Adriana Miranda Molinari

Aprendendo com Discursos: Uma Análise em Alta Dimensão da Ideologia e Polarização Política na Câmara dos Deputados Federais do Brasil

Brasília - DF, Brasil

Agosto de 2020

Adriana Miranda Molinari

Aprendendo com Discursos: Uma Análise em Alta Dimensão da Ideologia e Polarização Política na Câmara dos Deputados Federais do Brasil

Dissertação apresentada ao Curso de Mestrado Acadêmico em Economia, Universidade de Brasília, como requisito parcial para obtenção de título de Mestre em Economia.

Universidade de Brasília - UnB

Faculdade de Administração, Contabilidade e Economia - FACE

Departamento de Economia - ECO

Programa de Pós-Graduação

Orientador: Prof. Dr. Daniel Oliveira Cajueiro

Brasília - DF, Brasil

Agosto de 2020

Adriana Miranda Molinari

Aprendendo com Discursos: Uma Análise em Alta Dimensão da Ideologia e Polarização Política na Câmara dos Deputados Federais do Brasil/ Adriana Miranda Molinari. – Brasília - DF, Brasil, Agosto de 2020-

36p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Daniel Oliveira Cajueiro

Dissertação (Mestrado) – Universidade de Brasília - UnB
Faculdade de Administração, Contabilidade e Economia - FACE
Departamento de Economia - ECO
Programa de Pós-Graduação, Agosto de 2020.

1. Polarização Política 2 Ideologia 3. Aprendizagem de Máquina 4. Processamento de Linguagem Natural I. Orientador: Prof. Dr. Daniel Oliveira Cajueiro II. Universidade de Brasília. III. Faculdade de Administração, Economia e Contabilidade. IV. Departamento de Economia.

Adriana Miranda Molinari

Aprendendo com Discursos: Uma Análise em Alta Dimensão da Ideologia e Polarização Política na Câmara dos Deputados Federais do Brasil

Dissertação apresentada ao Curso de Mestrado Acadêmico em Economia, Universidade de Brasília, como requisito parcial para obtenção de título de Mestre em Economia.

Prof. Dr. Daniel Oliveira Cajueiro
Orientador

Prof. Dr. Bernardo Pinheiro Machado Mueller
Membro interno

Prof. Dr. Marcus Melo
Membro externo

Brasília - DF, Brasil
Agosto de 2020

Agradecimentos

Ao meu orientador, Daniel Oliveira Cajueiro, pelas contribuições a este trabalho e por todo aprendizado ao longo do mestrado.

Aos professores do Departamento de Economia da UNB pela seriedade, comprometimento e pela excelente formação que me proporcionaram.

Aos amigos que que me acompanharam ao longo dessa jornada. Certamente a estrada foi mais leve pela amizade de vocês.

Aos meus pais, José e Maria, que apesar de todas as dificuldades e sacrifícios sempre priorizaram a educação dos filhos. Aos meus irmãos, Paulo e Flávio, vocês são fonte de profunda inspiração, admiração e companheirismo.

Ao meu companheiro Thiago. Obrigada por ser fortaleza, pelo apoio incondicional e por escolher os caminhos que nos trouxe até aqui.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Este trabalho analisa os padrões ideológicos e qualifica a polarização política na Câmara dos Deputados do Brasil entre a 52ª e a 55ª legislatura. Para tanto, dispomos de uma base de dados com 317.980 discursos dos deputados federais, na qual utilizamos métodos de processamento de linguagem natural e aprendizagem supervisionada. Encontramos que esquerda, centro e direita usam palavras diferentes ao tratarem dos mesmos tópicos, além de divergirem no peso dado à determinadas palavras, o que mostra que a escolha dos tópicos também se relaciona com os padrões ideológicos. Ainda, utilizamos a acurácia de um modelo de classificação e a similaridade entre cossenos como medida de polarização. Neste caso, concluímos que houve um aumento da polarização política entre a 52ª e a 55ª legislatura.

Palavras-chave: 1. Polarização Política 2. Ideologia 3. Aprendizagem de Máquina 4. Processamento de Linguagem Natural

Abstract

This article analyzes the ideological patterns and the political polarization in the Brazil Chamber of Deputies, between the 52nd and 55th legislature. We have a database with 317,980 federal deputies speeches, and we use natural language processing and supervised learning. We found that left, center, and right use different words when handling with the same topics. Also, they diverge in the weight that they give to certain words, which shows that the topic choices are also related to ideological patterns. Moreover, we use the accuracy of a classification model and the similarity between cosines as a measure of polarization. In this case, we conclude that there was an increase in political polarization between the 52nd and 55th legislatures.

Keywords: 1. Political Polarization 2. Ideology 3. Machine Learning 4. Natural Processing Language

Lista de tabelas

Sumário

1 Introdução

A última década foi marcada por uma maior percepção da polarização política no debate público brasileiro. Entretanto, uma pergunta relevante é se de fato essa distância entre os polos ideológicos se traduziu em um aumento da divergência na linguagem empregada pelos políticos.

Medir as posições ideológicas que os partidos e políticos adotam é peça fundamental para construir e testar teorias relacionadas à questões intra-partidárias, representação, elaboração de políticas e polarização (Lauderdale e Herzog (2016)). Entretanto, esta sempre foi uma tarefa difícil. Ao contrário de afiliação partidária, ideologia não é uma variável diretamente observável. Consequentemente, pesquisadores já empregaram diferentes estratégias para mensurar posicionamento ideológico (Diermeier et al. (2012)).

Votos dos políticos e questionários são exemplos de técnicas utilizadas. Poole e Rosenthal (1985) desenvolveram uma medida de ideologia baseada em *roll call votes*[1] que é amplamente utilizada. Já Power e Jr (2009) utilizam questionários para examinar as preferências ideológicas da elite política brasileira. A maior disponibilidade de dados e recursos computacionais colocou novas estratégias à disposição. Diermeier et al. (2012) parte dos discursos do senado americano com o objetivo de entender o conteúdo político das diferentes correntes ideológicas. O objetivo dos autores não é atribuir uma classificação no espaço ideológico para os políticos, para tanto eles usam a medida criada por Poole e Rosenthal (1985), mas sim colocar luz sobre a estrutura interna das ideologias usando um método de aprendizagem supervisionada.

Abordagem semelhante é usada na disserta;'ao de mestrado de Figueredo (2019) com os discursos do senado brasileiro. Neste caso, o autor parte da hipótese que ideologias funcionam como um sistema de crenças, ou seja, os padrões que os políticos usam na comunicação atuam como uma restrição. De fato, a dissertação em questão mostra que senadores de esquerda e direita usam vocabulários distintos para abordar diferentes tópicos, como economia, direitos humanos e educação.

Dados textuais também motivaram diversos trabalhos que tentam medir polarização. Gentzkow, Shapiro e Taddy (2019) partem de discursos do congresso americano e tratam a questão como um problema de escolha de diferentes grupos em um conjunto de alta dimensão. Para os autores, discursos e *roll call votes* respondem a diferentes incentivos e restrições. Portanto, a análise de discursos pode revelar aspectos do cenário político que não ficam evidentes em análises que usam *roll call votes*. Os autores afirmam que o

[1] *Roll Call Vote* é um tipo de votação em que o congressista é chamado pelo nome e responde sim ou não. Portanto, a escolha do congressista é pública e fica gravada.

aumento da polarização no congresso dos Estados Unidos em anos recentes se deve mais ao fato de mudanças na linguagem empregada para discutir um tópico específico do que propriamente na escolha dos tópicos.

Peterson e Spirling (2018) treinam modelos de aprendizagem supervisionada com discursos do Parlamento Britânico, e em seguida usam os mesmos algoritmos para prever se novos discursos pertencem à políticos do partido trabalhista ou do partido conservador. A acurácia[2] das previsões é então usada como uma medida de polarização, ou seja, em momentos em que o parlamento está mais polarizado os discursos divergem e a acurácia dos modelos é mais alta. Já em momentos de maior convergência, os discursos se aproximam e a acurácias dos modelos diminui.

O objetivo deste trabalho é acessar os padrões ideológicos e a polarização na Câmara dos Deputados do Brasil. Para tanto, utilizamos uma base de dados composta por 317.980 discursos dos deputados federais entre a 52ª e a 55ª legislatura[3]. Por meio dos padrões subjacentes aos discursos, queremos entender os arquétipos associados às diferentes correntes ideológicas do espectro político brasileiro, além de qualificar a extensão da polarização política ao longo dos últimos anos.

Para acessar o conteúdo ideológico latente aos discursos, classificamos todos os discursos entre esquerda, centro e direita. A classificação é baseada no trabalho de Zucco e Power (2019). Em seguida, utilizamos um método de contagem de palavras (Term Frequency - Inverse Document Frequency - Tf-Idf) para construir vetores de vocabulários para cada corrente ideológica. Esta etapa se baseia especialmente nos trabalhos de Diermeier et al. (2012) e Figueredo (2019).

Ainda, utilizamos um método de aprendizagem supervisionada com o objetivo de validar os vocabulários formados e utilizar a acurácia do modelo como uma medida de polarização, seguindo Peterson e Spirling (2018). Por fim, este trabalho tenta inovar ao propor a utilização de uma medida de similaridade de vetores como um indicativo de polarização política.

Em linhas gerais, todas as etapas propostas indicam um aumento da polarização política na Câmara dos Deputados Federais entre a 52ª e a 55ª legislatura, sendo que a inflexão ocorre na 54ª legislatura. A análise dos vocabulários apontam para uma maior divergência dos discursos de esquerda e direita na 54ª e 55ª legislatura. Além disso, notamos uma aproximação do centro e a direita neste mesmo período. Este resultado é consistente com a acurácia do algoritmo de aprendizagem supervisionada, que aumenta significativamente a partir da 54ª legislatura, indicando uma cenário político mais polarizado.

[2] Acurácia: número de previsões corretas dividido pelo número total de previsões

[3] Período legislaturas: 52ª de 1º de fevereiro de 2003 à 31 de janeiro de 2007; 53ª de 1º de fevereiro de 2007 à 31 de janeiro de 2011; 54ª de 1º de fevereiro de 2011 à 31 de janeiro de 2015; 55ª de 1º de fevereiro de 2015 à 31 de janeiro de 2019

Por fim, o cosseno entre os vetores de vocabulários formados pelo Tf-Idf mostra um crescimento da distância entre esquerda e direita consistente em todo o período analisado, mas com um maior aprofundamento entre a partir 54ª legislatura. Além disso, segundo essa medida, centro e direita possuem vocabulário muito próximo, sendo que o centro começa a se distanciar da esquerda também a partir da 54ª legislatura.

Este trabalho segue a seguinte estrutura. No capítulo 2, fazemos uma breve revisão de literatura acerca de trabalhos que usam texto como dado. No capítulo 3, detalhamos a preparação da base de dados e as metodologias utilizadas para acessar os padrões ideológicos e qualificar a polarização política. No capítulo 4, apresentamos e comentamos os resultados. Por fim, no capítulo 5 fazemos uma conclusão.

2 Revisão de Literatura

Um dos primeiros trabalhos modernos que usa texto como dado remonta a Mosteller e Wallace (1963). Os autores tentam inferir a autoria de artigos escritos por Alexander Hamilton e James Madison na coleção *The Federalist Papers*. Mais recentemente, novas tecnologias contribuíram para que enormes quantidades de textos ficassem disponíveis digitalmente. Dessa forma, a informação contida nos textos tornou-se um rico complemento a dados mais estruturados e tradicionalmente usados em pesquisas.

Em economia podemos hoje encontrar uma variedade de aplicações do uso de texto como dado. Uma rica linha de pesquisa utiliza a comunicação dos bancos centrais para responder diversas questões. Alguns trabalhos focam em entender o impacto destas comunicações no retorno dos ativos, como Lucca e Trebbi (2009), que usam o conteúdo das declarações do *Federal Open Market Committee* (FOMC) para prever flutuações nos títulos do tesouro americano, e Born, Ehrmann e Fratzscher (2014), que a partir de relatórios de estabilidade financeira e discursos de 37 banco centrais constroem um índice de sentimentos a fim de entender o retorno de ativos financeiros.

Ainda no contexto de comunicação dos bancos centrais, Hansen, McMahon e Prat (2018) usam algoritmos de processamento linguagem para entender como a transparência afeta as deliberações dos membro do comitê de política monetária nos Estados Unidos. A abordagem dos autores parte de uma experimento natural, uma vez que até novembro de 1993 os encontros do FOMC eram secretos, mas um tempo depois os encontros passaram a ser públicos e as transcrições dos encontros anteriores foram disponibilizadas ao público. Os autores encontram uma grande mudança nos padrões de comunicação após a transaparência dos encontros. Se por um lado os membros do comitê começaram a se preocupar mais com a qualidade do debate e com a eficiência da comunicação, por outro passaram a ser mais cautelosos em relação à ideias inovadoras, o que leva a um viés em direção ao pensamento hegemônico.

Uma linha de pesquisa em ascensão utiliza dados de texto para estimativas em tempo real de diversas variáveis. Um dos trabalhos mais importantes deste contexto remonta a Scott e Varian (2015). Os autores usam termos buscados no Google para prever dados regionais. As buscas são agrupadas por semana e localização e em seguida agregadas em índices geográficos, os quais são utilizaddos para a previsão de diversas variáveis.

Os discursos políticos são uma fonte de pesquisas com diferentes perguntas. Entender como novas ideias surgem é a pergunta feita por Barron et al. (2018). Os autores utilizam transcrições de debates e discursos proferidos no parlamento francês durante a Revolução Francesa para compreender como as inovações irrompiam naquele contexto. Já Gennaro,

Lecce e Morelli (2019) aplicam análise textual nos discursos da campanha presidencial americana de 2016 e das eleições de meio de mandato de 2018 a fim entender em que medida os políticos se valem do populismo para maximizar o impacto das campanhas.

Ainda no contexto político, muitas pesquisas focam na comunicação de parlamentares. Neste sentido, Gentzkow, Shapiro e Taddy (2019) usam os discursos do congresso americano entre 1873 e 2016 para medir a polarização política. Os autores encontram que a polarização é maior em anos mais recentes do que no passado, além de ter aumentado acentuadamente no início dos anos 90, após permanecer baixa e relativamente constante ao longo do século 20.

Também com o objetivo de quantificar polarização, Peterson e Spirling (2018) utilizam discursos de 78 anos do Parlamento Britânico. O artigo propõe utilizar a acurácia dos modelos de classificação com uma medida de polarização política. Os autores encontram um aumento da polarização no parlamento entre o final dos anos 70 e o final dos anos 80. Ademais, os autores concluem que a acurácia como uma medida de polarização terá mais sentido em contextos em que os partidos discutem os mesmos tópicos mas usam vocabulários diferentes, além de propor utilização desta medida como uma variável informativa sobre questões relevantes nas ciências sociais.

Já Diermeier et al. (2012) usam discursos dos senado dos Estados Unidos para compreender os padrões ideológicos da política americana. Abordagem similar é feita por Figueredo (2019), entretanto utilizando discursos do senado brasileiro.

Este trabalho se baseia em diversos artigos da literatura citada, entretanto seguimos o padrão de Diermeier et al. (2012) e Figueredo (2019) com o objetivo de identificar os padrões ideológicos nos discursos da Câmara dos Deputados do Brasil. Além disso, repousamos na proposta de Peterson e Spirling (2018) ao tentar quantificar a polarização política na câmara. Por fim, propomos uma nova medida de polarização baseada na similaridade de cossenos, sendo que o arcabouço teórico consta em Jurafsky e Martin (2019).

3 Metodologia

Seguindo Diermeier et al. (2012) e Figueredo (2019), duas etapas foram propostas para a análise textual. A primeira consiste na formação de vetores de dicionários de palavras para o conjunto de discursos de cada corrente ideológica analisada em uma dada legislatura, centro, esquerda e direita. Os dicionários foram ordenados de acordo com a relevância das palavras, que é dada pela aplicação de um método de contagem, *term frequency – inverse document frequency*. Já a segunda etapa, propõe a utilização de um modelo de aprendizagem supervisionada, cujo objetivo é validar a relevância dos dicionários formados para explicar os padrões ideológicos do espectro político brasileiro e acessar acurácia do modelo como uma medida de polarização política.

Por fim, apresento uma terceira etapa a fim de quantificar a polarização no período analisado. O objetivo é usar a similaridade entre os vetores de dicionários formados na primeira etapa como medida de divergência entre as correntes ideológicas analisadas.

3.1 Base de Dados

A base de dados utilizada é composta por 317.980 discursos da Câmara dos Deputados Federais do Brasil, proferidos entre a 52ª e a 55ª legislatura. Embora seja uma fase da política brasileira definida pela presença quase única do PT na Presidência da República, com exceção dos anos entre 2016 e 2018, o período em questão foi marcado pela convergência e afastamento entre executivo e legislativo. De fato, a convergência entre os poderes observada nos anos Lula foi seguida de um maior afastamento nos anos Dilma, com um evidente aumento do acirramento político entre os polos ideológicos de esquerda e direita.

Para cada legislatura, o conjunto total de discursos é dado pelo documento D_l, tal que $l \in \{52, 53, 54, 55\}$, ou seja, l representa a legislatura em questão. Além disso, cada documento D_l é subdividido em documentos d_i, sendo que $d_i \in R^N$ e $i = 1, 2, ..., n$, onde N corresponde ao número total de palavras e n ao total de discursos da legislatura l.

3.1.1 Reduzindo a dimensionalidade dos dados

Para análise textual uma série de técnicas são utilizadas com o objetivo de pré processar o texto. A ideia é reduzir o número de elementos textuais considerados e, portanto, a dimensionalidade dos dados, o que, segundo Gentzkow, Kelly e Taddy (2019), proporciona uma grande benefício computacional, além de ser um ponto chave na construção de modelos com melhor ajuste e maior interpretabilidade.

O principal tratamento empregado foi a remoção de determinados elementos textuais. Nesse sentido, o primeiro passo foi remover pontuações, números e tags de HTML. Em seguida, removemos nomes dos deputados federais, siglas partidárias, siglas estaduais e combinações das siglas partidárias com as siglas estaduais, o que é de extrema relevância no contexto deste trabalho a fim de evitar algum tipo de viés do modelo de aprendizagem supervisionada.

Além disso, removemos palavras muito comuns, como preposições, artigos e formas de verbos usuais, como ser, estar e haver. Palavras muito comuns são usualmente chamadas de *stop words*, as quais possuem grande relevância para a a estrutura gramatical de sentenças, mas que guardam pouco significado se analisadas fora de contexto, de acordo com Gentzkow, Kelly e Taddy (2019). Dessa forma, também optamos por remover palavras muito frequentes nos textos analisados, como Sr, Sra, Exº, Exª, ordem e orador, além de palavras que aparecem em 85% dos documentos D_i de uma dada legislatura.

Ademais, ainda com o objetivo de reduzir a dimensionalidade dos dados, removemos as palavras cuja frequência é menor do que 50 considerando o conjunto de documentos de cada legislatura, dessa forma evitamos palavras com pouco significado ou com erros de digitação.

3.1.2 Classificando os partidos

Nos modelos de aprendizagem supervisionada cada conjunto de dados é uma coleção de exemplos categorizados. Dessa forma, classificamos cada documento d_i de acordo com a atribuição ideológica do deputado que protagonizou o discurso contido em d_i.

A classificação de cada deputado corresponde à classificação do seu partido no padrão estabelecido por Zucco e Power (2019). Assim, para cada documento d_i atribuímos uma classificação c_i[1], tal que $c_i \in \{-1, 0, 1\}$, sendo -1 para os partidos de esquerda, 0 centro e 1 direita.

3.2 Term Frequency - Inverse Document Frequency - Tf-Idf

O conjunto de documentos D_l de uma dada legislatura pode ser representado por uma matriz numérica C_l, sendo que essa representação pode ser feita usando a ponderação dada pelo Tf-Idf.

Para uma palavra j no documento d_i a *term frequency* (tf_{ij}) é a contagem (c_{ij}) de ocorrências de j em i. Já *Inverse document frequency*(idf_j) é o log de um sobre a proporção de documentos contendo j: $log(n/d_j)$, onde $d_j = \sum_i 1_{[c_{ij}>0]}$ e n é o total de documentos.

[1] A base inicial era composta por 350.224 discursos. Entretanto, apenas 19 partidos foram classificados, totalizando 317.980 discursos. Dessa forma, 90,8% da base inicial de discursos foi analisada neste trabalho.

Tf-Idf é o produto tf_{ij} x idf_j (Gentzkow, Kelly e Taddy (2019)). Palavras muito raras terão um baixo tf-idf, já que (tf_{ij}) será baixo. Palavras muito comuns também terão um baixo tf-idf, consequência do baixo idf_j.

3.3 Naive Bayes Classifier

Com o objetivo de validar a ralevância dos dicionários formados pelo Tf-Idf para explicar os padrões ideológicos do espectro político brasileiro, utilizamos um modelo de aprendizagem supervisionada para prever a classificação ideológica de cada discurso. Além disso, usamos a acurácia deste modelo como uma medida de polarização.

A demonstração do modelo segue Jurafsky e Martin (2019).

Para um documento d o algoritmo retorna a classe $\hat{c}$ que maximiza a probabilidade posterior dado um documento

$$\hat{c} = \arg\max_{c \in C} P(c|d) \tag{3.1}$$

A intuição por trás da classificação Bayesiana é usar a Regra de Bayes para transformar a equação (3.1) em outras probabilidades com propriedades interessantes. Dessa forma, a Regra de Bayes exposta em (3.2) nos permite quebrar qualquer probabilidade condicional $P(x|y)$ em outras três probabilidades:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{3.2}$$

Substituindo (3.2) em (3.1) temos:

$$\hat{c} = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} \frac{P(d|c)P(c)}{P(d)} \tag{3.3}$$

Queremos saber qual é a classe mais provável para o mesmo documento d. Portanto, $P(d)$ não muda para cada classe. Logo, podemos excluir $P(d)$ do denominador e simplificar a equação (3.3). Então, escolhemos a classe que maximiza a seguinte equação:

$$\hat{c} = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} P(d|c)P(c) \tag{3.4}$$

Computamos a classe mais provável $\hat{c}$ escolhendo a classe que retorna o maior produto de duas probabilidades: a probabilidade a priori da classse $P(c)$ (prior) e a probabilidade do documento $P(d|c)$

$$\hat{c} = \arg\max_{c \in C} \underbrace{P(d|c)}_{probabilidade} \underbrace{P(c)}_{prior} \tag{3.5}$$

Representando o documento d como um conjunto de características $f_1, f_2, \ldots, f_n$, temos:

$$\hat{c} = \arg\max_{c \in C} P(f_1, f_2, \ldots, f_n|c)P(c) \tag{3.6}$$

Para prosseguir, fazemos duas simplificações. A primeira delas, *bag-of-words*, assume que a posição das palavras não importa, ou seja, não faz difereça se uma palavra aparece na primeira posição ou na vigésima posição em um texto. Logo, assumimos que o conjunto de características $P(f_1, f_2, \ldots, f_n)$ apenas representa as palavras e não a posição.

A segunda suposição é comumente chamada de *Naive Bayes*: essa hipótese assume independência condicional, ou seja, as probabilidades $P(f_i|c)$ são independentes dadas as classes c e podem ser multiplicadas da seguinte maneira:

$$P(f_1, f_2, \ldots, f_n|c) = P(f_1|c) \cdot P(f_2|c) \ldots P(f_n|c) \tag{3.7}$$

A equação final para cada classe escolhida pelo Naive Bayes Classifiear é dada por:

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{f \in F} P(f|c) \tag{3.8}$$

3.3.1 Validação

Quando construímos um modelo de aprendizagem não queremos que este consiga prever apenas os eventos que o algoritmo já viu. Um algoritmo que apenas memoriza todos os exemplos de um conjunto de dados e usa essa memória para prever esses mesmos exemplos não cometerá erros, entretanto esse algoritmo não será proveitoso na prática. O que queremos é que o modelo consiga prever exemplos que o algoritmo ainda não viu, ou seja, o objetivo é obter uma boa performance em um conjunto de validação (Burkov (2019)).

Dessa forma, dividimos os documentos D_i de cada legislatura em um conjunto de treino $D_{i_{\text{treino}}}$ e em um conjunto de teste para a validação $D_{i_{\text{teste}}}$[2]. Após representar os conjuntos de treino e teste através dos valores de Tf-Idf, treinamos o Naive Bayes Classifier e procedemos para a previsão da classificação dos documentos no conjunto de teste.

3.4 Similaridade entre cossenos

A fim de entender a semelhança entre dois documentos, podemos utilizar a similaridade entre cossenos. A similaridade de documentos é utilizada para vários tipos de aplicações, como busca de informação, detecção de plágio, sistemas de recomendação e

[2] Usamos 80% dos discursos de D_i para compor o conjunto D_{i_treino} e os 20% restante para compor o conjunto D_{i_teste}

até para tarefas mais humanas, como comparar diferentes textos para ver quais são as semelhanças entre eles (Jurafsky e Martin (2019))

Considerando os vetores **v** e **w**, a semelhança entre eles pode ser calculada da seguinte maneira (Jurafsky e Martin (2019)):

$$cosine(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}.\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \tag{3.9}$$

Neste trabalho, calculamos a similaridade entre os cossenos dos vetores formados para cada corrente ideológica pelo Tf-Idf.

4 Resultados

4.1 Acurácia Naive Bayes Classifier

As tabelas abaixo apontam o total de discursos dividido por afiliação ideológica em cada legislatura. Considerando a participação de cada corrente ideológica, observamos uma predominância dos discursos dos políticos de centro ao longo de todo o período analisado, com destaque para o aumento da participação de discursos dos políticos de esquerda a partir da 54ª legislatura.

Tabela 1 – Discursos Total

	Legislatura			
	52ª	**53ª**	**54ª**	**55ª**
Esquerda	20.965	22.461	25.440	25.600
Centro	37.242	38.919	38.115	36.582
Direita	19.402	18.890	15.077	19.287
Total	77.609	80.270	78.632	81.469

Tabela 2 – Participação Por Corrente Ideológica no Total de Discursos

	Legislatura			
	52ª	**53ª**	**54ª**	**55ª**
Esquerda	27%	28%	32%	31%
Centro	48%	48%	48%	45%
Direita	25%	24%	19%	24%

Uma métrica comumente usada para avaliar modelos de classificação é a acurácia do modelo, a qual é dada pelo total de exemplos classificados corretamente dividido pelo total de observações na amostra. Considerando a acurácia, o modelo utilizado apresentou uma significativa melhora ao longo das legislaturas ao tentar prever a afiliação ideológica dos deputados.

Tabela 3 – Acurácia Naive Bayes Classifier

	Legislatura			
	52º	**53º**	**54º**	**55º**
Teste	0,54	0,58	0,63	0,62
Treino	0,73	0,78	0,78	0,77

Peterson e Spirling (2018) utilizam a acurácia de modelos de classificação como uma medida de polarização política no parlamento britânico. Segundo os autores, a acurácia indica uma medida de polarização relativa entre diferentes períodos. Traduzindo para o contexto deste trabalho, podemos afirmar que, segunda essa métrica, houve um aumento da polarização no período analisado, uma vez que a acurácia do conjunto teste passa de 0,54 na 52ª legislatura para 0,62 na 55ª. Este resultado condiz com os resultados observados nas duas próximas sessões.

4.2 Análise dos Vocabulários

Considerando os vocabulários formados pelo Tf-Idf, podemos entender os padrões dos discursos subjacentes às diversas correntes ideológicas. Para formar os dicionários, agrupamos os discursos entre centro, esquerda e direita e selecionamos as 15 primeiras palavras com o maior valor de Tf-Idf

Tabela 4 – Tf-Idf 52ª Legislatura

Esquerda	**Centro**	**Direita**
Servidores	Nordeste	Salário
Mulheres	Ensino	Tributária
Salário	Polícia	Agricultura
Polícia	Agricultura	Polícia
Violência	Crianças	Banco
Universidade	Banco	Nordeste
Famílias	Salário	Juros
Agricultura	Tributária	Previdência
Ensino	Universidade	Água
Energia	Violência	Produtores
Democracia	Água	Ensino
Banco	Mulheres	Violência
Previdência	Ambiente	Empregos
Corrupção	Juros	Jornal
Crianças	Tribunal	Tribunal

Tabela 5 – Tf-Idf 53ª Legislatura

Esquerda	**Centro**	**Direita**
Mulheres	Reforma	Polícia
Crise	Polícia	Democratas
Reforma	Mulheres	Dinheiro
Ensino	Crise	Reforma
Cultura	Ensino	Amazônia
Democracia	Tribunal	Ensino
Humanos	Nordeste	Crise
Violência	Dinheiro	Aposentados
Amazônia	Energia	Energia
Universidade	Amazônia	Tribunal
Jovens	Aposentados	Amazonas
Servidores	Crianças	Agricultura
Polícia	Empresa	Policiais
Empresa	Violência	Militares
Agricultura	Agricultura	Produtores

Analisando os vocabulários formados para a 52ª e a 53ª legislatura (tabela 4 e tabela 5) podemos observar alguns padrões. Em relação à economia, centro e direita mostram alguma convergência no peso dado a escolha das palavras, como em banco, juros, produtores, tributária e previdência - nestes casos uma possível alusão a essas reformas.

Assim como observado nos discursos do senado em Figueredo (2019), a esquerda trata de assuntos relacionados à família de maneira indireta ao usar palavras como mulheres, crianças e jovens. Já a direita não cita esses tópicas. Além disso, a esquerda se apropria de termos mais relacionadas às pautas progressistas, como universidade, democracia e cultura.

Tabela 6 – Tf-Idf 54ª Legislatura

Esquerda	**Centro**	**Direita**
Trabalhadores	Direitos	Direitos
Direitos	Empresas	Empresas
Mulheres	Produção	Mulheres
Lula	Trabalhadores	Produção
Reforma	Violência	Produtores
Violência	Polícia	Energia
Renda	Dinheiro	Voto
Democracia	Tribunal	Amazonas
Humanos	Reforma	Democratas
Produção	Empresa	Dinheiro

Continua na próxima página

Tabela 6 – Tf-Idf 54ª Legislatura - Continuação

Esquerda	**Centro**	**Direita**
Famílias	Nordeste	Ensino
Empresas	Mulheres	Tribunal
Universidade	Energia	Agricultura
Agricultura	Crianças	Violência
Mulher	Investimentos	Crescimento

Tabela 7 – Tf-Idf 55ª Legislatura

Esquerda	**Centro**	**Direita**
Trabalhadores	Dilma	Dilma
Lula	Reforma	Constituição
Direitos	Mulheres	Polícia
Reforma	Trabalhadores	Dinheiro
Dilma	Crise	Mulheres
Golpe	Dinheiro	Reforma
Mulheres	Polícia	Empresas
Democracia	Constituição	Crime
Presidenta	Economia	Crise
Luta	Direitos	Direitos
Previdência	Empresas	Economia
Sociais	Violência	Trabalhadores
Constituição	Crime	Tribunal
Violência	Econômica	Violência
Empresas	Tribunal	Fiscal

Considerando a 54ª e a 55ª legislatura (tabela 4 e tabela 5) fica evidente um maior distanciamento entre esquerda e direita, além de uma aproximação do centro com a direita.

Em relação às pautas econômicas, o vocabulário de centro e direita usa de diversos termos para abordar esse assunto, como dinheiro, empresas, produção, produtores, investimentos, crescimento, economia, econômica e fiscal - nestes dois últimos casos uma possível alusão às discussões sobre reformas econômicas e fiscal que marcaram o período em questão. Já os partidos de esquerda abordam as questões econômicas com um vocabulário menos diverso, usando apenas trabalhadores, renda, empresas e produção, sendo que estas duas últimas palavras em menor peso quando comparado ao centro e à direita.

Além disso, a 55ª legislatura foi marcada pelo impeachment da então presidente Dilma Roussef e, portanto, pelo acirramento da polarização política na câmara. Esse fato é evidenciado pela maior divergência entre o vocabulário formado pela esquerda e os vocabulários de centro e direita, os quais apresentam maior convergência.

De fato, no vocabulário da esquerda temos palavras muito marcantes no contexto do impeachment e associado à partidos deste espectro ideológico, como o PT e o PSOL, com destaque para os termos Lula, direitos, Dilma, golpe, democracia e lutas, o que também pode ser encontrado no trabalho de Figueredo (2019) que usa os discursos do senado. Já o centro e a direita fazem alusão ao processo de impeachment mas com um conjunto de termos diferente, como constituição, tribunal, crime e fiscal, sendo que os dois últimos termos estão relacionados aos crimes de responsabilidade fiscal cometidos pela então presidente Dilma Roussef.

4.3 Distância entre vetores como uma medida de polarização

Considerando que os vocabulários formados para as legislaturas são vetores com uma pontuação atribuída pelo Tf-Idf para cada palavra presente no vocabulário, podemos calcular o cosseno desses vetores como uma medida de similaridade entre eles. Dessa forma, em todas as legislaturas calculamos o cosseno entre os vetores de palavras das diferentes correntes ideológicas.

Tabela 8 – Cosseno Entre Vetores - Tf-Idf

Legislatura	**Esquerda-Direita**	**Esquerda-Centro**	**Direita-Centro**
52ª	0,94	0,95	0,98
53ª	0,92	0,95	0,98
54ª	0,90	0,93	0,97
55ª	0,84	0,87	0,97

Considerando a 52ª e a 53ª legislatura, os cossenos entre os vetores ficam mais próximos de 1, indicando uma maior similaridade entre os vocabulários. Entretanto,destaca-se a maior convergência de centro e direita, quando comprados com a similaridade entre centro e esquerda.

Já na 54ª e na 55ª legislatura, os cossenos entre os vetores de esquerda e direita apresentam uma redução relvante, evidenciando uma maior distância entre essas correntes ideológicas. Além disso, esquerda e centro também mostram um maior distanciamento, o que não ocorre entre direita e centro, já que o cosseno entre os vetores dessas correntes ideológicas permanece constante.

5 Conclusão

Essa dissertação teve como objetivos acessar os padrões ideológicos e qualificar a polarização política na Câmara dos Deputados do Brasil ao longo da 52ª e da 55ª legislatura. Para tanto, utilizamos uma base de dados composta por 317.980 discursos dos Deputados Federais

Em relação ao conteúdo das correntes ideológicas, a análise dos dicionários indica que esquerda e direita divergem na escolha das palavras para tratar dos mesmos tópicos, assim como evidenciado por Gentzkow, Shapiro e Taddy (2019). Um exemplo é dado pela divergência ao tratar de questões econômicas, sendo que a esquerda se debruça sobre palavras como “Renda” e “Trabalhadores” e a direita escolhe palavras como “Empresas” e “Produção”. Entretanto, no contexto analisado, fica evidente que as correntes políticas também divergem na escolha dos tópicos mais acessados. Por exemplo, a esquerda atribui um maior peso para palavras relacionadas ao campo mais progressista, como “Cultura”,“Universidade”, “Jovens” e “Luta”. Já o centro e a direita possuem um vocabulário muito mais diverso e dão maior peso à pautas econômicas, por exemplo ao escolherem palavras como “Crescimento”, “Investimento”, “Produção”, “Dinheiro” e “Economia”.

Além disso, os vocabulários indicam alguma inflexão na polarização a partir da 55ª legislatura. Neste caso, a polarização é marcada por questões relacionadas ao impeachment da então presidente Dilma Roussef, sendo a questão tratada pela esquerda com palavras como “Golpe” e “Luta”, e pela direita como “Consituição”, “Crime” e “Fiscal”.

Para qualificar a polarização, uma das abordagens partiu de um algoritmo de aprendizagem supervisionada. Usamos esse modelo para prever a afiliação ideológico dos Deputados Federais através dos seus discursos. Em seguida, usamos a acurácia do modelo como uma medida de polarização. De fato, houve uma melhora consistente da acurácia do modelo entre 52ª e a 55ª legislatura, indicando um aumento da polarização nesse período.

Por fim, calculamos a similaridade entre os cossenos dos vetores formados pelo Tf-Idf com o objetivo de propor uma medida de polarização política. Neste caso, observamos uma redução do cosseno entre os vetores de esquerda e direita na 54ª e na 55ª legislatura, evidenciando uma maior distância entre essas correntes ideológicas, o que é consistente com os demais resultados deste trabalho.

Como próximos passos, algumas melhorias poderiam se feitas para aprofundar a análise dos vocabulários, como relacionar os resultados à fatores históricos e a evolução do sistema de crenças no Brasil. Outro ponto relevante seria melhorar a acurácia do modelo de classificação. Um passo interessante poderia ser identificar as palavras como substantivos, adjetivos, pronomes e verbos e treinar modelos de classificação com esses subconjuntos.

Além disso, usar um recorte dos dados apenas com discursos de políticos mais extremados também poderia incrementar a acurácia.

Considerando a pluralidade de partidos políticos no Brasil, uma possível limitação deste trabalho é dificuldade em englobar os partidos em um posicionamento ideológico no conjunto esquerda, centro e direita, o que inclusive pode afetar o desempenho dos modelos ou levar a interpretações equivocadas dos conteúdos relacionados às correntes ideológicas. Entretanto, este pode ser um novo desafio de pesquisa, adaptar ou criar novas metodologias para o contexto político brasileiro.

Referências

BARRON, A. T. et al. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 115, n. 18, p. 4607–4612, 2018.

BORN, B.; EHRMANN, M.; FRATZSCHER, M. Central bank communication on financial stability. *The Economic Journal*, Wiley Online Library, v. 124, n. 577, p. 701–734, 2014.

BURKOV, A. *The hundred-page machine learning book*. [S.l.]: Andriy Burkov Quebec City, Can., 2019. v. 1.

DIERMEIER, D. et al. Language and ideology in congress. *British Journal of Political Science*, JSTOR, p. 31–55, 2012.

FIGUEREDO, F. C. d. Ideology as a belief system: a computational linguistic approach to brazilian senate. *Dissertação de Mestrado - UNB*, 2019.

GENNARO, G.; LECCE, G.; MORELLI, M. Intertemporal evidence on the strategy of populism. CEPR Discussion Paper No. DP13804, 2019.

GENTZKOW, M.; KELLY, B.; TADDY, M. Text as data. *Journal of Economic Literature*, v. 57, n. 3, p. 535–74, 2019.

GENTZKOW, M.; SHAPIRO, J. M.; TADDY, M. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, Wiley Online Library, v. 87, n. 4, p. 1307–1340, 2019.

HANSEN, S.; MCMAHON, M.; PRAT, A. Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, Oxford University Press, v. 133, n. 2, p. 801–870, 2018.

JURAFSKY, D.; MARTIN, J. H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Third Edition draft. 2019.

LAUDERDALE, B. E.; HERZOG, A. Measuring political positions from legislative speech. *Political Analysis*, JSTOR, p. 374–394, 2016.

LUCCA, D. O.; TREBBI, F. *Measuring central bank communication: an automated approach with application to FOMC statements*. [S.l.], 2009.

MOSTELLER, F.; WALLACE, D. L. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 58, n. 302, p. 275–309, 1963.

PETERSON, A.; SPIRLING, A. Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, Cambridge University Press, v. 26, n. 1, p. 120–128, 2018.

POOLE, K. T.; ROSENTHAL, H. A spatial model for legislative roll call analysis. *American Journal of Political Science*, JSTOR, p. 357–384, 1985.

POWER, T. J.; JR, C. Z. Estimating ideology of brazilian legislative parties, 1990-2005: a research communication. *Latin American Research Review*, JSTOR, p. 218–246, 2009.

SCOTT, S. L.; VARIAN, H. R. Bayesian Variable Selection for Nowcasting Economic Time Series. In: *Economic Analysis of the Digital Economy*. National Bureau of Economic Research, Inc, 2015, (NBER Chapters). p. 119–135. Disponível em: <https://ideas.repec.org/h/nbr/nberch/12995.html>.

ZUCCO, C.; POWER, T. J. Fragmentation without cleavages? endogenous fractionalization in the brazilian party system. *Endogenous Fractionalization in the Brazilian Party System (August 27, 2019). Forthcoming in Comparative Politics (ISSN 0010-4159)*, 2019.