UFRRJ

INSTITUTO MULTIDISCIPLINAR

PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

DISSERTAÇÃO

Metodologia de Determinação de Diferenciação Ideológica baseada em Análise por Tópicos

Alessandro Zelesco

2022

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

METODOLOGIA DE DETERMINAÇÃO DE DIFERENCIAÇÃO IDEOLÓGICA BASEADA EM ANÁLISE POR TÓPICOS

ALESSANDRO ZELESCO

Sob orientação de
Ricardo Cordeiro Corrêa

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências** no Programa de Pós-graduação Interdisciplinar em Humanidades Digitais, Área de Concentração em Análise Qualitativa e Quantitativa de Dinâmicas Sociais.

Nova Iguaçu, RJ
Abril de 2022

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

Zelesco, Alessandro, 1957-
ZZ49m METODOLOGIA DE DETERMINAÇÃO DE DIFERENCIAÇÃO IDEOLÓGICA BASEADA EM ANÁLISE POR TÓPICOS / Alessandro Zelesco. - Rio de Janeiro, 2022.
87 f.: il.

Orientador: Ricardo Cordeiro Corrêa.
Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, Programa de Pós-graduação Interdisciplinar em Humanidades Digitais, 2022.

1. Humanidades Digitais. 2. Análise automatizada por tópicos. 3. Posicionamento ideológico de agremiações partidárias. 4. Índice de Posicionamento Ideológico. 5. Teoria Marxista da Dependência. I. Corrêa, Ricardo Cordeiro, 1966-, orient. II Universidade Federal Rural do Rio de Janeiro. Programa de Pós-graduação Interdisciplinar em Humanidades Digitais III. Título.

Este documento foi criado usando o sistema LaTeX de preparação de documentos desenvolvido por Leslie Lamport a partir do sistema de formatação TeX criado por Donald Knuth.
O formato foi obtido usando a classe `UFRuralRJ`, uma adaptação livre das classes `mdtufsm` e `iiufrgs` para a formatação de documentos acadêmicos produzidos na Universidade Federal Rural do Rio de Janeiro (UFRRJ).

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

ALESSANDRO ZELESCO

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências** no Programa de Pós-graduação Interdisciplinar em Humanidades Digitais, Área de Concentração em Análise Qualitativa e Quantitativa de Dinâmicas Sociais.

DISSERTAÇÃO APROVADA EM 27/04/2022.

Ricardo Cordeiro Corrêa. Docteur UFRRJ (Orientador)
Leandro Guimarães Marques Alvim. D.Sc. UFRRJ
Antonio da Silveira Brasil Junior. D.Sc. UFRJ

ATA DE DEFESA DE TESE Nº 156/2022 - PPGIHD (11.39.00.16)

Nº do Protocolo: 23083.033333/2022-61

Nova Iguaçu-RJ, 01 de junho de 2022.

Visualize o documento original em https://sipac.ufrrj.br/public/documentos/index.jsp informando seu número: **156**, ano: **2022**, tipo: **ATA DE DEFESA DE TESE**, data de emissão: **01/06/2022** e o código de verificação: **91dd77dc4d**

(Assinado digitalmente em 01/06/2022 10:37)
LEANDRO GUIMARAES MARQUES ALVIM
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCC/IM (12.28.01.00.00.83)
Matrícula: ###008#2

(Assinado digitalmente em 01/06/2022 09:42)
RICARDO CORDEIRO CORREA
PROFESSOR DO MAGISTERIO SUPERIOR
PPGIHD (11.39.00.16)
Matrícula: ###07#4

(Assinado digitalmente em 01/06/2022 09:46)
ANTONIO DA SILVEIRA BRASIL JUNIOR
ASSINANTE EXTERNO
CPF: ###.###.537-##

Emitido em 05/05/2023

TERMO Nº 484/2023 - PPGIHD (11.39.00.16)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 06/05/2023 15:44)
RICARDO CORDEIRO CORREA
COORDENADOR CURS/POS-GRADUACAO - TITULAR
PPGIHD (11.39.00.16)
Matrícula: ###07#4

Visualize o documento original em https://sipac.ufrrj.br/documentos/ informando seu número: **484**, ano: **2023**, tipo: **TERMO**, data de emissão: **06/05/2023** e o código de verificação: **01970c8e5f**

AGRADECIMENTOS

Não poderia deixar de agradecer, primeiramente, aos amigos e camaradas Álvaro Carriello, Priscila Alencastre e Heitor Silva, que muito me incentivaram na decisão de encarar essa jornada. Agradeço também à Simone, esposa e companheira de vida, pela enorme paciência durante meus estudos e aos meus filhos Gabriel e Érica pelo incentivo recebido. E um agradecimento especial ao meu orientador Ricardo Corrêa por aceitar orientar-me no desbravamento do tema proposto.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Código de Financiamento 001. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

"A história do subdesenvolvimento latino-americano é a história do desenvolvimento do sistema capitalista mundial."

Subdesenvolvimento e Revolução, Ruy Mauro Marini, 1969

"Não devemos nos admirar de que os livre-cambistas não consigam compreender como um país pode se enriquecer a custa de outros, pois estes mesmos senhores também não querem compreender como, no interior de um país, uma classe pode se enriquecer a custa de outra classe."

Sobre a Questão do Livre-Câmbio, Karl Marx, 1848

"Não é porque foram cometidos abusos contra as nações não industriais que estas se tornaram economicamente débeis, é porque eram débeis que se abusou delas. Não é tampouco porque produziram além do necessário que sua posição comercial se deteriorou, mas foi a deterioração comercial o que as forçou a produzir em maior escala. Negar-se a ver as coisas dessa forma é mistificar a economia capitalista internacional, é fazer crer que essa economia poderia ser diferente do que realmente é. Em última instância, isso leva a reivindicar relações comerciais equitativas entre as nações, quando se trata de suprimir as relações econômicas internacionais que se baseiam no valor de troca."

Dialética da Dependência, Ruy Mauro Marini, 1973

RESUMO GERAL

ZELESCO, Alessandro. **Metodologia de Determinação de Diferenciação Ideológica baseada em Análise por Tópicos**. 2022. 86f. Dissertação (Mestrado em Humanidades Digitais). Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro, Nova Iguaçu, RJ, 2022.

Este trabalho procura contribuir no campo das humanidades digitais com a análise automatizada da atuação de partidos políticos apresentando nova abordagem para tratar as múltiplas dimensões programáticas, com um aspecto ideológico. No campo das ciências sociais produzimos inferências para analisar o conteúdo de documentos, mesmo complexos e multidimensionais como os que retratam a atuação dos partidos na sociedade. Porém, como realizar uma análise automatizada desses conteúdos partidários multidimensionais sob um aspecto ideológico de análise? O procedimento proposto para realizar medidas empíricas de distância sobre polarizações ideológicas ou diferenciações programáticas, necessárias para testar modelos espaciais de atuação política, passa ao largo da análise sintática e da anotação prévia de documentos. O objetivo do trabalho consiste no desenvolvimento de uma metodologia que utiliza técnicas de processamento de linguagem natural para extrair as múltiplas dimensões – da atuação partidária e da linha teórica sob a qual será feita a análise – varrendo integralmente a coleção digitalizada de documentos (corpora) para posterior comparação, a luz da teoria, do grau de afinidade entre as agremiações. A teoria marxista da dependência foi escolhida como base de comparação com as teses dos partidos de esquerda para a superação do subdesenvolvimento e da dependência do país. A sequência temática da dissertação consiste na apresentação de diferentes abordagens para análise de conteúdo de documentos partidários e coalizões de governo, das métricas utilizadas para esse fim e das aplicações no Brasil. Por fim, apresentamos uma nova metodologia baseada em análise por tópicos e o resultado preliminar alcançado.

Palavras-chave: Humanidades Digitais, Análise de conteúdo, Dimensões programáticas, Distância política, Métricas.

GENERAL ABSTRACT

ZELESCO, Alessandro. **An Ideological Differentiation Methodology Based on Topic Analysis**. 2022. 86p. Dissertation (Master in Digital Humanities). Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro, Nova Iguaçu, RJ, 2022.

This study seeks to contribute in the field of Digital Humanities with the automated analysis of the performance of political parties, presenting a new approach to deal with multiple programmatic dimensions, with an ideological aspect. In the field of social sciences, we produce inferences to analyze the content of documents, even complex and multidimensional ones such as those that portray the role of political parties in society. However, how to carry out an automated analysis of these multidimensional party contents under an ideological aspect of analysis? The procedure proposed for carrying out empirical measures of distance on ideological polarization or programmatic differentiation, necessary to test spatial models of political action, bypasses syntactic analysis and prior annotation of documents. The objective of this work is to develop a methodology that uses natural language processing techniques to extract the multiple dimensions – of the party's performance and the theoretical line under which the analysis will be carried out – by fully scanning the digitized collection of documents (corpora) to subsequent comparison, in the light of theory, the degree of affinity between the political parties. The marxist theory of dependency was chosen as a basis for comparison with the theses of left-wing parties to overcome the country's underdevelopment and dependency. The thematic sequence of the dissertation consists of the presentation of different approaches for content analysis of party documents and government coalitions, the metrics used for this purpose and applications in Brazil. Finally, we present a new methodology based on topic analysis and the preliminary result achieved.

Keywords: Digital Humanities, Content analysis, Programmatic dimensions, Political distance, Metrics..

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

O termo mineração de dados textuais refere-se genericamente a um conjunto de metodologias computacionais para análise de documentos escritos digitais. Uma característica dos métodos reside na busca automatizada de padrões que possam servir de referências para o resgate de informações não estruturadas ou a organização dos textos em coleções de forma a serem referenciados e manipulados com mais velocidade e precisão. O presente trabalho procura contribuir no campo das Humanidades Digitais com a análise automatizada da atuação de agremiações políticas. Esta atuação partidária, registrada sob a forma digital, está amplamente disponível nas redes sociais, nos programas eleitorais e em documentos próprios, como manifestos, programas e resoluções. O desafio é apresentar uma nova abordagem no processamento de linguagem natural para tratar as múltiplas dimensões da atuação partidária com um aspecto ideológico. No campo das Ciências Sociais usamos naturalmente nossa capacidade de produzir inferências para analisar o conteúdo de documentos complexos e multidimensionais como os que retratam a atuação de partidos políticos na sociedade. Porém, como realizar uma análise automatizada de conteúdos partidários multidimensionais sob um determinado aspecto ideológico de análise? O procedimento proposto passa ao largo da análise sintática ou gramatical e da anotação prévia de documentos. O foco consiste na utilização de técnicas de processamento de linguagem natural para extrair as múltiplas dimensões temáticas e de linhas teóricas diretamente dos documentos digitalizados.

Muitas teorias de política comparada dependem da capacidade dos pesquisadores de localizarem partidos políticos dentro de um espaço político. Apesar da importância dos posicionamentos partidários para o estudo da política comparada, localizar partidos em um espaço político ao longo do tempo não é uma tarefa fácil. As posições partidárias não são observáveis e, portanto, devem ser tratadas como uma variável latente no trabalho empírico. Os partidos revelam suas posições indiretamente por meio de uma variedade de atividades, desde a publicação de manifestos e programas eleitorais que definem seus objetivos políticos até declarações, discursos e a maneira como seus membros votam no parlamento. Dentre os principais métodos para estimar as posições latentes de partidos políticos figuram a codificação manual de documentos e a análise de manifestos baseada em computador. Nesta dissertação, primeiro apresentamos o trabalho desenvolvido pelos pesquisadores do Projeto Manifesto com a codificação manual de documentos e seu índice intrínseco de medição esquerda-direita. Depois, apresentamos o esforço de outros pesquisadores no desenvolvimento de novas métricas mantendo por base o mesmo esquema de codificação manual, envolvendo alguma forma de contagem das anotações feitas nos textos. Esse esforço mostrou-se relevante, mas ainda insuficiente na comparação de como as diferentes métricas propostas funcionariam num contexto de formação de coalizões partidárias. A partir de um fluxo constante de evidências pode-se estar bastante certo que as diferenças políticas entre os partidos desempenham um papel decisivo na formação de coalizões.

Partidos mais semelhantes têm maior probabilidade de acabar governando juntos. Se as várias estimativas de diferenças partidárias a disposição capturam as diferenças políticas reais entre os partidos em diferentes graus, então os modelos de previsão de formação de coalizões que utilizam uma métrica melhor, devem se ajustar melhor aos dados e prever

as coalizões reais com maior precisão. Todos os autores que propuseram novas métricas forneceram argumentos teóricos e metodológicos mais ou menos convincentes para seus índices, mas não há um estudo sistemático para compará-los a fim de ver qual realmente descreve partidos e interações de partidos melhor. Essa questão foge ao escopo deste trabalho, que se contenta em discutir formas automatizadas de se quantificar índices em duas abordagens distintas.

Pode-se pensar nas diferenças políticas entre os partidos como diferenças ideológicas ou diferenças programáticas. Temas de posições e diferenças ideológicas dizem respeito a temas com uma determinada divisão ideológica ou clivagem social. Sem esquecer que dentro de um mesmo campo ideológico, podem coexistir diferentes correntes teóricas. Já as diferenças programáticas são diretamente relacionadas aos tipos de perfis políticos que os partidos apresentam em seus manifestos. Novas propostas metodológicas para diminuir o esforço humano na codificação de documentos têm surgido de diversos especialistas em análise de conteúdo. A adoção de outra abordagem metodológica poderia aperfeiçoar o processo de análise. Nesse caso, análises léxicas automatizadas podem ser úteis, trazendo informações e subsídios relevantes para inferências, com a vantagem de poder envolver grande quantidade de textos em tempo muito inferior. A metodologia desta dissertação busca mostrar essa contribuição.

1.1 Motivação

A motivação intelectual para a pesquisa é fruto da reflexão sobre as bases teóricas da Esquerda Brasileira para retirar o País do subdesenvolvimento e da dependência. Qual arsenal teórico equipou a Esquerda que ressurgiu após o regime militar e por que esse arsenal mostrou-se incapaz de sustentar mudanças estruturantes para livrar o país do subdesenvolvimento e da dependência mesmo após treze anos de mandatos na máxima instância da república?

A confrontação de documentos históricos que refletem a atuação partidária de esquerda com uma teoria política voltada para a superação da dependência embasa a pesquisa em busca dessas respostas.

1.2 Contextualização Temática

Forjada no calor da luta de classes na América Latina dos anos 1960/1970 pelos brasileiros Ruy Mauro Marini, Vania Bambirra e Theotonio dos Santos, a Teoria Marxista da Dependência (TMD) é a síntese do encontro profícuo entre a teoria do valor de Marx e a teoria marxista do imperialismo de Lênin. Deste encontro nasceu o veio teórico em que se descobriram categorias originais para dar conta de explicar processos e tendências específicas no âmbito da totalidade integrante, mas também diferenciada, que é o capitalismo mundial.

Os três fundadores da TMD foram dirigentes da Organização Revolucionária Marxista Política Operária (ORM-Polop) no começo dos anos 1960, organização da esquerda brasileira que questionava as posições etapistas e dogmáticas que orientavam os partidos comunistas na região, como era o caso do Partido Comunista Brasileiro, hegemônico no pensamento de esquerda no período. Criticando as teses de que o subdesenvolvimento se devia à "insuficiência" de capitalismo, a Polop sustentava que a burguesia interna não

podia ter nenhuma veleidade anti-imperialista e que o caráter da revolução brasileira não seria democrático-burguesa, mas socialista.

Apesar de a TMD estar bastante disseminada na América Latina por ocasião da redemocratização do País no final do regime militar, seus fundamentos teóricos não foram levados em conta na reorganização dos partidos de esquerda que surgiram nesse período.

Passados quarenta anos, qual o grau de aderência do pensamento progressista brasileiro a essa teoria? E quanto consegue ser apropriada pelas organizações de esquerda contemporâneas para fundamentar seus diagnósticos que dirigirão suas ações politicas?

1.3 Descrição do Problema

O atual estágio da esquerda brasileira pode ser explicado pela maneira como ressurgiu após o regime militar? Devido ao seu caráter transformador a TMD foi perseguida pelo terror de Estado, combatida pelo dogmatismo teórico e também marginalizada pelo neoliberalismo acadêmico. E o exílio político de seus fundadores foi secundado pelo exílio teórico em seu país de origem. Porém, se nos anos oitenta e noventa do século XX a ofensiva do neoliberalismo na cena política e acadêmica tentou impor cadeados contra a crítica radical e o sentido transformador de que a TMD é portadora, hoje já não se pode dizer que ela siga sendo uma teoria exilada.

Com a marginalização da TMD na redemocratização do país, quatro teorias produzidas no ambiente da Escola de sociologia paulista conquistaram hegemonia no pensamento da esquerda brasileira. Teorias nascidas na USP que tentam, ainda hoje, explicar o Brasil sob uma perspectiva da dependência associada, quais sejam: a teoria da dependência propriamente dita, na vertente weberiana defendida por Fernando Henrique Cardoso; a teoria do populismo defendida por Francisco Weffort; a teoria do autonomismo, defendida por Eder Sader (Revista Desvios); e a teoria do autoritarismo, com vários autores.

A confrontação da influência dessas teorias uspianas no ressurgimento dos partidos de esquerda, que passam a adotar o discurso de uma democracia como valor universal e o consequente abandono de um programa para a Revolução Brasileira conforme defendido pela TMD, permanece em aberto para uma investigação.

A dissertação trabalha na hipótese que o abandono do marxismo e a insuficiência do quadro teórico uspiano para superar as condições da dominação burguesa contribuíram para o fracasso do projeto político da nova esquerda representada principalmente pelo PT. Porém, o aprofundamento do subdesenvolvimento e da dependência nesse período histórico atual move os partidos de esquerda para um novo diagnóstico e estratégia de mudança? Qual o grau de adesão dos partidos de esquerda brasileiros a uma teoria que explica o subdesenvolvimento e a dependência na sua origem, mas que não foi levada em conta na reorganização partidária e na adoção de um quadro teórico concorrente?

Tendo em vista o banimento da TMD na época do ressurgimento dos partidos de esquerda, qual o grau de adesão hoje a uma teoria que aponta para a superação do subdesenvolvimento e da dependência?

1.4 Premissas

O trabalho parte das seguintes premissas:

1) Imposição da realidade sobre a falsidade teórica, a teoria tem que explicar fenômenos da realidade.
2) Teoria politica como definidora de uma prática política consequente com a teoria, se falha a teoria, falha o projeto político praticado.
3) Existência de elementos comuns entre uma teoria política para explicação de fenômenos e análises de partidos políticos para interpretações de fenômenos.
4) Existência de correlação entre o léxico e o semântico; o padrão de ocorrência de palavras tem correlação com o significado que carregam.

1.5 Objetivo – Métricas a partir de Tópicos

Este trabalho intenta mostrar o uso da técnica de análise de conteúdo na pesquisa de diferenças programáticas entre agremiações partidárias e das métricas utilizadas para esse fim. Apresentaremos o Projeto Manifesto, importante base de dados de programas partidários codificados e largamente utilizada por pesquisadores em análises de conteúdo programático de partidos políticos, além de seu índice intrínseco – RILE – de medição unidimensional esquerda-direita. Ao lado das críticas sobre a imutabilidade da aplicação do índice RILE no tempo e espaço, e da utilização de menos da metade das categorias disponíveis, são apresentados índices alternativos desenvolvidos para superar essas deficiências. Dentre esses índices, duas abordagens conceituais contribuem de alguma forma com o escopo desta dissertação. A primeira abordagem ilustra a liberdade do pesquisador de, após definir as dimensões de seu espaço de análise, selecionar as métricas mais adequadas ao seu estudo. Isso pode ser visto no desenvolvimento do índice de similaridade (SIM), que avança sobre as dimensões não utilizadas pelo índice RILE e disponíveis no dataset do Projeto Manifesto. Outra abordagem inova ao buscar informações no próprio conjunto de dados para determinar o conteúdo das dimensões ideológicas. Em ambas as abordagens, procura-se realizar a análise de conteúdo para extrair as dimensões ideológicas nos documentos de partidos políticos varrendo integralmente o texto em busca dessas dimensões, seja por meio de uma codificação externa de categorias ou por uma busca interna no próprio documento. Particularmente, alinhamo-nos com aqueles pesquisadores que propõem métodos alternativos à codificação por notação de documentos; codificação externa guiada pela inferência abdutiva de um corpo de colaboradores. Além de trabalhosa, é uma codificação custosa pelo nível de treinamento e retreinamento necessários entre todos os colaboradores; única maneira para garantir que uma análise com aspectos semânticos e sintáticos muito fortes tenha um embasamento comum na notação dos documentos.

Além da grande intervenção humana por trás da metodologia do Projeto Manifesto, há uma imutabilidade de seu espaço de análise. Essa rigidez embute uma tendência à uniformização da análise, pois seu universo de categorias terá que ser sempre utilizado, independente do espaço e tempo. As incoerências nos resultados obtidos em certos países não são gratuitas. Quando o objeto de estudo são agremiações políticas, a análise da atuação dessas agremiações há que se fazer dentro de um espaço multidimensional que não pode ser estático, pois as sociedades evoluem. As sociedades atuais são muito complexas e a representação desta complexidade em constante evolução há que ser feita num espaço

multidimensional dinâmico; caso contrário há risco de não capturar todas as dimensões necessárias para formação do espaço de análise. Os partidos políticos congregam diferentes formas de demandas da sociedade, expressando suas relações econômicas, sociais e politicas. Suas atuações políticas registradas nos documentos também são multidimensionais, dada a complexidade da sociedade que representam. E como as sociedades evoluem, essas múltiplas dimensões não podem ser estáticas. Não há um espaço multidimensional único que dê conta de analisar tudo. Portanto é inapropriada uma metodologia que trata seu espaço de análise com dimensões inadequadas ao seu objeto de estudo. A rigor, podem ser definidos diversos espaços diferentes, todos multidimensionais. Espaços diferentes, que podem ter relação entre eles ou projeção dum noutro, mas não há razão para assumir certa definição espacial única para dar conta de tudo. Após definir a dimensão do espaço de análise, o corpus pode ser analisado com a escolha adequada das métricas. Diferente da metodologia do Projeto Manifesto, a abordagem para análise de conteúdo proposta nesta dissertação não envolve qualquer aspecto semântico ou sintático do texto. Por meio de técnicas de processamento de linguagem natural, mais especificamente por meio de análise por tópicos, pretendemos dar uma contribuição para extrair as dimensões do espaço de análise diretamente da varredura dos documentos. E a técnica permite avançar não apenas na extração das dimensões de um espaço de análise, mas também na extração de um viés ideológico a partir do qual o pesquisador deseja fazer a análise.

Portanto, a abordagem metodológica a ser apresentada introduz um caráter ideológico na análise, tal como é feito por humanos. O campo das Humanidades Digitais mostra-se então bastante apropriado para essa experiência. O trabalho a ser desenvolvido tem por base, de um lado, os registros digitais do objeto de estudo – a atuação de partidos políticos e, do outro, um arcabouço teórico-conceitual – uma teoria política, que será utilizada para fazer a análise. Não é um desafio fácil, mas acreditamos ser possível, com técnicas de PLN com enfoque majoritariamente léxico, parcialmente semântico e não sintático, realizar uma análise semiautomatizada de um corpus com um determinado viés ideológico. Tal como nas sociedades humanas, diferentes posições ideológicas vão definir diferentes espectros multidimensionais, que vão sempre, de alguma forma, interferir na análise.

A dissertação começa com uma abordagem dos fundamentos do processo de análise de conteúdo no Capítulo 2, onde também é apresentada a dinâmica do modelo de geração de textos que será usado nos experimentos. Os estudos sobre posicionamento ideológico de partidos políticos e índices associados serão abordados no Capítulo 3. Os dois índices propostos nesta dissertação – IDI e IPI – estão sendo apresentados no Capítulo 4 e no Capítulo 5 respectivamente, junto com os resultados de suas aplicações sobre o corpus selecionado para os experimentos. Todos os experimentos foram realizados por intermédio de codificação realizada pelo professor orientador na linguagem R. Por fim, no Capítulo 6, é apresentada a conclusão do trabalho.

2 ANÁLISE TEXTUAL EM CIÊNCIAS SOCIAIS

Textos são os registros observáveis da atuação e da construção ideológica das agremiações, tornando-se assim o objeto de investigação. Essa investigação segue uma certa metodologia de análise do conteúdo dos textos composta por uma série de ações efetuadas em etapas. A análise por tópicos, elemento primordial nos métodos estudados nesta dissertação, ocupam um lugar privilegiado na metodologia. Nas seções seguintes, apresentamos os fundamentos da metodologia geral, com destaque ao papel da análise por tópicos.

2.1 Fundamentos de Análise de Conteúdo

O principal objetivo da análise de conteúdo de textos nas Ciências Sociais é fazer inferências. Existem diferentes abordagens sobre como computadores, e os respectivos algoritmos, devem ser empregados nesse contexto (KRIPPENDORFF, 2018). Além do uso óbvio na geração de percepções qualitativas ou dados quantitativos, existe uma diferenciação de uso entre duas perspectivas. A primeira consiste no emprego como ferramenta de auxílio em uma codificação humana assistida de documentos, enquanto a segunda seria a sua atuação sem intervenção humana em uma análise automatizada dos textos. A verdadeira divisão qualitativa-quantitativa reside entre procedimentos auxiliados por computador e aqueles assumidos inteiramente por um programa uma vez que os parâmetros são definidos. Da mesma forma, há um modo intermediário de "abdução" – no sentido de ajuste interativo de padrões – que ocorre entre a dedução estrita e a indução, como pode ser visto na ilustração dos processos de codificação da Figura 2.1, adaptada de (KRIPPENDORFF, 2018). Nessa figura, são indicadas as etapas frequentemente envolvidas no fluxo operacional de um projeto de análise de conteúdo.

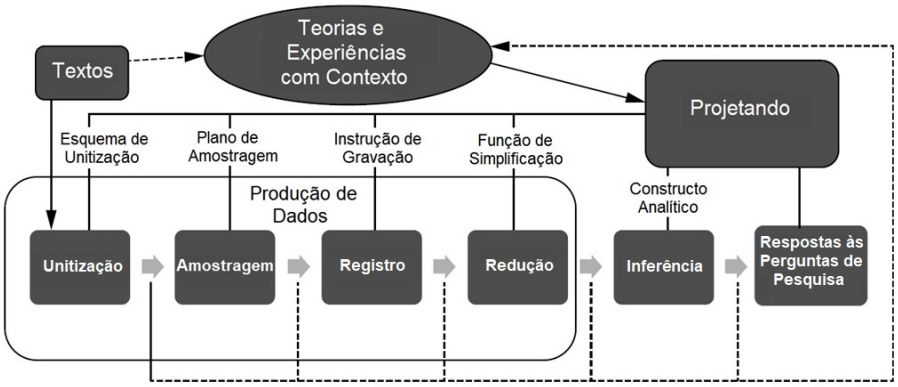


Figura 2.1: Processo de análise de conteúdo.

1. **Criação de dados**: inclui as etapas de **unitização**, **amostragem**, **registro** e **redução** que visam transformar os textos, digamos, brutos ou não editados em dados computáveis, ou seja, representados por estruturas apropriadas e manipuláveis por programas de análise dentro de um computador. Entendemos por textos brutos a fonte primária que são estruturados para compreensão humana, mas não estruturados para um direto tratamento computacional. O processo de transformação em dados envolve uma distinção sistemática de extratos de texto considerados relevantes para a análise. Um exemplo é mostrado na Tabela 2.1. Adicionalmente, pode-se

restringir o conjunto de extratos a um subconjunto gerenciável que seja estatística ou conceitualmente representativo do perfil semântico de textos. Os extratos assim selecionados são ainda tratados com o intuito de produzir representações eficientes para o propósito específico da análise. Um tal tratamento pode incluir a remoção de palavras pouco significativas, de declinações e de conjugações, conforme a ilustração da tabulação mostrada na Tabela 2.2, além de alteração nessa forma tabular para uma representação na forma de uma lista de tipos e de frequências de ocorrência, como ilustrado na Tabela 2.3.

2. **Inferência abdutiva de fenômenos contextuais**: consiste na análise de conteúdo propriamente, indo além da mera verificação de atributos representacionais dos textos. É nesta etapa que procedimentos computacionais são empregados para inferir hipóteses sobre perfis semânticos, e suas associações, a partir das representações dos textos. Nesse contexto, as inferências abdutivas – ao contrário das outras duas formas canônicas de inferência, a saber dedutivas ou indutivas – requerem construções analíticas que explicariam, em uma relação de causalidade, os perfis semânticos dos textos de uma forma plausível. O sentido dessa plausibilidade indica uma probabilidade da conclusão da inferência estar correta e não necessariamente a sua verdade. Os métodos de análise por tópicos objeto da Seção 2.2 enquadram-se nesta etapa. Esses métodos, como será visto, estabelecem como hipótese uma representação dos textos como uma mistura de vários tópicos, sendo cada tópico definido em termos de agrupamento ou coleção de elementos léxicos.
3. **Resposta à questão de pesquisa**: combinação dos passos 1 e 2 com a questão geral de pesquisa subjacente à análise. Esta etapa equivale a tornar os resultados da inferência abdutiva compreensíveis no âmbito da investigação e informativas para os objetivos dessa investigação. Visto de outra forma, isso corresponde a explicar o significado prático dos resultados ou as contribuições que eles aportam ao trabalho científico. O exemplo natural a mencionar aqui é a investigação das implicações que os perfis ideológicos podem ter sobre as relações das agremiações entre si na formação de coalizões eleitorais ou nas ações de governo.

2.2 Análise por Tópicos

Conceitualmente, a análise textual por tópicos é um ramo da área de processamento de linguagem natural cujo objetivo é estimar os perfis temáticos perceptíveis em um corpus constituído de textos semanticamente relacionados. Há na literatura uma vasta gama de métodos computacionais para esse fim, dentre os quais têm destaque os métodos estatísticos que descrevem propriedades das fontes geradoras dos textos analisados quanto à forma como essas fontes estabelecem os perfis temáticos")`[@jelodar_latent_2019]`.r corrigido("Nesse contexto, um perfil temático de um texto é a proporção de prevalência de certos tópicos nesse texto e, para cada tópico, o emprego das palavras ou expressões que definem o perfil semântico desse tópico. Assim sendo, é plausível supor que, embora as prevalências sejam estimadas a partir das frequências de ocorrência concomitante nos textos, exista algum nível de correlação entre tais informações essencialmente quantitativas e os perfis semânticos dos mesmos textos. Em outras palavras, presume-se que a composição de temas de um texto confere valor semântico a esse texto. Portanto, uma estimativa dos perfis temáticos, e o estabelecimento de associações entre esses perfis temáticos, é uma forma aproximada de extração das informações subjacentes mais relevantes

Tabela 2.1: Extratos de um texto de agremiação política.

id	2010PT01
2	2. Depois de duas décadas de estagnação ou avanços medíocres, a economia brasileira voltou a crescer. Mas esse crescimento obedece hoje a uma lógica distinta daquela do passado. Ele se faz com forte distribuição de renda, com inédito equilíbrio macroeconômico, com redução da vulnerabilidade externa e, sobretudo, com fortalecimento da democracia
3	Há mais de 70 anos os períodos de expansão da economia brasileira acabaram frustrando as expectativas da maioria da sociedade. Ora concentravam riqueza. Ora vinham acompanhados de surtos inflacionários. Ora produziam elevado endividamento externo e interno. Ora sufocavam a democracia 3. A partir dos anos 90 os investimentos produtivos foram reduzidos, o país sofreu restrições no seu parque industrial, a sua infra-estrutura foi comprometida, sobretudo na área de energia e transportes. Esta política debilitou as empresas estatais, e algumas fo...
4	5. No Governo Lula, o crescimento do PIB, a expansão do emprego formal, os aumentos reais do salário mínimo, as políticas de transferência de renda, o controle da inflação, a queda da taxa de juros, a ampliação do crédito, as medidas para a reforma agrária e apoio à agricultura familiar, o aumento do comércio exterior e a reconstrução da infraestrutura mudaram tudo isso
5	6. Essa mudança propiciou a formação de um grande mercado de bens de consumo popular, que nos protegeu dos efeitos devastadores da crise mundial desencadeada em setembro de 2008. Proteção reforçada pela saúde de nosso sistema bancário e pelas reservas internacionais acumuladas. 7. Diferentemente das crises externas anteriores -a mexicana, a asiática e a russa
11	No campo das políticas públicas, a implementação das ações afirmativas propiciaram oportunidades a esse segmento que foi esquecido desde a época da abolição da escravidão. 15. O sucesso alcançado por Lula permitirá que o futuro Governo seja não somente uma continuidade do até agora realizado 16. O Governo Lula criou as condições para um Projeto de Desenvolvimento Nacional Democrático Popular, sustentável e de longo prazo para o país. 17. O Brasil deixou de ser o eterno país do futuro. 18. O futuro chegou. E o pós-Lula é Dilma. o crescimento ...
12	19. A expansão e o fortalecimento do mercado de bens de consumo popular, que produziu forte impacto positivo sobre o conjunto do setor produtivo, se dará por meio da: a) preservação da estabilidade econômica, elevação dos investimentos e aumento da produtividade sistêmica, via desenvolvimento da infra-estrutura logística, energética e de comunicações; b) fortalecimento dos processos de produção, visando aumentar a competitividade nacional e agregar mais valor às exportações; c) ampliação do emprego formal; d) manutenção da política de valori...
15	Investimentos, crédito, ciência e inovação tecnológica a serviço de um novo desenvolvimento 20. As possibilidades abertas para a sociedade com os grandes avanços científicos e tecnológicos, combinadas com a necessidade de expansão do mercado interno e a dura competitividade global, que a crise acentuou, exigem uma profunda transformação do sistema produtivo
18	É necessário incentivar a participação privada por meio de empréstimos e o incentivo ao mercado de capitais; b) apoio à internacionalização das empresas brasileiras, garantido o interesse nacional e respeitada a soberania e as leis das nações; c) fortalecimento da EMBRAPA, priorizando a agricultura familiar e as suas atividades para estratégias da soberania alimentar e nutricional do país e para a cooperação cientifica no campo das pesquisas agropecuárias com os países em desenvolvimento; d) flexibilização da proteção a direitos relativos à ...
19	k) ampliação da desconcentração do sistema de ciência e tecnologia no território nacional; I) exercício do poder de compra do Estado para a indução da demanda nacional de ciência, tecnologia e inovação; m) implantação de projetos de desenvolvimento científico e tecnológico com países desenvolvidos e com os da América do Sul, África e outras regiões, a exemplo do que foi feito com a 1V Digital e do que vem sendo proposto na área de Defesa; n) construção de mecanismos para que os investimentos estrangeiros sejam vinculados à efetiva e inovador...
20	Infra-estrutura para impulsionar o desenvolvimento agrícola, industrial e comercial do país 22. A elevação das taxas de crescimento, que deverá marcar o Governo Dilma, exigirá a conclusão das obras do Plano de Aceleração do Crescimento. O PAC-l e o que estará no PAC-2 acentuarão a competitividade da economia brasileira mas, sobretudo, propiciarão consideráveis melhorias das condições de vida dos brasileiros
36	e) ressarcir o SUS por atendimentos públicos dispensados aos usuários de planos e seguros de saúde e fortalecer o monitoramento, avaliação, controle e regulação do setor, f) melhorar a gestão dos serviços do SUS por meio de novos métodos e tecnologias, principalmente para as unidades públicas de saúde, g) atender plenamente às necessidades qualitativas e quantitativas de recursos humanos do setor de saúde no Brasil, inclusive com a ampliação do aparelho formador, h) assegurar direitos trabalhistas e previdenciários aos trabalhadores do setor...
39	Pelos cálculos da Fundação Getúlio Vargas, 24 milhões de pessoas deixaram a pobreza entre 2003 de 2009. O mercado interno, mais fortalecido pelo poder de compra dos mais pobres, permitiu ao país enfrentar a crise econômica mundial de cabeça erguida, sem conseqüências mais graves para nossa economia 39. Esses resultados precisam ser mantidos a médio e longo prazo, para evitar retrocessos, e ainda temos de considerar o peso da dívida social acumulada por mais de 500 anos. Por isso, as políticas sociais precisam ser trabalhadas numa perspectiva...
69	Presença do Brasil no mundo 76. A Política Externa do Brasil tem profunda incidência em nosso Projeto Nacional de Desenvolvimento. Ela busca a defesa do interesse nacional e se nutre de valores como o multilateralismo, a paz, o respeito aos Direitos Humanos, a democratização das relações internacionais e a solidariedade com os países pobres e em desenvolvimento

Tabela 2.2: Extratos da Tabela 2.1 processados.

id	2010PT01
2	dois decada estagnacao avanco mediocre economia_brasileiro voltar crescer crescimento obedecer hoje logico distinto de aquele passado fazer forte distribuicao renda inedito equilibrio macroeconomico reducao vulnerabilidade externo sobretudo fortalecimento democracia
3	ano periodo expansao economia_brasileiro acabar frustrar expectativa maioria sociedade ora concentrar riqueza ora vir acompanhar surto inflacionario ora produzir elevar endividamento externo interno ora sufocar democracia partir ano investimento produtivo reduzir pai sofrer restricao parque industrial infraestrutura comprometer sobretudo area energia transporte politico debilitar empresa estatal algum privatizar programar brasil ser pai pequeno conformar existencia milhao homem mulher qual haver espaco acesso riqueza produzir
4	governo lulo crescimento pib expansao emprego formal aumento real salario minimo politico transferencia_renda controle inflacao queda taxa juro ampliacao credito medida reforma agrario apoio agricultura familiar aumento comercio exterior reconstrucao infraestrutura mudar tudo
5	mudanca propiciar formacao grande mercado bem consumo_popular proteger efeito devastador crise mundial desencadear setembro protecao reforcar saude sistema bancario reserva internacional acumular diferentemente crise externo anterior mexicano asiatico russo
11	campo politico publico implementacao acao afirmativo propiciar oportunidade segmento esquecer desde epoca abolicao escravidao sucesso alcancar lulo permitir futuro governo somente continuidade agora realizar governo lulo criar condicao projeto desenvolvimento nacional democratico popular sustentavel longo_prazo pai brasil deixar ser eterno pai futuro futuro chegar posl lua dilmo crescimento acelerar combate desigualdade racial social regional promocao sustentabilidade ambiental eixo ir estruturar desenvolvimento economico
12	expansao fortalecimento mercado bem consumo_popular produzir forte impacto positivo sobre conjunto setor produtivo dar meio preservacao estabilidade economico elevacao investimento aumento produtividade sistemico via desenvolvimento infraestrutura logistico energetico comunicacao b fortalecimento processo producao visar aumentar competitividade nacional agregar valor exportacao c ampliacao emprego formal d manutencao politico valorizacao salario minimo crescimento renda trabalhador aumento salarial eficiente politico publico educacao saude ...
15	investimento credito ciencia inovacao tecnologico servico novo desenvolvimento possibilidade aberto sociedade grande avanco cientifico tecnologico combinar necessidade expansao mercado_interno duro competitividade global crise acentuar exigir profundo transformacao sistema produtivo
18	necessario incentivar participacao privado meio emprestimo incentivo mercado capital b apoio internacionalizacao empresa brasileiro garantir interesse nacional respeitar soberania lei nacao c fortalecimento embrapo priorizar agricultura familiar atividade estrategia soberania alimentar nutricional pai cooperacao cientifico campo pesquisa agropecuario pai desenvolvimento d flexibilizacao protecao direito relativo propriedade intelectual sobre cultivar variedade vegetal ambito programa publico direcionar seguranca alimentar nutricional populac...
19	k ampliacao desconcentracao sistema ciencia tecnologia territorio nacional i exercicio poder_compra estar inducao demanda nacional ciencia tecnologia inovacao m implantacao projeto desenvolvimento cientifico tecnologico pai desenvolver americo sul africo outro regiao exemplo fazer v digital vir ser propor area defesa n construcao mecanismo investimento estrangeiro vincular efetivo inovador transferencia tecnologia poder promover atracao centro internacional pesquisa desenvolvimento brasil
20	infraestrutura impulsionar desenvolvimento agricola industrial comercial pai elevacao taxa crescimento dever marcar governo dilmo exigir conclusao obra plano aceleracao crescimento pacl estar pac acentuar competitividade economia_brasileiro sobretudo propiciar consideravel melhoria condicao vida brasileiro
36	ressarcir sus atendimento publico dispensar usuario plano seguro saude fortalecer monitoramento avaliacao controle regulacao setor f melhorar gestao servico sus meio novo metodo tecnologia principalmente unidade publico saude g atender plenamente necessidade qualitativo quantitativo recurso humano setor saude brasil inclusive ampliacao aparelho formador h assegurar direito_trabalhista previdenciario trabalhador setor reconhecer diversidade regional implantar novo carreira estrategico articulacao estado municipio criterio meritocratico seleca...
39	calculo fundacao getulio varga milhao pessoa deixar pobreza mercado_interno fortalecer poder_compra pobre permitir pai enfrentar crise economico mundial cabeca erguer consequencia grave economia resultado precisar ser manter medio longo_prazo evitar retrocesso ainda considerar peso divida social acumular ano politico social precisar ser trabalhar perspectiva medio longo_prazo estar presente forma enfatico programa dilmo presidente
69	presenca brasil mundo politica externo brasil profundo incidencia projeto nacional desenvolvimento busca defesa interesse nacional nutrar valor multilateralismo paz respeito direito humano democratizacao relacao_internacional solidariedade pai pobre desenvolvimento

Tabela 2.3: Frequências de ocorrência de certos elementos léxicos dos extratos da Tabela 2.1.

doc_id	brasil	politico	economia_brasileiro	crescimento	fortalecimento	pai	transferencia_renda	consumo_popular	saude	nacional	desenvolvimento	publico	longo_prazo	setor	mercado_interno	poder_compra
2	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	0	0	2	0	0	0	0	0	0	0	0	0	0
4	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
11	1	1	0	1	0	2	0	0	0	1	2	1	1	0	0	0
12	0	2	0	1	2	0	1	1	2	1	1	1	0	1	0	0
15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
18	0	0	0	0	4	2	0	0	0	1	5	1	0	2	0	0
19	1	0	0	0	0	1	0	0	0	2	2	0	0	0	0	1
20	0	0	1	2	0	1	0	0	0	0	1	0	0	0	0	0
36	1	0	0	0	0	0	0	0	5	2	0	4	0	4	0	0
39	0	1	0	0	0	1	0	0	0	0	0	0	2	0	1	1
69	2	0	0	0	0	1	0	0	0	2	2	0	0	0	0	0

e de comparação entre os textos do corpus. É nesse sentido que a análise por tópicos se enquadra no processo de análise de documentos de agremiações partidárias.

Nesta dissertação, utilizamos alguns métodos que são variações do pioneiro *Latent Dirichlet Allocation (LDA)* descrito originalmente em (BLEI; NG; JORDAN, 2003). O modelo probabilístico generativo bayesiano hierárquico apresentado a seguir reúne as características comuns a esses métodos. O essencial do modelo está no fato de que cada texto é visto como uma mescla de um conjunto de tópicos e cada tópico é, por sua vez, modelado como uma mescla de elementos léxicos. Além desses aspectos mais gerais, há alguns detalhes que tornam a descrição do modelo mais precisa. Ao longo desta seção, apresentamos os conceitos de análise por tópicos mais detalhadamente.

2.2.1 Corpus

Textos são constituídos de palavras de um certo vocabulário, sendo que cada uma dessas palavras transmite um (ou mais, dependendo do contexto) signficado elementar. Como uma extensão dessa ideia, um elemento léxico é uma expressão formada por uma sequência de uma ou mais palavras que transmitem, em conjunto, um significado elementar. Dessa forma, palavras isoladas são elementos léxicos e, além delas, expressões como

economia brasileira, consumo popular, poder compra, transferência renda, mercado interno, longo prazo

são exemplos de termos compostos que são entendidos como elementos léxicos. Esses exemplos foram escolhidos dentre os que ocorrem nos textos da Tabela 2.1 e que dão origem aos elementos léxicos

economia_brasileiro, transferencia_renda, consumo_popular, longo_prazo, mercado_interno, poder_compra

da Tabela 2.2. Tipicamente, algoritmos de detecção de ocorrências concomitantes e subsequentes são utilizados para determinar esses elementos léxicos (os quais estão além do escopo desta dissertação, mas vários deles podem ser encontrados em (MANNING; SCHÜTZE, 1999)).

A descrição de um corpus parte dos seus elementos léxicos, reunidos no conjunto $\mathbf{L} = \{\ell_1, \ldots, \ell_L\}$. O corpus é, então, definido pelo conjunto $\mathbf{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_D\}$ de D textos no qual cada $\mathbf{d}_i = \{w_1^i, \ldots, w_d^i\}$ é uma cesta de d elementos léxicos não necessariamente distintos de $\mathbf{L}$. O termo *cesta* visa indicar que a ordem na qual os elementos léxicos estão dispostos não é levada em consideração. A cada elemento léxico w_j^i de $\mathbf{d}_i$ é associada uma frequência de ocorrência f_j^i, indicando a quantidade de vezes que w_j^i aparece em $\mathbf{d}_i$. Cabe mencionar a existência de modelos propostos recentemente que adotam outras estruturas de codificação de textos capazes de capturar algumas de suas propriedades sintáticas ou semânticas (COSTA; ORTALE, 2021).

2.2.2 Tópicos

Os tópicos são estruturas com valor semântico intrínsico que formam grupos de elementos léxicos que relacionam-se, no sentido de seguir algum padrão de ocorrência concomitante, nos textos analisados. A identificação de tais grupos de elementos léxicos fornece indícios

Tabela 2.4: Relevância semântica de alguns elementos léxicos em uma análise com 5 tópicos.

	brasil	politico	economia_brasileiro	crescimento	fortalecimento	pai	transferencia_renda	consumo_popular	saude	nacional	desenvolvimento	publico	longo_prazo	setor	mercado_interno	poder_compra
β_1	0.0017	0.0104	0.0000	0.0000	0.0120	0.0052	0.0000	0.0000	0.0036	0.0128	0.0090	0.0069	0.000	0.0017	0.0000	0.0000
β_2	0.0055	0.0191	0.0012	0.0053	0.0042	0.0058	0.0000	0.0000	0.0055	0.0150	0.0065	0.0098	0.000	0.0000	0.0000	0.0000
β_3	0.0091	0.0082	0.0023	0.0047	0.0075	0.0164	0.0012	0.0000	0.0007	0.0118	0.0153	0.0070	0.000	0.0069	0.0012	0.0012
β_4	0.0068	0.0048	0.0000	0.0025	0.0046	0.0000	0.0000	0.0000	0.0130	0.0073	0.0012	0.0118	0.000	0.0060	0.0000	0.0000
β_5	0.0100	0.0268	0.0000	0.0029	0.0052	0.0092	0.0013	0.0027	0.0040	0.0096	0.0161	0.0042	0.004	0.0012	0.0013	0.0013

Tabela 2.5: Relevância tópica.

	Tópico 1	**Tópico 2**	**Tópico 3**	**Tópico 4**	**Tópico 5**
$\boldsymbol{\theta}_2$	7e-04	0.9970	0.0007	0.0007	0.0007
$\boldsymbol{\theta}_3$	3e-04	0.0003	0.9988	0.0003	0.0003
$\boldsymbol{\theta}_4$	6e-04	0.0006	0.9977	0.0006	0.0006
$\boldsymbol{\theta}_5$	7e-04	0.0007	0.0007	0.0007	0.9973
$\boldsymbol{\theta}_{11}$	3e-04	0.6283	0.0003	0.0003	0.3707
$\boldsymbol{\theta}_{12}$	2e-04	0.0002	0.0282	0.0002	0.9712
$\boldsymbol{\theta}_{15}$	7e-04	0.0007	0.9973	0.0007	0.0007
$\boldsymbol{\theta}_{18}$	1e-04	0.0001	0.9994	0.0001	0.0001
$\boldsymbol{\theta}_{19}$	4e-04	0.0004	0.9985	0.0004	0.0004
$\boldsymbol{\theta}_{20}$	6e-04	0.0006	0.9977	0.0006	0.0006
$\boldsymbol{\theta}_{36}$	1e-04	0.0001	0.0001	0.9996	0.0001
$\boldsymbol{\theta}_{39}$	4e-04	0.0004	0.0004	0.0004	0.9985
$\boldsymbol{\theta}_{69}$	7e-04	0.0007	0.0007	0.0007	0.9973

sobre a presença de um tema ou assunto na composição dos textos que, presume-se, são semanticamente relacionados. Denotamos por $\mathbf{T} = \{1, \ldots, T\}$ o conjunto de T tópicos que são combinados para a formação dos textos em $\mathbf{D}$, sendo que $\boldsymbol{\beta}_i : \mathbf{L} \to [0, 1]$, denominada *relevância semântica* do tópico i, quantifica o conteúdo semântico de i. Mais especificamente, $\boldsymbol{\beta}_i$ expressa a relevância dos elementos léxicos na forma de uma função massa de probabilidade sobre o vocabulário $\mathbf{L}$, atribuindo uma probabilidade de ocorrência $\boldsymbol{\beta}_i(\ell)$ a cada elemento léxico ℓ do vocabulário. Apesar de o vocabulário ser único para todos os tópicos, as respectivas relevâncias semânticas variam de tópico a tópico. Na Tabela 2.4 é apresentado um exemplo de valores típicos para os elementos léxicos mencionados na Tabela 2.3. Nesse exemplo, observa-se que o elemento léxico *politico* tem relevância semântica de 0.0268 no tópico 5, sendo portanto mais relevante para esse tópico do que para o tópico 4 (para o qual a relevância semântica é 0.0048). Um elemento léxico que não tenha qualquer relevância para um determinado tópico terá atribuída a ele uma probabilidade muito próxima ou igual a zero.

A quantidade de tópicos (denotada por T) é fixa e definida previamente à análise. A cada um desses T tópicos é atribuída uma *relevância tópica* com relação a cada texto consistindo na proporção em que o tópico em questão ocorre no texto. O conjunto de relevâncias tópicas com relação a um texto t forma uma função massa de probabilidade $\boldsymbol{\theta}_t : \mathbf{T} \to [0, 1]$. A título de ilustração, na Tabela 2.5 observamos que o texto 11 tem influência significativa dos tópicos 2, 5, fato indicado pelas relevâncias tópicas 0.6283, 0.3707, respectivamente. A proporção de ocorrência de cada elemento léxico em um texto depende das relevâncias tópicas dos tópicos que formam esse texto, assim como da sua relevância semântica nesses tópicos. Diferentes ocorrências de um mesmo elemento léxico em um mesmo texto podem estar associadas a diferentes tópicos.

2.2.3 Dinâmica do Modelo e Resultado da Análise

O objetivo de uma análise por tópicos é verificar a hipótese de abdução de que o corpus sendo analisado é uma amostra do que um certo modelo aleatório gerador de textos pode gerar. Nesse sentido, o modelo que estamos descrevendo seria uma explicação plausível para o perfil semântico dos textos em $\mathbf{D}$. A questão que se coloca, portanto, é de estimar os parâmetros latentes do modelo a partir dos textos do corpus, os quais, por hipóteses, formariam uma amostragem. Usamos aqui o termo latente para fazer referência a todos os elementos que interferem no funcionamento do modelo e que estão ocultos. Dentre esses elementos, estão as relevâncias semânticas e tópicas, as quais são estimadas por meio de métodos estatísticos aplicados sobre o padrão de ocorrência dos elementos léxicos nos textos. Os fundamentos dos métodos empregados nesta dissertação podem ser consultados em (BLEI; NG; JORDAN, 2003) e em outras referências citadas nas seções seguintes. Esses detalhes são omitidos por envolverem conceitos específicos de probabilidade e estatística que estão além do escopo deste trabalho.

A descrição da dinâmica de funcionamento do modelo na geração de textos consiste especificação dos passos efetuados para a geração de textos. Nessa apresentação, aparece mais um elemento latente importante. Tratando-se de um modelo hierárquico, as relevâncias tópica e semântica são empregadas respectivamente no segundo e no terceiro níveis da hierarquia. O primeiro nível é ocupada pela estratégia de atribuição de relevâncias tópicas. Sendo estas últimas funções massa de probabilidade, a sua estratégia de atribuição nada mais é do que a escolha do critério de escolha dessas funções. Essa estratégia é o elemento que diferencia os métodos apresentados a seguir. Vejamos a seguir o desenrolar do processo geral antes de descrever as especifidades dos métodos.

A geração de um texto no modelo geral poderia ser caracterizada figurativamente segundo a engenhosa descrição em (TOMAR, 2018) como o funcionamento de uma máquina que joga dados viciados. Nessa visão figurativa, um dado viciado é um objeto de diversas faces que, quando atirado, acaba por repousar com uma das faces expostas. Acontece que, em um dado viciado, não necessariamente todas as faces têm a mesma probabilidade de ocorrer como resultado de um lançamento. As chances das diferentes faces serem escolhidas em um experimento de jogar o dado são distintas, justificando a designação de dado viciado. São dois os tipos de dados viciados utilizados. O primeiro tipo, que denominamos *dado de tópicos*, tem em suas faces os diferentes tópicos. Trata-se, portanto, de uma dado com T faces e a chance de cada face é a relevância tópica do tópico correspondente. O segundo tipo é o *dado léxico* cujas faces são os elementos léxicos do vocabulário $\mathbf{L}$. Logo, esse tipo de dado viciado possui L faces, sendo a chance de escolha de cada face dada pela relevância semântica correspondente. O dado de tópico de cada texto t é regido por $\boldsymbol{\theta}_t$, enquanto o dado léxico de cada tópico i é regido por $\boldsymbol{\beta}_i$. A geração de um texto depende de um dado de tópicos (cada texto tem seu próprio dado de tópicos) e T dados léxicos, cada qual para um tópico. Os mesmos dados léxicos são empregados em todos os textos.

Relembrando da premissa de que um texto é, para o propósito de análise, plenamente caracterizado como uma cesta de elementos léxicos, a geração de um texto busca apenas definir os elementos léxicos que o constituem, com as respectivas frequências de ocorrência, visto que a disposição desses elementos léxicos não é relevante. Supondo que os dados viciados necessários estejam disponíveis, primeiro estabelece-se o tamanho do texto, ou seja, quantos elementos léxicos ele deve conter para, em seguida, gerar um elemento léxico após o outro da seguinte forma: atira-se o dado de tópicos para escolher um tópico e, ato

contínuo, atira-se o dado léxico do tópico escolhido para escolher o elemento léxico. Este é adicionado ao texto, repetindo o experimento iterativamente até completar o cesto com a quantidade estabelecida.

Vista a dinâmica de funcionamento do modelo gerador, vemos o papel desempenhado pelas relavâncias tópica e semântica no processo. São exatamente esses elementos que buscamos reconstruir. Uma reconstrução que, a rigor, é uma busca pelos parâmetros que fazem o modelo funcionar, os parâmetros que calibram os "dados viciados" que definem o modelo. Cabe esclarecer que o modelo é capaz de gerar diferentes textos, mas com um perfil semântico que guarda relação com as relevâncias tópica e semântica. A Figura 2.2 mostra o quadro esquemático do modelo, adaptado de (DOIG, 2015).

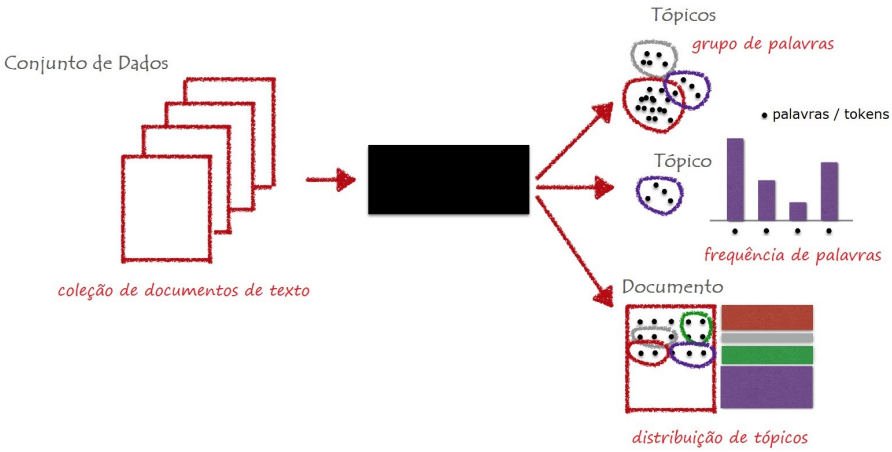


Figura 2.2: Modelo hierárquico de análise por tópicos.

Em suma, o processo de geração de um texto pelo modelo que acabamos de descrever a partir de um conjunto $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\}$ de T dados léxicos segue os seguintes passos:

1. Escolher, segundo a estratégia de atribuição de tópicos, a relevância tópica $\boldsymbol{\theta}$ para o dado de tópicos.
2. Escolher um valor N para o tamanho do texto.
3. Repetir os seguintes passos N vezes:
   - Escolher aleatoriamente um tópico i atirando o dado de tópicos.
   - Escolher aleatoriamente um elemento léxico w atirando o dado léxico correspondente ao tópico i.
   - Incluir w na cesta.

2.3 Estratégias de Atribuição de Tópicos

A conclusão da exposição de funcionamento do modelo de geração de textos dos métodos de análise por tópicos se faz estabelecendo o procedimento efetuado para construir os dados de tópicos e léxicos. Descrevemos a seguir dois métodos que se diferenciam, no essencial, no tipo de procedimento empregado.

2.3.1 Atribuição Semeada de Tópicos por *Dirichlet*

O método LDA Semeado de atribuição de tópicos é uma variação do LDA que envolve a utilização a indução de elementos léxicos, denominados como *sementes*. Visto que os

tópicos são induzidos usando um dicionário de sementes, trata-se de um modelo não totalmente supervisionado, mas semisupervisionado (LU et al., 2011). O papel das sementes é guiar a análise efetuada pelo método. No entanto, o perfil dos tópicos é ajustada mantendo algumas das sementes como elementos léxicos relevantes no respectivo tópico, mas outros elementos léxicos (além das sementes) podem também ser incluídos no rol dos elementos léxicos relevantes.

A estratégia de atribuição de tópicos aos documentos na etapa 1 é baseada na chamada distribuição de Dirichlet, assim como ocorre no método LDA. Como uma distribuição de distribuições, Dirichlet governa a escolha do dado de relevância tópica $\boldsymbol{\theta}$ usado para alocar as palavras do documento em diferentes tópicos PAULO FALEIROS; ANDRADE LOPES (2016). O processo possui as seguintes características:

- Dados de elementos léxicos são fixos para todos os documentos. A distribuição de palavras dentro de um tópico (ou seja, conteúdo do tópico) é estacionária no sentido em que o tópico i do documento $\mathbf{d}$ usa palavras idênticas ao tópico i dos demais documentos. Nesse sentido, os perfis dos documentos se diferenciam pelas correspondentes relevâncias tópicas.
- A geração dos dados de tópicos é feita usando a distribuição de distribuições Dirichlet, o que não leva em conta as possíveis correlações entre as relevâncias tópicas dos tópicos.
- O parâmetro da distribuição de Dirichlet determina o perfil dos dados gerados. Dependendo desse parâmetro, os dados podem ser mais inclinados a poucos tópicos ou mais igualitários sobre os tópicos. Um dado mais inclinado a um tópico indica que esse tópico prevalece sobre os demais na publicação correspondente. Por outro, um dado mais igualitário indica uma combinação de vários tópicos. Os principais parâmetros do modelo LDA são o α que indica de quantos tópicos os documentos são compostos e que permite ou não uma distribuição de tópicos por documento mais específica – quanto maior, maior a quantidade de tópicos, mais específica a distribuição; e o β que indica de quantas palavras os tópicos são compostos e que permite ou não uma distribuição de palavras por tópico mais específica – quanto maior, maior a quantidade de palavras, mais específica a distribuição.
- O ajuste no parâmetro da distribuição de Dirichlet depende do valor de T. Quanto maior esse valor, mais os tornam-se específicos, tendo como consequência uma maior participação de tópicos nas publicações.

2.3.2 STM – Atribuição de Tópicos por Estrutura

O *Modelo de Análise Estrutural por Tópicos* (cuja sigla é *STM*) é um método de análise por tópicos projetada especificamente com aplicações em Ciências Sociais em mente. Nesse sentido, a principal característica do STM é a possibilidade de incorporação flexível, na análise, de *metadados* associados aos textos. Esse é o aspecto que exploramos nos capítulos subsequentes desta dissertação. Na prática, informações conhecidas e identificadas sobre a origem dos textos podem ser expressas como valores de metadados, sendo plausível imaginar que as relevâncias tópicas estejam correlacionados com os metadados. Por exemplo, certas fontes podem ser mais propensas a escrever sobre política econômica e desigualdade, enquanto outras não entrem nesses mesmos temas. Os metadados podem incluir data de publicação, autor, publicação, curtidas nas mídias sociais ou qualquer número de variáveis categóricas ou numéricas sobre um texto.

Outra característica importante do STM que merece ser destacada é a aptidão em admitir que os tópicos podem ser abordados com diferentes nuances em textos distintos. Do ponto de vista da Ciência Política, dois documentos podem tratar sobre um mesmo conjunto de assuntos – digamos economia e relações de trabalho – mas abordar esses assuntos de ângulos bastante distintos. Tal diferença de abordagem, supõe-se, manifesta-se pelos aspectos a enfatizar e aqueles a negligenciar. Por exemplo, pode-se ter um texto que tenda a destacar "direitos trabalhistas" ou "empregos formais", enquanto outro se apoia sobre termos como "produtividade do trabalho" e "remuneração". Ambos os textos podem ser semelhantes quanto aos tópicos prevalentes, mas o conteúdo desses tópicos (ou seja, as palavras que os compõem) varia de texto para texto. Apesar de importante, essa última característica não é explorada ao longo do estudo levado a cabo aqui. Assim sendo, omitimos o seu detalhamento para evitar um excesso de formalismo. Mais detalhes podem ser encontrados em ROBERTS; STEWART; AIROLDI (2016), que são as principais fontes da descrição que segue.

A inovação do STM referente à maneira flexível de incorporar *metadados* advém de uma modelagem explícita das possíveis correlações existentes entre os tópicos com o intuito de tornar a compreensão desses tópicos mais realista. Tal propriedade de permitir que as proporções dos tópicos sejam correlacionadas apareceu inicialmente como uma evolução do LDA (no qual pressupõe-se que os tópicos sejam independentes) em (BLEI; LAFFERTY, 2007). Mais especificamente, as correlações intervêm na determinação dos dados de tópicos na forma da matriz de covariâncias $\mathbf{\Sigma}$ como parâmetro da distribuição normal logística AITCHISON (1982). Os metadados são então usados para que a média $\boldsymbol{\mu}_{\mathbf{d}}$ usada na distribuição normal logística deixe de ser única para todo documento $\mathbf{d}$. A mudança introduzida é que $\boldsymbol{\mu}_{\mathbf{d}}$ passa a ser definida dependente do grupo de textos que compartilham os mesmos valores de metadados que o documento $\mathbf{d}$. Assim sendo, se $\mathbf{d}_1$ e $\mathbf{d}_2$ são dois textos do mesmo grupo, $\boldsymbol{\mu}_{\mathbf{d}_1} = \boldsymbol{\mu}_{\mathbf{d}_2}$. Em caso contrário, as médias $\mathbf{d}_1$ e $\mathbf{d}_2$ podem diferir entre si. A relevância tópica do tópico i no texto t é dada por

$$\boldsymbol{\theta}_{\mathbf{d}}(i) = \frac{e^{\boldsymbol{\eta}_{\mathbf{d}}(i)}}{\sum_{j=1}^{T} e^{\boldsymbol{\eta}_{\mathbf{d}}(j)}} \tag{2.1}$$

sendo $\boldsymbol{\eta}_{\mathbf{d}} \sim \mathbf{N}_{T-1}(\boldsymbol{\mu}_{\mathbf{d}}, \mathbf{\Sigma})$ e $\boldsymbol{\eta}_{\mathbf{d}}(T)$ é fixado em zero a fim de tornar o modelo identificável. De (2.1) deduz-se que os textos de um mesmo grupo têm perfis ideológicos definidos pela mesma estratégia de atribuição de tópicos. Portanto, os metadados fornecem uma maneira de estruturar as atribuições de tópicos por grupo, injetando informações valiosas no procedimento de inferência. Embora esse método seja computacionalmente mais complexo, ele permite uma modelagem mais realista segundo aferições empíricas na literatura (BLEI; LAFFERTY, 2007).

O resultado de uma análise com STM inclui, além das relevâncias tópicas e semânticas, as médias e a matriz de covariâncias usadas na estratégia de atribuição de tópicos. A título de exemplo, vejamos o resultado de uma análise com STM do corpus formado pelos textos da Tabela 2.2, acrescidos dos extratos na Tabela 2.6. O metadados utilizado é a agremiação política, PT ou PSB, de cada extrato. Na Tabela 2.7, a matriz de covariância entre os tópicos é apresentada, com a qual se pode ver os tópicos que mais aparecem juntos e os que são mais independentes. Na Tabela 2.8, aparecem algumas das médias de textos de cada grupo, e pode ser observado que essas médias são idênticas por grupo, gerando as relevâncias tópicas da Tabela 2.9.

Tabela 2.6: Extratos adicionais à Tabela 2.2.

id	1995PSB01	Grupo
75	manifesto reorganizar partido socialista brasileiro psb quarenta ano apo fundacao animar mesmo proposito socialista democratico motivar instituidor partido reorganizar apo a guerra mundial vitoria sobre fascismo agora ressurgir apo vinte ano ditadura militar	PSB
76	ambo momento ditaduro enfrentar derrotar amplo legitimo frente democratico hoje bem passado vencer violencia autoritario impor organizacao todo forca politico partido dever revelar nitidez programa pratica programa adotar fundador partido dramatico atualidade quarenta ano pai ver prisioneiro mesmo forma exploracao ainda agravar brutalidade ditadura militar programa si denuncia caber vida partidario incorporar programa denuncia combate antigo forma exploracao agora bom identificar comprovar discriminacao racial opressao minoria mulher crianca violencia contra manifestacao cultural alternativo degradacao qualidade vida depredacao meio ambiente genocidio nacao indigena	PSB
77	haver lugar moderno declaracao direito ser humano contemplar efetivo garantia cidadania face controle exercer grande corporacao estatal privado mediante uso informatica meio comunicacao massa agreguar direito individual tradicional direito social educacao saude transporte publico habitacao saneamento basico direito vizinhanca segurodesemprego novo forma organizacao social comunitario direito privacidade acesso informacao controle atividade estatal amplo participacao politico	PSB
78	finalmente partido socialista moderno estar aberto descentralizacao completo poder interferencia sistematica cidadao tempo buscar valorizar soberania popular mediante controle legislativo atividade estar economia progressivamente socializar partido porque socialista conformar apena programa democratico organizacao democratico avesso maquina partidario clientela oligarquia plano externo psb lutar principio autodeterminacao povo fortalecimento organismo internacional contra todo forma imperialismo colonialismo belicismo em ele incluir proposta hegemonico grande potencia organizacao pai terceiro mundo grande entendimento nacao latinoamericano lutar comum afirmacao soberano interesse nacional inclusive negociacao profundo divida externo contraidar governo ilegitimo	PSB
79	psb partido abrir vontade vontade militante execucao programa convocar todo setor movimento popular democratico socialista defesa regime civil liberdade publico dispor aliar todo brasileiro assembleia nacional constituinte momento decisivo reorganizacao democratico estado brasileiro convocar todo socialista participar eleicao em ela cumprir papel socialismo liberdade	PSB

Tabela 2.7: Matriz de covariâncias.

	Tópico 1	**Tópico 2**	**Tópico 3**	**Tópico 4**
Tópico 1	12.1888	6.2579	7.0777	7.0391
Tópico 2	6.2579	11.8521	7.2790	3.7908
Tópico 3	7.0777	7.2790	15.0672	7.7444
Tópico 4	7.0391	3.7908	7.7444	13.0523

Tabela 2.8: Médias.

	μ_{65}	μ_{66}	μ_{67}	μ_{68}	μ_{69}	μ_{70}	μ_{71}	μ_{72}	μ_{73}	μ_{74}	μ_{75}	μ_{76}	μ_{77}	μ_{78}
Tópico 1	-0.4227	-0.4227	-0.4227	-0.4227	-0.4227	-0.4227	-0.4227	-0.4227	-0.4227	5.0525	5.0525	5.0525	5.0525	5.0525
Tópico 2	0.3406	0.3406	0.3406	0.3406	0.3406	0.3406	0.3406	0.3406	0.3406	-0.1776	-0.1776	-0.1776	-0.1776	-0.1776
Tópico 3	0.1076	0.1076	0.1076	0.1076	0.1076	0.1076	0.1076	0.1076	0.1076	-0.4207	-0.4207	-0.4207	-0.4207	-0.4207
Tópico 4	0.7453	0.7453	0.7453	0.7453	0.7453	0.7453	0.7453	0.7453	0.7453	0.5373	0.5373	0.5373	0.5373	0.5373

Tabela 2.9: Relevâncias tópicas.

	Tópico 1	**Tópico 2**	**Tópico 3**	**Tópico 4**	**Tópico 5**
$\boldsymbol{\theta}_{65}$	0.9673	0.0106	0.0050	0.0112	0.0059
$\boldsymbol{\theta}_{66}$	0.8603	0.0570	0.0459	0.0231	0.0138
$\boldsymbol{\theta}_{67}$	0.0108	0.0262	0.9408	0.0156	0.0067
$\boldsymbol{\theta}_{68}$	0.2368	0.3644	0.0500	0.0176	0.3312
$\boldsymbol{\theta}_{69}$	0.0110	0.9517	0.0170	0.0088	0.0115
$\boldsymbol{\theta}_{70}$	0.0069	0.0315	0.9406	0.0144	0.0065
$\boldsymbol{\theta}_{71}$	0.0076	0.0064	0.0195	0.9582	0.0083
$\boldsymbol{\theta}_{72}$	0.0044	0.9726	0.0126	0.0039	0.0065
$\boldsymbol{\theta}_{73}$	0.0051	0.0203	0.9524	0.0156	0.0066
$\boldsymbol{\theta}_{74}$	0.9890	0.0024	0.0017	0.0044	0.0025
$\boldsymbol{\theta}_{75}$	0.9936	0.0016	0.0010	0.0025	0.0014
$\boldsymbol{\theta}_{76}$	0.8484	0.0048	0.0046	0.0612	0.0810
$\boldsymbol{\theta}_{77}$	0.9936	0.0016	0.0010	0.0023	0.0015
$\boldsymbol{\theta}_{78}$	0.9896	0.0023	0.0015	0.0039	0.0027

3 POSICIONAMENTO IDEOLÓGICO DE AGREMIAÇÕES PARTIDÁRIAS

Medidas empíricas de posicionamento e afinidade ideológicos são elementos usados para testar modelos de disputa política, não obstante as inúmeras questões de difícil enfretamento envolvidas. Trata-se aqui de contemplar a superestrutura composta por diversas representações que compõem a consciência e que regem formulações e atuações na cena política. Afinal, como se devem extrair as dimensões ideológicas nos documentos de partidos políticos e coalizões de governo? Como representar a multidimensionalidade do espectro de atributos para análise de afinidade ou disparidade política das agremiações brasileiras? Nesse contexto, definir e auscultar medidas que sejam efetivamente métricas válidas e confiáveis têm se tornado grandes desafios na análise da atuação de partidos políticos nas Ciências Sociais. Ao longo deste capítulo, reunimos alguns dos métodos encontrados na literatura que corroboram com essa afirmação.

3.1 Natureza de Documentos Textuais

Análises quantitativas são feitas sobre registros concretos de atividades políticas. Há diversas formas de registro (discursos, projetos de lei, redes sociais, manifestos, programas etc), mas nesta dissertação trataremos de documentos textuais específicos, como manifestos, teses e programas de governo. Além de outras razões, essas formas de registro permitem a automatização de muitas etapas de análise e podem ser encontradas em forma digital na rede mundial de computadores. Nesse sentido, a análise efetivamente realizada é do posicionamento expresso nos documentos. Logo, trata-se de uma análise indireta das agremiações envolvidas. Em razão dessa observação, as referências feitas nos textos aos posicionamentos das agremiações políticas dizem respeito ao que se encontra expresso nos documentos utilizados na análise.

Em geral, a análise de partidos políticos é feita por abordagens voltadas para a distinção entre a heterogeneidade programática e a polarização ideológica (FRANZMANN, 2011). Diferenças programáticas são diretamente relacionadas aos tipos de perfis políticos que os partidos apresentam em seus manifestos e programas de governo. Temas de posições e diferenças ideológicas dizem respeito a temas com uma certa divisão ideológica ou clivagem social. Uma medida de posição ideológica visa dizer algo substantivo sobre a localização do partido isoladamente no contexto social mais amplo dos sistemas partidários.

Apesar da existência dessas duas abordagens de análise, neste trabalho a atenção é concentrada na análise ideológica. No entanto, documentos de cunho programático também são incluídos na análise sob a hipótese que esses também carregam aspectos ideológicos relevantes. Também optamos pelo termo genérico “documento” tanto para designar programas de governo ou plataformas eleitorais e resoluções políticas ao longo do tempo, quanto para designar manifestos registrados no momento de fundação dos partidos. Todos esses documentos são igualmente valiosos como fontes de análise de atuação política.

3.2 Modelos de Diferenciação Ideológica

Como um roteiro para exposição dos métodos de análise automatizada de documentos para diferenciação ideológica utilizamos o quadro geral mostrado na Figura 3.1, inspirado e adaptado do quadro em (GRIMMER; STEWART, 2013). Em nossa apresentação, as técnicas empregadas na litaratura podem ser divididas em dois grandes grupos: as de classificação categórica e as de classificação estatística. Cada um desses grupos reúne abordagens que levam a índices representativos, cada qual a sua maneira, das diferenças ideológicas entre as agremiações. Os métodos de classificação categórica envolvem uma fase de estabelecimento de um espaço multidimensional de categorias definidoras de perfis ideológicos, conforme discutidos nas sebseções a seguir. Já as técnicas de classificação estatística oferecem posições espaciais mensuradas a partir da análise estatística dos documentos. Os métodos em ambos os grupos se baseiam fortemente na suposição de que a ideologia domina a linguagem usada nos documentos analisados. Quando essa suposição é satisfeita, os modelos podem ter um bom desempenho. Quando essa premissa é violada, o modelo colocará os atores em um espaço diferente e não ideológico. Esse espaço pode ser útil, mas validações são necessárias para estabelecer seu significado.

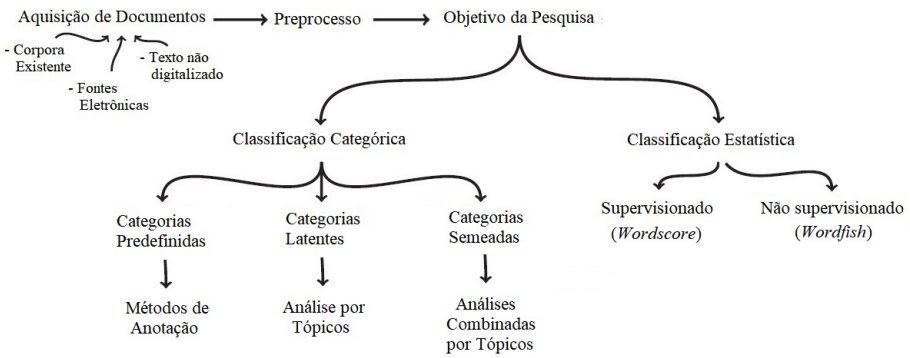


Figura 3.1: Visão geral de métodos de texto como dados.

3.2.1 Posicionamento, Afinidade e Disparidade Ideológicas

Uma característica fundamental das complexas sociedades atuais é a sua multidimensionalidade das formas de organização dos seus indivíduos. Como um corpo em permanente transformação, grupos sociais de uma sociedade moldam as suas demandas dependendo de uma multidude de categorias morais, econômicas, jurídicas, culturais, ecológicas, religiosas, entre outras. As diversas formas de associações de representações dessas categorias, com todas as suas contradições, definem perfis ideológicos cujas transformações reconfiguram as relações políticas constantemente. Amiúde, fatos relevantes como, por exemplo, grandes escândalos ou tragédias, podem causar descontinuidades expressivas. Enquanto agremiações catalizadoras das diversas grupos de uma sociedade, o comportamento dos perfis ideológicos das agremiações reflete a multidimensionalidade e o caráter dinâmico dessa mesma sociedade, influenciando a organização das suas estruturas internas e a formulação das suas estratégias de atuação externa.

Tendo em vista a associação existente entre os perfis ideológicos presentes tanto na sociedade quanto nas agremiações, um dos aspectos primordiais da análise política consiste na identificação dos fatores que regem o comportamento dinâmico, ao longo do tempo, das alianças e entendimentos das agremiações políticas como resposta às demandas da socie-

dade. Nesse contexto, diferenciar ideologicamente as agremiações é um dos elementos que permitem predições mais ou menos precisas, dependendo de diversos fatores, da evolução das alianças políticas essenciais para a hierarquização das pautas em debate em campanhas eleitorais e para constituição da linha de atuação posterior tanto do governo quanto das oposições (GROSSMAN; GUINAUDEAU, 2021). Duas abordagens de análise de documentos de agremiações presentes no grupo de classificação categórica, nomeadamente a análise por tópicos em categorias latentes e a análise combinada por tópicos em categorias semeadas, podem ser identificadas seguindo critérios que convêm ao estudo apresentado nos capítulos subsequentes, embora fortemente inspirados na literatura, sobretudo em (GRIMMER; STEWART, 2013).

A nomenclatura que adotamos para a primeira dessas abordagens reflete o objetivo de situar cada agremiação individualmente em um espectro de perfis ideológicos. Assim sendo, o posicionamento ideológico é uma metodologia frequentemente adotada que consiste no agrupamento das agremiações segundo padrões. Uma subdivisão dessa metodologia corresponde a diferenciar as formas de estabelecer as categorias com as quais avaliar os padrões de perfis ideológicos. Em todos os casos, métodos automatizados podem minimizar a quantidade de trabalho necessária para efetuar a classificação. De uma forma geral, apontamos três possíveis abordagens, segundo as categorias sejam predefinidas, latentes ou semeadas. No primeiro caso, as categorias são definidas previamente à análise e busca-se identificar a aderência dos documentos a cada categoria. As formas empregadas para esse fim usam frequências de palavras-chave ou anotações dos documentos, este último utilizado no Projeto Manifesto que será apresentado ainda neste capítulo. Alternativamente, é possível extrair as categorias latentes dos próprios documentos, o que pode ser automatizado por meio de métodos de análise por tópicos, também discutido adiante neste capítulo. Finalmente, uma abordagem intermediária consiste em induzir a extração das categorias do texto, direcionando-a a partir de categorias que se espera favorecer na análise (WATANABE; ZHOU, 2020). Esta é a abordagem que tratamos em detalhes no método apresentado no Capítulo 5.

Ao lado do posicionamento ideológico, há também a abordagem de estimar a diferença entre os partidos políticos, abordagem que denominamos afinidade ou diversidade ideológica. Neste caso, busca-se a posição relativa entre as agremiações, e não a posição absoluta de cada uma. Assim sendo, uma forma de estimar a diversidade em um sistema partidário não seria primeiro obtendo uma representação das posições ideológicas de partidos, mas derivar uma estimativa de quão diferentes quaisquer dois partidos são um do outro. Apesar dessa distinção, a comparação entre os perfis ideológicos das agremiações é levada a cabo tendo as categorias como referência. Assim sendo, as mesmas subdivisões quanto ao estabelecimento das categorias a considerar na análise estão presentes. Em particular, a abordagem de categorias latentes é empregada no método apresentado no Capítulo 4.

Um comentário importante quanto às duas abordagens mencionadas acima é a não existência de equivalência entre as duas. Essa afirmação se deve ao fato de o posicionamento ideológico, via de regra, não ser resultante da proximidade absoluta das agremiações a cada uma das categorias. Ao contrário, a noção de proximidade ideológica é relativa dentro de cada categoria no sentido que uma diferenciação observada de uma agremiação perante uma certa categoria depende de fatores próprios a essa categoria, como por exemplo o seu impacto face às demandas da sociedade em um dado momento. A consequência imediata é que um posicionamento ideológico próximo de duas agremiações indica que

ambas possuem um grau significativo de afinidade política, mas que pode ser maior ou menor em cada uma das dimensões do espectro de categorias. Particularmente no caso do Brasil, vários são os trabalhos que demonstram as diferenças de atuação, ao longo do tempo, dos partidos situados em um mesmo lado da classificação esquerda-direita. A análise comparativa levada a efeito em (BAMBIRRA, 1981) mostra que existem diferenças substantivas entre os partidos e, mesmo aqueles posicionados ideologicamente como de esquerda têm afinidade muito pronunciada somente em algumas dimensões do espectro de categorias.

3.2.2 Multidimensionalidade Categórica e Índices Unidimensionais

Apesar da multidimensionalidade categórica dos perfis ideológicos, o tema da análise de diferenciação partidária, seja ela por classificação, afinidade ou disparidade, busca aferir como os partidos se diferenciam através de uma medida estimada frequentemente em um espectro unidimensional. A classificação segundo o espectro ideológico esquerda-direita é o exemplo que supera em quantidade todas as outras escalas de classificação de agremiações políticas na literatura. Refletindo a localização dos assentos na Assembleia Constituinte após a Revolução Francesa, essa metáfora posicional impôs-se como a mais resiliente estrutura conceitual de classificação por posicionamento ideológico. Um refinamento dessa classificação inclui também outras classes em um degradé entre os pólos de direita e esquerda, e até mesmo além deles, embora ainda mantendo a unidimensionalidade. Há, portanto, a necessidade de uma transformação de dimensionalidade, partindo de uma representação do perfil ideológico num espaço categórico de múltiplas dimensões, para um índice ideológico senão unidimensional, pelo menos em dimensionalidade menor. Alguns índices importantes encontrados na litaratura são resumidos na Seção 3.4.

As opções para a construção de qualquer índice estão intrinsecamente relacionadas à dimensionalidade do fenômeno subjacente. A dimensionalidade é o número de fontes separadas e interessantes de variações que existem entre os objetos que são analisados. No contexto da construção de índices, ou seja, na elaboração de uma única medida para caracterizar um fenômeno, essa questão se resume a unidimensionalidade e escalabilidade unidimensional - selecionando dados que podem ser demonstrados para corresponder a uma única dimensão. Em um sentido amplo, tal correspondência pode ser considerada como um teste de validade da medição e o argumento neste caso é que mesmo que um índice seja construído a priori em base teórico-ideológica, a validade do índice ainda deve ser testada em relação aos dados em que se baseia.

A redução de dimensionalidade proporciona a vantagem evidente de tornar a análise mais simples e condensada. A análise qualitativa mais detalhada pode então ser orientada por uma quantidade limitada de indicadores. No entanto, vale ressaltar que essa motivação não ocorre sem perda de informação, a qual deve ser levada em conta na análise qualitativa posterior. Em (ALBRIGHT, 2010), o autor nota que não há evidência que uma única dimensão política existente possa descrever adequadamente a variação entre partidos. Sua análise das principais categorias mostrou que apenas uma pequena proporção delas são suficientes para explicar a maior parte da variância nos índices. Essa observação empírica sugere que uma única dimensão esquerda-direita definida de maneira invariante para diversos países ao longo de um período prolongado do tempo é injustificada. Sugere ainda que a escolha do conjunto de dimensões categóricas inicial afeta o valor e a interpretação do índice final.

3.3 Categorias do Projeto Manifesto

Na pesquisa sobre diferenciações entre agremiações partidárias e das métricas utilizadas para esse fim tem se sobressaído os recursos acumulados pelo Projeto Manifesto (WERNER et al., 2021). Trata-se de uma importante base de dados, principalmente centrada em programas partidários, usada por uma grande quantidade dos pesquisadores no mundo em análises de agremiações políticas. A base de dados que o Projeto Manifesto disponibiliza para análise de preferências políticas cobre mais de 1000 partidos de 1945 até hoje em mais de 50 países em cinco continentes. Nele, documentos são entendidos como sendo indicadores das preferências políticas dos partidos em um determinado momento. Por esse motivo, esses documentos são escolhidos como objeto de análise quantitativa de conteúdo. Um esquema de classificação categórica unificado com um método de anotação são respeitados com o objetivo de tornar os documentos comparáveis. O projeto envolve a coleta e a análise comparativa de conteúdo de manifestos de partidos com o apoio de classificadores humanos em diversos países.

3.3.1 Categorias

O Projeto Manifesto desenvolveu um sistema de categorias pelo qual cada declaração política de cada documento é classificada por classificadores humanos em não mais que uma das 56 categorias desse sistema. Para esse fim, as declarações políticas, que não são mais do que uma frase, devem ser previamente identificadas. A grande maioria das declarações devem ser classificadas, embora alguns casos excepcionais podem não encontrar uma categoria adequada. Por seu lado, as categorias estão agrupadas em sete grandes domínios. O principal objetivo da definição das categorias é o seu caráter universalizante no sentido de permitir a comparação entre agremiações em um grande espectro de países e eleições ao longo do tempo. Além disso, 12 das 56 categorias têm a particularidade de serem divididas em duas ou mais subcategorias que capturam aspectos específicos das respectivas categorias. A classificação das declarações políticas nessas categorias particulares deve especificar a subcategoria adequada ao caso.

A Tabela 3.2, adaptada de (WERNER et al., 2021), mostra as categorias, subcategorias e domínios em uso pelo Projeto Manifesto.

Uma etapa adicional deve ser realizada no que diz respeito a uma análise ideológica mais específica. Essa etapa consiste na aplicação de um critério para determinar um subconjunto das 56 categorias que expressem os aspectos ideológicos que são objetos da análise.

3.3.2 Método de Anotação

O fluxo de trabalho para constituição da base de dados do Projeto Manifesto conta com a intervenção humana para a codificação dos textos. Decisões humanas sobre qual trecho de texto combina com qual categoria de codificação são, em alguns aspectos, qualitativas, mesmo que no final gerem dados quantitativos.

A característica especial do Projeto Manifesto em comparação com as abordagens analíticas de conteúdo comuns é sua concentração no processo de produção de dados. Conforme observado, ele fornece uma infraestrutura de dados adequada para uso em uma ampla

Domínio 1: Relações Externas

101 Relações Especiais Estrangeiras: Positivo
102 Relações Especiais Estrangeiras: Negativo
103 Antiimperialismo: Positivo
103.1 Antiimperialismo Centrado no Estado
103.2 Influência Financeira Estrangeira
104 Militar: Positivo
105 Militar: Negativo
106 Paz: Positivo
107 Internacionalismo: Positivo
108 Integração Europa/A.Latina: Positivo
109 Internacionalismo: Negativo
110 Integração Europa/A.Latina: Negativo

Domínio 2: Liberdade e Democracia

201 Liberdade e Direitos Humanos: Positivo
201.1 Liberdade
201.2 Direitos Humanos
202 Democracia
202.1 Geral: Positivo
202.2 Geral: Negativo
202.3 Democracia Representativa: Positivo
202.4 Democracia Direta: Positivo
203 Constitucionalismo: Positivo
204 Constitucionalismo: Negativo

Domínio 3: Sistema Político

301 Descentralização: Positivo
302 Centralização: Positivo
303 Eficiência Governamental e Administrativa: Positivo
304 Corrupção Política: Negativo
305 Autoridade Política: Positivo
305.1 Autoridade Política: Competência Partidária
305.2 Autoridade Política: Competência Pessoal
305.3 Autoridade Política: Governo forte
305.4 Elites pré-democráticas: Positivo
305.5 Elites pré-democráticas: Negativo
305.6 Reabilitação e Compensação

Domínio 4: Economia

401 Economia de Mercado Livre: Positivo
402 Incentivos: Positivo
403 Regulamento do Mercado: Positivo
404 Planejamento Econômico: Positivo
405 Corporativismo: Positivo
406 Protecionismo: Positivo
407 Protecionismo: Negativo
408 Metas Econômicas
409 Gerencia de Demanda Keynesiana: Positivo
410 Crescimento Econômico
411 Tecnologia e Infraestrutura: Positivo
412 Economia Controlada: Positivo
413 Nacionalização: Positivo
414 Ortodoxia Econômica: Positivo
415 Análise Marxista: Positivo
416 Economia Anticrescimento: Positivo
416.1 Economia Anticrescimento: Positivo
416.2 Sustentabilidade: Positivo

Domínio 5: Bem-estar e Qualidade de Vida

501 Proteção Ambiental: Positivo
502 Cultura: Positivo
503 Igualdade: Positivo
504 Expansão do Estado de Bem-Estar
505 Limitação do Estado de Bem-Estar
506 Expansão da Educação
507 Limitação da Educação

Domínio 6: Tecido da Sociedade

601 Modo de Vida Nacional: Positivo
601.1 Geral
601.2 Imigração: Negativo
602 Modo de Vida Nacional: Negativo
602.1 Geral
602.2 Imigração: Positivo
603 Moralidade Tradicional: Positivo
604 Moralidade Tradicional: Negativo
605 Lei e Ordem
605.1 Lei e Ordem: Positivo
605.2 Lei e Ordem: Negativo
606 Mentalidade Cívica: Positivo
606.1 Geral
606.2 Ativismo de baixo para cima
607 Multiculturalismo: Positivo
607.1 Geral
607.2 Integração de Imigrantes: Diversidade
607.3 Direitos indígenas: Positivo
608 Multiculturalismo: Negativo
608.1 Geral
608.2 Integração de Imigrantes: Diversidade
608.3 Direitos indígenas: Negativo

Domínio 7: Grupos Sociais

701 Grupos de Trabalho: Positivo
702 Grupos de Trabalho: Negativo
703 Agricultura e Agricultores
703.1 Agricultura e Agricultores: Positivo
703.2 Agricultura e Agricultores: Negativo
704 Classe Média e Grupos Profissionais: Positivo
705 Grupos Minoritários: Positivo
706 Grupos Demográficos Não Econômicos: Positivo

000 Nenhuma categoria significativa se aplica

Figura 3.2: Categorias e subcategorias nos sete domínios políticos.

gama de questões de pesquisa. A maioria das inferências analíticas de conteúdo feitas no Projeto Manifesto estão ocultas no processo de codificação e gravação humana.

No Projeto Manifesto, a adequação das codificações só pode ser garantida por treinamento intensivo dos codificadores e comunicação intensiva entre diferentes codificadores. Igualmente importante é ter uma teoria de medição claramente definida subjacente a todo o processo de fabricação de dados, a fim de produzir as instruções de codificação que permitam a todos os codificadores referir-se a um terreno comum ao fazer as inferências abdutivas.

Cada parte textual do corpo de um documento partidário precisa ser unitizada e codificada. Na preparação do corpus para a codificação, certas partes do documento devem ser ignoradas, mas não eliminadas (para fins de documentação). Os codificadores humanos são solicitados a codificar de acordo com certo procedimento de anotação do corpus. A unidade de codificação é uma quase-sentença. Uma quase-sentença contém exatamente uma declaração ou "mensagem". Em muitos casos, os partidos fazem uma declaração por sentença, o que resulta em uma quase-sentença igual a uma sentença completa. Portanto, a regra básica da unitização é que uma sentença é, no mínimo, uma quase-sentença. Em nenhum caso duas ou mais sentenças podem formar uma quase-sentença. Há, entretanto, casos em que uma sentença natural contém mais de uma quase-sentença. Nesse caso, somente se a sentença natural contiver mais de um argumento único, esta sentença deve ser dividida. Há duas possibilidades de argumentos únicos: 1) uma frase contém duas afirmações que não estão totalmente relacionadas; ou 2) uma frase contém duas afirmações que estão relacionadas (por exemplo, elas vêm do mesmo campo político), mas tratam de aspectos diferentes de uma política maior.

As pistas para declarações únicas podem ser 1) ponto-e-vírgula; 2) a possibilidade de dividir a frase em uma lista de pontos significativos; 3) pistas gerais a partir de códigos. Quanto a este último ponto, é provável que a frase inclua duas afirmações únicas se contiver códigos de dois ou mais domínios (ver Tabela 3.3). Um exemplo seria:

> "*Precisamos abordar nossos estreitos laços com nossos vizinhos (107), bem como os desafios únicos enfrentados pelos proprietários de pequenas empresas nesta época de dificuldades econômicas. (402)*"

Como encontrar o código certo para uma quase-sentença? Cada decisão de anotação para atribuir uma quase-sentença particular a uma categoria específica é em si uma espécie de formulação e teste de hipótese. A hipótese é que a quase-sentença 54 no texto pertence à categoria 203. Esta coexiste com outras hipóteses de que pertence a outras categorias. Indo e voltando entre a frase observada e a definição das categorias nas instruções (reconhecimento de padrão e ajuste) decide-se a favor da hipótese de que pertence a 203 com uma probabilidade maior do que as outras. Essa é, na prática, a execução do referido processo abdutivo. Dessa forma, a anotação (codificação do texto) é um processo contínuo de formulação e teste de hipóteses de baixo nível. Um dos pontos sensíveis na manutenção do projeto é a necessidade de manter uma significativa rede de codificadores humanos permanentemente treinados. Como pode ser constatado, o esforço humano em todo o processo de manutenção e atualização da base de dados do Projeto Manifesto é considerável, além de custoso, pois essa metodologia envolve, em vários níveis, análises semânticas de textos por especialistas.

3.4 Índices Unidimensionais

Em pesquisas de agremiações políticas geralmente os partidos são medidos pelo índice disponível do conjunto de dados do Projeto Manifesto que é aparentemente fácil de interpretar e facilmente disponível. Paralelamente ao uso generalizado desse índice intrínseco do projeto, muitos pesquisadores apontaram problemas e colocaram a validade do índice em dúvida. Assim, não é de surpreender que outros tenham se debruçado no assunto e várias métricas alternativas de posicionamento ideológico tenham sido propostas com base no conjunto de dados do Manifesto. Os índices resumidos a seguir constituem uma ampla variedade de abordagens e de várias sofisticações metodológicas para alcançar um objetivo semelhante a partir do mesmo ponto de partida. A essência da questão de definição de índices unidimensionais é a forma como se transforma o espaço multidimensional de categorias em um índice unidimensional com representatividade do posicionamento ou das afinidade ou disparidade ideológicas. Dentre os índices elencados, há os que levam em consideração todo ou quase todo o esquema de codificação do Projeto Manifesto e há os que se concentram apenas em determinadas categorias temáticas específicas. Adicionalmente, há os que permitem que o significado das dimensões ideológicas mude entre os países e o tempo, e, contrariamente, índices que não o permitem (como o RILE descrito a seguir). No entanto, todas as abordagens compartilham o fato que uma transformação é aplicada para alterar um tipo de dado em outro. Os posicionamentos ideológicos refletidos nesses índices retratam perfis ideológicas sob distintas perspectivas tendo como base exatamente o mesmo conjunto de dados, compartilhando assim todas as suas possíveis falhas decorrentes do processo de codificação, seleção de documentos etc.

3.4.1 Índices de Posicionamento Ideológico

Antes de chegar a uma estimativa de posição ideológica, os dados devem ser transformados. É necessário aplicar um mecanismo para determinar quais das 56 categorias pertencem à ideologia em que estamos interessados e como obter delas uma estimativa da dimensão que representa esta ideologia. Puras diferenças programáticas seriam calculadas diretamente do conjunto de dados brutos do Manifesto e de sua estrutura dimensional de 56 categorias. Esse é o principal elemento que caracteriza os índices de posicionamento ideológico.

Índice RILE

Junto com a oferta de uma base de dados de manifestos partidários codificados e em constante atualização, o Projeto Manifesto também disponibiliza uma métrica intrínseca, o índice RILE de medição unidimensional esquerda-direita. O índice RILE tem sido a métrica de escolha para a maioria das pesquisas que precisam de estimativas de posicionamento ou diferença partidária, incluindo pesquisas de coalizões. Embora as primeiras análises dos dados do Manifesto partissem de uma perspectiva indutiva e voltada para os dados, a formulação final do índice é mais orientada para suposições. O índice é baseado em dois subconjuntos de 13 categorias de codificação do conjunto de dados do Manifesto que devem pertencer aos polos opostos da dimensão esquerda-direita, conforme mostrado na Tabela 3.3, reproduzida e adaptada de (MÖLDER, 2016).

Esquerda		**Direita**	
código	nome	código	nome
103	Anti-imperialismo: positivo	104	Forças Armadas: positivo
105	Forças Armadas: negativo	201	Liberdades e Direitos Humanos
106	Paz: positivo	203	Constitucionalismo: positivo
107	Internacionalismo: positivo	305	Autoridade Política
202	Democracia	401	Livre Iniciativa
403	Regulação do Mercado	402	Incentivos Econômicos
404	Planejamento Econômico	407	Protecionismo: negativo
406	Protecionismo: positivo	414	Ortodoxia econômica
412	Economia controlada	505	Limitação do *Welfare State*
413	Nacionalização	601	Nacionalismo: positivo
504	Expansão do *Welfare State*	603	Moralidade tradicional: positivo
506	Expansão da Educação	605	Lei e Ordem
701	Classes trabalhadoras: positivo	606	Harmonia Social

Figura 3.3: Conjuntos das categorias "esquerda"e "direita"do índice RILE.

Conjuntos das categorias "esquerda" e "direita" do índice RILE

O índice RILE é calculado da seguinte forma: Em primeiro lugar, as proporções dos manifestos partidários cobertas pelos conjuntos de categorias esquerda e direita são determinadas separadamente pela soma dos valores das categorias e, em segundo lugar, a proporção esquerda é subtraída da proporção direita. Isso resulta numa medida que poderia em princípio variar de -100 (o manifesto inteiro é dedicado às categorias esquerda) a +100 (o manifesto inteiro é dedicado às categorias direita). O cálculo do índice é mostrado na equação (3.1), onde N_R e N_L se referem às contagens do total de declarações à esquerda e à direita e N se refere ao número total de declarações políticas num manifesto ou programa partidário.

$$RILE = \frac{N_R - N_L}{N} \tag{3.1}$$

O fato de uma métrica construída dessa forma poder ser aplicada comparativamente ao longo do tempo e dos países foi visto como um grande trunfo do índice RILE. A isso se opõe as métricas "derivadas empiricamente e contingentes" (KLINGEMANN et al., 2006), que mudam conforme os padrões nos dados mudam e que, portanto, não são "aplicáveis" entre países e ao longo do tempo. De fato, os autores do índice RILE estão bem cientes de que esta métrica deve representar um fenômeno unidimensional subjacente. Não apenas foi sugerido que os conjuntos de categorias esquerda-direita em que o índice se baseia devem se opor, o que implica uma correlação negativa entre os conjuntos. Também foi destacado que, para que o índice seja significativo, as categorias de codificação da esquerda e as categorias da direita devem "sair juntas," além da oposição entre os conjuntos (BUDGE; MEYER, 2013).

A análise da validade do índice RILE começa com uma discussão sobre a natureza da dimensão esquerda-direita, já que esta é a base conceitual do índice. Para a definição do índice, supõe-se que a dimensão esquerda-direita é significativamente invariante através do tempo e do espaço. Polarizações dentro de espaços políticos, porém, não são as mesmas hoje como eram no passado nas sociedades e sistemas políticos ocidentais. Evidência da natureza mutável da esquerda e da direita também pode ser vista nas pesquisas de coalizão, onde a maneira padrão que a dimensão esquerda-direita é definida, não parece funcionar em países da Europa Central e Oriental ((SAVAGE, 2012)). Além disso, há evidências que as relações entre as áreas políticas que são usadas para definir esquerda e direita se relacionam de maneiras diferentes, dependendo se olhamos para partidos na

esquerda ou partidos na direita ((COCHRANE, 2011)). Tudo isso, diz (MÖLDER, 2016), é uma evidência contra uma invariável conceituação de esquerda-direita que está no centro do índice RILE.

O índice RILE tem recebido uma quantidade proporcional de críticas ao seu uso. Entre outras coisas, conforme dito acima, tem sido apontado que a estrutura da dimensão esquerda-direita que supostamente ele mede nem sempre está presente nos dados (MÖLDER, 2016). Para que o índice RILE seja válido, esse pesquisador defende que os padrões de associação presumidos pela lógica do índice devem estar presentes nos dados utilizados para calculá-lo. Para demonstrar, ele aplicou uma análise de correlação canônica (CCA) para testar estas associações, dentro e entre os conjuntos de variáveis que formam o índice esquerda-direita de RILE, resultando que, para países que não experimentaram um passado comunista a relação estava presente, embora substancialmente fraca. Porém, o mais importante é que, para países pós-comunistas, as associações requeridas nos dados claramente não estavam presentes. Concluiu então que o índice é uma métrica inválida de posicionamento esquerda-direita para partidos desse conjunto de países.

Uma crítica recorrente recai sobre o índice embutido no Projeto Manifesto para avaliação de diferenças programáticas entre partidos. Este índice (RILE) utiliza apenas uma parte da codificação para determinar sua inferência de distanciamento político. O alcance real é mais limitado, já que os conjuntos de categorias esquerda e direita cobrem menos da metade do conjunto total de categorias codificadas. Mölder alega que a interpretabilidade do índice, ou seja, o que é possível saber sobre o que ele mede pelos valores que ele atribui aos casos, depende de como os dois supostos compostos polares de esquerda e direita do índice estão associados nos dados (MÖLDER, 2016). Se eles estiverem empiricamente relacionados positivamente, então a suposição teórica de que eles representam extremos opostos de uma dimensão é violada e os valores são desprovidos de sentido. Se não existe uma relação substantivamente significativa entre eles, então um valor do índice, especialmente na faixa média da escala, pode resultar de uma mistura indeterminada de posições esquerda e direita e não é possível dizer de forma significativa o quanto um partido é esquerda ou direita. Somente se houver uma notável correlação negativa entre os conjuntos seria possível concluir que um partido no lado esquerdo ou direito do índice era esquerda ou direita de acordo com o significado presumido pelo índice.

Devido ao tamanho do acervo programático em permanente expansão do Projeto Manifesto várias métricas alternativas de posicionamento ideológico foram propostas ainda com base no conjunto de dados do Manifesto. Dentre os índices alternativos desenvolvidos, alguns utilizam a totalidade de códigos previamente disponíveis no Projeto Manifesto para realizar a análise de conteúdo dos programas partidários, enquanto outros abandonam a codificação prévia e fazem a busca de informações no próprio conjunto de dados para determinar o conteúdo das dimensões ideológicas.

Índices Derivados

Existem inúmeras alternativas que foram propostas que alteraram o método usado pelo índice RILE. Todos eles compartilham uma característica fundamental em comum, que remete às propriedades do conceito de polarização que foi mencionado acima. Dentre esses índices, podemos destacar os desenvolvidos por:

- KIM; FORDING (2002): baseado nos mesmos subconjuntos de categorias para

esquerda-direita como o índice RILE, mas emprega uma lógica diferente para calcular a posição de um documento partidário. Enquanto o índice RILE normaliza as contagens de declarações políticas que pertencem a qualquer categoria à esquerda ou à direita em relação ao número total de declarações políticas em um documento, a medida proposta por Kim e Fording o faz em relação ao número total de declarações políticas nos dois subconjuntos de esquerda e direita. A vantagem desta medida é que avalia posição a respeito da parte supostamente ideológica do documento e não o documento como um todo. Portanto, não é dependente de categorias de questões que são ideologicamente irrelevantes, que é um dos problemas do índice RILE.

- (LOWE et al., 2011): uma alteração em relação à última medida aborda o problema do efeito marginal de declarações adicionais. Este índice, definido no contexto esquerda-direita, assume a forma de um logaritmo da razão do número de afirmações da direita para as afirmações da esquerda (0,5 é adicionado a cada contagem por razões metodológicas (ibid., p. 132)). Esta escala logit não tem pontos finais predefinidos e qualquer posição nela é teoricamente possível, dada uma contagem extrema em uma das categorias. Eles propuseram 13 escalas diferentes para vários pares de questões no conjunto de dados do Manifesto.
- (JAHN, 2011): em uma abordagem muito semelhante para estimar as posições dos partidos em uma dimensão esquerda-direita, este índice usa a explicação teórica de Norberto Bobbio da dimensão esquerda-direita (BOBBIO, 1996) para determinar a priori as questões centrais que se relacionam com cada um dos polos desta dimensão. O escalonamento multidimensional (MDS) é então usado para determinar a localização dessas questões consideradas centrais na dimensão. Depois disso, Jahn usa regressão para selecionar os elementos específicos de tempo e país de esquerda e direita (aqueles que se correlacionam altamente com a dimensão central). Outro MDS é aplicado a este novo conjunto de questões para determinar a localização dessas questões na forma final da dimensão. As posições dos partidos na dimensão esquerda-direita são construídas como a soma dos valores das categorias de questões no conjunto de dados do manifesto, cada uma multiplicada por suas localizações determinadas pelo MDS.
- (KÖNIG; MARBACH; OSNABRÜGGE, 2017): este índice incorpora informações adicionais de pesquisas de especialistas sobre o processo de estimativa. Uma de suas principais preocupações é a comparabilidade entre países. Eles usam uma estrutura bayesiana para incorporar informações anteriores (avaliações de especialistas) sobre a forma do espaço político latente e uma transformação logit ((LOWE et al., 2011)) dos dados.

Índices Alternativos

Outros pesquisadores propuseram formas inteiramente novas de usar os dados para determinar as posições dos partidos.

- Franzmann e Kaiser (FRANZMANN; KAISER, 2006): repetindo o enfoque em como as categorias de questões ideológicas devem ser selecionadas de todo o conjunto de dados, este índice na dimensão esquerda-direita muda de acordo com o país e a época é calculado como uma regressão linear para cada uma das 56 categorias de codificação, com os valores das categorias como a variável dependente e "modelos de partido" como variáveis independentes para selecionar as categorias que mais

distinguem partidos. Dependendo de quais partidos enfatizam quais questões, eles podem ser classificados como esquerda ou direita. Categorias que não diferenciam entre partidos são classificadas como questões de valência. A posição final na dimensão esquerda-direita é calculada de forma semelhante ao índice RILE - a soma das pontuações da posição direita menos a soma das pontuações da posição esquerda é dividida pela soma das pontuações de posição mais as pontuações de valência (ibid., p. 173). As próprias pontuações são calculadas a partir dos valores das categorias no conjunto de dados subtraindo a mais baixa pontuação para uma categoria entre os partidos em uma eleição. A posição ideológica de um partido numa eleição é uma média móvel com as duas eleições adjacentes levadas em consideração. Usando este método, as posições partidárias também foram calculadas em dimensões econômicas e sociais separadas.

- Elff (ELFF, 2013): a partir da constatação do fato de que as posições em um espaço ideológico são diferentes das frequências de palavras ou frases em um documento partidário, propõe um modelo de medição capaz de levar essa diferença em conta e vá do último para o primeiro. Neste índice, pressupõe-se que os partidos, a menos que sejam novas agremiações, não estabelecem suas posições tabula rasa, mas usam sua posição anterior como um ponto de partida. Ele emprega este modelo em um conjunto bastante restrito de categorias de codificação para estimar as posições partidárias em um espaço econômico unidimensional e num espaço bidimensional liberal-autoritário - permissividade-tradicionalismo. As últimas duas dimensões são tratadas separadamente.
- Prosser (PROSSER, 2014): considera a escala logarítmica como a forma mais válida de construir posições esquerda-direita dos dados do Manifesto, mas se concentra em um método para selecionar as categorias apropriadas do conjunto de dados. Ele constrói uma escala geral esquerda-direita, bem como escalas econômicas e sociais separadas.

3.4.2 Análise de Agremiações Brasileiras

No tocante à América Latina, desde 2011 o Projeto Manifesto vem realizando esforços na ampliação do banco de dados. Seria interessante avaliar a aplicabilidade do índice RILE em nossa região, em especial naqueles países que, conceitualmente, poderiam ter tido uma experiência "pós-socialista" como Chile ou Nicarágua. Em 2020, o projeto já disponibilizava 192 documentos referentes a seis países dessa região (Argentina, Bolívia, Brasil, Chile, México e Uruguai). Como consequência dessa ampliação, surgiram pesquisas para comparar as posições políticas dos partidos brasileiros nas disputas eleitorais. Uma recente contribuição nesse sentido foi feita por Vladimyr L. Jorge et al. (JORGE; FARIA; SILVA, 2020), que utilizou o índice RILE para refletir sobre sua aplicabilidade em nosso País. Realizou o experimento para mapear o posicionamento dos principais partidos políticos na dimensão esquerda-direita ao longo de todas as eleições presidenciais realizadas sob a égide da Constituição de 1988.

O resultado gráfico obtido permite ver a posição ocupada por cada partido em um pleito específico e a variação de um pleito para outro conforme mostrado na Figura 3.4.

De acordo com a metodologia utilizada, o gráfico indica que o Brasil teve candidatos que assumiram propostas extremistas em três eleições, sendo estas, portanto, as mais polarizadas do atual "período democrático": os pleitos de 1989 (PT), 2018 (PSL) e, em

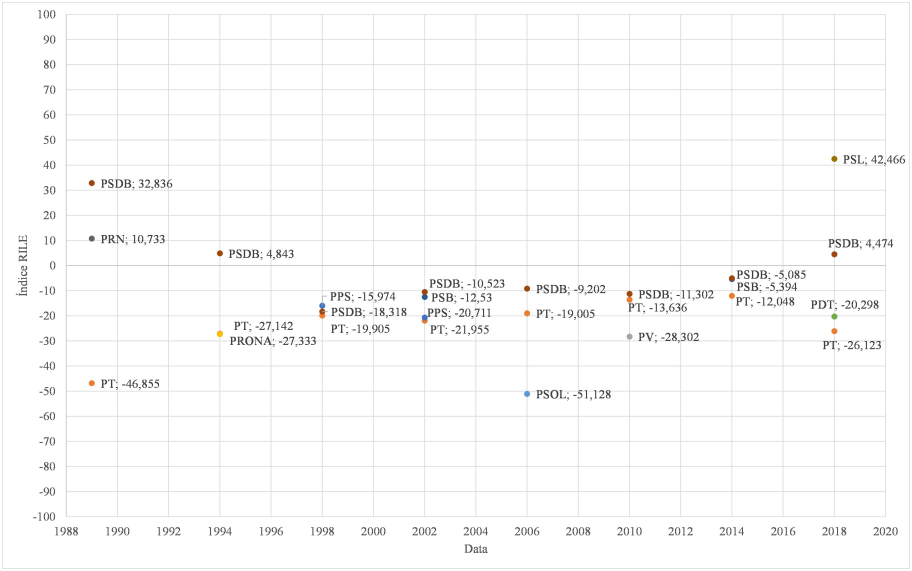


Figura 3.4: Posicionamento dos principais partidos políticos nas eleições brasileiras.

menor grau, o de 2006 (PSOL). Os pesquisadores reconheceram a aplicabilidade do índice RILE no contexto brasileiro e, nesse esforço, começaram a

> "*refletir sobre os problemas e as vantagens da aplicação de uma métrica comum a um universo de casos marcado pela diversidade. Desse modo, embora reconheçamos a pertinência das críticas a essa metodologia, consideramos, ainda assim, que ela permite identificar o posicionamento ideológico dos partidos políticos latino-americanos e, em particular, dos brasileiros em uma campanha eleitoral. Essa opção indica que não consideramos nem o Brasil, nem a América Latina como casos excepcionais, cujas singularidades internas impediriam o uso de uma metodologia aplicável a outros países e regiões. Nossa intenção é justamente a de empregar tal metodologia para que seja possível inserir o Brasil em um rol de indicadores que propiciam a comparação com outros casos.*"

No entanto, outros pesquisadores brasileiros, como Tarouco e Madeira (TAROUCO; MADEIRA, 2013), apontam que é preciso considerar as limitações da codificação proposta, que tem levado a classificações ideológicas deficientes, senão esdrúxulas, dos partidos brasileiros. Alegam que não é possível aplicar a codificação ideológica proposta pelo Projeto Manifesto ao caso brasileiro – ao menos, não sem alterações. Essa conclusão foi apontada pelos pesquisadores em 2013, quando relembraram que, no Brasil, não seria a questão da igualdade que dividiria esquerda e direita, mas sim os meios para alcançá-la. Para a direita, isso se daria através do reforço da ordem via Estado, enquanto, para a esquerda, seria por meio da demanda de políticas estatais distributivas. Nesse sentido, eles reforçam a necessidade de adequação da codificação ao Brasil, considerando tais ambiguidades no que se refere às expectativas relativas ao papel do Estado:

> "*Por conta da experiência histórica da ditadura militar e da transição, a esquerda incorpora reinvindicações que o Projeto Manifesto identifica como de direta: liberdade, direitos humanos e constitucionalismo, por exemplo. Da mesma forma, as defesas da paz como meta geral e do internacionalismo, que nunca chegaram a constituir uma bandeira das esquerdas brasileiras, constam, naquela escala, entre as categorias indicativas de posicionamento à esquerda.*"

A identificação desse desajuste entre a escala RILE e do que, de fato, distinguiria a esquerda da direita no Brasil deram origem a adaptações dessa categorização. Tarouco e Madeira consideraram que, dentro do esquema de codificação do Projeto Manifesto, as

categorias indicativas da direita seriam: (1) Forças Armadas: Positivo; (2) Economia de livre mercado; (3) Incentivos; (4) Ortodoxia econômica; (5) Limitação do Estado de Bem-Estar; (6) Classe média e grupos profissionais. A posição final dos partidos na escala foi dada pelo total de menções nessas categorias menos o total obtido nas categorias consideradas de esquerda: (1) Regulação do mercado; (2) Planejamento econômico; (3) Controle da economia; (4) Análise marxista; (5) Expansão do Estado de Bem-Estar; (6) Classes trabalhadoras: Positivo. Essa nova medida, no entanto, não resultou classificação muito distinta da anterior: apenas o programa do PFL, de 2005 e de 1995, e do PSDB, de 2001, foram categorizados como de direita.

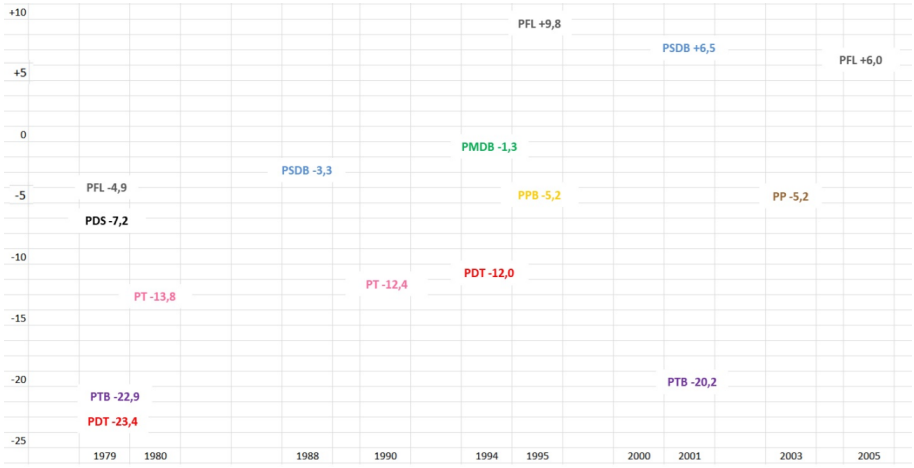


Figura 3.5: Classificação dos partidos brasileiros com o ajuste das categorias do índice RILE.

A classificação completa entre a extrema esquerda (-100) e a extrema direita (+100) dos partidos brasileiros com o ajuste das categorias do índice RILE é mostrada na Figura 3.5. Apesar dos ajustes efetuados, o quadro mostra que, segundo a metodologia do Projeto Manifesto e seu índice intrínseco, praticamente não existiria Direita no Brasil.

3.4.3 Índices de Afinidade

Conforme mencionado, o conceito de polarização está fundamentalmente relacionado ao pensamento sobre partidos em termos de uma única dimensão ideológica com dois pólos opostos, tradicionalmente chamados de esquerda e direita, e geralmente mediados pelas clivagens sociais. É um conceito importante, mas muitas vezes insuficiente, para medir a multiplicidade política de um sistema partidário. Alguns pesquisadores, Mölder por exemplo, preferem usar o termo "diversidade política" em vez disso, pois é livre de conotações esquerda-direita e, portanto, mais confortavelmente aplicável também em um contexto em que não estamos necessariamente pensando junto a uma dimensão.

Nessa linha de pesquisa de buscar índices condizentes com as representações multidimensionais dos sistemas partidários, Franzmann sugeriu o Índice de Similaridade (SIM) para medir as diferenças partidárias, ainda com base nos dados do Manifesto (VOLKENS et al., 2013). O SIM é calculado como a soma das diferenças absolutas entre todas as 56 categorias de codificação para um par partidário. O índice de similaridade tem uma interpretação intuitiva e direta - pode ser dimensionado para variar de 0 a 100 e interpretado como uma proporção sobreposta entre os perfis políticos de dois partidos. Matematicamente, este índice é equivalente a uma medida da distância entre blocos num espaço de

56 dimensões e pode ser representado da seguinte forma:

$$SIM = 100 - \frac{\sum_{i=1}^{56} |c_{i,2} - c_{i,2}|}{2}$$

onde SIM denota a sobreposição programática ou semelhança entre um par de partidos e $c_{i,1}$ e $c_{i,2}$ referem-se às proporções de documentos de dois partidos que foram dedicadas a cada uma das 56 posições políticas (categorias) definidas no esquema de codificação do Manifesto.

Embora todas as outras abordagens tenham tentado reduzir a dimensionalidade deste espaço, o índice de similaridade depende da dimensionalidade inicial. A métrica é uma medida de sobreposição ou semelhança entre dois documentos (programas) e, portanto, entre dois partidos. Os valores para as 56 categorias representam o perfil político de um partido. Assim, a soma das diferenças entre essas pontuações é equivalente à diferença entre todos os perfis políticos dos dois partidos representados em seus manifestos.

Não é menos razoável supor que esta similaridade se traduza no comportamento dos partidos da mesma forma que a diferença entre eles assumida ou derivada de uma dimensão esquerda-direita, pois, afinal de contas, é uma transformação dos mesmos dados. Embora em um caso estejamos falando de diferenças ideológicas e, no outro caso, de diferenças programáticas. Portanto, o Índice de Similaridade pode ser utilizado indistintamente em todas as análises que tradicionalmente têm se baseado em estimativas de diferenças políticas derivadas de posições ideológicas.

3.5 Classificação Estatística

Duas abordagens de análise de documentos de agremiações presentes no grupo de classificação estatística, nomeadamente a análise supervisionada executada com o método *Wordscore* e a análise não supervisionada executada com o método *Wordfish*, também podem ser utilizadas para determinação de uma classificação ideológica. Este último método (*Wordfish*), utilizado num recente estudo sobre posicionamento ideológico dos partidos políticos brasileiros é também apresentado a seguir, a título de comparação. Esses métodos podem estimar a localização dos atores no espaço da política ou produzir uma escala. O *Wordscore* depende da orientação de textos de referência para situar outros atores políticos em um espaço. Já o *Wordfish* explora uma suposição sobre como a ideologia afeta o uso de elementos léxicos. Ambos podem ser utilizados para superar os limites dos métodos de anotação manual de textos por categorizações na análise automatizada de texto.

Nara Salles (SALLES, 2021), buscando alternativas ao esquema de codificação manual elaborado pelo Projeto Manifesto realizou um exercício de análise de 889 plataformas eleitorais registradas por candidatos a cargos executivos no Brasil nos três níveis de disputa desde 2010. Em seu trabalho, faz um exercício de análise dos programas de governo a partir de uma técnica automatizada de análise de texto (*Wordfish*) que atribui posições espaciais a partir da frequência das palavras usadas pelos candidatos. Escrito na linguagem R, o programa *Wordfish* é capaz de extrair posições políticas de documentos de texto. As frequências de palavras são usadas para colocar os documentos em uma única dimensão. O Wordfish não é um método, mas sim uma técnica de dimensionamento que não precisa de nenhum documento de ancoragem para realizar a análise. Em vez disso, ele se baseia em um modelo estatístico de contagem de palavras (distribuição de Poisson).

Embora essa não seja a única técnica possível e possua limitações, a pesquisadora alega que sua utilização se justifica pelo fato de ideologia ser, intrinsecamente, um conceito espacial. A análise dos programas de governo de candidatos a cargos executivos no Brasil desde 2010 ofereceu um panorama mais complexo – ainda que imperfeito – quando comparado aos estudos que empregaram a codificação do Projeto Manifesto. A pesquisadora utilizou espaços de análise distintos para distribuir numa escala esquerda-direita os programas de governos nos três níveis de disputa (presidência, governadores e prefeitos). A maioria dos partidos brasileiros foi considerada de esquerda ou centro-esquerda. O resultado obtido para as disputas presidenciais em (SALLES, 2021) é reproduzido na Figura 3.6.

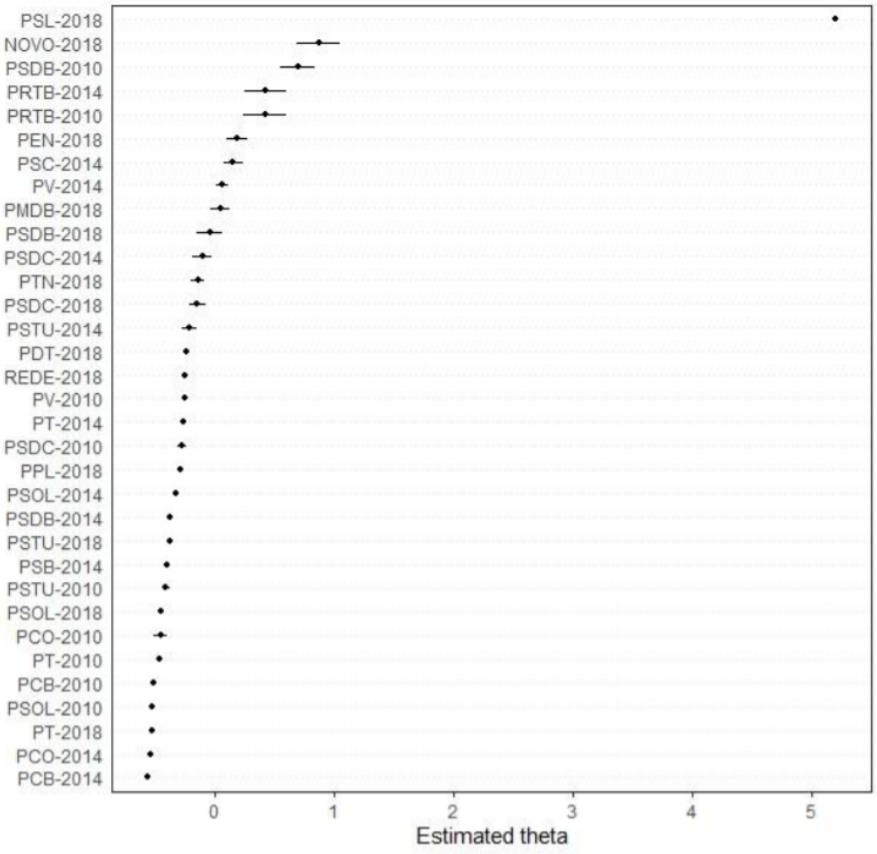


Figura 3.6: Competição programática presidencial (2010-2018).

4 DISPARIDADE ENTRE AGREMIAÇÕES PARTIDÁRIAS

A metodologia proposta nesta dissertação não envolve uma codificação externa aplicada a um texto, nem a atribuição de posição espacial a partir da frequência de termos, mas a extração endógena das dimensões do espaço de análise diretamente do próprio documento do partido. Este capítulo tem um duplo objetivo: apresentar um índice e resultados experimentais sobre esse índice. Um índice para tratar da disparidade ideológica, que é um termo complementar da afinidade ideológica. Foi escolhido o termo disparidade porque quanto mais parecidos são dois textos – ou grupos de textos, menor é o índice. Quando o índice é pequeno, há pouca disparidade; quando cresce, significa que aumenta a disparidade. Os dois conceitos – disparidade e afinidade – andam em par. Neste capítulo, esse índice de disparidade, considerado relativo, procura situar grupos de textos entre si, um com relação ao outro. A metodologia que está associada ao índice também será apresentada, pois envolve a maneira que os dados serão tratados previamente à análise por tópicos. Trata-se de fazer a análise nos dados e, em cima dessa análise, fazer um cálculo e obter um índice. Essa é a razão do duplo objetivo deste capítulo: a apresentação de um índice com sua metodologia associada junto com os dados usados nos experimentos, cujos resultados também serão mostrados. Esses experimentos foram realizados por intermédio de codificação realizada pelo professor orientador na linguagem R.

4.1 Categorias como Tópicos

A extração endógena das dimensões do espaço de análise significa dizer que as categorias da análise serão obtidas do perfil ideológico dos próprios documentos, não sendo, portanto, previamente definidas. No Projeto Manifesto, a definição prévia de categorias permite estabelecer uma base de comparação uniforme para todos os contextos. Esse princípio é aqui questionado na medida em que usamos os tópicos extraídos dos próprios documentos como categorias de análise. O conteúdo dos documentos direcionando as categorias. Abordagem que faz sentido visto que o objetivo do capítulo é a obtenção de um índice relativo entre documentos específicos. Não se trata, portanto, de um índice absoluto. Esta abordagem é a que mais fortemente se baseia na suposição de que a ideologia domina a linguagem usada nos documentos analisados. Quando essa suposição é satisfeita, os modelos podem ter um bom desempenho, mas validações são necessárias para estabelecer seu significado em caso contrário. São essas "representações explícitas" que usaremos como base comparativa entre documentos de agremiações políticas.

As opções para a construção de qualquer índice estão intrinsecamente relacionadas à dimensionalidade do fenômeno subjacente. A dimensionalidade é o número de fontes separadas e interessantes de variação que existem entre os objetos que são analisados. No contexto da construção de índices, ou seja, na elaboração de uma única medida para caracterizar um fenômeno, essa questão se resume a unidimensionalidade e escalabilidade unidimensional - selecionando dados que podem ser demonstrados para corresponder a uma única dimensão. Em um sentido amplo, tal correspondência pode ser considerada como um teste de validade da medição e o argumento neste caso é que mesmo que um índice seja previamente construído em base teórico-ideológica, a validade do índice ainda deve ser testada em relação aos dados em que se baseia.

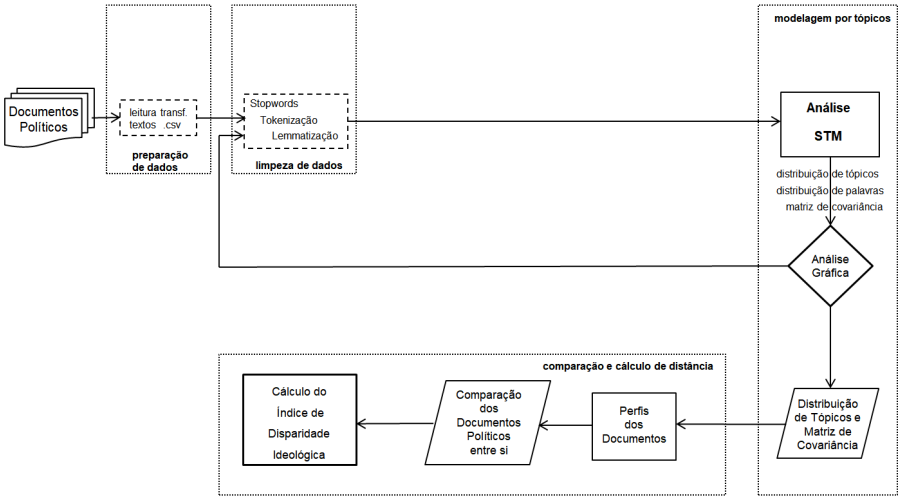


Figura 4.1: Visão geral da metodologia de análise chegando ao índice de disparidade ideológica.

A metodologia desenvolvida neste capítulo se subdivide em quatro estágios ilustrados em linhas pontilhadas no quadro acima. O primeiro é responsável pela leitura e transformação dos documentos das agremiações. Neste estágio são identificados e extraídos os campos (parágrafos) dos documentos a serem analisados. Em seguida são carregados para processamento na forma de arquivos .csv com os fragmentos de textos separados por virgulas. Estes arquivos constituirão os corpora do experimento. O segundo estágio, denominado de limpeza de dados, é responsável pela decomposição do corpus em termos (tokenização). Neste processo são eliminados os símbolos e caracteres de controle de arquivo ou de formatação, bem como os sinais de pontuação, números e datas. Em seguida inicia-se o processo de limpeza para a retirada das *stopwords*, que são compostas por preposições, artigos, advérbios, números, pronomes, conjunções, interjeição e pontuação. No terceiro estágio, denominado "modelagem por tópicos", é realizado o processo de identificação dos "topic model" com a utilização do modelo STM. Por fim, no quarto estágio, denominado "comparação e cálculo de distância", é realizada a extração dos perfis para comparação dos documentos entre si e cálculo do índice de disparidade ideológica.

4.2 Conjunto de Documentos

A fonte primária de dados para o experimento deste trabalho são os documentos partidários. Documentos que, ao serem emitidos pelos partidos ou coalizões, encontram-se obviamente estruturados para compreensão humana, mas não estruturados para um direto tratamento computacional. É necessária uma maneira de compreender, organizar e rotular esses dados para tomar decisões informadas.

4.2.1 Aquisição

Para um retrato do posicionamento político dos partidos de esquerda no Brasil, basicamente aqueles situados na parte inferior esquerda da Figura 3.6, foram selecionados textos e programas partidários produzidos, em sua grande maioria, após a redemocratização do País e relacionados no Apêndice A. Textos excessivamente grandes foram repartidos em documentos menores com o critério de eliminar partes com eventuais pautas específicas

que em nada contribuiriam no resultado do experimento. Exemplos deste caso foram os documentos originais nomeados 2017PSB01 e 2020PT02.

Como o escopo da análise a ser desenvolvida é verificar a similaridade política ou o grau de afinidade ideológica das organizações e partidos de esquerda atuais, foram selecionados documentos não apenas de partidos, mas também de tendências ou correntes que atuam organizadas dentro desses partidos, publicados a partir de 1990. Nesse sentido, podemos relacionar:

- em relação à agremiação política:
  - Partido Democrático Trabalhista – PDT
  - Partido Socialista Brasileiro – PSB
  - Partido dos Trabalhadores – PT (organizado em tendências)
  - Partido Socialismo e Liberdade – PSOL (organizado em correntes)
  - Revolução Brasileira – RB (corrente do PSOL, organização marxista)

A organização RB foi tratada como agremiação independente e não como uma corrente interna do PSOL para fins de análise visto que essa organização, em tese, é guiada pela TMD.

- em relação ao evento associado à elaboração de documentos:
  - Congresso: evento de discussão e deliberação de questões fundamentais da política partidária.
  - Encontro: evento ordinário ou extraordinário para discussão e deliberação da política partidária no período.
  - Tese: documento com determinada visão política para ser submetido ao debate em espaços decisórios – congressos e encontros.
  - Manifesto: declaração formal para a transmissão de intenções, decisões e ideias com o objetivo principal de expor determinado ponto de vista publicamente.
  - Programa Partidário: descreve a linha ideológica e os objetivos políticos que norteiam a atuação do partido.
  - Programa Eleitoral: documento de apresentação de projetos de governo numa eleição.
- em relação ao ano de divulgação do documento (principais marcos políticos):
  - 1995 – fundação de novos partidos
  - 2002 – documento eleitoral PT
  - 2004 – documento fundacional PSOL
  - 2010 – eleição presidencial
  - 2014 – eleição presidencial
  - 2017 – VI Congresso PT
  - 2018 – eleição presidencial
  - 2019 – VII Congresso PT
  - 2020 – Reunião Diretório Nacional PT e VII Congresso PSOL

A Tabela 4.1 contém um resumo do quantitativo dos documentos por agremiação e por evento. O elevado número de documentos relacionados com teses de congressos/encontros no PT e PSOL reflete a maneira como esses partidos se organizam, com diversas tendências internas disputando a hegemonia da direção partidária justamente nesse tipo de evento. Os programas eleitorais são referentes às eleições presidenciais após 2009, ano que o TSE obrigou o registro do programa junto com a candidatura.

Tabela 4.1: Quantidade de documentos por agremiação e evento.

	MANIFESTO	PROGRAMA ELEITORAL	PROGRAMA PARTIDÁRIO	TESE CONGRESSO
1995PSB01	5	0	0	0
1995PSB02	0	0	15	0
2002PT01	24	0	0	0
2004PSOL01	0	0	84	0
2010PSOL01	0	10	0	0
2010PT01	0	74	0	0
2014PSB01	0	213	0	0
2014PSOL01	0	58	0	0
2014PT01	0	139	0	0
2017PSB01_1	0	0	107	0
2017PSB01_2	0	0	129	0
2017PSB01_3	0	0	142	0
2017PSB01_4	0	0	48	0
2017PT01	0	0	0	175
2017PT02	0	0	0	89
2017RB01	52	0	0	0
2018PDT01	0	186	0	0
2018PSOL01	0	822	0	0
2018PT01	0	403	0	0
2019PSB01	72	0	0	0
2019PT01	0	0	0	33
2019PT02	0	0	0	46
2019PT03	0	0	0	34
2019PT04	0	0	0	30
2019PT05	0	0	0	30
2019PT06	0	0	0	53
2019PT07	0	0	0	46
2019PT08	0	0	0	38
2020PSB01	34	0	0	0
2020PSOL01	0	0	0	66
2020PSOL02	0	0	0	62
2020PSOL03	0	0	0	69
2020PSOL04	0	0	0	63
2020PSOL06	0	0	0	62
2020PSOL07	0	0	0	62
2020PSOL08	0	0	0	48
2020PSOL09	0	0	0	64
2020PSOL10	0	0	0	80
2020PSOL11	0	0	0	60
2020PSOL12	0	0	0	60
2020PT01	8	0	0	0
2020PT02_1	0	0	19	0
2020PT02_2	0	0	136	0
2020PT02_3	0	0	74	0
2020PT02_4	0	0	377	0
2020PT02_5	0	0	53	0
2020PT04	0	0	0	33
2020PT05	0	0	0	32
2020PT06	0	0	0	9
2020PT07	0	0	0	36
2020PT08	0	0	0	26
2020PT09	0	0	0	32
2020PT10	0	0	0	35
2020PT12	0	0	0	25
2020PT13	0	0	0	9
2020RB01	0	0	0	48

4.2.2 Segmentação em Extratos

Os textos originais são segmentados em extratos, cada qual definido sequencialmente a partir do início como o menor segmento do texto que satisfaça os seguintes critérios:

- pelo menos 2 frases
- pelo menos 280 caracteres (incluindo espaços em branco)
- pelo menos 50 palavras

O objetivo é que o extrato seja constituído de uma mescla de uma quantidade reduzida de tópicos (tipicamente, 1 ou 2). Cada extrato deve ter um tamanho restrito com o objetivo de induzir uma prevalência tópica expressiva. Os extratos são dispostos sequencialmente em linhas de planilhas salvas em formato CSV ("valores separados por vírgulas").

O quantitativo de extratos construídos com os critérios acima aparecem na Tabela 4.2.

4.2.3 Limpeza de Dados

As etapas anteriores transformam documentos textuais em dados. Esta etapa suplementar consiste em um tratamento na representação desses dados a fim de torná-los mais apropriados para as análises subsequentes. Antes do processamento foram eliminados os elementos léxicos com pouco sentido semântico ou irrelevantes para a análise, assim chamados de stopwords, como preposições e artigos. Nesta etapa também foram aplicadas técnicas de "normalização de palavras" que podem ser entendidas como a redução ou a simplificação de elementos léxicos pela qual várias formas diferentes do mesmo elemento são mapeados para uma única forma, chamada de forma raiz ou forma básica. Em termos mais técnicos, a forma raiz é chamada de "lema." Ao reduzir o número de formas que um elemento léxico pode assumir, garantimos que reduzimos nosso espaço de dados e que não precisamos verificar todas as formas do elemento. Isso nos ajuda a ignorar variações morfológicas em um único elemento léxico.

Em resumo, dentro de cada texto efetuaram-se as seguintes etapas para a limpeza dos dados:

- Remoção de pontuação: os sinais foram descartados para melhorar a classificação das palavras nos demais passos;
- Lematização: o processo foi utilizado para deflexionar palavras, reduzindo-as ao seu lema;
- Remoção de *stopwords*: foram removidas palavras de interrupção comuns à linguagem, que não possuem conteúdo tópico e não são interessantes ao modelo.

4.2.4 Agrupamentos

Para evitar que o texto analisado contenha muita informação é recomendável diminuí-lo para que ele seja dominado por um ou, no máximo, dois tópicos. Feito isso, esses extratos são então agrupados segundo critérios associados à fonte dos documentos. A análise envolve também os metadados e possíveis associações. Um exempo de metadados e associações de valores está representado na Tabela 4.3. A tabela mostra a quantidade de documentos por agrupamento. Na primeira linha, existe um documento que tem agremiação (PDT), tem evento (programa eleitoral) e tem o ano de publicação (2018) com

Tabela 4.2: Quantidade de extratos por agremiação e evento.

Agrem	Categoria	Quant
1995PSB01	MANIFESTO	5
1995PSB02	PROGRAMA PARTIDÁRIO	15
2002PT01	MANIFESTO	24
2004PSOL01	PROGRAMA PARTIDÁRIO	84
2010PSOL01	PROGRAMA ELEITORAL	10
2010PT01	PROGRAMA ELEITORAL	74
2014PSB01	PROGRAMA ELEITORAL	213
2014PSOL01	PROGRAMA ELEITORAL	58
2014PT01	PROGRAMA ELEITORAL	139
2017PSB01_1	PROGRAMA PARTIDÁRIO	107
2017PSB01_2	PROGRAMA PARTIDÁRIO	129
2017PSB01_3	PROGRAMA PARTIDÁRIO	142
2017PSB01_4	PROGRAMA PARTIDÁRIO	48
2017PT01	TESE CONGRESSO	175
2017PT02	TESE CONGRESSO	89
2017RB01	MANIFESTO	52
2018PDT01	PROGRAMA ELEITORAL	186
2018PSOL01	PROGRAMA ELEITORAL	822
2018PT01	PROGRAMA ELEITORAL	403
2019PSB01	MANIFESTO	72
2019PT01	TESE CONGRESSO	33
2019PT02	TESE CONGRESSO	46
2019PT03	TESE CONGRESSO	34
2019PT04	TESE CONGRESSO	30
2019PT05	TESE CONGRESSO	30
2019PT06	TESE CONGRESSO	53
2019PT07	TESE CONGRESSO	46
2019PT08	TESE CONGRESSO	38
2020PSB01	MANIFESTO	34
2020PSOL01	TESE CONGRESSO	66
2020PSOL02	TESE CONGRESSO	62
2020PSOL03	TESE CONGRESSO	69
2020PSOL04	TESE CONGRESSO	63
2020PSOL06	TESE CONGRESSO	62
2020PSOL07	TESE CONGRESSO	62
2020PSOL08	TESE CONGRESSO	48
2020PSOL09	TESE CONGRESSO	64
2020PSOL10	TESE CONGRESSO	80
2020PSOL11	TESE CONGRESSO	60
2020PSOL12	TESE CONGRESSO	60
2020PT01	MANIFESTO	8
2020PT02_1	PROGRAMA PARTIDÁRIO	19
2020PT02_2	PROGRAMA PARTIDÁRIO	136
2020PT02_3	PROGRAMA PARTIDÁRIO	74
2020PT02_4	PROGRAMA PARTIDÁRIO	377
2020PT02_5	PROGRAMA PARTIDÁRIO	53
2020PT04	TESE CONGRESSO	33
2020PT05	TESE CONGRESSO	32
2020PT06	TESE CONGRESSO	9
2020PT07	TESE CONGRESSO	36
2020PT08	TESE CONGRESSO	26
2020PT09	TESE CONGRESSO	32
2020PT10	TESE CONGRESSO	35
2020PT12	TESE CONGRESSO	25
2020PT13	TESE CONGRESSO	9
2020RB01	TESE CONGRESSO	48

um valor para cada um dos metadados; só há um documento que satisfaz essa associação entre valores de metadados.

4.3 Índice de Disparidade Ideológica

Tomando por base uma análise por tópicos dos documentos das agremiações, o perfil ideológico, por definição do modelo usado na análise, é representado por uma conjunção das relevâncias tópica e semântica. No entanto, a disparidade existente nesses perfis ideológicos se expressa na relevância tópica em decorrência do fato de as relevâncias semânticas dos elementos léxicos nos tópicos serem idênticas em todos os documentos. Posto de outra forma, as disparidades entre os perfis ideológicos se expressam pelas formas em que os tópicos contribuem para cada documento (ROBERTS et al., 2014). Assim sendo, uma forma natural de quantificar essa disparidade é através do uso de uma distância estatística entre as respectivas funções massa de probabilidade. A medida que adotamos como índice de disparidade ideológica é a *distância de Mahalanobis* entre médias de agrupamentos definidos segundo valores específicos de metadados. Visto que os métodos de análise por tópicos que adotamos são hierárquicos, a estratégia de atribuição de tópicos tem o papel de regularizador das relavâncias tópicas, servindo portanto de referência para o estabelecimento de medidas de distância. Além de definir em mais detalhes o índice de disparidade ideológica, também avaliamos a seguir como duas medidas estatísticas conhecidas na literatura que, quando definidas em cenários mais gerais são distintas da *distância de Mahalanobis* mas que, quando aplicadas ao cenário mais espcífico do resultado de uma análise por tópicos com o método STM de textos acabam se mostrando equivalente à *distância de Mahalanobis*. A apresentação inicial desses aspectos é feita para texto genérico para, em seguida, analisarmos a aplicação para análise dos extratos dos documentos das agremiações.

4.3.1 Distância de Mahalanobis

Tomemos, então, como ponto de partida o resultado de uma análise por tópicos de um corpus $\mathbf{D}$ com STM usando M metadados que, de forma genérica, denotamos por $m_1, \ldots, m_M$. Adotamos a seguinte notação para os possíveis agrupamentos de textos segundo os valores desses metadados. Para todo texto $\mathbf{d}$ do corpus $\mathbf{D}$ analisado, o valor do metadado m_i de $\mathbf{d}$ é denotado por $m_i(\mathbf{d})$. Um *subgrupo* é um conjunto $\boldsymbol{\kappa}_i(x) = \{\mathbf{d} \in \mathbf{D} \mid m_i(\mathbf{d}) = x\}$ de textos definido por um mesmo valor x do metadado m_i. Um *grupo*, que é uma noção já utilizada na Seção 2.3.2, é denotado por $\boldsymbol{\gamma}(x_1, \ldots, x_M) = \boldsymbol{\kappa}_1(x_1) \cap \ldots \cap \boldsymbol{\kappa}_M(x_M)$ como conjunto de textos definido pela associação específica de valores dos metadados. Mais precisamente, fazem parte de $\boldsymbol{\gamma}(x_1, \ldots, x_M)$ os textos que têm x_1 como valor do metadado m_1, x_2 como o valor de m_2, e assim por diante. Observamos que um subgrupo $\boldsymbol{\kappa}_i(x)$ é ele mesmo particionado em subgrupos definidos por metadados m_j distintos de m_i.

Além de permitir a definição de agrupamentos de textos na forma de subgrupos e grupos, o resultado da análise por STM fornece tanto a matriz de covariâncias entre os tópicos $\boldsymbol{\Sigma}$ quanto as médias $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{\mu}_{\mathbf{d}}$, para todo texto $\mathbf{d} \in \boldsymbol{\gamma}$ e todo grupo $\boldsymbol{\gamma}$ definido pelos metadados. A Tabela 2.7 e a Tabela 2.8 apresentadas no Capítulo 2 são exemplos desses valores que compõem o resultado da análise com STM. Adicionalmente, as distribuições $\boldsymbol{\eta} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dos grupos também fazem parte do resultado. A partir dos resultados por

Tabela 4.3: Quantidade de documentos por agremiação e evento.

Agrem	Categoria	Ano	Quant
1995PSB01	MANIFESTO	1995	5
1995PSB02	PROGRAMA PARTIDÁRIO	1995	15
2002PT01	MANIFESTO	2002	24
2004PSOL01	PROGRAMA PARTIDÁRIO	2004	84
2010PSOL01	PROGRAMA ELEITORAL	2010	10
2010PT01	PROGRAMA ELEITORAL	2010	74
2014PSB01	PROGRAMA ELEITORAL	2014	213
2014PSOL01	PROGRAMA ELEITORAL	2014	58
2014PT01	PROGRAMA ELEITORAL	2014	139
2017PSB01_1	PROGRAMA PARTIDÁRIO	2017	107
2017PSB01_2	PROGRAMA PARTIDÁRIO	2017	129
2017PSB01_3	PROGRAMA PARTIDÁRIO	2017	142
2017PSB01_4	PROGRAMA PARTIDÁRIO	2017	48
2017PT01	TESE CONGRESSO	2017	175
2017PT02	TESE CONGRESSO	2017	89
2017RB01	MANIFESTO	2017	52
2018PDT01	PROGRAMA ELEITORAL	2018	186
2018PSOL01	PROGRAMA ELEITORAL	2018	822
2018PT01	PROGRAMA ELEITORAL	2018	403
2019PSB01	MANIFESTO	2019	72
2019PT01	TESE CONGRESSO	2019	33
2019PT02	TESE CONGRESSO	2019	46
2019PT03	TESE CONGRESSO	2019	34
2019PT04	TESE CONGRESSO	2019	30
2019PT05	TESE CONGRESSO	2019	30
2019PT06	TESE CONGRESSO	2019	53
2019PT07	TESE CONGRESSO	2019	46
2019PT08	TESE CONGRESSO	2019	38
2020PSB01	MANIFESTO	2020	34
2020PSOL01	TESE CONGRESSO	2020	66
2020PSOL02	TESE CONGRESSO	2020	62
2020PSOL03	TESE CONGRESSO	2020	69
2020PSOL04	TESE CONGRESSO	2020	63
2020PSOL06	TESE CONGRESSO	2020	62
2020PSOL07	TESE CONGRESSO	2020	62
2020PSOL08	TESE CONGRESSO	2020	48
2020PSOL09	TESE CONGRESSO	2020	64
2020PSOL10	TESE CONGRESSO	2020	80
2020PSOL11	TESE CONGRESSO	2020	60
2020PSOL12	TESE CONGRESSO	2020	60
2020PT01	MANIFESTO	2020	8
2020PT02_1	PROGRAMA PARTIDÁRIO	2020	19
2020PT02_2	PROGRAMA PARTIDÁRIO	2020	136
2020PT02_3	PROGRAMA PARTIDÁRIO	2020	74
2020PT02_4	PROGRAMA PARTIDÁRIO	2020	377
2020PT02_5	PROGRAMA PARTIDÁRIO	2020	53
2020PT04	TESE CONGRESSO	2020	33
2020PT05	TESE CONGRESSO	2020	32
2020PT06	TESE CONGRESSO	2020	9
2020PT07	TESE CONGRESSO	2020	36
2020PT08	TESE CONGRESSO	2020	26
2020PT09	TESE CONGRESSO	2020	32
2020PT10	TESE CONGRESSO	2020	35
2020PT12	TESE CONGRESSO	2020	25
2020PT13	TESE CONGRESSO	2020	9
2020RB01	TESE CONGRESSO	2020	48

grupo, podemos calcular parâmetros para cada subgrupo como uma média ponderada sob a hipótese de distribuição uniforme de grupos sobre os textos. Em particular, a formulação matemática para as médias e para as distribuições são dadas por

$$\boldsymbol{\mu}_{\boldsymbol{\kappa}_i(x)} = \frac{\sum_{\mathbf{d}\in\boldsymbol{\kappa}_i(x)} \boldsymbol{\mu}_{\mathbf{d}}}{|\boldsymbol{\kappa}_i(x)|} \text{ e } \boldsymbol{\eta}_{\boldsymbol{\kappa}_i(x)} \sim \mathbf{N}(\boldsymbol{\mu}_{\boldsymbol{\kappa}_i(x)}, \boldsymbol{\Sigma}).$$

As medidas de distância que discutimos são definidas entre as médias de subgrupos normalizada pelas covariâncias entre os tópicos. Ambas, dadas as características específicas da relevância tópica, têm relação com a *distância de Mahalanobis* expressa por

$$IDI(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2) = (\boldsymbol{\mu}_{\boldsymbol{\kappa}_1} - \boldsymbol{\mu}_{\boldsymbol{\kappa}_2})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\boldsymbol{\kappa}_1} - \boldsymbol{\mu}_{\boldsymbol{\kappa}_2}). \tag{4.1}$$

4.3.2 Divergência de Kullback–Leibler

A primeira medida de distância é a medida de entropia relativa denominada *divergência de Kullback–Leibler*, $D_{KL}(\boldsymbol{\kappa}_1 \parallel \boldsymbol{\kappa}_2)$, expressando como a distribuição das relevâncias tópicas do subgrupo $\boldsymbol{\kappa}_1$ difere daquela de $\boldsymbol{\kappa}_2$. Em outras palavras, é a quantidade de informação perdida quando $\boldsymbol{\kappa}_2$ é usado para aproximar $\boldsymbol{\kappa}_1$ (BURNHAM; ANDERSON, 2002). Considerando que ambos os subgrupos são definidos para o mesmo conjunto de tópicos $\mathbf{T}$,

$$D_{KL}(\boldsymbol{\kappa}_1 \parallel \boldsymbol{\kappa}_2) = \sum_{i\in\mathbf{T}} \boldsymbol{\eta}_{\boldsymbol{\kappa}_1}(i) \log \frac{\boldsymbol{\eta}_{\boldsymbol{\kappa}_1}(i)}{\boldsymbol{\eta}_{\boldsymbol{\kappa}_2}(i)}.$$

Em outras palavras, $D_{KL}(\boldsymbol{\kappa}_1 \parallel \boldsymbol{\kappa}_2)$ é o valor esperado, segundo $\boldsymbol{\eta}_{\boldsymbol{\kappa}_1}$, da diferença logarítmica entre as relevâncias tópicas $\boldsymbol{\eta}_{\boldsymbol{\kappa}_1}$ e $\boldsymbol{\eta}_{\boldsymbol{\kappa}_2}$. No caso simples, uma entropia relativa de 0 indica que as duas distribuições em questão possuem quantidades idênticas de informação. Embora esse índice seja uma distância assimétrica, não sendo portanto uma métrica estatística, a matriz de covariância entre os tópicos é a mesma nas duas relevâncias tópicas, o que torna o índice simétrico no seguinte sentido.

Theorem 4.1. *Dado que* $\boldsymbol{\eta}_{\boldsymbol{\kappa}_1} \sim \mathbf{N}(\boldsymbol{\mu}_{\boldsymbol{\kappa}_1}, \boldsymbol{\Sigma})$ *e* $\boldsymbol{\eta}_{\boldsymbol{\kappa}_2} \sim \mathbf{N}(\boldsymbol{\mu}_{\boldsymbol{\kappa}_2}, \boldsymbol{\Sigma})$*, então a divergência de Kullback-Leibler das relevâncias tópicas dos subgrupos* $\boldsymbol{\kappa}_1$ *e* $\boldsymbol{\kappa}_2$ *é dada por*

$$D_{KL}(\boldsymbol{\kappa}_1 \parallel \boldsymbol{\kappa}_2) = \frac{1}{2} IDI(\boldsymbol{\kappa}_1 \parallel \boldsymbol{\kappa}_2).$$

4.3.3 Coeficiente Bhattacharyya

A segunda medida é o *coeficiente Bhattacharyya*, o qual é uma medida aproximada da quantidade de sobreposição entre duas amostras estatísticas. A sua definição, na comparação das relevâncias tópicas dos subgrupos $\boldsymbol{\kappa}_1$ e $\boldsymbol{\kappa}_2$, é

$$BC(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2) = -\ln\left(\sum_{i\in\mathbf{T}} \sqrt{\boldsymbol{\eta}_{\boldsymbol{\kappa}_1}(i)\boldsymbol{\eta}_{\boldsymbol{\kappa}_2}(i)}\right).$$

Mais uma vez, as particularidades das relevâncias tópicas levam o coeficiente Bhattacharyya a convergir para a distância de Mahalanobis.

Theorem 4.2. *Dado que* $\boldsymbol{\eta}_{\boldsymbol{\kappa}_1} \sim \mathbf{N}(\boldsymbol{\mu}_{\boldsymbol{\kappa}_1}, \boldsymbol{\Sigma})$ *e* $\boldsymbol{\eta}_{\boldsymbol{\kappa}_2} \sim \mathbf{N}(\boldsymbol{\mu}_{\boldsymbol{\kappa}_2}, \boldsymbol{\Sigma})$*, então a distância Bhattacharyya das relevâncias tópicas dos textos* $\boldsymbol{\kappa}_1$ *e* $\boldsymbol{\kappa}_2$ *é dada por*

$$BC(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2) = \frac{1}{8} IDI(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2).$$

4.4 IDI das Agremiações Partidárias

As seções anteriores estabeleceram a metodologia para a obtenção de um índice relativo para comparações entre pares de documentos. Uma vez apresentado o IDI – Índice de Disparidade Ideológica, chegou o momento de aplicá-lo. E será aplicado para fazer três tipos de comparações entre grupos de documentos, que são os seguintes:

- pares de documentos;
- grupos definidos por agremiação, evento, ano;
- grupos definidos por agremiação e ano.

4.4.1 Pares de Documentos

A primeira avaliação é realizada para comparar pares de documentos de agremiações na Figura 4.2. Como cada documento foi desmembrado em vários extratos, esses extratos passam a ser o "novo" conjunto de dados sobre o qual será feita a análise. O índice, porém, será usado para comparar documentos, não extratos. Então novo agrupamento de extratos de um mesmo documento é realizado para fazer a comparação e calcular o índice.

Algumas regiões de afinidade são ressaltadas na diagonal principal, principalmente entre documentos das seguintes agremiações:

- 2020PSOL01-12 – retrata uma forte afinidade entre as teses de diversas correntes do PSOL apresentadas para discussão no VII Congresso do partido.
- 2020PT04-08 – mostra a afinidade entre os documentos das tendências Movimento PT, Articulação de Esquerda, CNB e Rui Falcão com a proposta da Secretaria Geral do DN-PT para a reunião do Diretório Nacional do partido em abril/2020.
- 2014PT01, 2010PT01, 2002PT01 – este agrupamento mostra forte correlação entre os programas eleitorais do PT para as eleições presidenciais de 2010 e 2014 com a Carta aos Brasileiros, de 2002.

Outros agrupamentos podem ser observados na aproximação entre os documentos da RB publicados em 2017 e 2020 mostrando uma forte afinidade entre seus tópicos que perpassa o resultado eleitoral do período (2018).

Todos esses grupos foram posicionados próximos entre si, isto é, com reduzida disparidade ideológica. Por outro lado, um resultado com significativa disparidade entre documentos pode ser observado na comparação do documento 1995PSB02 com os demais, refletido nas pálidas gradações de cores associadas pelo método a esse documento.

Cabe um esclarecimento em relação aos documentos do Partido Socialista Brasileiro utilizados nesta dissertação. Conforme mencionado, foi efetuado um corte no corpus para privilegiar documentos mais recentes, eliminando aqueles originados antes de 1990. O Partido Socialista Brasileiro foi fundado em 1947 e extinto em 1965 por força do Ato Institucional n° 2 do regime militar. Em 1985, com o fim da ditadura e a redemocratização do País, foi fundado um novo Partido Socialista Brasileiro, resgatando o mesmo programa apresentado em 1947. O partido obtém o registro definitivo em 1988, antes portanto da entrada em vigor da nova lei dos partidos políticos em 1995 quando diversos outros partidos são fundados no Brasil. Ocorre que um erro foi involuntariamente cometido, mas que, ao fim e ao cabo, mostrou-se útil. Na busca de documentos partidários na Internet,

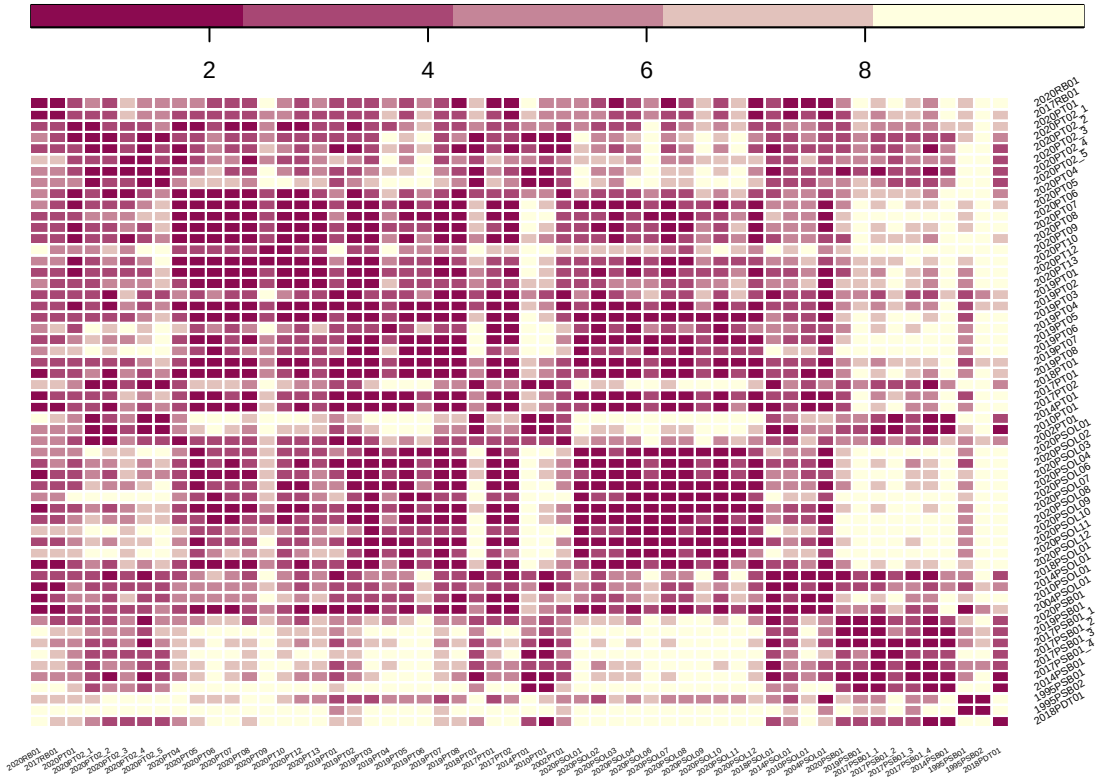


Figura 4.2: IDI entre pares de documentos de agremiações, normalizado usando z-score.

dois documentos fundacionais do PSB – o Manifesto e o Programa, foram encontrados não datados em formato PDF. Eles foram baixados e salvos associados erroneamente ao ano de 1995, quando na realidade deveriam ter sido associados ao ano de 1985, ano da refundação do partido. Como foram associados ao ano de 1995, conseguiram passar pelo corte acima mencionado de documentos gerados antes de 1990 e entraram na análise deste experimento.

A disparidade observada entre dois documentos de um mesmo partido emitidos juntos pode ser explicada pelos diferentes termos usados nesses documentos:

Programa (1947) – documento 1995PSB02	desigualdades sociais profundas; predomínio de umas nações sobre as outras; soluções socialistas; aspirações socialistas do povo brasileiro; despertar no operariado uma consciência política; eliminação de um regime econômico de exploração do homem pelo homem; transformação da estrutura da sociedade; socialização dos meios de produção; abolição de todos os privilégios de classe; estabelecimento de um regime socialista; abolição do antagonismo de classe; socialização da riqueza; nacionalização do crédito; satisfação das necessidades coletivas; comércio exterior sob controle do Estado

Manifesto (1985) – documento 1995PSB01	discriminação racial; opressão às minorias, às mulheres e às crianças; violência contra manifestações culturais alternativas; degradação da qualidade de vida; depredação do meio ambiente; genocídio das nações indígenas; moderna declaração dos direitos do ser humano; garantias de cidadania; uso da informática e dos meios de comunicação de massa; direitos individuais tradicionais; direito social à educação, à saúde, ao transporte público, à habitação e ao saneamento básico; direito de vizinhança, ao seguro-desemprego, e às novas formas de organização social e comunitária; direito à privacidade; acesso à informação; controle das atividades estatais; ampla participação política; descentralização mais completa do poder; interferência sistemática dos cidadãos; soberania popular; controle, pelo Legislativo, das atividades do Estado em uma economia progressivamente socializada

O PSB na sua refundação prestou homenagem à tradição histórica mantendo o Programa de 1947 com seus termos originais, mas ressalvou em seu Manifesto de 1985 a necessidade de uma atuação partidária em sintonia com as demandas atuais da sociedade brasileira. A radicalidade programática expressa no léxico do documento 1995PSB02 não passará desapercebida pelo método na análise efetuada no próximo capítulo.

4.4.2 Grupos Definidos por Agremiação, Evento e Ano

A segunda comparação é realizada juntando os metadados do nome da agremiação, do evento e do ano de produção do documento. Em cada grupo, portanto, estão todos os extratos que tem o mesmo valor desses três metadados, ou seja, todo extrato que está na mesma agremiação, mesmo tipo de evento e mesmo ano. O resultado pode ser visto na Figura 4.3.

O quadro resultante apresenta uma granularidade maior, isto é, a comparação é mais grosseira do que a anterior. Não deixa de ser curioso, porém, o agrupamento provocado pelo método ao detectar forte afinidade entre documentos do PT como um todo ao longo dos anos (programas de 2010, 2014, 2018 e 2020), concomitantemente com elevada disparidade desses documentos com os documentos de suas tendências internas no mesmo período (teses congressuais). O resultado pode ser entendido na provável imutabilidade da hegemonia da tendencia majoritária sobre o partido ao longo do tempo, com as demais tendencias não conseguindo influencia-lo com suas propostas.

Além da afinidade entre os documentos da RB na extremidade superior esquerda da diagonal, logo abaixo nota-se forte afinidade entre as teses das tendências Mensagem ao Partido, Avante S21 e Militância Socialista apresentadas em 2017 no VI Congresso do PT com as teses congressuais apresentadas dois anos depois no VII Congresso (2019).

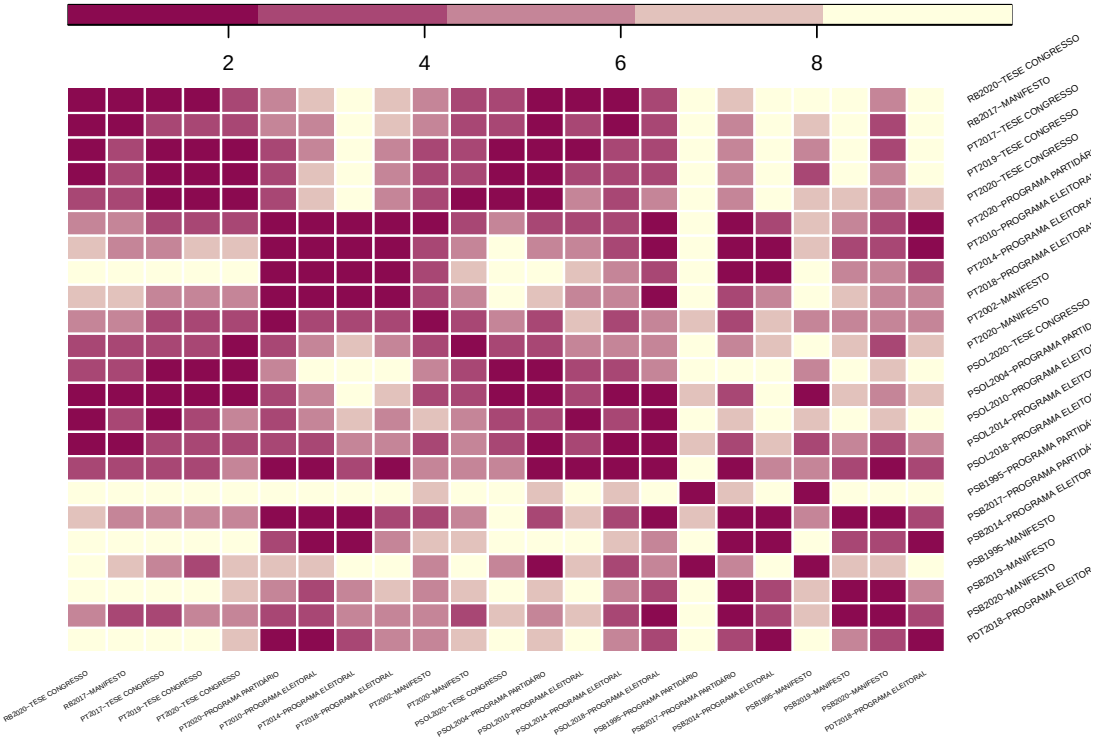


Figura 4.3: IDI entre grupos definidos pelos metadados Agremiação, Evento e Ano, normalizado usando z-score.

4.4.3 Grupos Definidos por Agremiação e Ano

Por fim, uma terceira comparação envolvendo a agremiação e o ano do documento, cujo objetivo inicial é a comparação de agremiações entre si.

O resultado, que pode ser visto na Figura 4.4, apresenta uma forte afinidade no agrupamento realizado no centro do quadro com os programas eleitorais de 2018 do PDT, PSOL, PT e PSB (este de 2017). O tom escuro e a proximidade espacial indicam que os programas desses partidos estavam bem parecidos naquela ocasião.

Todos esses resultados abrem uma nova perspectiva no estudo da análise automatizada de conteúdo guiada por categorias não previamente identificadas, mas retiradas como tópicos dos próprios documentos. Não apenas com os tipos de comparações feitas nos experimentos acima, mas principalmente para buscar eventuais coerências entre agremiações comparando tipos diferentes de eventos, por exemplo, o que se fala ou se escreve num congresso em comparação com um programa eleitoral; ou comparações entre diferentes programas eleitorais de um mesmo partido ao longo do tempo ou até comparações entre diferentes resoluções congressuais como medida da própria metamorfose de um partido. Todas essas comparações, no entanto, são relativas, pois comparam um documento em relação a outro. A introdução de uma base comum de comparação, representada por um índice absoluto, será feita no próximo capítulo.

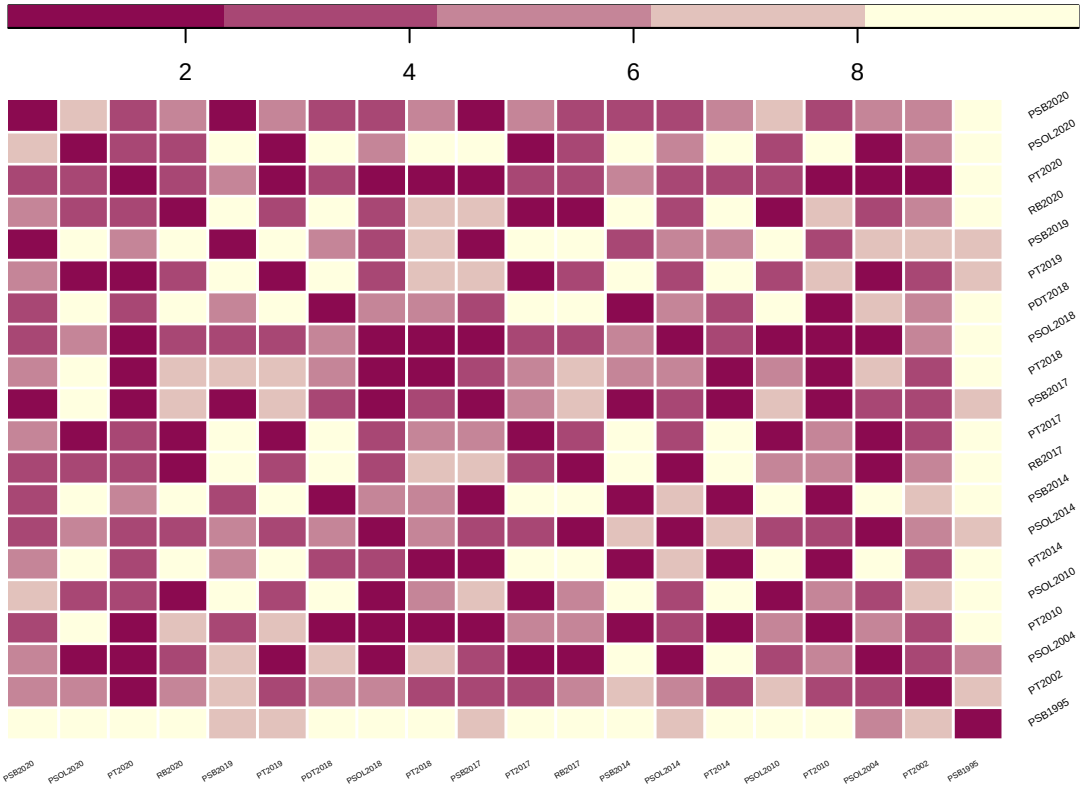


Figura 4.4: IDI entre grupos definidos pelos metadados Agremiação e Ano, normalizado usando z-score.

5 ÍNDICE DE POSICIONAMENTO IDEOLÓGICO

Este capítulo descreve a metodologia para cálculo de um índice que avalia os documentos de agremiações políticas face a uma referência comum. Foi escolhida uma base teórica, de cunho marxista, para realização de uma série de experimentos de aplicação do índice na análise de documentos de agremiações com perfil ideológico reconhecidamente de esquerda. Esses experimentos foram realizados por intermédio de codificação realizada pelo professor orientador na linguagem R.

5.1 Metodologia

O mecanismo para determinar as categorias que pertencem à ideologia vem da indução das categorias por obras teóricas. E a medida da distância será entre as bases ideológicas dos textos a analisar com as obras teóricas de referência.

A metodologia em questão, dentro do escopo deste trabalho, é apresentada sucintamente no fluxograma da Figura 5.1 e se constitui dos seguintes elementos essenciais:

- Escolha de textos teóricos que formalizam as categorias de uma corrente ideológica;
- Análise automatizada desses textos a fim de identificar os extratos que tratam de cada uma das categorias;
- Escolha dos documentos produzidos através de processos coletivos e que expressem as teses predominantes resultantes dos processos de debates e negociações políticas internas às agremiações;
- Análise automatizada dos documentos das agremiações em conjunto com os teóricos, estabelecendo medidas de divergência entre os primeiros e as diferentes categorias do segundo.

Em relação ao método do Capítulo 4, a metodologia acima apresenta semelhanças nas duas primeiras – preparação e limpeza de dados – e na última etapa – cálculo do índice de disparidade ideológica. A novidade aqui é a utilização de dois métodos de modelagem por tópicos em sequência. O primeiro – LDA Semeado – para obtenção da distribuição de elementos léxicos e o segundo – STM – para determinação da matriz de covariancia e consequente estimativa de disparidade ideológica. Tal como no método anterior, as categorias da análise são obtidas do perfil ideológico dos próprios documentos, não sendo, portanto, previamente definidas como no Projeto Manifesto. Conforme apresentado, a definição prévia de categorias feita no Projeto Manifesto permite estabelecer uma base de comparação uniforme para todos os contextos. Na metodologia apresentada neste capítulo e na medida em que usamos os tópicos extraídos dos próprios documentos como categorias de análise, a comparação se faz fixando as obras teóricas como base uniforme para a obtenção dos índices.

A metodologia desenvolvida neste capítulo se subdivide em cinco estágios ilustrados em linhas pontilhadas no quadro acima. O primeiro é responsável pela leitura e transformação dos textos (obras teóricas e documentos de agremiações). Neste estágio são identificados e extraídos os campos (parágrafos) dos documentos a serem analisados. Em seguida são carregados para processamento na forma de arquivos .csv com os fragmentos de textos separados por vírgulas. Estes arquivos constituirão os corpora do experimento. O segundo

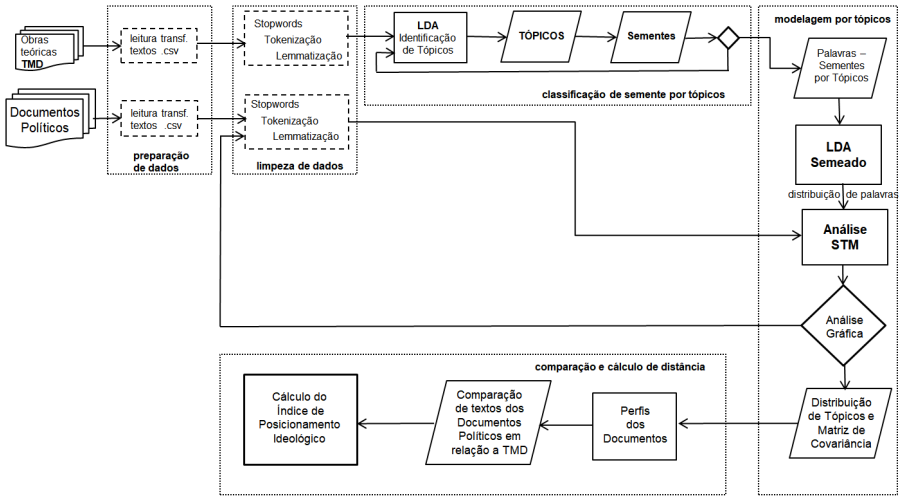


Figura 5.1: Visão geral da metodologia de análise chegando ao índice de posicionamento ideológico.

estágio, denominado de limpeza de dados, é responsável pela decomposição do corpus em termos (tokenização). Neste processo são eliminados os símbolos e caracteres de controle de arquivo ou de formatação, bem como os sinais de pontuação, números e datas. Em seguida inicia-se o processo de limpeza para a retirada das *stopwords*, que são compostas por preposições, artigos, advérbios, números, pronomes, conjunções, interjeição e pontuação. O terceiro estágio, denominado "classificação de semente por tópicos" será responsável por categorizar termos relevantes associando-os às respectivas frequências de ocorrência no corpus analisado, possibilitando assim inferência sobre suas afinidades, disparidades e elementos léxicos relacionados nos tópicos. Esses termos relevantes serão usados como semente no estágio seguinte. No quarto estágio, denominado "modelagem por tópicos", em função da constituição do *Document Term Matrix (DTM)*, é realizado o processo de identificação dos *topic model* com a utilização do modelo LDA Semeado. Por fim, no quinto estágio, denominado "comparação e cálculo de distância", é realizada a extração dos perfis para comparação dos documentos em relação à TMD e cálculo do índice de posicionamento ideológico.

5.1.1 Obras de Fundamentos Teóricos

Uma vez mais o índice está intrinsecamente relacionado à dimensionalidade do fenômeno subjacente, mas a novidade são os novos textos inseridos para formar a referência comum de comparação.

A dimensionalidade é o número de fontes separadas e interessantes de variação que existem entre os objetos que são analisados. No contexto da construção de índices, ou seja, na elaboração de uma única medida para caracterizar um fenômeno, essa questão se resume a unidimensionalidade e escalabilidade unidimensional - selecionando dados que podem ser demonstrados para corresponder a uma única dimensão. Em um sentido amplo, tal correspondência pode ser considerada como um teste de validade da medição e o argumento neste caso é que mesmo que um índice seja construído a priori em base teórico-ideológica, a validade do índice ainda deve ser testada em relação aos dados em que se baseia.

A determinação das categorias e o uso do LDA semeado inclui estimativa das sementes com metodologia inspirada em (WATANABE; ZHOU, 2020).

5.1.2 Documentos das Agremiações

O conjunto de dados é formado pela junção das duas categorias de documentos discutidos até aqui. Os documentos de obras de fundamentos teóricos participam do conjunto de dados com o intuito de se constituir na referência comum da análise. Os documentos a serem efetivamente analisados, quais sejam os documentos das agremiações, o são com base na referência comum formada pelos documentos das obras teóricas.

5.2 Teoria Marxista de Dependência

Como caso de estudo, é adotada uma base teórica representante de uma corrente de pensamento político que tem estado em evidência no Brasil, a Teoria Marxista da Dependência (abreviada pelo acrônimo TMD). O objetivo do experimento é a investigação da aplicação da metodologia como instrumento capaz de contribuir em uma análise alternativa à outras já existentes feitas por diversos autores com sustentação na TMD. Apesar de essa teoria estar bastante disseminada na América Latina por ocasião da redemocratização do País no final do regime militar, seus fundamentos teóricos não foram levados em conta na reorganização dos partidos de esquerda que surgiram nesse período. Passados quarenta anos, qual o grau de aderência do pensamento progressista brasileiro a essa teoria? Quanto dessa teoria consegue ser apropriada pelas organizações de esquerda contemporâneas para fundamentar seus diagnósticos que dirigirão suas ações politicas?

A premissa seria a incapacidade de a esquerda brasileira de propor uma alternativa socialista contra a dependência e o subdesenvolvimento por ter se afastado do marxismo. O objetivo não é comprovar se a premissa está correta ou não, mas, a partir do uso de uma metodologia com certa base teórica, obter de maneira automatizada as informações que darão sustentação ou não à premissa.

O corpus das obras teóricas foi sendo modificado ao longo do trabalho. De meros resumos de obras teóricas dispersas procurou-se trabalhar então com um corpus mais conciso e contendo apenas o essencial em torno da teoria. A principal alteração introduzida foi na categorização da TMD tomando por base a maneira como foi estruturado o ensaio fundante (MARINI, 2017). Porém, em substituição ao *Post Scriptum* como último capítulo da obra original, foi incluído o tema "imperialismo e poder global" inspirado na temática de (GARCIA; MARTINS; MENEZES, 2021). A alteração é justificada pelo enfoque radical e dialético de Marini de sempre articular dois planos indissociáveis da realidade: as análises de classe e a inserção internacional do Brasil, incompreensíveis uma sem a outra. Cabe aqui uma explicação. A inserção no capitalismo internacional de forma dependente faz com que a burguesia brasileira, para melhorar suas condições de competitividade e incapacitada de se valer da produtividade, implemente uma superexploração do trabalho nas relações de produção; o que acentua o caráter deformado das condições de reprodução social na periferia e, não podendo, em consequência, contar com o mercado interno de consumo popular, concentre então a realização da produção na esfera alta do consumo e na exportação. O resultado, segundo Marini, é um capitalismo *sui generis* brasileiro, cujo desenvolvimento se assemelha ao crescimento de um anão deformado. O desenvol-

Tabela 5.1: Obras Teóricas

Obra	Autor	Ano
Dialética da Dependência	Ruy Mauro Marini	1973
Desenvolvimento e Dependência	Ruy Mauro Marini	1992
Dependência e Superexploração da Força de Trabalho no Desenvolvimento Periférico	Marcelo Carcanholo	2005
O Atual Resgate Crítico da Teoria Marxista da Dependência	Marcelo Carcanholo	2013
O Legado de Ruy Mauro Marini para as Ciências Sociais: a Economia Política do Capitalismo Dependente	Carlos Eduardo Martins	2016
Hipótese a Respeito da Extensão da Superexploração do Trabalho no Capitalismo Avançado desde a Perspectiva da Teoria Marxista da Dependência	Adrián Sotelo Valencia	2016
A Teoria Marxista da Dependência à Luz de Marx e do Capitalismo Contemporâneo	Carlos Eduardo Martins	2018
Teoria Marxista da Dependência: Problemas e Categorias - Uma Visão Histórica Cap4	Mathias Seibel Luce	2018
Um Marco para o Fim de um Longo Exílio	Hugo F. Corrêa	2019
Resenha - Teoria Marxista da Dependência: Problemas e Categorias. Uma Visão Histórica	Maíra Machado Bichir	2020

vimento, em suma, do subdesenvolvimento. Cada obra teórica usada para a análise é formada pela união do seu título com o seu conteúdo. Uma versão reduzida das obras teóricas é elencada na Tabela 5.1.

Os principais elementos léxicos encontrados nas obras teóricas, com pelo menos 2 aparições são mostrados na nuvem de palavras da Figura 5.2.

5.3 Categorias da TMD via LDA Semeado

A técnica para encontrar boas sementes para tópicos é uma das peças centrais para a precisão do índice de posicionamento ideológico. A dificuldade em fazer um dicionário de sementes adequado é um obstáculo para os pesquisadores empregarem modelos semisupervisionados. Por esse motivo, experimentos relatados na literatura asseveram que os dicionários de semente devam ser construídos com base em uma mescla de análise humana com elementos léxicos frequentes para produzir bons resultados de classificação. De qualquer maneira, como é de se esperar, a inclusão de sementes falsas em um dicionário de semente prejudica o desempenho do classificador (WATANABE; ZHOU, 2020).

Posto de forma sintética, deve-se lidar com categorias correlatas e sementes que procuram destacar diferenças, pois as obras teóricas tratam categorias sem separação bem estabelecida. No entanto, é possível a identificação e a associação de extratos com as categorias. Para alcançar esse objetivo, seguimos uma linha inspirada em (WATANABE; ZHOU, 2020). Entretanto, substituímos a etapa de consulta a especialistas para selecionar as melhores sementes para formar um dicionário de indução no modelo por uma seleção au-

Figura 5.2: Quantidades relativas de ocorrência de elementos léxicos nas obras teóricas.

tomatizada de palavras temáticas extraídas das obras teóricas. Posteriormente, buscam-se os elementos léxicos que melhor designam conceitos das categorias de que são constituídas as obras teóricas para figurar como sementes, cada qual na categoria respectiva.

No presente caso de estudo, os elementos léxicos relevantes constituintes dos tópicos dos resumos de obras teóricas foram extraídos por meio de um procedimento em duas etapas. A primeira etapa é uma análise por tópicos com o método LDA dos documentos que dão a fundamentação teórica usando a quantidade de categorias da TMD como quantidade de tópicos. Os elementos léxicos de maior relevância semântica dos tópicos obtidos são interpretados e associados às categorias da TMD. A segunda etapa consiste em agregar à lista de elementos léxicos de cada categoria outros elementos léxicos advindos de uma leitura interpretativa das obras teóricas. As listas de elementos léxicos por categoria assim formadas são então consideradas como dicionário de sementes para análise por tópicos por meio do LDA Semeado.

Na identificação e associação de extratos das obras teóricas com as categorias da TMD, a distribuição de elementos léxicos dentro de um tópico é estacionária no sentido em que relevância semântica dos elementos léxicos em cada tópico é a mesma em todos os documentos. Há uma etapa de seleção dos extratos que devem participar da análise para determinação do índice de posicionamento ideológico. Os extratos selecionados são aqueles alocados a alguma categoria. Alguns extratos são excluídos da análise porque o método detectou que eles não se enquadram em nenhuma dessas categorias. O método trabalha com um limiar de 0,4 de relevância tópica, isto é, o documento da obra teórica tem que ter uma proporção de pelo menos 40% de uma categoria para ele ser mantido na análise, caso contrário é descartado. A classificação de extratos selecionados com relevância tópica acima de limiar de alguma das categorias está mostrada na Tabela 5.4.

Tabela 5.2: Sementes das categorias.

Integração ao mercado mundial	Segredo da troca desigual	Superexploração do trabalho	Ciclo do capital na economia dependente	Processo de industrialização	Novo anel da espiral	Imperialismo e poder global
alimento	apropriacao_valor	abaixo	acumulacao	acumulacao_capital	abundancia_recurso	abertura
bem_primario	aumento	ampliar	acumulacao_capital	ampliacao	anel	anexacao
capacidade_produtivo	bem_salario	avancado	atraso	atividade_subordinar	anel_espiral	brecha
carater	capitalista	baixo	barateamento_mercadoria	aumento_maisvalio	aproximacao	capital_financeiro
contraditorio	ciencia	capital	base	baratear_mercadoria	capital_estrangeiro	capital_transnacional
dependencia	compensacao	capitalismo_dependente	capacidade_produtivo	bem_superfluo	composicao_importacao	capitalista
dependencia_brasileiro	democracia	ciclo	ciclo_capital	camada_medio	compressao	consenso_washington
desenvolvimento	depreciacao	circulacao	circulacao	capacidade	configuracao	contrainsurgencia
desenvolver	desigual	classe	circulacao_capitalista	comercio	contemporaneo	controle
direito_trabalhista	deterioracao	compensacao	cisao_consumo	compressao_permanente	corporacao_imperialista	cooperacao_antagonica
divisao_internacional_trabalho	dinamico	critico	comercio_mundial	condicao_producao	desemprego	desafio
especializacao	forca_trabalho	dependencia	composicao	consumo_individual	destruir	desestabilizacao
especializacao_produtivo	forma	dependente	consumidor	consumo_popular	divisao_internacional	desregulamentacao
exploracao	grande	economia	consumo	crise	esfera	destruicao
exportacao	grau	economia_brasileiro	consumo_individual	crise_comercial	esfera_circulacao	ditadura
exportador	incremento	especificidade	contradicao	demanda_preexistente	espiral	divida_publica
ficticio	intercambio	expropriacao	demanda	depreciacao	estratificacao	divisao
fluxo_capital	internacional	extensao	dependencia	dificil	estrutura_circulacao	dominante
fluxo_mercadoria	lei	extensivo	dependencia_brasileiro	dois_esfera	estrutura_producao	doutrina
importacao	lucro	extracao	depressao	eixo	estrutural	geopolitica
integracao	maisvalio_extraordinario	extracao_maisvalio	descolonizacao	elevacao	etapa_inferior	globalizacao
interno	maisvalio_relativo	forca_produtivo	determinacao	elevar	exercito_reserva	golpe
latinoamericana	mecanismo	formacao	dialetica	esfera_alto_circulacao	financiamento	golpe_brando
latinoamericano	mercado	fundamento	dilaceracao	esfera_alto_consumo	fluxo_capital	hegemon
manufatura	monopolio	intensificacao	dilaceracao_economia	esfera_consumo	formacao_social	imperativos
materia_primo	nacao	intensivo	duplo_carater	generalizar_consumo	hierarquizacao	imperialista
mercado	pais_industrial	jornada	economia_condicionado	industria	industria_periferico	insercao_externa
mercadoria	pobre	maisvalio	economia_dependente	industria_debil	industrializacao_brasileiro	integração
mundial	popular	massa	economia_exportacao	industrial	inflacao	midiatico
oferta	preco	mecanismo	economia_exportador	lentidao	intervencao_estatal	monopolio
operar	preco_relativo	mercado_mundial	economia_industrial	longo_prazo	investimento_direto	monroe
periferico	produtividade	modo_circulacao	essencia	maisvalio_acumular	luta_classe	multilateral
producao	realidade	mundial	estratificacao	manufaturar	maquinario	mundo
producao_brasileiro	relacao_internacional	nivel	estrutura_social	mecanismo	nivel_vida	nacionalizacao
produtivo	segredo	organizacao	exigencia	modo_acumulacao	novo	neoliberalismo
relacao	social	organizacao_interno	fluxo_circulacao	nivel_salarial	pais_dependente	nova_ordem
reproducao_ampliar	termos	perda_maisvalio	fluxo_producao	obstruir_transicao	poder_compra	oligarquico
subalterno	transferencia	producao	formacao_social	oferta_externo	posguerra	oligopolio
subdesenvolvimento	transferencia_maisvalio	producao_capitalista	historico	operario	producao_industrial	ordem_mundial
subordinacao	transferencia_valor	prolongacao	independencia	processo_industrializacao	produto_semielaborar	pai_central
taxa	troca	remuneracao	industrializacao	produtividade_trabalho	progresso_tecnologico	passiva
tecnologico	troca_desigual	reproducao	inflexao	realizacao	redefinicao	perifericos
trabalhador	valor_realizado	salario	insercao	tecnologia_estrangeiro	semielaborar	potencia
trabalho	valor_trocado	sindicato	interdependencia	transformacao	setor_industrial	regional
alimento	apropriacao_valor	superexploracao	marxismo	valor	transferencia_renda	revolucao_nacional
bem_primario	aumento	superexploracao_trabalho	mercado_interno	acumulacao_capital	abundancia_recurso	revolucionario
capacidade_produtivo	bem_salario	trabalho	pais_central	ampliacao	anel	sistema
carater	capitalista	trabalho_excedente	periferia	atividade_subordinar	anel_espiral	soberania
contraditorio	ciencia	trabalho_necessario	producao	aumento_maisvalio	aproximacao	socialismo
dependencia	compensacao	transferencia_valor	producao_brasileiro	baratear_mercadoria	capital_estrangeiro	socios

Tabela 5.3: Elementos léxicos mais relevantes por categoria como resultado da análise através de LDA semeado.

Integração ao mercado mundial	Segredo da troca desigual	Superexploração do trabalho	Ciclo do capital na economia dependente	Processo de industrialização	Novo anel da espiral	Imperialismo e poder global
desenvolvimento	forma	capital	dependencia	industria	novo	sistema
exploracao	grande	economia	economia_dependente	valor	estrutural	capitalista
mundial	lei	valor	teoria	elevacao	esfera	pai_central
mercadoria	forca_trabalho	trabalho	relacao	crise	capital_estrangeiro	mundo
taxa	social	dependente	consumo	operario	contemporaneo	imperialista
producao	preco	superexploracao_trabalho	historico	elevar	configuracao	dominante
relacao	aumento	producao	producao	acumulacao_capital	formacao_social	controle
dependencia	capitalista	formacao	situacao	produtividade_trabalho	luta_classe	monopolio
interno	internacional	capitalismo_dependente	base	mecanismo	investimento_direto	regional
trabalho	lucro	mundial	acumulacao	comercio	setor_industrial	globalizacao
trabalhador	produtividade	nivel	superexploracao	transformacao	desemprego	destruicao
latinoamericano	mercado	superexploracao	composicao	capacidade	aproximacao	divisao
produtivo	nacao	dependencia	trabalhador	processo_industrializacao	producao_industrial	contrainsurgencia
desenvolver	transferencia_valor	classe	demanda	eixo	fluxo_capital	socialismo
tecnologico	intercambio	mercado_mundial	contradicao	realizacao	financiamento	abertura
carater	realidade	producao_capitalista	periferia	ampliacao	compressao	neoliberalismo
operar	transferencia	massa	acumulacao_capital	consumo_individual	esfera_circulacao	golpe
ficticio	dinamico	maisvalio	industrializacao	manufaturar	poder_compra	capital_financeiro
mercado	grau	salario	determinacao	consumo_popular	destruir	desregulamentacao
periferico	desigual	reproducao	ciclo_capital	longo_prazo	nivel_vida	soberania
subdesenvolvimento	troca	baixo	economia_industrial	aumento_maisvalio	maquinario	ditadura
exportador	mecanismo	ciclo	mercado_interno	condicao_producao	progresso_tecnologico	revolucionario
divisao_internacional_trabalho	monopolio	abaixo	economia_exportador	esfera_alto_circulacao	exercito_reserva	brecha
materia_primo	democracia	jornada	produtor	bem_superfluo	etapa_inferior	doutrina
integracao	troca_desigual	transferencia_valor	circulacao	camada_medio	hierarquizacao	desestabilizacao
contraditorio	ciencia	critico	essencia	tecnologia_estrangeiro	estratificacao	potencia
exportacao	maisvalio_relativo	mecanismo	marxismo	dois_esfera	anel_espiral	capital_transnacional
importacao	incremento	especificidade	formacao_social	maisvalio_acumular	composicao_importacao	terror
subordinacao	popular	forca_produtivo	insercao	esfera_alto_consumo	produto_semielaborar	midiatico
oferta	compensacao	ampliar	realizacao	nivel_salarial	corporacao_imperialista	multilateral
capacidade_produtivo	relacao_internacional	uso	interdependencia	dificil	abundancia_recurso	revolucao_nacional
alimento	deterioracao	circulacao	resgate	depreciacao	redefinicao	ordem_mundial
reproducao_ampliar	segredo	extensao	dialetica	atividade_subordinar	inflacao	consenso_washington
manufatura	maisvalio_extraordinario	trabalho_excedente	separacao	oferta_externo	transferencia_renda	desafio
especializacao_produtivo	valor_trocado	organizacao	independencia	compressao_permanente	estrutura_producao	golpe_brando

Tabela 5.4: Extratos selecionados por categoria.

Integração ao mercado mundial	Segredo da troca desigual	Superexploração do trabalho	Ciclo do capital na economia dependente	Processo de industrialização	Novo anel da espiral	Imperialismo e poder global
Marcelo Carcanholo-02	Carlos Eduardo Martins-02	Carlos Eduardo Martins-01	Marcelo Carcanholo-01	Ruy Mauro Marini-06	Carlos Eduardo Martins-018	Carlos Eduardo Martins-05
Marcelo Carcanholo-03	Carlos Eduardo Martins-03	Carlos Eduardo Martins-04	Ruy Mauro Marini-029	Ruy Mauro Marini-08	Carlos Eduardo Martins-019	Carlos Eduardo Martins-06
Marcelo Carcanholo-07	Carlos Eduardo Martins-07	Carlos Eduardo Martins-08	Ruy Mauro Marini-030	Ruy Mauro Marini-010	Carlos Eduardo Martins-020	Carlos Eduardo Martins-021
Marcelo Carcanholo-013	Carlos Eduardo Martins-09	Carlos Eduardo Martins-014	Ruy Mauro Marini-033	Ruy Mauro Marini-011	Ruy Mauro Marini-044	Carlos Eduardo Martins-022
Ruy Mauro Marini-02	Carlos Eduardo Martins-010	Carlos Eduardo Martins-015	Ruy Mauro Marini-035	Ruy Mauro Marini-015	Ruy Mauro Marini-046	Carlos Eduardo Martins-023
Ruy Mauro Marini-03	Carlos Eduardo Martins-011	Marcelo Carcanholo-04	Ruy Mauro Marini-054	Ruy Mauro Marini-021	Ruy Mauro Marini-055	Carlos Eduardo Martins-024
Ruy Mauro Marini-04	Carlos Eduardo Martins-012	Marcelo Carcanholo-05	Ruy Mauro Marini-059	Ruy Mauro Marini-022	Carlos Eduardo Martins-044	Carlos Eduardo Martins-025
Ruy Mauro Marini-05	Carlos Eduardo Martins-013	Marcelo Carcanholo-06	Ruy Mauro Marini-061	Ruy Mauro Marini-025	Carlos Eduardo Martins-045	Carlos Eduardo Martins-026
Ruy Mauro Marini-07	Carlos Eduardo Martins-016	Marcelo Carcanholo-08	Ruy Mauro Marini-062	Ruy Mauro Marini-026	Mathias Seibel Luce-014	Carlos Eduardo Martins-027
Ruy Mauro Marini-012	Carlos Eduardo Martins-017	Marcelo Carcanholo-09	Marcelo Carcanholo-015	Ruy Mauro Marini-034	Mathias Seibel Luce-015	Carlos Eduardo Martins-028
Ruy Mauro Marini-013	Ruy Mauro Marini-09	Marcelo Carcanholo-010	Marcelo Carcanholo-016	Ruy Mauro Marini-036	Mathias Seibel Luce-017	Carlos Eduardo Martins-029
Ruy Mauro Marini-023	Ruy Mauro Marini-014	Marcelo Carcanholo-011	Marcelo Carcanholo-017	Ruy Mauro Marini-037	Mathias Seibel Luce-018	Carlos Eduardo Martins-030
Ruy Mauro Marini-024	Ruy Mauro Marini-016	Marcelo Carcanholo-012	Marcelo Carcanholo-018	Ruy Mauro Marini-039	Mathias Seibel Luce-019	Ruy Mauro Marini-01
Ruy Mauro Marini-027	Ruy Mauro Marini-017	Marcelo Carcanholo-014	Marcelo Carcanholo-019	Ruy Mauro Marini-041	Mathias Seibel Luce-032	Ruy Mauro Marini-045
Ruy Mauro Marini-031	Ruy Mauro Marini-018	Ruy Mauro Marini-020	Marcelo Carcanholo-035	Ruy Mauro Marini-042	Hugo F. Corrêa-02	Ruy Mauro Marini-053
Ruy Mauro Marini-032	Ruy Mauro Marini-019	Ruy Mauro Marini-028	Marcelo Carcanholo-036	Ruy Mauro Marini-043		Carlos Eduardo Martins-038
Adrián Sotelo Valencia-04	Ruy Mauro Marini-038	Ruy Mauro Marini-040	Carlos Eduardo Martins-031	Ruy Mauro Marini-047		Carlos Eduardo Martins-042
Mathias Seibel Luce-05	Ruy Mauro Marini-060	Ruy Mauro Marini-057	Maíra Machado Bichir-01	Ruy Mauro Marini-048		Carlos Eduardo Martins-043
Mathias Seibel Luce-010	Marcelo Carcanholo-020	Ruy Mauro Marini-058	Mathias Seibel Luce-01	Ruy Mauro Marini-049		Carlos Eduardo Martins-046
	Marcelo Carcanholo-023	Adrián Sotelo Valencia-01	Mathias Seibel Luce-02	Ruy Mauro Marini-050		Carlos Eduardo Martins-047
	Marcelo Carcanholo-024	Adrián Sotelo Valencia-02	Mathias Seibel Luce-03	Ruy Mauro Marini-051		Carlos Eduardo Martins-048
	Marcelo Carcanholo-026	Adrián Sotelo Valencia-03	Mathias Seibel Luce-04	Ruy Mauro Marini-052		Carlos Eduardo Martins-049
	Marcelo Carcanholo-028	Adrián Sotelo Valencia-05	Mathias Seibel Luce-06	Ruy Mauro Marini-056		Carlos Eduardo Martins-050
	Marcelo Carcanholo-034	Adrián Sotelo Valencia-06	Mathias Seibel Luce-07	Adrián Sotelo Valencia-07		Carlos Eduardo Martins-051
	Marcelo Carcanholo-037	Adrián Sotelo Valencia-08	Mathias Seibel Luce-08	Mathias Seibel Luce-016		Carlos Eduardo Martins-052
	Carlos Eduardo Martins-032	Marcelo Carcanholo-021	Mathias Seibel Luce-09			Carlos Eduardo Martins-053
	Carlos Eduardo Martins-033	Marcelo Carcanholo-022	Mathias Seibel Luce-011			Carlos Eduardo Martins-054
	Carlos Eduardo Martins-034	Marcelo Carcanholo-025	Mathias Seibel Luce-012			Carlos Eduardo Martins-055
	Carlos Eduardo Martins-035	Marcelo Carcanholo-027	Mathias Seibel Luce-020			Carlos Eduardo Martins-056
	Carlos Eduardo Martins-036	Marcelo Carcanholo-029	Mathias Seibel Luce-021			Carlos Eduardo Martins-057
	Carlos Eduardo Martins-037	Marcelo Carcanholo-030	Mathias Seibel Luce-022			Carlos Eduardo Martins-058
	Carlos Eduardo Martins-039	Marcelo Carcanholo-031	Mathias Seibel Luce-024			Carlos Eduardo Martins-059
	Carlos Eduardo Martins-040	Marcelo Carcanholo-032	Mathias Seibel Luce-025			Carlos Eduardo Martins-060
	Carlos Eduardo Martins-041	Marcelo Carcanholo-033	Mathias Seibel Luce-026			Carlos Eduardo Martins-061
	Maíra Machado Bichir-03	Maíra Machado Bichir-02	Mathias Seibel Luce-028			Carlos Eduardo Martins-062
	Mathias Seibel Luce-023	Mathias Seibel Luce-013	Mathias Seibel Luce-029			Carlos Eduardo Martins-063
	Mathias Seibel Luce-033	Mathias Seibel Luce-027	Mathias Seibel Luce-030			Carlos Eduardo Martins-064
	Mathias Seibel Luce-034	Mathias Seibel Luce-031	Mathias Seibel Luce-035			
	Hugo F. Corrêa-03	Mathias Seibel Luce-036	Hugo F. Corrêa-01			
		Mathias Seibel Luce-037				

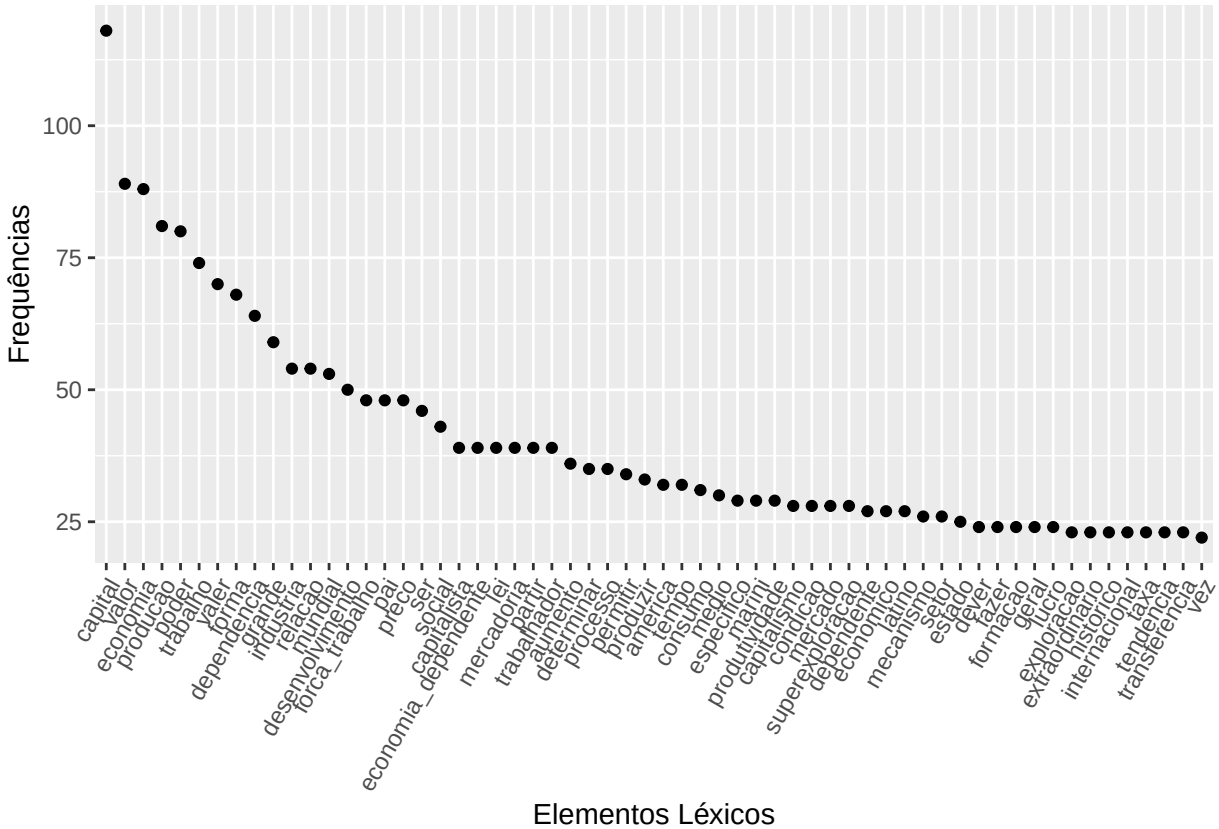


Figura 5.3: Frequências de ocorrência de elementos léxicos nos extratos selecionados.

5.4 Índices de Agremiações Políticas

Em conjunto com os extratos selecionados das obras teóricas classificados, são identificados os tópicos constitutivos, calculando-se um índice entre cada documento e os textos de cada categoria teórica conforme (4.1). O corpus **D** analisado contém, então, os extratos selecionados das obras teóricas e todos os extratos dos documentos de agremiações partidárias. Algumas premissas estão relacionadas a tal abordagem. No contexto da análise proposta nesta dissertação:

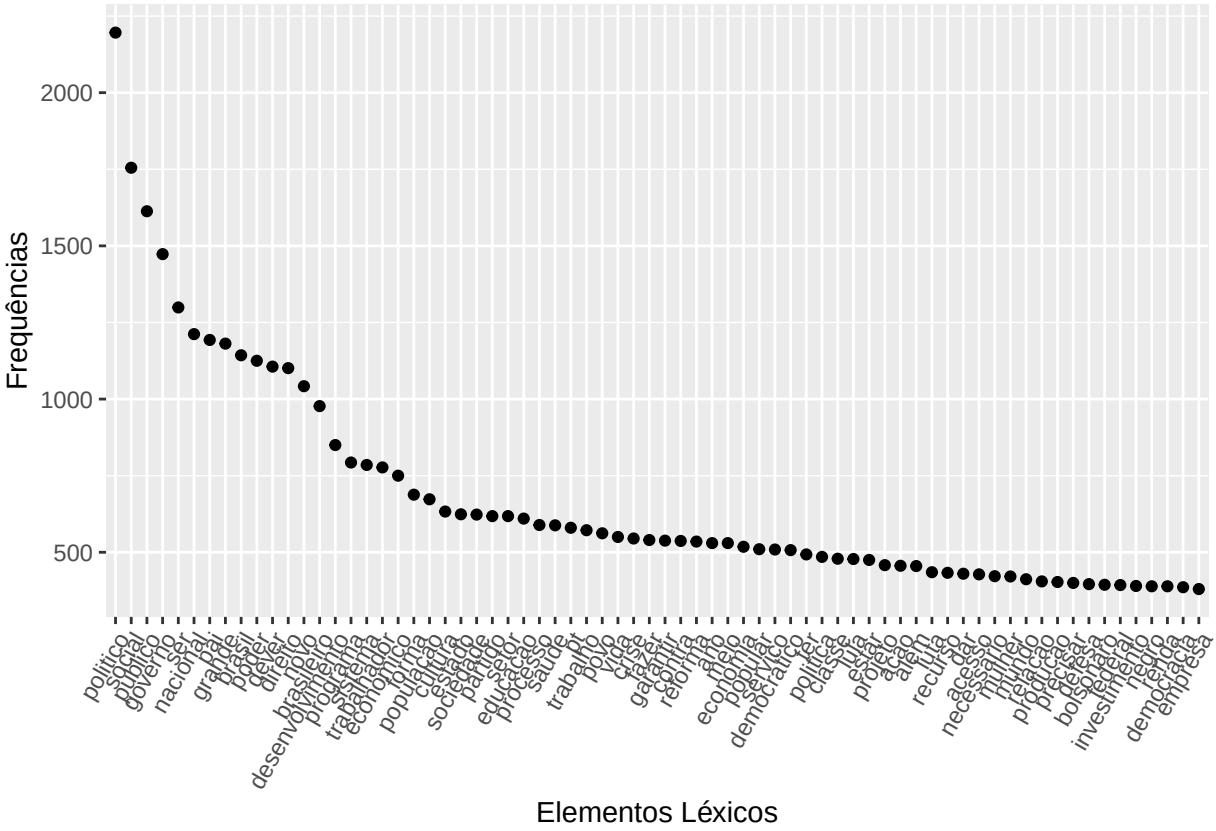


Figura 5.4: Frequências de ocorrência de elementos léxicos nos extratos do corpus completo.

- em relação a tópicos correlatos: os tópicos abordados nos documentos não são independentes uns dos outros. Ao contrário, supõe-se a existência de alguma correlação entre esses tópicos. Essa é uma premissa razoável dada a ausência de fronteira clara entre as categorias da TMD, favorecendo portanto algum padrão de ocorrências concomitantes entre elas nos documentos.
- em relação a fonte de documentos conhecida e identificada: grupos de documentos são identificados por valores específicos de metadados associados aos documentos. O agrupamento que leva em conta outras informações podendo considerar que o perfil ideológico está correlacionado com metadados sobre o documento. Por exemplo, os metadados podem incluir data de publicação, agremiação ou categoria do documento.

Aplicação do STM

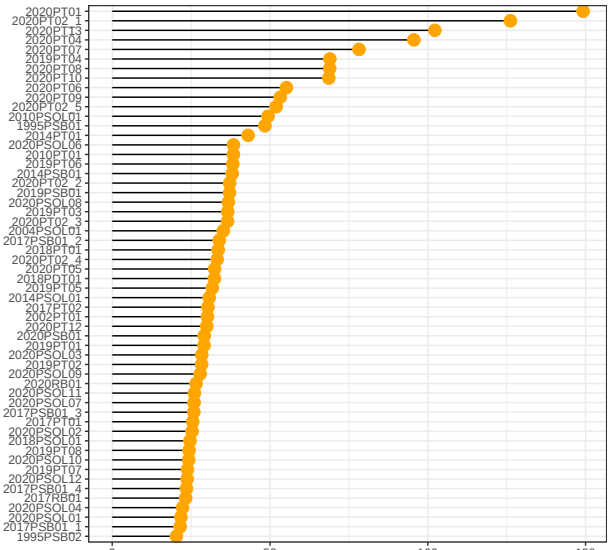
Uma alteração introduzida na análise foi considerar a variabilidade dos "dados de palavras" conforme o grupo de documentos. Esses diferentes "dados" para um mesmo tópico (um de cada grupo) são variações em torno de um mesmo perfil, que é o perfil do tópico.

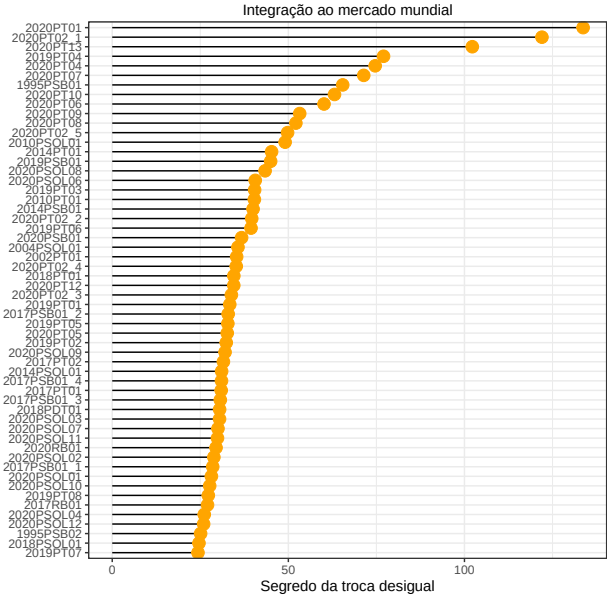
Tabela 5.5: Medidas de divergência ideológica entre os documentos das agremiações e as categorias.

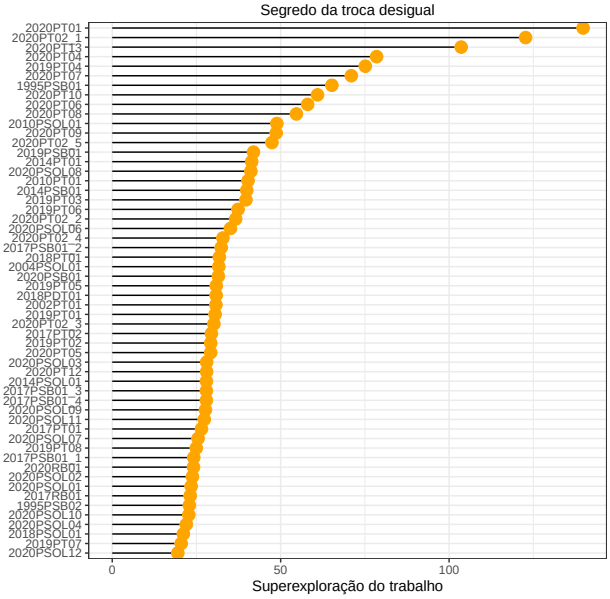
	Integração ao mercado mundial	Segredo da troca desigual	Superexploração do trabalho	Ciclo do capital na economia dependente	Processo de industrialização	Novo anel da espiral	Imperialismo e poder global
1995PSB01	48.37859	65.43418	65.20763	50.98599	57.59687	65.23435	41.31077
1995PSB02	20.35355	25.09286	22.90188	18.62376	31.68854	31.76340	20.02449
2002PT01	30.18040	35.32705	30.78156	29.07419	38.52218	32.00872	23.18396
2004PSOL01	35.19543	35.69950	31.64388	32.71936	51.58784	22.13273	27.23019
2010PSOL01	49.38138	49.09080	48.87664	46.14445	60.91237	39.96416	27.95846
2010PT01	38.40254	40.35791	40.31528	37.08858	49.13534	34.07305	25.23363
2014PSB01	38.02517	39.98701	39.94007	34.73164	44.66397	46.73832	30.18167
2014PSOL01	30.74864	31.05889	27.96440	27.12823	44.52337	24.25092	18.03177
2014PT01	43.07674	45.24633	41.40224	41.02490	54.72693	38.41767	29.91643
2017PSB01_1	21.53077	28.53011	24.20078	19.94159	30.90187	36.04827	17.54512
2017PSB01_2	33.92598	32.92070	32.36663	29.32345	44.74292	37.38588	27.36962
2017PSB01_3	25.86291	30.68806	27.95899	23.55236	36.31297	35.41148	20.01930
2017PSB01_4	23.52416	31.02518	27.95193	22.06384	31.43559	39.89696	17.79901
2017PT01	25.51782	30.94036	26.49904	23.71665	35.83735	30.30589	18.25908
2017PT02	30.32562	31.56726	29.45864	28.58278	43.73281	24.20722	17.01645
2017RB01	23.27857	27.05588	23.13950	22.34582	37.66510	21.56266	19.42471
2018PDT01	32.39374	30.49167	30.87136	30.30018	43.06113	25.22613	17.89376
2018PSOL01	24.64140	24.62943	21.11209	20.45331	36.06962	25.04387	15.26443
2018PT01	33.58282	34.52334	31.79977	30.55157	44.25130	31.89470	19.73487
2019PSB01	37.20027	44.97563	41.93877	34.71472	45.59444	49.16707	30.87441
2019PT01	29.11011	33.36769	30.55348	26.86968	40.43405	34.78831	18.87178
2019PT02	28.33443	32.37535	29.24631	25.66654	40.17335	28.58289	17.27108
2019PT03	36.61177	40.42905	39.69723	34.07670	50.50245	34.13817	23.72200
2019PT04	68.97471	77.02716	75.13292	66.30014	85.28425	65.18189	52.90971
2019PT05	31.69965	32.82936	30.89028	27.80510	46.21267	26.67569	18.66521
2019PT06	38.23126	39.40971	37.36110	34.78185	54.83358	26.34167	26.41496
2019PT07	23.84503	24.34735	20.47872	20.72490	36.16776	21.65759	14.81546
2019PT08	24.37137	27.27994	24.95939	22.43305	34.15419	25.59526	13.11537
2020PSB01	29.15600	36.73332	31.52155	26.11093	36.81915	44.16240	23.81031
2020PSOL01	21.78328	28.15785	23.43705	19.29103	33.99129	26.65683	18.08976
2020PSOL02	25.31050	28.85715	23.80979	23.29399	36.58683	23.08516	17.95711
2020PSOL03	28.37757	30.46872	28.02199	27.84053	41.36182	17.92890	16.63922
2020PSOL04	22.29355	26.12956	21.99827	20.38907	32.43979	25.12311	14.66956
2020PSOL06	38.42273	40.60165	35.09911	34.37941	54.34046	27.64961	28.86656
2020PSOL07	25.94114	30.02087	25.47682	23.64475	38.26720	27.25580	17.38638
2020PSOL08	36.79960	43.40579	41.15125	34.88169	49.50942	36.25171	24.78836
2020PSOL09	27.85068	32.03415	27.67584	24.87989	39.61627	28.33882	20.19443
2020PSOL10	24.23675	27.66946	22.73640	21.60853	35.72843	25.25539	16.96345
2020PSOL11	26.05761	29.88759	27.32692	22.92414	37.97842	27.20611	16.82327
2020PSOL12	23.75577	25.95006	19.47295	21.71511	37.30566	16.86506	21.14735
2020PT01	149.21998	133.68498	139.78330	140.89376	175.95803	98.11709	117.13026
2020PT02_1	126.19509	121.99686	122.71994	118.30309	152.03519	93.40005	98.74592
2020PT02_2	37.22242	39.60543	36.64125	34.27965	49.33401	34.03455	23.44971
2020PT02_3	36.55542	33.82981	30.18057	30.39031	46.11291	29.54628	22.04599
2020PT02_4	33.27175	35.21454	32.85842	30.16357	43.97802	32.64637	20.30531
2020PT02_5	51.94594	49.78721	47.40801	48.28079	64.90744	38.61757	35.20415
2020PT04	95.60904	74.65810	78.48709	87.50778	115.43492	41.27788	65.44807
2020PT05	32.44465	32.65556	29.23411	27.40291	44.27631	25.49960	17.40751
2020PT06	55.21302	60.13323	58.01606	52.73259	71.56392	43.76578	38.04704
2020PT07	78.22143	71.40482	70.97542	73.03496	97.89717	37.75590	56.11755
2020PT08	68.92699	52.13943	54.67735	62.20562	86.87538	28.94707	41.70702
2020PT09	53.25138	53.22678	48.67721	44.86894	70.69008	44.88051	40.10279
2020PT10	68.63658	63.10750	60.94759	62.34636	85.95301	42.48115	50.77189
2020PT12	29.96420	34.50903	28.02134	26.02502	39.98401	32.58446	19.88521
2020PT13	102.23000	102.22784	103.59989	96.62864	126.84530	72.31171	78.70071
2020RB01	26.58142	29.47576	24.13964	25.59039	41.79506	17.65112	22.84531

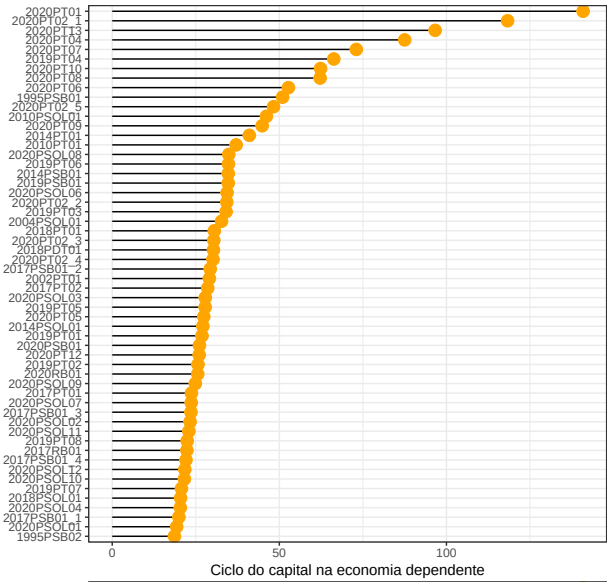
A rigor, é o conjunto de elementos léxicos que caracteriza um mesmo tópico, o mesmo tema, mas a ênfase ou o ângulo que se usa para um tópico pode ser diferente de outro. Os elementos léxicos com mais ênfase num tópico não são exatamente os mesmos de outro, mas variações em torno de um mesmo perfil, o tópico é o mesmo.

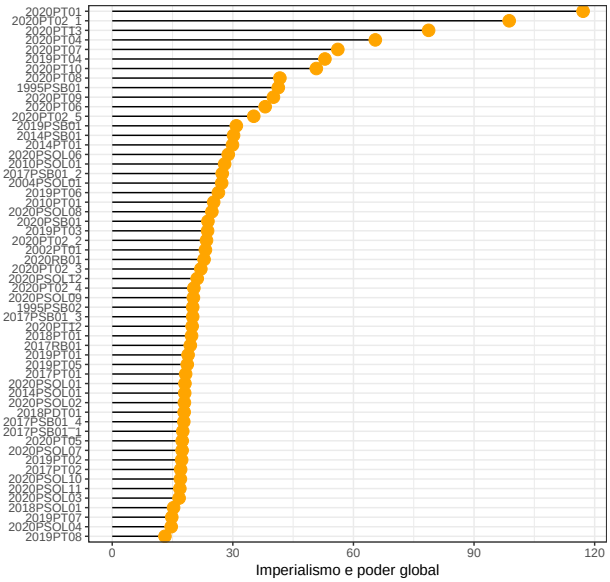
O novo corpus, formado pela junção dos extratos selecionados das obras teóricas e de todos os extratos dos documentos de agremiações partidárias, foi submetido neste experimento ao STM com variabilidade da distribuição de elementos léxicos. O resultado pode ser visto nos gráficos abaixo, por categoria.


2020PT01
2020PT02_1
2020PT13
2020PT04
2020PT07
2019PT04
2020PT08
2020PT10
2020PT06
2020PT09
2020PT02_5
2010PSQL01
1995PSB01
2014PT01
2020PSQL06
2010PT01
2019PT06
2014PSB01
2020PT02_2
2019PSB01
2020PSQL08
2019PT03
2020PT02_3
2004PSQL01
2017PSB01_2
2018PT01
2020PT02_4
2020PT05
2018PDT01
2019PT05
2014PSQL01
2017PT02
2002PT01
2020PT12
2020PSB01
2019PT01
2020PSQL03
2019PT02
2020PSQL09
2020RB01
2020PSQL11
2020PSQL07
2017PSB01_3
2017PT01
2020PSQL02
2018PSQL01
2019PT08
2020PSQL10
2019PT07
2020PSQL12
2017PSB01_4
2017RB01
2020PSQL04
2020PSQL01
2017PSB01_1
1995PSB02
0
50
100
150
Integração ao mercado mundial


2020PT01
2020PT02_1
2020PT13
2019PT04
2020PT04
2020PT07
1995PSB01
2020PT10
2020PT06
2020PT09
2020PT08
2020PT02_5
2010PSQL01
2014PT01
2019PSB01
2020PSQL08
2020PSQL06
2019PT03
2010PT01
2014PSB01
2020PT02_2
2019PT06
2020PSB01
2004PSQL01
2002PT01
2020PT02_4
2018PT01
2020PT12
2020PT02_3
2019PT01
2017PSB01_2
2019PT05
2020PT05
2019PT02
2020PSQL09
2017PT02
2014PSQL01
2017PSB01_4
2017PT01
2017PSB01_3
2018PDT01
2020PSQL03
2020PSQL07
2020PSQL11
2020RB01
2020PSQL02
2017PSB01_1
2020PSQL01
2020PSQL10
2019PT08
2017RB01
2020PSQL04
2020PSQL12
1995PSB02
2018PSQL01
2019PT07
0
50
100
Segredo da troca desigual


2020PT01
2020PT02_1
2020PT13
2020PT04
2019PT04
2020PT07
1995PSB01
2020PT10
2020PT06
2020PT08
2010PSQL01
2020PT09
2020PT02_5
2019PSB01
2014PT01
2020PSQL08
2010PT01
2014PSB01
2019PT03
2019PT06
2020PT02_2
2020PSQL06
2020PT02_4
2017PSB01_2
2018PT01
2004PSQL01
2020PSB01
2019PT05
2018PDT01
2002PT01
2019PT01
2020PT02_3
2017PT02
2019PT02
2020PT05
2020PSQL03
2020PT12
2014PSQL01
2017PSB01_3
2017PSB01_4
2020PSQL09
2020PSQL11
2017PT01
2020PSQL07
2019PT08
2017PSB01_1
2020RB01
2020PSQL02
2020PSQL01
2017RB01
1995PSB02
2020PSQL10
2020PSQL04
2018PSQL01
2019PT07
2020PSQL12
0
50
100
Superexploração do trabalho


2020PT01
2020PT02_1
2020PT13
2020PT04
2020PT07
2019PT04
2020PT10
2020PT08
2020PT06
1995PSB01
2020PT02_5
2010PSOL01
2020PT09
2014PT01
2010PT01
2020PSOL08
2019PT06
2014PSB01
2019PSB01
2020PSOL06
2020PT02_2
2019PT03
2004PSOL01
2018PT01
2020PT02_3
2018PDT01
2020PT02_4
2017PSB01_2
2002PT01
2017PT02
2020PSOL03
2019PT05
2020PT05
2014PSOL01
2019PT01
2020PSB01
2020PT12
2019PT02
2020RB01
2020PSOL09
2017PT01
2020PSOL07
2017PSB01_3
2020PSOL02
2020PSOL11
2019PT08
2017RB01
2017PSB01_4
2020PSOL12
2020PSOL10
2019PT07
2018PSOL01
2020PSOL04
2017PSB01_1
2020PSOL01
1995PSB02
0
50
100
Ciclo do capital na economia dependente
2020PT01
2020PT02_1
2020PT13
2020PT04
2020PT07
2020PT08
2020PT10
2019PT04
2020PT06
2020PT09
2020PT02_5
2010PSOL01
1995PSB01
2019PT06
2014PT01
2020PSOL06
2004PSOL01
2019PT03
2020PSOL08
2020PT02_2
2010PT01
2019PT05
2020PT02_3
2019PSB01
2017PSB01_2
2014PSB01
2014PSOL01
2020PT05
2018PT01
2020PT02_4
2017PT02
2018PDT01
2020RB01
2020PSOL03
2019PT01
2019PT02
2020PT12
2020PSOL09
2002PT01
2020PSOL07
2020PSOL11
2017RB01
2020PSOL12
2020PSB01
2020PSOL02
2017PSB01_3
2019PT07
2018PSOL01
2017PT01
2020PSOL10
2019PT08
2020PSOL01
2020PSOL04
1995PSB02
2017PSB01_4
2017PSB01_1
0
50
100
150
Processo de industrialização
2020PT01
2020PT02_1
2020PT13
1995PSB01
2019PT04
2019PSB01
2014PSB01
2020PT09
2020PSB01
2020PT06
2020PT10
2020PT04
2010PSOL01
2017PSB01_4
2020PT02_5
2014PT01
2020PT07
2017PSB01_2
2020PSOL08
2017PSB01_1
2017PSB01_3
2019PT01
2019PT03
2010PT01
2020PT02_2
2020PT02_4
2020PT12
2002PT01
2018PT01
1995PSB02
2017PT01
2020PT02_3
2020PT08
2019PT02
2020PSOL09
2020PSOL06
2020PSOL07
2020PSOL11
2019PT05
2020PSOL01
2019PT06
2019PT08
2020PT05
2020PSOL10
2018PDT01
2020PSOL04
2018PSOL01
2014PSOL01
2017PT02
2020PSOL02
2004PSOL01
2019PT07
2017RB01
2020PSOL03
2020RB01
2020PSOL12
0
25
50
75
100
Novo anel da espiral



Cada gráfico mostra a medida de divergência ideológica entre os documentos das agremiações em cada uma das sete categorias estipuladas para a TMD. Denominamos a média dessas medidas de Índice de Posicionamento Ideológico – IPI para retratar a aproximação dos documentos das agremiações com a TMD.

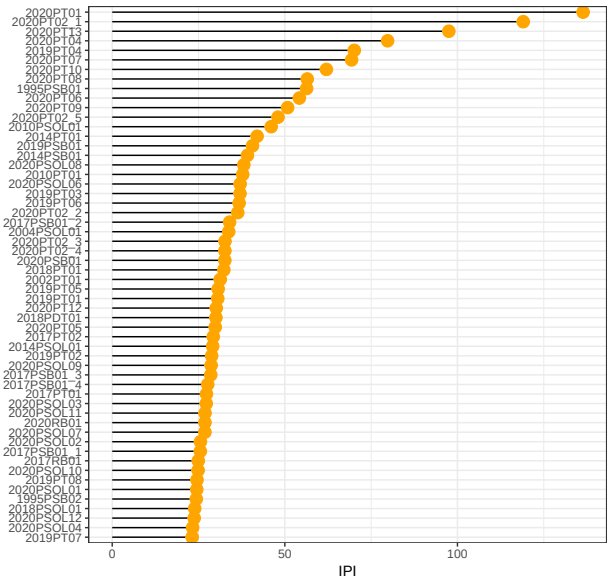


Figura 5.5: Média por categoria.

Nos gráficos por categoria, um valor elevado reflete a prevalência do tópico dentro do conjunto de documentos. No gráfico do IPI, que é a média das medidas obtidas por um documento nas sete categorias da TMD, um valor elevado significa que o documento se afasta dessa teoria. Nesse sentido, podem ser elencados como documentos dentre os mais distantes da TMD aqueles produzidos pelas agremiações mostradas na Tabela 5.6.

Tabela 5.6: Agremiações cujos documentos estão mais distantes das obras da TMD.

Identificação	**Título**	**Organização, Tendência ou Corrente**
2010PSOL01	Uma alternativa socialista: nossas tarefas e diretrizes	Partido Socialismo e Liberdade

2020PT01	2020 Texto de Emenda “Sobre a conjuntura política imediata” - Tendências CNB e MPT ao DN - PT	CNB e MPT
2020PT04	2020 Texto de Contribuição “PT na Luta pela Vida e pela Dignidade” - PT	Movimento PT
2020PT13	2020 Proposta de Texto - Coletivo PT de Todas as Lutas para Reunião DN Abril 2020 -PT	Coletivo PT de Todas as Lutas

Por outro lado, o método determinou medidas de maior proximidade com a TMD nos documentos produzidos pelas agremiações da Tabela 5.7.

Tabela 5.7: Agremiações cujos documentos estão mais próximos das obras da TMD.

Identificação	**Título**	**Organização, Tendência ou Corrente**
1995PSB02	1995 Programa PSB	Partido Socialista Brasileiro
2017PSB01	2017 Plano Estratégico de Desenvolvimento Nacional - PSB	Partido Socialista Brasileiro
2017RB01	2017 Manifesto pela Revolução Brasileira - RB	Revolução Brasileira
2020PSOL12	2020 VII Congresso PSOL Tese Coletivo Alicerce - PSOL	Coletivo Alicerce
2020RB01	2020 Contribuição da RB ao 7º Congresso Nacional do PSOL - RB	Revolução Brasileira

O método detectou, nos cinco documentos acima relacionados com razoável conteúdo classista, extratos de textos com relevância tópica similar a TMD. Um desses documentos, 2020PSOL12, do Coletivo Alicerce, efetivamente traz elementos comuns com a TMD como pode ser visto nesses parágrafos:

> *“Há algumas décadas o pensamento crítico brasileiro já havia desvelado a natureza do subdesenvolvimento. O subdesenvolvimento da periferia aparece como condição do desenvolvimento do centro, não se tratando de uma etapa anterior e/ou necessária para alcançarmos os padrões de produção e de vida das nações desenvolvidas. É uma relação capitalista moderna caracterizada pelos padrões de superexploração e subordinação, da qual decorre um modelo de urbanização - a suburbanidade. Apesar do crescimento econômico, dos momentos de redução relativa da pobreza, da modernização via inundação de mercadorias importadas, não resolvemos os problemas mais básicos da sobrevivência e reprodução da sociedade, como saneamento básico, mobilidade urbana, educação e saúde, moradia, condições dignas de trabalho.”*

> *“Os problemas do Brasil não se resumem a um ou outro governo, o que não significa que o comando do país não faça diferença, mas sim que acumulamos questões não resolvidas desde a nossa formação. O tipo de colonização e desenvolvimento econômico fraturou o país num abismo social e racial com repercussão atual. Mesmo estando entre os países mais ricos do mundo, ocupamos os últimos lugares em igualdade e justiça social. Não se trata apenas de uma questão de distribuição da renda, são problemas estruturais que há cinco séculos seguem sem resolução.”*

Em relação ao PSB, partido caracterizado por uma atuação política vacilante e muitas vezes até contraditória com os princípios do socialismo que carrega no nome, cabe lembrar a radicalidade programática de seu documento fundacional de 1947 – notado como 1995PSB02, que é resgatado em muitas passagens no seu documento mais recente,

2017PSB01, que também foi bem posicionado pelo método. Um exemplo desse resgate do passado para atualizar uma nova roupagem discursiva adotada pelo partido pode ser visto neste parágrafo em sua página 28:

> *"A socialização dos meios de produção, afirmada quando de sua fundação em 1947, é hoje entendida dentro do princípio constitucional que determina a "função social da propriedade" ampliado para toda atividade econômica e social. O liberal-capitalismo cultivou o espírito do individualismo, da competição, do consumismo alienado, os interesses individuais e corporativos e o sistema hierárquico e patronal nas atividades econômicas. O socialismo, assumindo ser dimensão ética e humanista a "função social" de toda atividade humana, adota outros valores quais a solidariedade, a cooperação, a equidade, a justiça, a igualdade, e, desta forma, a superação das civilizações baseadas na "exploração do homem pelo homem". Trabalho é a atividade tipicamente humana pelas suas dimensões de inteligência e éticas. É no trabalho que se desenvolvem tanto a valorização como a exploração dos seres humanos. O socialismo tem como um dos maiores desafios revolucionar a concepção e a prática da produção humana e do trabalho, que não podem ser submetidos aos princípios do mercado, mas que devem contribuir para desenvolver os valores da solidariedade e da cooperação, uma nova concepção de desenvolvimento individual e social, assumir o princípio da sustentabilidade que é a luta pela qualidade da vida e pela permanência de tudo o que é vida, para hoje e para o futuro."*

Na análise dos resultados por categoria, chama a atenção o posicionamento dos próprios documentos da teoria, agregados sob a sigla TMD que, na maioria dos casos, serviram de referência comparativa na parte inferior dos gráficos, mas que saltaram para a extremidade superior apenas na categoria "imperialismo e poder global". Lembrando que essa categoria foi criada artificialmente em complemento aos pilares da obra fundante da TMD, esse deslocamento pode ser interpretado pela origem de suas palavras sementes, que não vieram das obras teóricas. A criação forçada do tópico mostrou então que o mesmo não era prevalente na busca por semelhança com os documentos das agremiações. De fato, não foram escolhidos textos da TMD que tratem desse tema. Essa é a origem do problema desvelado pelo método no penúltimo gráfico. Ao realizar a análise semeada dos documentos, tentando forçar "imperialismo e poder global" como tópico, o método faz uma análise que não é realista e, na comparação desse tópico com o conjunto de documentos, ele o posiciona distante da base comum de comparação.

Apesar de contar com dois documentos dentre os cinco mais bem posicionados com a TMD, a organização Revolução Brasileira não mostrou destaque significativo em relação aos demais tendo em vista que utiliza os fundamentos dessa teoria para guiar sua praxis política. Ao conferir na lista de elementos léxicos mais relevantes em cada categoria, quantos desses elementos léxicos aparecem nos dois documentos da RB, obtêm-se os seguintes percentuais, por categoria:

Integração ao mercado mundial	Segredo da troca desigual	Superexploração do trabalho	Ciclo do capital na economia dependente	Processo de industrialização	Novo anel da espiral	Imperialismo e poder global
49%	49%	51%	63%	26%	14%	54%

Levando em conta a base teórica escolhida para atuar como elemento indutor da análise

dos documentos das agremiações, concluímos que a TMD está fracamente perceptível nesses documentos da RB. Outra base teórica – outras obras e/ou autores da TMD – poderia ter sido escolhida com, eventualmente, melhor resultado.

6 CONCLUSÃO

A metodologia apresentada procurou avançar no desenvolvimento da análise automatizada da atuação de partidos políticos sob um determinado aspecto ideológico. A trajetória começou com a determinação da linha ideológica e a escolha dos textos teóricos que vão fundamentar a linha. A partir do conhecimento dessa linha teórica, se estabeleceram os temas que formam os pilares, as categorias, dessa teoria. Os textos teóricos foram então fragmentados e alocados dentro das categorias com o LDA Semeado. Depois, os textos teóricos categorizados foram formar a base para a análise dos documentos das agremiações. Essa análise conjunta gerou uma medida de similaridade entre os documentos que diz o quanto há de comum entre os grupos de texto.

Há que deixar claro que não se está a buscar a base teórica que inspirou os documentos, a busca visa encontrar elementos comuns de fundo ideológico nos documentos e numa teoria. Qualquer teoria.

O que a análise automatizada de textos pode levantar é uma distância entre a base teórica inspiradora do documento e a teoria. Se estão próximos significa que há elementos comuns entre a ideologia que baseou o documento da agremiação e uma certa categoria da teoria. Isso está longe de dizer que a teoria baseou o documento. Há, porém, aspectos que se tocam: os fenômenos.

Uma teoria pretende explicar fenômenos. Não é de surpreender, portanto, que nos documentos de uma agremiação que estiver fazendo uma análise política, apareça ali alguma teoria. Ela, a agremiação, precisa de elementos teóricos para fazer uma análise. Se aquela teoria é um guia para a agremiação explicar os fenômenos reais, então os elementos dela podem (devem!) aparecer ali. O que há em comum são os fenômenos. Tanto as agremiações estão fazendo análise e interpretações de fenômenos, quanto a teoria pretende ser uma base para explicação desses fenômenos. Existe então uma possibilidade de existir elementos comuns entre eles.

Há, porém, outros desafios. Se num documento a metodologia aponta uma medida não coerente com ele, a razão teria que ser buscada no início do processo, na escolha das palavras sementes e das categorias. Em outros termos, se essas categorias estão bem caracterizadas, o resultado final tem que ser coerente. Outro foco de discrepância, de afastamento ou aproximação, seria o vocabulário utilizado. Como os documentos partidários não necessariamente vão se pautar, não só pelos temas, como também pelo vocabulário da TMD, há que levar-se em conta esse grau de incerteza na análise. Mas os desafios não param aqui e antes de apontar causas de uma eventual fraqueza do método, outros fatores tem que ser levados em consideração.

No percurso ao longo deste trabalho, a questão levantada sobre o grau de aderência dos partidos políticos de esquerda à teoria marxista da dependência permaneceu em aberto. Algo coerente, porém, foi observado nos primeiros experimentos, onde similaridades eram detectadas ainda sem o ferramental para medição de distâncias de proximidade dos últimos experimentos. Foram esses os casos de similaridades entre agremiações com manifestos de forte conteúdo classista, mesmo estando bastante espaçados no tempo (POLOP de 1967 com PCR de 2008, PCO de 1995 com RB de 2017). O resultado reforça a convicção que, pelo menos dentro do espaço de análise de manifestações político-partidárias no

tocante ao tema da conquista e manutenção do Poder, a mudança evolutiva da linguagem ou vocabulário político não se alterou significativamente; claro que isso não ocorre com outros temas. Outros casos, em outros experimentos, confirmam que o método detectou aproximações de cunho classista, mas isso não seria uma surpresa visto que foi induzido por uma teoria marxista. Nesse sentido, o método acertou, mas esse não era o nosso objetivo específico. Isso era até, de certa forma, esperado. Supõe-se que partidos de esquerda, em suas análises de conjuntura política da sociedade capitalista, recorram ao marxismo para uma análise das classes em conflito dessa sociedade.

O escopo da análise, no entanto, não era a afinidade com o marxismo, essa "insuperável filosofia de nosso tempo" como dizia Sartre. O escopo era levantar o grau de proximidade ideológica entre organizações da esquerda brasileira em relação aos temas tratados pela TMD, uma teoria marxista aplicada à realidade latinoamericana ou, mais genericamente, à realidade periférica em relação aos países capitalistas centrais.

Geralmente, métodos estatísticos de PLN fazem uma análise puramente léxica, sem olhar o significado das palavras, mas contam com uma hipótese fundamental: a existência de uma correlação entre o léxico e o semântico. Essa hipótese tem que existir, tem que se confirmar. A hipótese de que, captando o padrão de ocorrência de palavras, esse padrão tenha uma correlação com o significado delas. Então, captando esse padrão, consegue-se extrair por correlação alguma semântica.

Se o objetivo da dissertação tivesse sido alcançado, poderíamos afirmar que a hipótese da correlação léxico-semântica dentro de um método induzido pela TMD funcionou.

Porém, se o método não alcançou o resultado esperado, duas possibilidades de explicação se abrem. A primeira é obvia: houve uma falha na correlação léxico-semântica, o que compromete o método. A segunda possibilidade é que não houve falha nessa correlação, mas, simplesmente, que o léxico é outro, o que implicaria numa correlação diferente da esperada. Embora a agremiação tenha utilizado a mesma semântica (significado da teoria), usou nos documentos outro léxico, outro vocabulário para representar a mesma semântica. A teoria é a mesma, mas externada com outras palavras nos documentos. Isso é um problema porque o método aqui descrito não olha a semântica. O método olha o léxico esperando uma correlação com a semântica. Se o léxico entre o indutor e o induzido for diferente, não há como ter correlação, mas isso não compromete o método.

Outra possibilidade, bastante plausível, seria que a teoria não estivesse contemplada nos documentos. Não há como detectar o que não existe. Isso também não comprometeria o método. Essa trágica possibilidade da ausência de uma teoria a guiar uma práxis desvelaria o pântano que a esquerda brasileira hoje está imitida.

Levando em conta a alta probabilidade de ocorrência dessa última possibilidade foi levantado um foco sobre a organização "Revolução Brasileira," pois é fato que a RB pauta sua análise política sob a influência da TMD. Esperava-se então que essa teoria, supostamente presente nos documentos da organização, fosse perceptível pelo método. A TMD, porém, não está claramente perceptível nos documentos da RB selecionados, há outra terminologia nesses documentos. O texto fundante da TMD é um pequeno ensaio de denso conteúdo marxista, com um linguajar que provavelmente não seria usado numa comunicação externa eficaz da posição de um partido. Não é uma questão nova, o problema da comunicação da vanguarda com a classe. Em analogia com o campo da ciência da computação, faltaria aqui uma compilação, a transformação de uma linguagem fonte

numa linguagem alvo equivalente. Nessa analogia, a vanguarda iria atender a etapa de análise do código fonte e a retaguarda com a síntese do código fonte em linguagem alvo. Não houve preocupação com uma eventual mediação, nem se isso seria possível. A título de não interferir no conteúdo do corpus, foi tomado o léxico do ensaio em "estado bruto" como elemento indutor. Mas poderia ter sido feito de outra forma com sucesso?

Apesar do resultado alcançado, que não comprometem o método, a linha de pesquisa do trabalho revelou-se bastante promissora nas metodologias desenvolvidas. A aplicação em sequência de dois métodos de análise mostrou um potencial caminho de desenvolvimento. A indução de palavras sementes de uma teoria no LDA Semeado resultou numa distribuição de elementos léxicos que foi usada no STM para fazer a análise, determinar os tópicos e sua distribuição nos documentos. Vimos depois como o STM pode incorporar de forma visível várias formas de informação em nível de documento. Cada grupo de documentos foi posicionado então em relação à mesma referência teórica para efeito de comparação. Ao longo dos experimentos e dos ajustes no método outras possibilidades foram detectadas porém não desenvolvidas. É o caso da possibilidade de extração de sementes, não apenas das obras teóricas, mas também dos documentos das agremiações que estão sendo analisados. Em outras palavras, detectar e trazer influências dos documentos políticos para dentro da análise junto com as obras teóricas. Isso abre uma nova e interessante linha de investigação.

Esperamos que os experimentos encorajem outros analistas a seguir com a linha aberta nesta dissertação; em especial, de encontrar elementos comuns da teoria marxista da dependência nos manifestos de agremiações de esquerda em nosso país. A ausência desses elementos aponta para a urgente necessidade da retomada do pensamento marxista como instrumento de conhecimento e transformação social. O marxismo, definitivamente, não perdeu validade no século XXI.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AITCHISON, J. The Statistical Analysis of Compositional Data. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 44, n. 2, p. 139–177, 1982.

AITCHISON, J.; SHEN, S. M. Logistic-Normal Distributions: Some Properties and Uses. **Biometrika**, v. 67, n. 2, p. 261–272, 1980.

ALBRIGHT, J. J. The Multidimensional Nature of Party Competition. **Party Politics**, v. 16, n. 6, p. 699–719, 2010.

BAMBIRRA, V. **Os Programas dos Partidos Políticos no Brasil**. [s.l.] Assembleia Legislativa do Estado do Rio Grande do Sul, 1981.

BLEI, D. M.; LAFFERTY, J. D. A correlated topic model of Science. **The Annals of Applied Statistics**, v. 1, n. 1, jun. 2007.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. **The Journal of Machine Learning Research**, v. 3, p. 993–1022, mar. 2003.

BOBBIO, N. **Left and Right - The Significance of a Political Distinction**. [s.l.] University of Chicago Press, 1996.

BUDGE, I.; MEYER, T. M. Understanding and Validating the Left-Right Scale (RILE). Em: VOLKENS, A. (Ed.). **Mapping policy preferences from texts**. Oxford: Oxford University Press, 2013. v. 3p. 85–106.

BURNHAM, K. P.; ANDERSON, D. R. **Model Selection and Multi-Model Inference**. [s.l.] Springer, 2002.

COCHRANE, C. The asymmetrical structure of left/right disagreement: Left-wing coherence and right-wing fragmentation in comparative party policy. **Party Politics**, v. 19, n. 1, p. 104–121, 2011.

COSTA, G.; ORTALE, R. Jointly modeling and simultaneously discovering topics and clusters in text corpora using word vectors. **Information sciences**, v. 563, p. 226–240, 2021.

DOIG, C. **Introduction to Topic Modeling in Python**PyGotham 2015. **Anais**...Continuum Analytics, 2015Disponível em: <<https://chdoig.github.io/pygotham-topic-modeling/#/>>

ELFF, M. A Dynamic State-Space Model of Coded Political Texts. **Political Analysis**, v. 21, n. 2, p. 217–232, 2013.

FRANZMANN, S. Competition, contest, and cooperation: the analytic framework of the issue market. **Journal of Theoretical Politics**, v. 23, n. 3, p. 317–343, 2011.

FRANZMANN, S.; KAISER, A. Locating Political Parties in Policy Space: A Reanalysis of Party Manifesto Data. **Party Politics**, v. 12, n. 2, p. 163–188, mar. 2006.

GARCIA, A. S.; MARTINS, C. E.; MENEZES, R. G. **Revista do Laboratório de Estudos sobre Hegemonia e Contra-Hegemonia**. [s.l.] IRID/UFRJ e PEPI/UFRJ, 2021. v. 1

GRIMMER, J.; STEWART, B. M. 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.

GROSSMAN, E.; GUINAUDEAU, I. **Do elections (still) matter?** [s.l.] Oxford University Press, 2021.

JAHN, D. Conceptualizing Left and Right in comparative politics: Towards a deductive approach. **Party Politics**, v. 17, n. 6, p. 745–765, nov. 2011.

JORGE, V. L.; FARIA, A. M. T. DE; SILVA, M. G. DA. Posicionamento dos partidos políticos brasileiros na escala esquerda-direita: dilemas metodológicos e revisão da literatura. **Revista Brasileira de Ciência Política**, v. 33, 2020.

KIM, H.; FORDING, R. C. Voter ideology in Western Democracies, 1946– 1989. **European Journal of Political Research**, v. 33, n. 1, p. 73–97, jan. 1998.

KIM, H.; FORDING, R. C. Government partisanship in Western democracies, 1945–1998. **European Journal of Political Research**, v. 41, n. 2, p. 187–206, 2002.

KLINGEMANN, H.-D. et al. **Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD 1990-2003**. [s.l.] Oxford University Press, 2006.

KÖNIG, T.; MARBACH, M.; OSNABRÜGGE, M. Estimating Party Positions across Countries and Time—A Dynamic Latent Variable Model for Manifesto Data. **Political Analysis**, v. 21, n. 4, p. 468–491, 2017.

KRIPPENDORFF, K. **Content Analysis: An Introduction to Its Methodology**. [s.l.] SAGE Publishing, 2018.

LOWE, W. et al. Scaling Policy Preferences from Coded Political Texts. **Legislative Studies Quarterly**, v. 36, n. 1, p. 123–155, 2011.

LU, B. et al. **Multi-Aspect Sentiment Analysis with Topic Models**: ICDMW '11.USA: IEEE Computer Society, 2011

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. [s.l.] The MIT Press, 1999.

MARINI, R. M. Dialética da Dependência. **Germinal: marxismo e educação em debate**, v. 9, n. 3, p. 325–356, 2017.

MÖLDER, M. The validity of the RILE left–right index as a measure of party policy. **Party Politics**, v. 22, n. 1, p. 37–48, 2016.

PAULO FALEIROS, T. DE; ANDRADE LOPES, A. DE. **Modelos Probabilísticos de Tópicos: Desvendando o Latent Dirichlet Allocation**. [s.l.] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

PROSSER, C. Building policy scales from manifesto data: A referential content validity approach. **Electoral Studies**, v. 35, p. 88–101, 2014.

ROBERTS, M. E. et al. Structural Topic Models for Open-Ended Survey Responses. **American Journal of Political Science**, v. 58, n. 4, p. 1064–1082, out. 2014.

ROBERTS, M. E.; STEWART, B. M.; AIROLDI, E. M. A Model of Text for Experimen-

tation in the Social Sciences. **Journal of the American Statistical Association**, v. 111, n. 515, p. 988–1003, jul. 2016.

SALLES, N. Ideologia e Partidos no Brasil: reflexão e prática a partir dos programas de governo. **Revista Política Hoje**, v. 0, n. Early View, 2021.

SAVAGE, L. M. Who gets in? Ideology and government membership in Central and Eastern Europe. **Party Politics**, 2012.

TAROUCO, G. DA S.; MADEIRA, R. M. Partidos, programas e o debate sobre esquerda e direita no Brasil. **Revista de Sociologia e Política**, v. 21, n. 45, 2013.

TOMAR, A. **Topic modeling using Latent Dirichlet Allocation(LDA) and Gibbs Sampling explained!** [s.l.] Analytics Vidhya, 2018. Disponível em: <<https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d10

VOLKENS, A. et al. From Data to Inference and Back Again: Perspectives From Content Analysis. Em: VOLKENS, A. et al. (Eds.). **Mapping Policy Preferences From Texts: Statistical Solutions for Manifesto Analysts**. [s.l.] Oxford University Press, 2013.

WATANABE, K.; ZHOU, Y. Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches. **Social Science Computer Review**, p. 1–21, 2020.

WERNER, A. et al. **Manifesto Coding Instructions: 5th fully revised edition**. [s.l.] Manifesto Project (MRG/CMP/MARPOR), 2021.

8 APÊNDICE

APÊNDICE A – DOCUMENTOS DAS AGREMIAÇÕES POLÍTICAS

Tabela 8.1: Identificação dos documentos das agremiações políticas

Identificação	Título	Organização, Tendência ou Corrente
1964PTB01	1964 Discurso de João Goulart no Comício da Central do Brasil - PTB	Partido Trabalhista Brasileiro
1967POLOP01	1967 Programa Socialista para o Brasil - POLOP	Organização Revolucionária Marxista Política Operária
1979PT01	1979 Carta de Princípios do Partido dos Trabalhadores - PT	Partido dos Trabalhadores
1980PDT01	1980 Manifesto PDT	Partido Democrático Trabalhista
1980PDT02	1980 Programa PDT	Partido Democrático Trabalhista
1980PT01	1980 Manifesto de Fundação do Partido dos Trabalhadores - PT	Partido dos Trabalhadores
1983PDT01	1983 Carta de Mendes - PDT	Partido Democrático Trabalhista
1987PT01	1987 Resoluções Politicas - V Encontro - PT	Partido dos Trabalhadores
1995PCO01	1995 Programa do PCO	Partido da Causa Operária
1995PSB01	1995 Manifesto PSB	Partido Socialista Brasileiro
1995PSB02	1995 Programa PSB	Partido Socialista Brasileiro
2002PT01	2002 Carta ao povo brasileiro - Lula	Partido dos Trabalhadores
2003PCR01	2003 O PCR e a Revolução Brasileira - PCR	Partido Comunista Revolucionário
2003PCR02	2003 Programa da Revolução Brasileira - PCR	Partido Comunista Revolucionário
2004PSOL01	2004 Programa PSOL	Partido Socialismo e Liberdade
2008PCR01	2008 O caráter socialista da revolução - PCR	Partido Comunista Revolucionário
2009PCdB01	2009 Programa PCdoB	Partido Comunista do Brasil
2010PCB01	Programa anticapitalista e antiimperialista para o Brasil	Partido Comunista Brasileiro
2010PCO01	Diretrizes para o programa eleitoral do Partido da Causa Operária	Partido da Causa Operária
2010PSOL01	Uma alternativa socialista: nossas tarefas e diretrizes	Partido Socialismo e Liberdade
2010PSTU01	O Brasil precisa de uma segunda independência	Partido Socialista dos Trabalhadores Unificado
2010PT01	Diretrizes do Programa 2011-2014	Partido dos Trabalhadores
2011PT01	2011 Manifesto da Esquerda Popular Socialista - PT	Partido dos Trabalhadores
2014PCB01	Construindo o Poder Popular por um Brasil Socialista	Partido Comunista Brasileiro
2014PCO01	Programa do PCO para as eleições 2014	Partido da Causa Operária
2014PSB01	Programa de Governo - PSB	Partido Socialista Brasileiro
2014PSOL01	Diretrizes Gerais para Programa de Governo nas Eleições de 2014	Partido Socialismo e Liberdade
2014PSTU01	16 Propostas para construir um Brasil para os trabalhadores	Partido Socialista dos Trabalhadores Unificado
2014PT01	Programa de Governo - PT	Partido dos Trabalhadores
2016PT01	2016 Manifesto EPS e Novo Rumo ao VI Congresso do PT	Partido dos Trabalhadores
2017PSB01	2017 Plano Estratégico de Desenvolvimento Nacional - PSB	Partido Socialista Brasileiro

2017PT01	2017 Tese Avante Militância Socialista ao VI Congresso Nacional do PT	Avante S21 e Militância Socialista
2017PT02	2017 Tese da Mensagem ao Partido - VI Congresso PT	Mensagem ao Partido
2017RB01	2017 Manifesto pela Revolução Brasileira - RB	Revolução Brasileira
2018PDT01	2018 Diretrizes Para Uma Estrategia Nacional de Desenvolvimento Para o Brasil - PDT	Partido Democrático Trabalhista
2018PPL01	2018 Distribuir a Renda Superar a Crise e Desenvolver o Brasil - PPL	Partido Pátria Livre
2018PSOL01	2018 Propostas de Governo PSOL	Partido Socialismo e Liberdade
2018PSTU01	2018 16 Pontos de um Programa Socialista para o Brasil contra a Crise - PSTU	Partido Socialista dos Trabalhadores Unificado
2018PT01	2018 Plano LULA de Governo - PT	Partido dos Trabalhadores
2019PSB01	2019 O Desafio - A Autorreforma do PSB e o Projeto Civilizatório para o Brasil - PSB	Partido Socialista Brasileiro
2019PT01	2019 Tese “Na Luta, Ruas e Redes – LulaLivre” Chapa Nacional - PT	Representantes: Henrique Donin, Lourival Casula e Ricardo Hott Junior
2019PT02	2019 Tese “Lula Livre Para Mudar o Brasil!” - PT	Representantes: Gleide Andrade, Francisco Rocha e Mônica Valente
2019PT03	2019 Tese “Lula Livre: Resistência Socialista!” - PT	Representantes: Paulo Teixeira, Paulo Pimenta e Camila Moreno
2019PT04	2019 Tese “Repensar o PT para enfrentar o retrocesso, defender a democracia e os direitos do povo” - PT	Representantes: Jacy Afonso, Ricardo Berzoini e Letícia Espíndola
2019PT05	2019 Tese “Diálogo e Ação Petista - 7º Congresso” - PT	Representantes: Markus Sokol, Luiz Eduardo Greenhalgh e Misa Boito
2019PT06	2019 Tese “Em tempos de guerra, a esperança é vermelha“ - PT	Representantes: Natália de Sena, Valter Pomar e Patrick Campos Araújo
2019PT07	2019 Tese “Fora Bolsonaro - DS - VII Congresso” - PT	Representantes: Carlos Árabe, Renato Simões e Vilson Oliveira
2019PT08	2019 Tese “Partido é para todos e todas” Chapa Lula Livre – PT	Representantes: Romênio Pereira, Marcos Lemos e Saulo Dias
2020PSB01	2020 A Autoreforma do PSB - PSB	Partido Socialista Brasileiro
2020PSOL01	2020 VII Congresso PSOL Tese Comuna - PSOL	Comuna e militantes independentes
2020PSOL02	2020 VII Congresso PSOL Tese Primavera Socialista - PSOL	Primavera Socialista, Coletivo Lutas, A Esquerda e militantes independentes
2020PSOL03	2020 VII Congresso PSOL Tese Coletivo 1o de Maio - PSOL	Coletivo 1º de Maio
2020PSOL04	2020 VII Congresso PSOL Tese Fortalecer o PSOL - Brigadas Populares - PSOL	Fortalecer o PSOL, Brigadas Populares, Coletivo Travessia, Coletivo Kaàweé, Nova Práxis, Trampolim Socialista/Parnamirim, Avança PSOL e Movimento 9 de Maio.
2020PSOL06	2020 VII Congresso PSOL Tese Construção Socialista e Independentes - PSOL	Construção Socialista e independentes
2020PSOL07	2020 VII Congresso PSOL Tese Bloco Esquerda Radical - PSOL	Bloco Esquerda Radical
2020PSOL08	2020 VII Congresso PSOL Tese Raiz Popular - PSOL	Raiz Popular
2020PSOL09	2020 VII Congresso PSOL Tese MES-TLS-Barulho - PSOL	MES, TLS, Barulho, “Construção pela base”, “Coletivo Direito para quem” e independentes

2020PSOL10	2020 VII Congresso PSOL Tese Insurgencia - LSR - Resistencia - Subverta - C.Portinho - 4_Novembro e Independentes - PSOL	Insurgência, Coletivo Carmen Portinho, LSR, Resistência, Subverta, Coletivo 4 de Novembro e militantes independentes
2020PSOL11	2020 VII Congresso PSOL Tese APS, PSOL da Resistência e Independentes - PSOL	Ação Popular Socialista, PSOL da Resistência e independentes
2020PSOL12	2020 VII Congresso PSOL Tese Coletivo Alicerce - PSOL	Coletivo Alicerce
2020PT01	2020 Texto de Emenda “Sobre a conjuntura política imediata” - Tendências CNB e MPT ao DN - PT	CNB e MPT
2020PT02	2020 Plano de Reconstrução do Brasil - PT	Partido dos Trabalhadores
2020PT03	2020 Carta PT-DF ao DN - Fora Bolsonaro - PT	PT-DF
2020PT04	2020 Texto de Contribuição “PT na Luta pela Vida e pela Dignidade” - PT	Movimento PT
2020PT05	2020 Resolução da AE ao DN do PT - PT	Articulação de Esquerda
2020PT06	2020 Projeto de Emendas ao Texto-base Rui Falcão - Reunião DN Abril 2020 - PT	Rui Falcão e outros
2020PT07	2020 Proposta CNB ao Texto Conjuntura DN Abril 2020 - PT	CNB
2020PT08	2020 Proposta de Texto - Reunião do DN - Secretaria Geral - Abril 2020 - PT	DN-PT
2020PT09	2020 Proposta de Texto - Tendência Avante para a Reunião DN Abril 2020 - PT	Avante
2020PT10	2020 Proposta de Texto - Tendência DS para o DN Abril 2020 - PT	Democracia Socialista
2020PT11	2020 Proposta de Texto - Tendência DAP Para a Reunião DN Abril 2020 - PT	Diálogo e Ação Petista
2020PT12	2020 Proposta de Texto - Tendência ESQUERDA POPULAR SOCIALISTA para a Reunião DN Abril 2020- PT	Esquerda Popular Socialista
2020PT13	2020 Proposta de Texto - Coletivo PT de Todas as Lutas para Reunião DN Abril 2020 -PT	Coletivo PT de Todas as Lutas
2020RB01	2020 Contribuição da RB ao 7º Congresso Nacional do PSOL - RB	Revolução Brasileira
2021PCdB01	2021 Comemoração 99 anos - PCdoB indispensável à democracia - PCdoB	Partido Comunista do Brasil

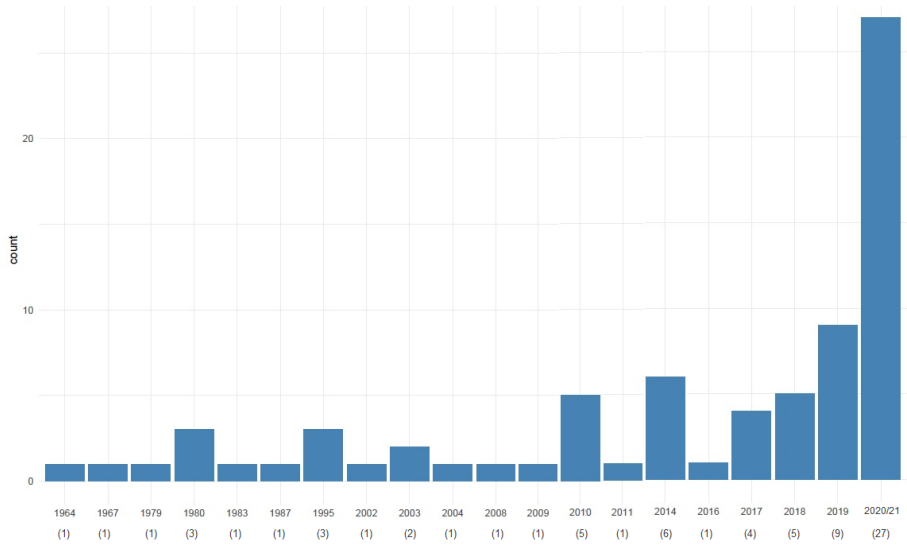


Figura 8.1: Documentos de agremiações políticas por ano de publicação.

it	documento	ano	sigla	evento
1	1964PTB01	1964	PTB	**Manifesto**
2	1967POLOP01	1967	POLOP	**Programa Partidário**
3	1979PT01	1979	PT	**Manifesto**
4	1980PDT01	1980	PDT	**Manifesto**
5	1980PDT02	1980	PDT	**Programa Partidário**
6	1980PT01	1980	PT	**Manifesto**
7	1983PDT01	1983	PDT	**Manifesto**
8	1987PT01	1987	PT	**V Encontro do PT**
9	1995PCO01	1995	PCO	**Programa Partidário**
10	1995PSB01	1995	PSB	**Manifesto**
11	1995PSB02	1995	PSB	**Programa Partidário**
12	2002PT01	2002	PT	**Manifesto**
13	2003PCR01	2003	PCR	**Manifesto**
14	2003PCR02	2003	PCR	**Programa Partidário**
15	2004PSOL01	2004	PSOL	**Programa Partidário**
16	2008PCR01	2008	PCR	**Manifesto**
17	2009PCdB01	2009	PCdB	**Programa Partidário**
18	2010PCB01	2010	PCB	**Programas Eleitorais (2010)**
19	2010PCO01	2010	PCO	
20	2010PSOL01	2010	PSOL	
21	2010PSTU01	2010	PSTU	
22	2010PT01	2010	PT	
23	2011PT01	2011	PT	**Manifesto**
24	2014PCB01	2014	PCB	**Programas Eleitorais (2014)**
25	2014PCO01	2014	PCO	
26	2014PSB01	2014	PSB	
27	2014PSOL01	2014	PSOL	
28	2014PSTU01	2014	PSTU	
29	2014PT01	2014	PT	
30	2016PT01	2016	PT	**VI Congresso do PT**
31	2017PSB01	2017	PSB	**Manifesto**
32	2017PT01	2017	PT	**VI Congresso do PT**
33	2017PT02	2017	PT	
34	2017RB01	2017	RB	**Manifesto**
35	2018PDT01	2018	PDT	**Programas Eleitorais (2018)**
36	2018PPL01	2018	PPL	
37	2018PSOL01	2018	PSOL	
38	2018PSTU01	2018	PSTU	
39	2018PT01	2018	PT	
40	2019PSB01	2019	PSB	**Manifesto**
41	2019PT01	2019	PT	**VII Congresso do PT**
42	2019PT02	2019	PT	
43	2019PT03	2019	PT	
44	2019PT04	2019	PT	
45	2019PT05	2019	PT	
46	2019PT06	2019	PT	
47	2019PT07	2019	PT	
48	2019PT08	2019	PT	
49	2020PSB01	2020	PSB	**Manifesto**
50	2020PSOL01	2020	PSOL	**VII Congresso do PSOL**
51	2020PSOL02	2020	PSOL	
52	2020PSOL03	2020	PSOL	
53	2020PSOL04	2020	PSOL	
54	2020PSOL06	2020	PSOL	
55	2020PSOL07	2020	PSOL	
56	2020PSOL08	2020	PSOL	
57	2020PSOL09	2020	PSOL	
58	2020PSOL010	2020	PSOL	
59	2020PSOL011	2020	PSOL	
60	2020PSOL012	2020	PSOL	
61	2020PT01	2020	PT	**Reunião DN-PT Abril 2020**
62	2020PT02	2020	PT	**Manifesto**
63	2020PT03	2020	PT	**Reunião DN-PT Abril 2020**
64	2020PT04	2020	PT	
65	2020PT05	2020	PT	
66	2020PT06	2020	PT	
67	2020PT07	2020	PT	
68	2020PT08	2020	PT	
69	2020PT09	2020	PT	
70	2020PT010	2020	PT	
71	2020PT011	2020	PT	
72	2020PT012	2020	PT	
73	2020PT013	2020	PT	
74	2020RB01	2020	RB	**Manifesto**
75	2021PCdB	2021	PCdB	**Manifesto**

Figura 8.2: Documentos de agremiações políticas por evento.