **EconomiA**, Amsterdam, v. 15, n. 3, p. 275-293, 2014.

FIGUEIREDO, Argelina C.; LIMONGI, Fernando M. P. **Executivo e Legislativo na nova ordem constitucional**. Rio de Janeiro: Editora FGV, 1999.

FIGUEIREDO, Argelina C.; LIMONGI, Fernando. Poder de agenda na democracia brasileira: desempenho do governo no presidencialismo pluripartidário. *In*: SOARES, Glaucio A. D.; RENNÓ JUNIOR, Lucio R. (orgs.). **Reforma política**: lições da história recente. Rio de Janeiro: Fundação Getúlio Vargas, 2006. p. 249-280.

GENTZKOW, M.; KELLY, B.; TADDY, M. Text as data. **NBER Working Paper**, n. 23276, 2017.

GENTZKOW, M.; SHAPIRO, J.; TADDY, M. Measuring group differences in high-dimensional choices: method and application to congressional speech. **NBER Working Paper**, n. 22423, 2019.

GRIMMER, J.; STEWART B. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. **Political Analysis**, v. 21, n. 3, p. 267-297, 2013.

HARE, C.; POOLE, K. The polarization of contemporary American politics. **Polity**, v. 46, n. 3, p. 413-429, 2014.

IZUMI, M.Y. Governo e oposição no Senado Brasileiro (1989-2010). **Dados**, Rio de Janeiro, v. 59, n. 1, p. 91-138, 2016.

IZUMI, M.Y.; MEDEIROS, D. B. Government and opposition in legislative speechmaking: using text-as-data to estimate Brazilian political parties' policy positions. **Latin American Politics and Society**, v. 63, n. 1, p. 145-164, 2021.

JURAFSKY, D.; MARTIN, J. **Speech and language processing**: an introduction to natural language processing, computational linguistics, and speech recognition. Third Edition draft, 2020.

KIM, I.; LONDREGAN, J.; RATKOVIC, M. Estimating spatial preferences from votes and text. **Political Analysis**, v. 26, n. 2, p. 210-229, 2018.

LAMOUNIER, Bolivar. **Partidos e utopias?** O Brasil no limiar dos anos 90. São Paulo: Loyola, 1989.

LAUDERDALE, B.; HERZOG, A. Measuring political positions from legislative speech. **Political Analysis**, v. 24, n. 3, p. 374-394, 2016.

LAVER, M.; BENOIT, K.; GARRY J. Extracting policy positions from political texts using words as data. **American Political Science Review**, v. 97, n. 2, p. 311-331, 2003.

LE, Q.; and MIKOLOV, T. Distributed representations of sentences and documents. *In*: XING, E. P.; JEBARA, T. (eds.). Proceedings of the 31st International Conference on Machine Learning, v. 32. PMLR, 22-24 jun. 2014. p. 1188-1196. Disponível em: http://proceedings.mlr.press/v32/le14.html Acesso em: 01 jan. 2021.

LEONI, Eduardo. Ideologia, democracia e comportamento parlamentar: a Câmara dos Deputados (1991-1998). **Dados**, Rio de Janeiro, v. 45, n. 3, p. 361-386, 2002.

LEVITSKY, S.; ZIBLATT, D. **How democracies die**. New York: Crown, 2018.

LIMONGI, Fernando M. P.; FIGUEIREDO, Argelina C. A crise atual e o debate institucional. **Novos Estudos CEBRAP**, São Paulo, v. 36, n. 3, p. 79-97, 2017.

MAINWARING, Scott; SCULLY, Timothy (eds.). **Building democratic institutions**: Party systems in Latin America. Stanford University Press, 1995.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge, UK: Cambridge University Press, 2008.

MCCARTY, N. M.; POOLE, K. T.; ROSENTHAL H. **Income redistribution and the realignment of American politics**. AEI Press, 1997.

MOREIRA, D. Com a palavra os nobres deputados: ênfase temática dos discursos dos parlamentares brasileiros. **Dados**, Rio de Janeiro, v. 63, n. 1, p. 1-37, 2020.

MORGENSTERN, Scott. **Patterns of legislative politics**: roll-call voting in Latin America and the United States. Cambridge, UK: Cambridge University Press, 2003.

NEIVA, R.P.R. Disciplina partidária e apoio ao governo no bicameralismo brasileiro. **Revista de Sociologia e Política**, Florianópolis, v. 19, n. 39, p. 183-196, jun., 2011.

NEIVA, P.R.P.; SOARES, M.M. Senado: casa federativa ou partidária? **Revista Brasileira de Ciências Sociais**, São Paulo, v. 28, n. 81, p. 97-115, fev. 2013.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. **Foundations and Trends® in Information Retrieval**: v. 2, n. 1-2, p. 1-135, 2008.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.

PEREIRA, Carlos; MUELLER, Bernardo. Comportamento estratégico em presidencialismo de coalizão: as relações entre Executivo e Legislativo na elaboração do orçamento brasileiro. **Dados**, Rio de Janeiro, v. 45, n. 2, p. 265-301, 2002.

PEREIRA, Carlos; MUELLER, Bernardo. Partidos fracos na arena eleitoral e partidos fortes na arena legislativa: a conexão eleitoral no Brasil. **Dados**, Rio de Janeiro, v. 46, n. 4, p. 735-771, 2003.

PETERSON, A.; and SPIRLING, A. Classification accuracy as a substantive quantity of interest: measuring polarization in Westminster systems. **Political Analysis**, v. 26, n. 1, p. 120-128, 2018.

POOLE, K. Changing minds? Not in congress. **Public Choice**, v. 131, n. 3-4, p. 435-451, 2007.

POOLE, K.; and ROSENTHAL, H. A spatial model for legislative roll call analysis. **American Journal of Political Science**, v. 29, n. 2, p. 357-384, 1985.

POWER, T.; ZUCCO JR., C. Estimating ideology of Brazilian legislative parties, 1990-2005. **Latin American Research Review**, v. 44, n. 1, p. 218-246, 2009.

QUINN, K. et al. How to analyze political attention with minimal assumptions and costs. **American Journal of Political Science**, v. 54, n. 1, p. 209-228, 2010.

SAMUELS, David. **Ambition, federalism, and legislative politics in Brazil**. Cambridge, UK: Cambridge University Press, 2003.

SHUGART, Matthew Soberg; MAINWARING, Scott. Presidentialism and democracy in Latin America: rethinking the terms of the debate. *In:* MAINWARING, Scott**;** SHUGART, Matthew Soberg (eds.). **Presidentialism and democracy in Latin America**. Cambridge, UK: Cambridge University Press, 1997. p. 12-54.

SHUGART, Matthew Soberg; CAREY, John M. **Presidents and assemblies**: constitutional design and electoral dynamics. Cambridge, UK: Cambridge University Press, 1992.

SLAPIN, J.B.; PROKSCH, S. A scaling model for estimating time-series party positions from texts. **American Journal of Political Science**, v. 52, n. 3, p. 705-722, 2008.

YU, B.; KAUFMANN, S.; DIERMEIER, D. Classifying party affiliation from political speech. **Journal of Information Technology Politics**, v. 5, n. 1, p. 33-48, 2008.

A natural language measure of ideology in the Brazilian Senate

Abstract: We estimate a measure of political ideology using as data a corpus of over two decades of speeches delivered by Brazilian Federal Senators across five legislatures. We employ a computational technique that analyses political speech by extracting the dictionaries that best translate the content of each ideology. Through this supervised learning method, we calculate the classification accuracy over these political texts and show that polarization is increasing across the legislatures. The method also reveals the evolving patterns of political ideologies over a period of deep change in Brazilian society. We further investigate the political dynamic across legislatures by comparing our results with current approaches to ideology in the political science literature.

Keywords: ideology, language, machine learning, polarization.

Uma medida de linguagem natural de ideologia no Senado Brasileiro

Resumo: Este trabalho usa um *corpus* de mais de duas décadas de discursos proferidos por senadores federais brasileiros em cinco legislaturas para estimar uma medida de ideologia política. Nós empregamos uma técnica computacional que extrai do discurso político os dicionários que melhor traduzem o conteúdo de cada ideologia. Por meio desse método de aprendizado supervisionado, calculamos a precisão da classificação sobre esses textos políticos e medimos o aumento da polarização nas legislaturas. O método também é usado para contribuir para o debate sobre a mensuração de ideologia partidária no Brasil ao investigar a evolução dos padrões de ideologias políticas durante um período de profundas mudanças na sociedade brasileira.

Palavras-chave: ideologia, linguagem, aprendizado por máquina, polarização.

Submetido em 15 de dezembro de 2020
Aprovado em 19 de novembro de 2021.