CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

UMA ANÁLISE DE PARTIDOS POLÍTICOS BASEADA EM DISCURSOS NO CONGRESSO NACIONAL BRASILEIRO

Willian Pitter Cardoso Lima

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de Mestre em Ciência da Computação.

Orientadores:
Laura Silva de Assis
Douglas O. Cardoso

Rio de Janeiro,

17 de Janeiro de 2024

Uma Análise de Partidos Políticos Baseada em Discursos no Congresso Nacional Brasileiro

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Willian Pitter Cardoso Lima

Aprovada por:

Prof.Laura Silva de Assis, D.Sc. (orientadora)

Prof. Douglas O. Cardoso, D.Sc. (coorientador)

Prof. Eduardo Bezerra da Silva, D.Sc.

Prof. Rafael Lima de Carvalho, D.Sc.

Rio de Janeiro,

17 de Janeiro de 2024

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

L732 Lima, Willian Pitter Cardoso
Uma análise de partidos políticos baseada em discursos no Congresso Nacional Brasileiro / Willian Pitter Cardoso Lima. — 2024.
49f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2024.
Bibliografia : f. 44-49
Orientadora: Laura Silva de Assis
Coorientador: Douglas O. Cardoso

1. Aprendizado de Máquina. 2. Processamento de linguagem natural (Computação). 3. Ciência política. 4. Redes sociais on-line. I. Assis, Laura Silva de. (Orient.). II. Cardoso, Douglas O. (Coorient.). III. Título.

CDD 006.31

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

RESUMO

Uma Análise de Partidos Políticos Baseada em Discursos no Congresso Nacional Brasileiro

Willian Pitter Cardoso Lima

Orientadores:
Laura Silva de Assis
Douglas O. Cardoso

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Discursar é parte intrínseca do trabalho dos parlamentares, onde eles expõem fatos, pontos de vista e opiniões sobre assuntos diversos. Este trabalho tem como objetivo analisar as relações entre parlamentares de acordo com os discursos proferidos por membros da Câmara dos Deputados do Brasil. O período considerado no presente estudo compreende o mandato entre 2011 e 2015. Para atingir esse objetivo, a metodologia proposta baseada em técnicas de Processamento de Linguagem Natural, *Term Frequency–Inverse Document Frequency* e *Universal Sentence Encoder* foi utilizada com intuito de avaliar as relações pareadas entre congressistas, as quais foram analisadas sob a ótica de Redes Complexas. Neste trabalho, para a representação do problema em estudo, foi construído um grafo completo em que cada nó representa um deputado, e os pesos associados às arestas que conectam estes nós representam as semelhanças entre os seus posicionamentos políticos, obtidos a partir de seus respectivos discursos. O agrupamento de nós foi utilizado para avaliar múltiplas medidas de distância baseadas nos discursos entre cada par de congressistas, bem como a coesão resultante de seus partidos políticos. Os resultados experimentais mostraram que uma das medidas propostas neste trabalho, que é baseada na agregação de semelhanças entre cada par de discursos, se mostrou superior a uma alternativa previamente estabelecida na literatura, a qual considera concatenações dos discursos relativos a cada indivíduo com o objetivo de agrupar os parlamentares organicamente.

Palavras-chave:
Aprendizado de Máquina, Processamento de Linguagem Natural, Política, Redes Sociais.

Rio de Janeiro,

17 de Janeiro de 2024

ABSTRACT

Uma Análise de Partidos Políticos Baseada em Discursos no Congresso Nacional Brasileiro

Willian Pitter Cardoso Lima

Advisors:
Laura Silva de Assis
Douglas O. Cardoso

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Speeching is an intrinsic part of the work of congressmen, as they expose facts as well as their points of view and opinions on several subjects. This paper aims to analyze relations between congressmen according to the speeches given by members of the lower house of the national congress of Brazil. The period considered in this study comprises the mandate between 2011 and 2015. In order to accomplish this goal, the proposed methodology based on Natural Language Processing, Term Frequency–Inverse Document Frequency and Universal Sentence Encoder were used to assess pairwise relationships between congressmen which were then analyzed from the perspective of Complex Networks. In this work, to represent the problem under study, a complete graph was constructed in which each node represents a deputy, and the weights associated with the edges that connect these nodes represent the similarities between their political positions. Node clustering was used to evaluate multiple speech-based measures of distance between each pair of congressmen, as well as the resulting cohesion of their political parties. Experimental results showed that one of the proposed measures, based on aggregating similarities between each pair of speeches, is superior to a previously established alternative of considering concatenations of these elements relative to each individual when targeting to group parliamentarians organically.

Key-words:
Machine Learning, Natural Language Processing, Politics, Social Networks.

Rio de Janeiro,

17 de Janeiro de 2024

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

BOW	Bag-of-words	17
CNN	Convolutional Neural Network	7
IR	Information Retrieval	5
LSTM	Long Short-Term Memory	17
MLL	Modelo Largo De Linguagem	31
NLTK	Natural Language Toolkit	5, 22
PLN	Processamento De Linguagem Natural	2, 3, 4, 5, 6, 14, 16, 22, 26
PMDB	Partido Do Movimento Democrático Brasileiro	38
PP	Progressistas	38
PSB	Partido Socialista Brasileiro	38
PSDB	Partido Da Social Democracia Brasileira	38
PSOL	Partido Socialismo E Liberdade	38
PT	Partido Dos Trabalhadores	38
PTDOB	Partido Trabalhista Do Brasil	38
TF-IDF	Term Frequency–Inverse Document Frequency	2, 6, 7, 14, 22, 24, 26, 27, 28, 31, 32, 33, 34, 35, 39, 42
UAD	User-as-Document	22, 23, 24, 30
USE	Universal Sentence Encoder	2, 7, 17, 24, 26, 30, 31, 32, 33, 34, 35, 36, 39, 40, 41, 42

Sumário

Capítulo I Introdução

A Câmara dos Deputados, uma das casas integrantes do Congresso Nacional do Brasil, é composta por 513 deputados federais eleitos por voto popular para um mandato de quatro anos. A principal função da Câmara dos Deputados é a elaboração de leis, que são discutidas, emendadas e votadas pelos deputados. A Câmara também tem a responsabilidade de fiscalizar as ações do Poder Executivo, bem como de julgar as contas do Presidente da República. No cenário político brasileiro, a Câmara dos Deputados é uma instituição de grande importância, já que é responsável por definir as políticas públicas e legislar sobre temas fundamentais para a sociedade, como saúde, educação, segurança, meio ambiente, dentre outros.

Os discursos dos deputados da Câmara dos Deputados proferidos durante as sessões na Câmara podem representar diferentes pontos de vista, interesses e ideologias, dependendo da posição política e das convicções pessoais de cada parlamentar. Em geral, os deputados usam seus discursos para expor suas opiniões sobre os temas em discussão e defender suas propostas e projetos de lei. Além disso, os discursos dos deputados também são capazes de refletir as posições dos partidos políticos a que pertencem. Os partidos têm orientações ideológicas e programáticas que guiam suas ações e decisões, e os discursos dos deputados podem expressar essas posições [Van Dijk, 2011]. Assim, os discursos tem o potencial de representar a visão geral do partido sobre questões políticas, econômicas, sociais, ambientais e culturais.

Redes complexas são um campo interdisciplinar de estudo que se concentra na análise de estruturas complexas, representadas por um conjunto de elementos interconectados. Esses elementos podem ser pessoas, empresas, moléculas, neurônios, proteínas, páginas da Web, entre outros [Dias et al., 2019; Alanis et al., 2021]. Uma rede pode ser definida como um grafo no qual há um conjunto de nós representando os objetos em estudo, e um conjunto de arestas que conectam esses nós. As arestas representam uma relação existente entre dois nós de acordo com o contexto em que estes estão inseridos [Metz et al., 2007].

Utilizando os discursos proferidos nos debates da Câmara, este trabalho propõe uma metodologia com o objetivo de analisar as relações entre deputados brasileiros de acordo com seus discursos. Caracterizar as interações entre os parlamentares é uma atividade interessante para a democracia, pois fornece evidências aos eleitores sobre potenciais associações e conflitos de interesses. Para isso o conceito de redes complexas, assim como suas técnicas, podem ser utilizados de forma eficaz [Ba-

rabási and Pósfai, 2016].

Neste trabalho foi utilizado um modelo de rede complexa para analisar a similaridade entre os posicionamentos políticos de deputados. Para isso, foi considerado um grafo completo, em que cada nó do grafo representa um deputado. As arestas do grafo são usadas para conectar os nós e são ponderadas de acordo com as semelhanças entre os posicionamentos políticos dos deputados.

Para quantificar a proximidade entre dois parlamentares, o trabalho utiliza e compara duas abordagens, o *Term Frequency–Inverse Document Frequency (TF-IDF)* e o USE. O TF-IDF é uma técnica de Processamento de Linguagem Natural (PLN) usada para avaliar a importância de uma palavra em um documento, com base em sua frequência e no número de documentos em que aparece. Nesse caso, os vetores TF-IDF são gerados a partir dos discursos dos deputados, e a distância de cosseno é usada para medir a similaridade entre esses vetores. A distância de cosseno é uma medida de similaridade que mede o ângulo entre dois vetores em um espaço n-dimensional. Quanto menor o ângulo, maior a similaridade entre os vetores. Assim, essa abordagem permite avaliar a proximidade entre dois deputados com base em seus discursos [Dezembro, 2019]. A segunda abordagem utiliza o algoritmo *Universal Sentence Encoder* Multilinguagem [Yang et al., 2019], dado que estes modelos multilinguagem são treinados de forma robusta, e treinados exclusivamente na língua portuguesa, os mesmos são treinados de forma menos robusta e são raramente encontrados na literatura.

O USE, por sua vez, é um modelo de aprendizado profundo que pertence à classe dos codificadores de sentenças universais. Ele foi desenvolvido com o objetivo de mapear frases ou sentenças de texto para representações numéricas de alta dimensionalidade e densas, também conhecidas como vetores de *embeddings*. Essas representações capturam informações semânticas e contextuais das sentenças, permitindo uma compreensão mais rica e abrangente do significado dos textos. Os modelos USE recebem como entrada sentenças e produzem como saída um *embedding* dimensional com a representação da sentença. Através da utilização desses *embeddings*, torna-se viável efetuar uma comparação da similaridade semântica existente entre dois discursos distintos. Aferindo-se a similaridade, a mesma é quantificada em uma escala contida no intervalo de 0 a 1, na qual valores mais próximos de 1 indicam uma maior semelhança entre os discursos analisados. Essa abordagem fundamenta-se na hipótese de que *embedding* gerados por modelos de aprendizado profundo, como o *Universal Sentence Encoder*, capturam informações semânticas significativas, permitindo a realização de tais comparações de forma eficiente e eficaz.

A partir da criação da rede utilizando os deputados e seus discursos é possível analisar a coesão dos partidos aos quais estes pertencem. A agregação dos deputados representa as afinidades políticas entre si, pois membros de um mesmo grupo ideológico tendem a ter discursos semelhantes [Halberstam and Knight, 2016; Chen et al., 2017; Elejalde et al., 2017; Van Dijk, 2003; Cristiani et al., 2020; Pastor-Galindo et al., 2020; Caetano et al., 2018]. O fato de cada deputado pertencer a

um único partido político nos permite realizar uma comparação mais direta entre os relacionamentos dos deputados, de acordo com seus partidos políticos e seus discursos. Investigar a coesão partidária é então possível pela avaliação de como o alinhamento de discursos dos deputados é distribuído pela rede. Esta informação é importante pois viabiliza uma análise quantitativa da conformidade dos discursos de um deputado com os discursos dos deputados pertencentes ao seu partido assim como, também promove esta análise com deputados de outros partidos. Portanto, possibilitando detectar o alinhamento do discurso do deputado com os partidos existentes. Tais análises podem destacar até mesmo secessões dentro de um mesmo partido. Para além da categorização com equivalente número de partidos originais, é também factível conduzir uma avaliação dos agrupamentos utilizando o método do cotovelo, com o propósito de observar qual seria a quantidade ideal de partidos conforme a métrica da silhueta.

Neste trabalho foi realizado um estudo sobre métodos de PLN para quantificar as arestas de uma rede de deputados, visando analisar tal contexto deste ponto de vista inédito. As principais contribuições desta pesquisa são: *(i)* estabelecer novas métricas de agregação que permitem resumir em uma única medida as distâncias entre os pares de discursos de dois parlamentares quaisquer *(ii)* utilizar agrupamento hierárquico aglomerativo para estimar a qualidade de tais medidas; *(iii)* avaliar a coesão partidária de uma perspectiva de redes complexas a partir das medidas consideradas *(iv)* avaliar uma possível reestruturação na quantidade ideal de partidos de acordo com os agrupamentos gerados.

O restante deste trabalho está organizado da seguinte forma: O Capítulo II examina o arcabouço teórico que fundamenta este trabalho. O Capítulo III apresenta trabalhos relacionados ao problema de medida de similaridade textual. A metodologia proposta é descrita no Capítulo IV. O Capítulo V apresenta os resultados obtidos através dos experimentos computacionais realizados e a respectiva discussão destes. O Capítulo VI realiza as conclusões sobre esta pesquisa.

Capítulo II Referencial Teórico

O presente capítulo tem como objetivo apresentar conceitos relevantes para o entendimento e desenvolvimento deste trabalho, a fim de contextualizar os conceitos em relação às principais perspectivas teóricas e debates existentes neste trabalho. E para isso, as sub-sessões seguintes buscam identificar e analisar os principais conceitos, teorias e métodos presentes na metodologia deste trabalho, que permitam estabelecer uma base sólida para a construção do conhecimento.

O PLN é uma subárea da Inteligência Artificial que estuda a capacidade dos computadores lidarem com informações textuais, buscando resolver problemas relacionados à geração e compreensão automática de dados contendo tais informações [Manning et al., 2008].

Para tornar possível a cognição automática de textos, o PLN combina linguística computacional com modelos estatísticos, permitindo que os computadores processem a linguagem humana, utilizando diversas técnicas de pré-processamento e representação de texto voltadas para computação [Camacho-Collados and Pilehvar, 2017].

Para modelar os dados e possibilitar que um algoritmo de inteligência artificial compreenda melhor o texto e faça melhores associações, são necessários pré-processamentos que abstraiam a estrutura da língua e tornem os dados adequados para análise, idealmente preservando apenas o que é informação relevante. As técnicas de pré-processamento utilizadas neste trabalho são brevemente descritas nas Seções II.1, II.2 e II.3.

II.1 Tokenização

Geralmente textos são representados computacionalmente por uma sequência, potencialmente longa, de caracteres. Para diversos tipos de processamento linguístico é necessário identificar e categorizar individualmente as palavras de um texto.

Determinar o que é uma palavra é um passo importante para a separação do texto em partes individuais. Separar palavras por espaço é algo que parece trivial, porém pode acarretar em erros, tais como a palavra *d'água* que seria entendida como apenas uma palavra.

A tokenização é o processo de quebrar a sequência de caracteres em um texto localizando o limite de cada palavra, ou seja, os pontos onde uma palavra termina e outra começa [Indurkhya and Damerau, 2010]. A tokenização pode ser realizada de várias maneiras, dependendo do objetivo

e do idioma do texto, como tokenização por espaço, por pontuação ou por expressões regulares, no presente trabalho foi utilizado a tokenização por espaço. As palavras assim identificadas são frequentemente chamadas de *tokens*. Para auxiliar e lidar com os detalhes da tokenização, existem bibliotecas para a linguagem Python, como por exemplo *nltk.tokenize*, que recebe o texto como parâmetro e retorna o texto tokenizado.

II.2 Stemming

O *stemming* é o processo de reduzir a inflexão das palavras à sua forma raiz, como mapear um grupo de palavras para a mesma raiz, mesmo que a própria raiz não seja uma palavra válida na língua [Lovins, 1968a]. O *stemming* reduz as palavras flexionadas sem compromisso em garantir que a raiz da palavra pertença ao idioma. Um exemplo de uma possível redução que não leva a raiz original do idioma seria reduzir com *stemming* as palavras *meninas*, *meninos*, *menina*, *menino* para a raiz a*menin*.

O objetivo do *stemming* é reduzir diferentes formas de uma palavra a uma única forma comum, que pode ser usada como representação genérica da palavra. O processo de *stemming* é usado em muitas tarefas de processamento de linguagem natural, como análise de sentimentos, classificação de texto e recuperação de informações. Ele pode ajudar a reduzir a complexidade do texto e melhorar a eficácia de algoritmos de PLN, diminuindo a variação nas palavras usadas e capturando melhor o significado geral do texto.

II.3 Remoção de Stop Words

No processamento de linguagem natural, uma *stop word* é uma palavra cuja frequência no texto é muito elevada, fazendo com que ela não seja uma palavra característica o suficiente para ser relevante para a decisão do algoritmo em classificar as frases [Saif et al., 2014]. Palavras com pouco significado como artigos definidos e indefinidos, e outras palavras tais como “aquele”, “esse” e “disso”, são exemplos de *stop words*.

Não há um padrão de palavras a ser considerado como *stop words*. Diversas listas de *stop words* estão disponíveis, como a lista da biblioteca *Natural Language Toolkit (NLTK)* Bird [2006], e a escolha de uma lista ou outra pode afetar o resultado final, como no cálculo da precisão em métodos de *Information Retrieval (IR)*, conforme mostrado em [Dolamic and Savoy, 2010]. Outra abordagem possível é a de criar listas de *stop words* de acordo com o número de palavras encontradas no texto, como apresentado em [Alajmi et al., 2012].

A remoção de *stop words* tem o objetivo de reduzir o ruído no texto e simplificar a análise do conteúdo significativo. Isso é especialmente importante em tarefas como análise de sentimento, classificação de documentos e reconhecimento e de tópicos, onde o objetivo é identificar os temas

principais e a intenção do texto. Pelo fato destas palavras não agregarem significativamente para o aprendizado do algoritmo, a remoção destas palavras diminui a quantidade de *tokens* que serão analisados futuramente, atenuando o espaço vetorial analisado e aumentando a velocidade de execução do algoritmo de aprendizado.

II.4 Vetorização

No contexto de PLN Camacho-Collados and Pilehvar [2017] mostram que a análise de uma coleção de documentos pode ser beneficiada por uma representação apropriada dos seus *tokens*. Um método comumente utilizado é a geração de uma matriz na qual cada linha é um vetor que se refere a um dos documentos em análise e cada coluna é relativa a um *token*. Um *token* é uma sequência de caracteres contíguos que desempenham um determinado papel em uma linguagem escrita, como palavras ou até mesmo partes destas (Seção II.1). Cada entrada da matriz é calculada utilizando uma medida amplamente empregada chamada TF-IDF [Beel et al., 2016]. Esta é uma medida estatística que tem o intuito de indicar a importância de uma palavra em um documento em relação a uma coleção de documentos ou em um corpus linguístico. O objetivo de utilizar o TF-IDF é reduzir o impacto de *tokens* que ocorrem numa grande maioria ou em pouquíssimos documentos, sendo pouco úteis para diferenciá-los.

O cálculo do TF-IDF para um termo t de um documento d em um conjunto de documentos é dada pela Equação (II.1), sendo que $\mathrm{tf}(t, d)$ é a frequência do termo t no documento d Luhn [1957].

$$\mathrm{tfidf}(t, d) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t). \tag{II.1}$$

O $\mathrm{idf}(t)$ representa a raridade de uma palavra em toda a coleção de documentos é calculado como mostrado na Equação II.2, sendo n o número total de documentos no conjunto de documentos, e $\mathrm{df}(t)$ é a frequência de documentos em que t ocorre. A frequência do documento é o número de documentos no conjunto de dados que contém o termo t Jones [2004].

$$\mathrm{idf}(t) = \log\left(\frac{n}{\mathrm{df}(t)}\right) + 1. \tag{II.2}$$

Para dois elementos de um espaço vetorial qualquer, o seu produto escalar é proporcional ao cosseno do ângulo entre os mesmos. Tal valor pode ser interpretado como uma indicação de semelhança entre esses vetores numa escala que varia, caso estes sejam unitários, de -1 (diametralmente opostos) a 1 (coincidentes). Utilizando a matriz gerada pelo TF-IDF, pode-se usar a similaridade de cosseno para medir a semelhança entre pares de documentos, medida que pode variar de 0 a 1

dado que todas as entradas são não-negativas. Já para o cálculo de distâncias entre documentos, é utilizado o complemento da similaridade. Sendo a e b vetores distintos da matriz TF-IDF a distância entre estes vetores é dada como apresentado na Equação (II.3).

$$d(a, b) = 1 - (\langle a, b \rangle)/(||a|| \cdot ||b||). \tag{II.3}$$

II.5 Universal sentence encoder

Os Sentence Level Embeddings, como o USE, têm como objetivo representar sentenças inteiras de uma vez. Eles são capazes de capturar o significado geral e o contexto de toda a sentença, em vez de apenas focar em palavras isoladas [Cer et al., 2018].

Para gerar esses embeddings, são utilizadas arquiteturas de redes neurais, como as redes neurais utilizando *transformers*, que são pré-treinadas em grandes conjuntos de dados não rotulados [Lin et al., 2017]. O USE é baseado em *transformer* e constrói *embeddings* de frases usando o subgrafo de codificação da arquitetura *transformer* [Vaswani et al., 2017]. Este subgrafo emprega a atenção para computar representações contextualmente sensíveis das palavras em uma sentença, levando em consideração tanto a sequência quanto o significado de todas as outras palavras. As representações contextualmente reconhecidas das palavras são transformadas em um vetor de codificação de frase de comprimento uniforme através da soma elemento por elemento das representações em cada posição da palavra. O codificador aceita como entrada uma sequência de caracteres e produz um vetor de 512 dimensões como a representação incorporada da frase.

Durante o pré-treinamento, o modelo aprende a codificar sentenças de diferentes comprimentos em representações de dimensionalidade fixa. A principal vantagem dos *embeddings* de nível de sentença é que eles podem lidar com sentenças de tamanhos variados e, ao contrário dos *word embeddings*, não requerem uma média ou agregação das representações de palavras individuais. Isso permite uma representação mais abrangente do significado da sentença como um todo.

O *Multilingual Universal Sentence Encoder for Semantic Retrieval* [Yang et al., 2019] utilizado neste trabalho, apresenta dois modelos de codificação de sentenças multilíngues focados na recuperação semântica pré-treinados, respectivamente baseados no *Transformer* e na arquitetura *Convolutional neural network (CNN)*. Os modelos incorporam texto de 16 idiomas em um único espaço semântico utilizando um codificador duplo treinado para várias tarefas que aprende representações vinculadas usando tarefas de tradução. Os modelos oferecem desempenho competitivo com o estado da arte em: recuperação semântica, tradução, e respostas e perguntas.

II.6 Teoria de grafos e Redes Complexas

A teoria dos grafos é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto, tais estruturas são chamadas de grafos ou redes [Bondy and Murty, 1976]. Um grafo é uma estrutura de abstração que auxilia na representação e solução de problemas computacionais, por representarem a relação de interdependência entre elementos de um conjunto. Um grafo $G = (V, E)$ é uma estrutura definida por um conjunto V de nós, que se relacionam através de um conjunto E de arestas. Em grafos ponderados, cada aresta possui um peso, sendo que tal peso é frequentemente referido como o custo da aresta. Um grafo completo com n vértices é um grafo simples $G = (V, E)$, onde cada par de vértices distintos em V é conectado por uma única aresta em E.

Nas aplicações, o peso pode ser uma medida do comprimento de uma rota, a capacidade de uma linha, a energia necessária para se mover entre os locais ao longo de uma rota, etc. Existem diversas características/propriedades em grafos que podem ser destacadas. São salientados a seguir dois conceitos importantes para a metodologia proposta neste trabalho, os quais são apresentados a seguir.

II.6.1 Grafo bipartido

Um grafo bipartido $G = (V, E)$ é um grafo não direcionado em que o conjunto de vértices V pode ser dividido em dois conjuntos disjuntos, V_1 e V_2, sendo $V_1 \cup V_2 = V$ e $V_1 \cap V_2 = \emptyset$. Assim, cada aresta $(i, j) \in V$ conecta um vértice de um dos conjuntos a um vértice do outro conjunto, isto é, $i \in V_1$ e $j \in V_2$. Podemos representar um grafo bipartido como um diagrama de duas colunas, em que os vértices de $V1$ estão na coluna da esquerda e os vértices de $V2$ estão na coluna da direita, e as arestas são linhas que conectam vértices localizados em colunas diferentes. Um grafo bipartido completo é um grafo bipartido para o qual cada vértice do conjunto V_1 possui uma aresta o conectando a cada vértice do conjunto V_2. A Figura IV.4 ilustra um grafo bipartido completo.

II.6.2 Árvore Geradora Mínima

Uma árvore T é um grafo conexo, isto é, existe pelo menos um caminho entre quaisquer dois de seus nós, e acíclico (este caminho deve ser único). Caso o grafo seja acíclico mas não conexo, ele é dito ser uma floresta. Uma floresta também pode ser definida como uma união disjunta de árvores. Todo grafo conexo $G(V, E)$ possui pelo menos uma árvore geradora $T(V', E')$ associada, a qual é composta de todos os seus nós ($V' = V$) e um subconjunto de suas arestas ($E' \subseteq E$). Esta árvore é denominada árvore geradora. Uma árvore geradora $T \subseteq G$ é mínima se não existe uma árvore geradora $T' \subseteq G$ tal que seu custo é menor que o custo de T. Analogamente, uma árvore geradora

$T \subseteq G$ é máxima se não existe alguma outra árvore geradora $T' \subseteq G$ tal que seu custo é maior que o custo de T. O custo de um subgrafo não-direcionado é a soma dos custos associados às arestas que o compõe.

II.6.3 Coeficiente de Agrupamento

No contexto de redes complexas [Wasserman and Faust, 1994], o coeficiente de agrupamento (*clustering coefficient*) mede o grau com que os nós de um grafo tendem a agrupar-se. Em definição, para certo nó u, o coeficiente de agrupamento representa a frequência relativa de triângulos formados por u e seus vizinhos.

Para grafos ponderados, existem várias maneiras de realizar tal avaliação [Saramäki et al., 2007]. Uma delas é baseada na média geométrica dos pesos das arestas de cada triângulo [Onnela et al., 2005], conforme descrito pela Equação (II.4).

$$ca_u = \frac{\sum_{v,z \in N^u} (\hat{w}_{uv}.\hat{w}_{uz}.\hat{w}_{vz})^{1/3}}{deg(u).(deg(u)-1))}, \tag{II.4}$$

onde $deg(u)$ é o grau do nó u, os nós v e z são vizinhos (diretamente conectados pela mesma aresta) de u, logo N^u é a vizinhança deste nó, e $\hat{w}_{xy}$ é o peso da aresta $(x, y) \in E$. Evidências sugerem que na maioria das redes do mundo real, e especialmente em redes sociais, os nós tendem a formar grupos fortemente conexos caracterizados por uma densidade relativamente alta de laços. Esta probabilidade propende a ser maior do que a probabilidade média de um laço ser estabelecido aleatoriamente entre dois nós [Watts and Strogatz, 1998]. A clusterização é uma propriedade muito comum nas redes sociais, referindo-se aos círculos de amigos ou conhecidos onde os seus membros se conhecem, formando, assim, um grupo na rede.

II.7 Agrupamento Aglomerativo

Ainda no contexto da análise da organização de coleções em grupos, além da existência de medidas como o coeficiente de agrupamento, há também métodos que se propõem a realizar o agrupamento de itens em classes de afinidade. Com isso, é possível subdividir textos, indivíduos ou quaisquer entidades. O objetivo dos algoritmos de clusterização é descobrir automaticamente agrupamentos de dados não rotulados, fazendo desta, uma estratégia não-supervisionada. Tais grupos podem ser baseados em uma medida de similaridade tal que as semelhanças entre os indivíduos do mesmo grupo devem ser altas, enquanto as semelhanças entre indivíduos de grupos diferentes devem ser baixas. Um *cluster* ideal terá seu conjunto de pontos denso e isolado dos demais, sinalizando uma boa segregação dos dados.

Há diversas formas de realizar uma clusterização, e uma delas é o agrupamento hierárquico. Os

métodos de agrupamento hierárquico criam uma decomposição de um conjunto de dados na forma de árvore, dividindo-a recursivamente em conjuntos de dados menores. Os métodos hierárquicos, podem ser construídos de duas formas, divisiva ou aglomerativa, e constroem uma estrutura que descreve uma hierarquia de agrupamentos sobre os dados. Um algoritmo hierárquico é, então, a especificação de ações que resultam nessa estrutura. Na abordagem divisiva, o algoritmo inicia com todos os objetos em um único grupo, o qual vai sendo dividido sucessivamente até que cada grupo contenha um único elemento. Neste trabalho foi preferida uma abordagem aglomerativa, na qual inicialmente cada grupo contém um único elemento, e iterativamente, os dois grupos mais similares são concatenados até que, ao final, a quantidade de grupos seja igual a quantidade de partidos.

Para definir o critério de clusterização do agrupamento aglomerativo, é utilizado o hiperparâmetro *linkage*. O critério de *linkage* determina qual distância considerar entre os conjuntos de observação. O algoritmo irá mesclar os pares de *clusters* que minimizam este critério. Os possíveis valores de *linkage* são:

- ***Single linkage***: a similaridade entre dois *clusters* é definida pela menor distância de qualquer ponto de um *cluster* para qualquer ponto de um outro *cluster*.
- ***Average linkage***: a similaridade entre dois *clusters* é definida pela média das distâncias de todos os pontos de um *cluster* em relação aos pontos de um outro *cluster*.
- ***Complete linkage***: a similaridade entre dois *clusters* é definida pela maior distância de qualquer ponto de um *cluster* para qualquer ponto de um outro *cluster*.

II.8 *V-Measure*

O *V-measure* é uma métrica de avaliação utilizada em tarefas de agrupamento, que mede a similaridade entre os rótulos de classe verdadeiros e os rótulos de classe atribuídos pelo algoritmo de agrupamento. Ao organizar os itens de uma coleção em grupos, há dois critérios que podem ser utilizados para avaliação do resultado do agrupamento: *i)* homogeneidade (*homogeneity*, h) e *ii)* completude (*completeness*, c). Ambos dependem do estabelecimento *a priori* de um agrupamento considerado ideal para os dados que estão sendo processados, indicando para cada item sua respectiva classe. Um agrupamento possui um valor máximo de homogeneidade se todos os seus grupos contiverem apenas amostras que são originalmente membros da mesma classe, sem mesclar elementos que deveriam estar separados. Já a completude é máxima se todas as amostras que são membros de uma determinada classe são elementos do mesmo grupo, não tendo sido indevidamente alocadas em grupos distintos.

A homogeneidade e a completude de um agrupamento são de certa forma concorrentes. O agrupamento trivial em que cada item está em um grupo unitário possui máxima homogeneidade

porém mínima completude, enquanto que este cenário se inverte se todos os itens forem colocados em um único grupo. O *V-Measure* mede o sucesso dos critérios de homogeneidade e completude de forma combinada [Rosenberg and Hirschberg, 2007] como mostrado na Equação (II.5)

$$V_\beta = \frac{((1+\beta)hc)}{(\beta h + c)}. \tag{II.5}$$

O *V-Measure* é uma medida calculada como a média harmônica dos valores de homogeneidade e completude, a qual pode ser ajustada para atribuir diferentes pesos às contribuições de homogeneidade ou completude através do parâmetro $\beta \geq 0$. Normalmente é usado um valor *default* igual a 1. Se β estiver no intervalo $[0, 1[$ a homogeneidade tem maior relevância na ponderação. Entretanto se β assumir valores no intervalo $(1, \infty)$ a completude tem maior peso no cálculo. No caso de $\beta = 1$ ocorre o equilíbrio entre os dois critérios. Quanto mais próximo de 1 é o resultado alcançado pelo *V-Measure*, melhor é seu desempenho.

As medidas de homogeneidade e completude podem ser aplicadas a partir de qualquer solução de clusterização, independentemente do número de itens, classes ou *clusters*. As Equações (II.6) e (II.7) formalizam o cálculo de tais medidas.

$$h = \frac{1 - H(C|K)}{H(C)} \tag{II.6}$$

$$c = \frac{1 - H(K|C)}{H(K)} \tag{II.7}$$

No caso, $H(C|K)$ é a entropia condicional das classes em C segundo as atribuições dos itens aos *clusters* em K. Tal medida é calculada como mostrado na Equação (II.8), considerando $|C|$ como o número de classes, $|K|$ como o número de *clusters*, n_c como número de itens da classe c, n_k como número de itens atribuídos ao grupo k, e $n_{c,k}$ o número de itens da classe c no grupo k. Já $H(C)$ é a entropia das classes, que é dada pela Equação (II.9). A entropia condicional dos *clusters* dadas as classes $H(K|C)$ e a entropia dos *clusters* $H(K)$ são definidas de forma análoga.

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|} \frac{n_{c,k}}{n} \log\left(\frac{n_{c,k}}{n_k}\right) \tag{II.8}$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \log\left(\frac{n_c}{n}\right) \tag{II.9}$$

Para a validação da consistência do processo de clusterização pode-se utilizar uma medida chamada *coeficiente de silhueta*. O valor do coeficiente de silhueta é uma medida de semelhança entre um objeto e seu próprio *cluster* em comparação com outros *clusters*. O coeficiente de silhueta é

calculado usando a distância média intra-cluster e a distância média do cluster mais próximo, utilizando a Equação (II.10), onde quanto mais próximo de 1 é o valor de silhueta, melhor foi o processo de clusterização. Valores próximo de -1 indicam que a clusterização teve menor sucesso. Valores próximos a 0 indicam *clusters* sobrepostos. Valores negativos geralmente indicam que uma amostra foi atribuída erroneamente ao *cluster*, pois um *cluster* diferente seria mais semelhante.

$$silhouette = \frac{(b-a)}{max(a,b)} \tag{II.10}$$

Onde a é a distância *intra-cluster* e b é a distância entre uma amostra e o *cluster* mais próximo do qual a amostra não faz parte. Para o cálculo da distância *intra-cluster*, para uma amostra $i \in C_i$ (amostra i no *cluster* C_i), é utilizada a Equação (II.11):

$$a_i = \frac{1}{|C_j - 1|} \sum_{j \in C_j, i \neq j} d(i,j) \tag{II.11}$$

Onde $d(i,j)$ é a distância do ponto i ao ponto j no *cluster* C_j. A medida de distância *intra-cluster* indica a atribuição da amostra i ao seu *cluster*, quanto menor o valor, melhor a atribuição. Para o cálculo de b, definimos a distância da amostra i para cada *cluster* C_k, sendo C_k os *clusters* que $i \notin C_k$. Para a amostra $i \in C_i$, é utilizada a Equação (II.12).

$$b(i) = \min_{k \neq i} \frac{1}{C_k} \sum_{j \in C_k} d(i,j) \tag{II.12}$$

Onde $b(i)$ é a distância média de i a todos os pontos em um *cluster*, do qual i não é membro. O *cluster* com a menor distância média é considerado o *cluster* mais próximo de i.

Capítulo III Trabalhos Relacionados

Este capítulo apresenta trabalhos relacionados a análise de relacionamentos e coesão partidária de deputados. Foi realizada uma revisão sistemática para entendermos o estado atual da pesquisa nessa área. Além disso, a técnica de *Snowballing* foi utilizada para abranger estudos correlatos que podem não terem sido identificados na revisão sistemática. Alguns trabalhos relacionados são apresentados e discutidos no presente capítulo, considerando tais métodos de busca na literatura. A busca por trabalhos relacionados, foi realizada com as seguintes *strings* de busca para realização das pesquisas: '*bipartite projections*','*cosponsorship networks*','*legislative networks*','*voting networs*','*análise de discursos parlamentares NLP*','*complex network to political analysis*', '*automatic evaluation of the political tendency*','*media bias*','*sentiment analysis*','*social networks*','*political ideology detection*' e '*political discourse and ideology*', '*universal sentence encoder political*'.

Historicamente, os cientistas políticos têm dedicado muita atenção ao papel que as instituições e os atores políticos tem em uma variedade de fenômenos deste contexto. Recentemente, no entanto, tem ocorrido uma forte tendência para um ponto de vista baseado nos fundamentos relacionais da política, dando origem a várias redes políticas e metodologias para caracterizar suas estruturas [Neal, 2014; Lee et al., 2016; Kirkland and Gross, 2014]. No trabalho de Cardoso et al. [2023] os autores propõem uma metodologia para uma avaliação de centralidade no contexto das redes de votação. A metodologia utilizada no artigo consiste em construir um grafo bipartido não ponderado que relaciona os parlamentares às votações. Cada nó também mantém a informação de qual foi seu voto (sim, não, abstenção ou obstrução) em determinada votação. Com isso, é então possível projetar uma rede com todos os parlamentares, cujos links indicam pelo menos uma participação conjunta nas votações consideradas, com o objetivo de realizar um cálculo probabilístico para o alinhamento entre os extremos das arestas.

Alguns dos mais recentes trabalhos relacionados a este tema utilizam diversas informações para associar parlamentares entre si, por exemplo: votações, doações de campanha, e participação em eventos. Em [Bursztyn et al., 2020] são utilizadas redes complexas para avaliar a relação entre as doações recebidas por parlamentares eleitos em 2014 e seus comportamentos de voto durante 2015 e 2016. Os autores examinam a homofilia e a coesão dos parlamentares nas redes criadas com respeito aos seus partidos políticos e regiões eleitorais. Em [Dal Maso et al., 2014] o artigo analisou a rede de relações entre os membros do parlamento com base no seu comportamento eleitoral e explora

a estrutura comunitária no que diz respeito às coligações políticas e às alianças governamentais, Os autores desenvolvem novas métricas, como a densidade intra-cluster, uma métrica que mede a tendência de um partido votar como uma entidade única; para polarização partidária, coesão interna da coalizão e força governamental usando análises de redes complexas. O estudo centra-se na Câmara dos Deputados do Parlamento italiano e os métodos também podem ser aplicados a outros cenários. O trabalho [Lima et al., 2023] utiliza PLN e Redes Complexas para avaliar a coesão entre os parlamentares em seus partidos, utilizado TF-IDF e distância de cosseno para correlacionar os parlamentares de acordo com seus discursos.

A determinação do perfil temático dos deputados federais, através do processamento dos textos obtidos de seus discursos e proposições é realizada em [Fernandes, 2017]. O trabalho apresenta técnicas de processamento de linguagem natural utilizadas para a análise dos discursos dos deputados, que incluem a remoção de palavras com pouco significado semântico (*stop words*), redução dos termos às suas raízes morfológicas (*stemming*), representação computacional dos textos (*bag-of-words*) e utiliza o modelo de aprendizado de máquina supervisionado *Naive Bayes* para a classificação temática dos discursos e proposições. Já o trabalho de Menini and Tonelli [2016] se concentra na tarefa de comparação automatizada de pontos de vista entre políticos, especificamente no contexto de campanhas eleitorais. Os autores realizaram uma anotação manual dos componentes envolvidos na detecção de concordância e discordância entre discursos. O estudo usa esses dados anotados, junto com outras características lexicais, para treinar dois classificadores, uma CNN e SVM, que classificam com sucesso a concordância e a discordância com boa precisão.

O desafio de relacionar os deputados, pela similaridade de seus votos, utilizando as votações nas quais os parlamentares participam foi abordado em [Brito et al., 2020]. Essa abordagem pode ser aplicada em diversos cenários políticos, visto que a maioria dos processos legislativos atuais possuem votações para a aprovação de propostas. No trabalho de Schwarz et al. [2017] os autores abordam um método comparativo entre votos e discursos, para estimar as preferências intrapartidárias. Segundo os autores, as votações podem ser enviesadas pois sofrem de uma forte disciplina partidária que tende a tornar a votação uma indicação estratégica e não uma indicação sincera de preferências. Já os discursos tendem a ser menos restritos pelos partidos, podendo fornecer informações importantes sobre suas verdadeiras preferências políticas. O artigo conclui que os discursos revelam maiores diferenças intrapartidárias do que os votos nominais, indicando que os discursos legislativos fornecem uma indicação mais irrestrita e sincera das preferências em comparação com o comportamento eleitoral.

Em [Cherepnalkoski et al., 2016] os autores analisam os padrões de co-votação e o comportamento nas redes sociais dos membros da Parlamento Europeu, bem como a interação entre estes dois sistemas. Os autores utilizam dois conjuntos de dados no estudo. O primeiro conjunto de análise

diz respeito aos padrões de co-voto e, por outro, o seu comportamento de *retweets*. O primeiro é o conjunto de dados de votação nominal, onde a coesão é considerada como a tendência de co-votar dentro de um partido, e uma coalizão é formada quando os membros de vários partidos exibem um alto grau de acordo de co-voto sobre um assunto. O segundo conjunto de dados tem como base *retweets*, o conjunto representa o padrão de *retweets* dos membros do Parlamento Europeu e implica coesão (*retweets* dentro do mesmo grupo) e coalizões (*retweets* entre grupos). O artigo emprega duas metodologias diferentes para analisar coesão e coalizões: a confiabilidade Alfa de Krippendorff e os Modelos de Grafos Aleatórios Exponenciais. Os resultados da pesquisa indicam um nível considerável de correlação entre esses dois sistemas complexos.

Gomes Ferreira et al. [2018] examina a criação e a evolução de comunidades ideológicas em sistemas partidários políticos, usando dados de votação pública do Brasil e dos EUA ao longo de um período de 15 anos. Os autores propõem uma metodologia que envolve a criação de grafos ponderados e não direcionados, que são construídos com base na semelhança das posições de voto entre os membros, para modelar a dinâmica das comunidades ideológicas nas sessões de votação. No Brasil, apesar de um sistema partidário fragmentado, a polarização pode ser observada até certo ponto em comunidades menores e fortemente vinculadas. No entanto, os valores médios da disciplina partidária são altos, indicando alta disciplina partidária entre a maioria dos membros do partido. Já nos EUA, quase todos os membros e comunidades estão altamente polarizados, indicando uma polarização partidária forte e estável.

Em Gerrish and Blei [2011] os autores desenvolvem vários modelos preditivos que conectam sentimento legislativo ao texto legislativo. Baseados em modelos de tópicos supervisionados, o trabalho visa prever padrões de votação com base no conteúdo dos projetos de lei e inferir as inclinações políticas dos legisladores. Utilizando dados de 12 anos legislativos, o trabalho obteve êxito em padrões de votação específicos, obtendo uma acurácia de 96%. No trabalho de Nay [2017] os autores desenvolveram uma abordagem de aprendizado de máquina para prever a probabilidade de um projeto de lei se tornar lei. Utilizando modelo de *embending*, o trabalho investiga sobre quais palavras aumentam a probabilidade de promulgação de algum tópico. O modelo que pontua o texto completo do projeto de lei usando representações da linguagem aprendida pelo *word2vec*, foi usado para prever quais frases de um projeto de lei contribuem para sua promulgação.

Balahur et al. [2009] investiga diferentes abordagens para classificar opiniões e descobrir fontes de opinião a partir do texto, usando o léxico de afeto, opinião e atitude, aplicado em dados de discursos do Congresso Americano. Os autores propõem três métodos para classificar a opinião no nível do segmento de fala, incluindo medidas de similaridade com léxicos, análise de dependência e aprendizado de máquina SVM. Eles também estudam o impacto de considerar a fonte de opinião e propõem métodos para classificar a opinião no nível de intervenção do palestrante. Os resultados

apresentaram melhorias em relação à classificação individual de segmentos de texto, e o trabalho mostra que sua abordagem funciona melhor do que classificadores treinados em dados específicos. Esses métodos foram aplicados aos debates no Congresso, e os resultados mostraram melhorias na classificação de segmentos individuais de fala e nas intervenções dos palestrantes.

Uma abordagem de técnicas de redes para relacionar diferentes portais de notícia de acordo com o viés ideológico das notícias publicadas é apresentada em [Aires et al., 2020]. Para analisar a relação entre características estruturais das redes e o viés político, isto é, se a ideologia dos portais reflete-se em propriedades de redes que modelam citações (*hiperlinks*) entre eles, foi feito o uso desses *hiperlinks* para desenvolver um método de classificação automática. Nos trabalhos de [Zhitomirsky-Geffet et al., 2016] e [Dallmann et al., 2015] foram utilizadas técnicas de PLN e aprendizado de máquina para determinar o viés de notícias, utilizando textos abertamente políticos para avaliação totalmente automática da tendência política de sites de notícias *online*.

Outra grande área de pesquisa, apresenta trabalhos voltados para análise de redes sociais, como abordado por Elejalde et al. [2017], que utilizou *tweets* para verificar automaticamente a orientação política e socioeconômica de portais de notícias chilenos. Já o estudo de Tumasjan et al. [2010], utiliza mais de 100.000 publicações na rede social *Twitter* durante as eleições para o Governo Federal Alemão para investigar se o a plataforma é utilizada para expressão política e se as mensagens publicadas na rede social espelham acertadamente o sentimento político *offline*. Foram analisadas mensagens no *Twitter* mencionando partidos ou políticos antes das eleições federais alemãs de 2009. Os resultados demonstraram que o *Twitter* é uma plataforma amplamente empregada como plataforma de deliberação política. O número de *tweets* refletem as preferências dos eleitores e se aproximam pesquisas eleitorais tradicionais, enquanto o sentimento do *twitter* mensagens corresponde estreitamente aos programas políticos, perfis de candidatos e evidências da cobertura da mídia da trilha da campanha.

Em [Cristiani et al., 2020] foram aplicadas técnicas de análise de sentimentos para avaliar a relação entre a opinião dos usuários do Twitter escritos na língua portuguesa e o resultado final das eleições. O mesmo ocorre para outros países e línguas como pode ser observado nos seguintes trabalhos [Halberstam and Knight, 2016; Chen et al., 2017; Pastor-Galindo et al., 2020; Caetano et al., 2018; Van Dijk, 2003], os quais consideram estudos sobre identificação de posicionamento e ideologia política através de dados textuais. Apesar de que uma maior abundância de trabalhos sobre PLN estão voltados para aprendizado supervisionado, há também na literatura aqueles nesta temática utilizando aprendizado não-supervisionado. Por exemplo, como os trabalhos [Wives, 1999; Dias et al., 2019] sobre métodos de agrupamento de objetos textuais, onde os tais objetos são organizados automaticamente em grupos similares e é realizado um estudo comparativo de algoritmos de agrupamento aplicados ao agrupamento de tais objetos. Já em [Greene and Cross, 2017] é realizada

uma modelagem de tópicos de discursos no parlamento da União Europeia.

Em [Rao and Spasojevic, 2016], é aplicado *word embeddings* e redes neurais com Long Short-Term Memory (LSTM) ao problema de classificação de texto, uma classificação de mensagens com respeito à inclinação política foi realizada. As mensagens de mídia social foram classificadas como democratas ou republicanas, considerando os partidos políticos dos EUA. O modelo foi capaz de classificar as mensagens com uma precisão de 87,57%.

No trabalho de Abercrombie and Batista-Navarro [2018] os autores aplicam métodos de mineração de opinião, em transcrições dos debates dos parlamentares do Reino Unidos, para classificar a polaridade de sentimento dos oradores como sendo positiva ou negativa em relação aos temas propostos nos debates. O artigo avalia o desempenho dos classificadores SVM e MLP na análise de sentimentos e compara o uso de recursos textuais de n-gramas e recursos de metadados contextuais, sugerindo que os rótulos anotados manualmente refletem mais de perto o sentimento dos palestrantes, enquanto os metadados contextuais podem ser altamente preditivos dos votos por divisão, destacando possíveis entendimentos sobre o processo democrático parlamentar e as opiniões dos membros do Parlamento fornecidos pela análise de sentimentos dos debates parlamentares.

Em [Biessmann, 2016] o autor automatiza o processo de detecção de viés político utilizando um conjunto de dados de políticos alemães em discursos parlamentares e declarações de manifesto. Foi utilizado o modelo de classificação Regressão Logística com vetores de recursos Bag-of-words (BOW). A classificação se deu entre 5 partidos alemães de acordo com os dados dos discursos. No trabalho de Kummervold et al. [2021], é mostrado o processo de construção de um conjunto de treinamento em larga escala a partir de acervos digitais e digitalizados em uma biblioteca, um classificador foi treinado para realizar a detecção de filiação partidária usando *transformers* e o modelo de linguagem *NB-BERT* utilizando os discursos dos dois maiores partidos do Parlamento Norueguês.

Há diversos trabalhos que utilizam o USE em contextos políticos. Em [Silva et al., 2021] o USE foi empregado para avaliar modelos de tópicos na política utilizando discursos sobre projetos de lei da Câmara de Deputados do Brasil. Os tópicos obtidos são comparados de forma automatizada com tópicos anotados por um especialista. Em [dos Santos and Siqueira, 2019] os autores apresentam um modelo baseado em USE para classificar a opinião de *tweets* em português sobre o tema da Reforma da Previdência no contexto político brasileiro. Utilizando o *Multilingual Universal Sentence Encoder for Semantic Retrieval*, obtiveram uma acurácia média de 82% na classificação das duas opções, a favor ou contra a Reforma da Previdência. Já em [Saligrama, 2019] os autores propõem um esquema de classificação para detectar viés político em textos longos como artigos de jornais. Os autores treinam um USE com *tweets* e executam a classificação em textos maiores.

Diferente dos trabalhos apresentados, que visam classificar e relacionar notícias e pessoas de

acordo com o viés ideológico, este estudo busca trabalhar com discursos dos deputados da Câmara, com o objetivo de relacionar a proximidade dos discursos de diferentes deputados e observar a coesão de discursos entre os partidos políticos utilizando aprendizado não-supervisionado.

A Tabela 2.1 mostra os artigos selecionados pelos métodos utilizados, listados com informações resumidas do objetivo, fonte de dados da análise e categoria da resolução.

#	**Referência**	**Problema**	**Dataset**	**Categoria**
1	[Bursztyn et al., 2020]	Relação entre as doações recebidas	Doações	Agrupamento
2	[Dal Maso et al., 2015]	Relações entre os membros do parlamento com base no seu comportamento eleitoral	Discurso	Agrupamento
3	[Fernandes, 2017]	Determinação do perfil temático	Discurso	Classificação
4	[Menini and Tonelli, 2016]	Comparação automatizada de pontos de vista entre políticos em campanhas eleitorais	Discurso	Classificação
5	[Brito et al., 2020]	Relacionar os deputados, pela similaridade de seus votos, utilizando as votações dos parlamentares	Votos	Agrupamento
6	[Schwarz et al., 2017]	Método comparativo entre votos e discursos, para estimar as preferências intrapartidária	Votos e Discursos	Classificação
7	[Cherepnalkoski et al., 2016]	Análise dos padrões de co-votação e comportamento nas redes sociais dos membros do Parlamento Europeu	Votos e Discursos	Alfa de Krippendorff e os Modelos de Grafos Aleatórios Exponenciais
8	[Gomes Ferreira et al., 2018]	Análise da criação e evolução de comunidades ideológicas em sistemas partidários políticos	Discursos	Grafos Ponderados
9	[Gerrish and Blei, 2011]	Predição de padrões de votação com base no conteúdo dos projetos	Discursos	Classificação
10	[Nay, 2017]	Probabilidade de um projeto de lei se tornar lei.	Discursos	Word2Vec
11	[Balahur et al., 2009]	Classificação de opinião através do discurso em debates no Congresso	Discursos	Classificação
12	[Aires et al., 2020]	Classificação de viés ideológico	Portais de notícias	Classificação
13	[Zhitomirsky-Geffet et al., 2016]	Classificação de viés ideológico	Portais de notícias	Classificação
14	[Dallmann et al., 2015]	Classificação de viés ideológico	Portais de notícias	Classificação
15	[Elejalde et al., 2017]	Verificar automaticamente a orientação política e sócio-econômica	Redes Sociais	Classificação
16	[Tumasjan et al., 2010]	Comparação entre expressão política online e offline	Redes Sociais	Classificação
17	[Cristiani et al., 2020]	e Identificação de posicionamento e ideologia política através de dados textuais	Redes Sociais	Classificação
18	[Halberstam and Knight, 2016]	e Identificação de posicionamento e ideologia política através de dados textuais	Redes Sociais	Classificação

#	Referência	Problema	Dataset	Categoria
19	[Chen et al., 2017]	e Identificação de posicionamento e ideologia política através de dados textuais	Redes Sociais	Classificação
20	[Pastor-Galindo et al., 2020]	e Identificação de posicionamento e ideologia política através de dados textuais	Redes Sociais	Classificação
21	[Caetano et al., 2018]	e Identificação de posicionamento e ideologia política através de dados textuais	Redes Sociais	Classificação
22	[Van Dijk, 2003]	e Identificação de posicionamento e ideologia política através de dados textuais	Redes Sociais	Classificação
23	[Wives, 1999]	e Organização em grupos de similaridades	Redes Sociais	Agrupamento
24	[Dias et al., 2019]	Organização em grupos de similaridades	Redes Sociais	Agrupamento
25	[Dias et al., 2019]	Organização em grupos de similaridades	Redes Sociais	Agrupamento
26	[Greene and Cross, 2017]	Modelagem de tópicos	Discursos	Agrupamento
27	[Rao and Spasojevic, 2016]	Análise de inclinação política	Redes Sociais	Classificação
28	[Abercrombie and Batista-Navarro, 2018]	Polaridade de sentimento	Discursos	Classificação
29	[Biessmann, 2016]	Detecção de viés político	Discursos	Classificação
30	[Kummervold et al., 2021]	Detecção de filiação partidária	Acervos digitais	Classificação
31	[Silva et al., 2021]	Modelos de tópicos na políticas	Redes Sociais	Classificação
32	[dos Santos and Siqueira, 2019]	Classificação de opinião de *tweets*	Redes Sociais	Classificação
33	[Saligrama, 2019]	Detecção de viés político em textos longos	Artigos de jornais	Classificação

Tabela III.1: Lista de artigos da revisão sistemática com informações gerais resumidas.

Capítulo IV Metodologia

Neste Capítulo é apresentada a metodologia proposta para analisar a similaridade entre o posicionamento político de deputados através de seus discursos. Neste processo serão descritos: *(i)* o conjunto de dados utilizado, *(ii)* uma abordagem para resolver o problema em estudo baseada em trabalhos prévios, *(iii)* o método proposto nesta pesquisa para analisar similaridade entre o posicionamento político de deputados e *(iv)* a medida usada para avaliação dos *clusters* gerados a partir das alternativas consideradas.

IV.1 Descrição dos dados

Os dados dos parlamentares utilizados neste trabalho são dados abertos, disponíveis no Serviço de Dados Abertos da Câmara dos Deputados [dos Deputados, 2021], uma aplicação do poder Legislativo onde são disponibilizadas funcionalidades que permitem o acesso direto aos dados legislativos produzidos na Câmara dos Deputados. Neste serviço, diversos tipos de dados são disponibilizados, tais como dados sobre deputados, órgãos legislativos, proposições, sessões plenárias e reuniões de comissões. Os dados utilizados nesta pesquisa contém informações relevantes sobre cada discurso dos parlamentares, como: *(i)* tipo, *(ii)* data e hora da proclamação, *(iii)* sumário e *(iv)* transcrição completa, como ilustrado pela Tabela IV.1.

Tipo	**Data**	**Sumário**	**Transcrição**
PELA ORDEM	15/04/2019	Protesto contra a decisão do Governo$\cdots$	Sr. Presidente, Jair Bolsonaro acaba de queimar 19 reais$\cdots$
PELA ORDEM	16/02/2012	Parecer, pela Comissão Especial, ao Projeto de Lei 382 de 2011 $\cdots$	Obrigado, Presidente Inocêncio Oliveira. Em nome da Câmara dos Deputados, quero dizer$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
GRANDE EXPEDIENTE	14/04/2015	Realização de evento do Fórum Sindical Sul$\cdots$	Sr. Presidente, Sras. e Srs. Deputados, quero apenas destacar$\cdots$

Tabela IV.1: Exemplo dos dados utilizados nesta pesquisa.

Para a realização das análises, foram recuperados do Serviço de Dados Abertos da Câmara dos Deputados os dados dos discursos dos 627 parlamentares ativos durante a 54^a legislatura, que compreende os anos entre 2011 a 2015. A legislatura analisada contém 8.378 discursos ao total,

o tamanho médio dos discursos antes e depois da aplicação das técnicas de pré-processamento: *i) tokenização* (Seção II.1), *ii) stemming* (Seção II.2), e *iii)* remoção de *stop words* (Seção II.3) é de 3.556 e 2.580 caracteres, respectivamente. Cada parlamentar discursou pelo menos uma vez no período analisado, tendo em média $13,25$ discursos por parlamentar e no máximo 15 discursos de um único parlamentar.

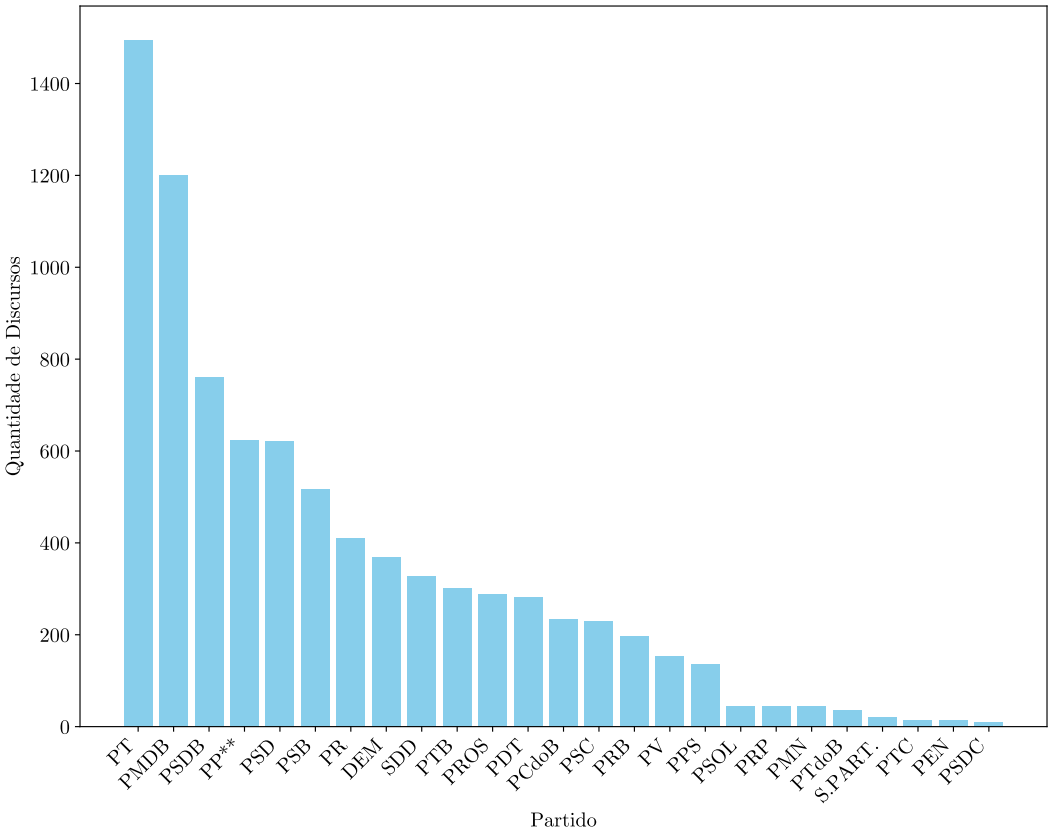

Figura IV.1: Quantidade de Discursos por Partido.

O gráfico de barras apresentado na Figura IV.1 ilustra a distribuição da quantidade de discursos por partido político. Cada barra no gráfico representa um partido específico, enquanto as informações mostradas no eixo Y indicam o número de discursos associados a cada partido. Esta representação visual é uma ferramenta eficaz para a análise inicial da participação discursiva dos diferentes partidos políticos no contexto em estudo.

A inspeção do gráfico revela uma variação notável na produção discursiva entre os partidos. Destaca-se partidos como o PT e o PMDB que exibem as maiores frequências de discursos, indicando um nível significativo de atividade discursiva associada a estes partidos. Por outro lado, os partidos PSDC, PEN e PTC apresentam uma frequência comparativamente menor de discursos, sugerindo uma presença discursiva inferior em relação aos demais partidos, devido a sua baixa presença em quantidade de parlamentares. Esta observação inicial indica uma discrepância na participação

discursiva entre os partidos políticos analisados.

IV.2 Abordagem Preliminar User-as-Document com TF-IDF

Para definir a distância entre os parlamentares baseada em seus discursos, foi utilizada a modelagem *User-as-Document (UaD)* [Cossu et al., 2016]. Esta consiste em concatenar todos os discursos de um parlamentar, formando assim, um único documento que o representa. Então, a similaridade entre dois parlamentares é calculada utilizando a semelhança entre seus respectivos documentos.

Para a criação destes documentos, são utilizadas duas técnicas de pré-processamento de PLN: *(i)* remoção de *stop-words* (Seção II.3) e outros *tokens* (Seção II.1) considerados menos relevantes, e *(ii) stemming* (Seção II.2) [Lovins, 1968b]. A lista de *stop-words* teve como base a lista disponível em português no pacote NLTK [Bird, 2006] com a adição de palavras muito comuns ao contexto político, como, por exemplo, os pronomes de tratamento "sra", "v. exa" entre outros. Para a execução do *stemming*, o módulo do NLTK *SnowballStemmer* foi aplicado no corpus. Após esta etapa os documentos são representados de forma vetorial utilizando o TF-IDF e as similaridades entre estes são calculadas pela distância de cosseno. A Figura IV.2 mostra o fluxograma que ilustra o processamento de diversos discursos de um parlamentar concatenado em um único documento, o qual passa por todas as etapas da metodologia e por fim entrega.

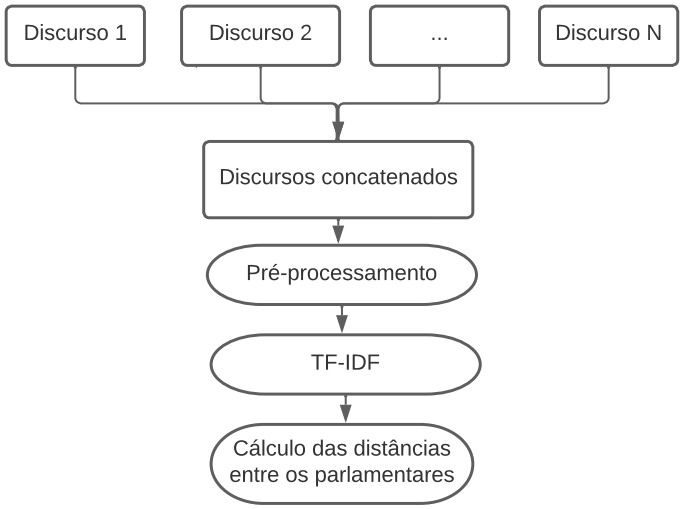

Figura IV.2: Fluxograma de processamento para criação do *User-as-Document*.

Ainda sobre o pré-processamento, o ajuste fino de seus hiperparâmetros é uma ação que pode auxiliar na análise da similaridade. O objetivo é fazer com que a avaliação do agrupamento com base na similaridade entre parlamentares reflita simultaneamente a estrutura partidária vigente assim como evidências que contrastem com tal estrutura. Um exemplo seria avaliar a coesão partidária e

a diversidade de discursos intra-partidário.

A Figura IV.3 exemplifica uma rede com dados como os utilizados neste trabalho. Os nós representam os parlamentares, as arestas são determinadas pela similaridade entre eles de acordo com seus discursos. Os vetores localizados abaixo de cada nó representam o conjunto de discursos de cada parlamentar, os quais estão classificados pelo assunto/ideia através das cores. Os valores, apresentados em cada vetor, mostram a quantidade de discursos proferidos de cada tipo. O cálculo das similaridades nada mais é que o produto escalar desses vetores. Ele é realizado através da multiplicação da quantidade de discursos, de cada tipo, de um deputado com o de outro deputado. Esta multiplicação ocorre apenas entre a quantidade de discursos do mesmo tipo, ou seja, situados na mesma coloração. A soma dessas multiplicações resulta no valor da ponderação de cada aresta, isto é, quantificando desta forma a similaridade entre eles. Considerando esta abordagem, quanto maior for esse valor, maior será a similaridade entre os parlamentares.

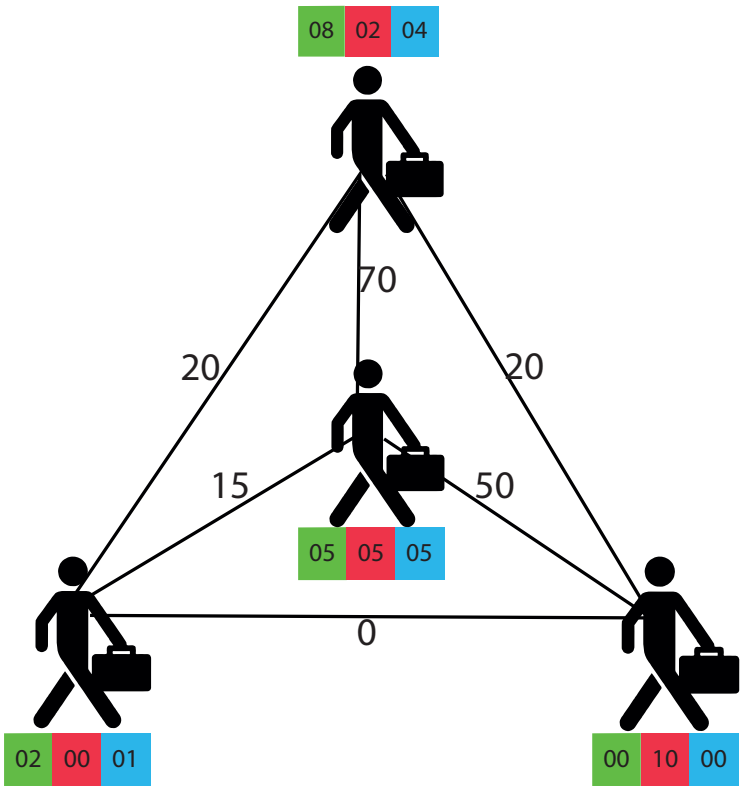

Figura IV.3: Ilustração da representação e avaliação de similaridade entre parlamentares através de seus discursos.

IV.3 Abordagem Proposta

No contexto dos discursos de parlamentares, a utilização da abordagem UaD pode acarretar em algumas dificuldades no cálculo das distâncias. Uma das razões disto é a descaracterização do número de documentos no corpus, o que reflete diretamente no cálculo das similaridades utilizando

TF-IDF e USE. Outro problema relacionado é o cruzamento de termos de diferentes discursos, o que pode levar a uma extrapolação exagerada na caracterização do parlamentar.

Uma abordagem alternativa para avaliar o alinhamento entre dois parlamentares é calcular a distância entre cada par dos discursos destes, e então agregar tal coleção de distâncias em um único valor que representa a distância entre os parlamentares. Dessa maneira, a distribuição dos *tokens* no corpus gerado seria mais realista em comparação àquela correspondente a abordagem UaD. Além disso, o impacto do cruzamento de diferentes termos entre os discursos pode ser minimizado utilizando métodos distintos de agregação.

Existem diversos métodos de agregação que permitem resumir em uma única medida as distâncias entre os pares de discursos de dois parlamentares quaisquer. Neste trabalho são propostos seis tipos de agregação. Todos estes tem como premissa a formação de um grafo bipartido completo determinado de tal forma que cada nó corresponde a um discurso, e cada partição dos nós corresponde a um parlamentar. O valor associado a uma aresta representa a distância entre os discursos em que esta incide. A Figura IV.4 ilustra um grafo bipartido completo como os que são utilizados nas agregações, onde cada cor representa um parlamentar, cada nó de uma partição representa um discurso deste parlamentar e os pesos associados às arestas representam a semelhança entre os discursos.

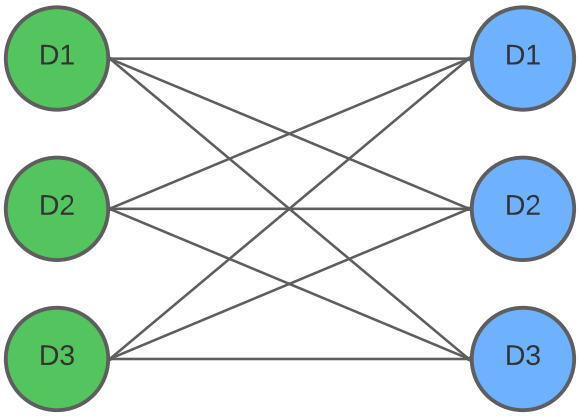

Figura IV.4: Grafo bipartido completo representando os discursos.

Os tipos de agregação propostos, implementados e avaliados são sucintamente descritos a seguir:

1. **Média das distâncias**: Utiliza a média de todos os valores de distâncias entre pares de discursos, o que corresponde ao peso médio das arestas do grafo bipartido, como descrita na Equação (IV.1). Onde $|E|$ é o número total de arestas no grafo e w_{ij} representa o peso associado à aresta $(i, j) \in E, i \in V, j \in V$.

$$\mu = \frac{1}{|E|} \sum_{(ij) \in E} w_{ij}. \tag{IV.1}$$

2. **Mínimo de distâncias**: O menor valor de distância entre cada par de discursos corresponde ao valor mínimo dentre todos os pesos das arestas do grafo bipartido completo. O valor mínimo pode ser obtido por meio da Equação (IV.2), onde w_{min} representa tal valor, $|E|$ é o número total de arestas no grafo e w_{ij} representa o peso da aresta $(i, j) \in E, i \in V, j \in V$.

$$w_{min} = \min_{\forall (i,j) \in E} w_{ij}. \tag{IV.2}$$

3. **Máximo de distâncias**: Refere-se ao maior valor de distância entre cada par de discursos, o que corresponde ao maior peso dentre todas as arestas do grafo bipartido. Para determinar a maior distância entre cada par de discursos no grafo bipartido completo, é necessário determinar o maior valor dentre os pesos das arestas do grafo G. O valor máximo pode ser obtido através da Equação (IV.3), onde w_{max} é o maior valor de distância, w_{ij} é o peso da aresta $(i, j) \in E, i \in V, j \in V$.

$$w_{max} = \max_{\forall (i,j) \in E} w_{ij}. \tag{IV.3}$$

4. **Média das menores distâncias (*MédiaMin*)**: Medida relativa à média das menores distâncias de cada discurso. Essa medida pode ser obtida pela Equação (IV.4), onde μ_{min} representa a medida relativa à média das menores distâncias de cada discurso, N é o número total de discursos, w_{ij} representa a distância entre os discursos $(i, j) \in E, i \in V, j \in V$, e a expressão $\min_{\forall j \in V, j \neq i} w_{ij}$ retorna a menor distância entre o discurso i e todos os outros discursos j diferentes de i. A equação calcula a média dessas menores distâncias para cada discurso e, por isso, fornece uma medida relativa à média das menores distâncias em todo o conjunto de discursos.

$$\mu_{min} = \frac{1}{N} \sum_{i=1}^{N} \min_{\forall j \in V,\ j \neq i} w_{ij}. \tag{IV.4}$$

5. **Média das maiores distâncias (*MédiaMax*)**: É a medida relativa à média das maiores distâncias entre cada par de discursos, e pode ser obtida por meio da Equação (IV.5), em

que μ_{max} representa essa medida, N é o número total de discursos, w_{ij} é a distância entre os discursos $(i,j) \in E, i \in V, j \in V$, e $\max_{\forall j \in V, j \neq i} w_{ij}$ é a maior distância entre o discurso i e todos os outros discursos j distintos de i. A equação calcula a média dessas maiores distâncias para cada discurso, fornecendo assim uma medida relativa à média das maiores distâncias em todo o conjunto de discursos.

$$\mu_{max} = \frac{1}{N} \sum_{i=1}^{N} \max_{\forall j \in V,\ j \neq i} w_{ij}. \tag{IV.5}$$

6. **Árvore geradora mínima (*AGMin*)**: s_T é a soma das arestas da árvore geradora mínima $T(V,E)$ obtida através do grafo bipartido. A árvore pode ser obtida de acordo com o descrito na Seção II.6.2. A Equação (IV.6) mostra como realizar este cálculo.

$$s_T = \sum_{(i,j) \in T} w_{ij} \tag{IV.6}$$

Onde $(i,j) \in E$ é uma aresta da árvore geradora mínima T e w_{ij} é o peso associado a esta aresta. Esta equação representa a soma dos pesos de todas as arestas pertencentes à T, que é uma árvore que contém todos os vértices do grafo original, conectando-se com o menor custo possível.

IV.4 Validação dos Resultados Utilizando TF-IDF

Com objetivo de legitimar o uso de grafos e PLN para determinar grupos orgânicos de parlamentares foi idealizado o uso de um critério de avaliação objetivo para o TF-IDF. Tornando possível assim, identificar a configuração ótima referente ao tipo de agregação e aos hiperparâmetros da fase de pré-processamento. Ademais, é válido observar que o USE dispensa a necessidade de validação de hiperparâmetros, uma vez que não possui ajustes configuráveis. Este critério idealmente indicaria o quão bem um agrupamento reflete a estrutura partidária vigente. Todavia cabe ressaltar que isto não impossibilita que tal agrupamento inclua evidências que contrastem com a estrutura mencionada, pela similaridade (ou dissemelhança) entre parlamentares.

Neste trabalho os grupos de parlamentares são definidos conforme suas distâncias dois-a-dois, ou seja, de acordo com a distância entre cada par de parlamentar. A quantidade de grupos é definida pela quantidade de partidos existentes, considerando os dados usados neste trabalho tem-se 24 grupos. Dadas tais condições, foi utilizada a abordagem de agrupamento hierárquico aglomerativo (Seção II.7, pelo uso de similaridades pré-definidas ao invés do uso de coordenadas de pontos cuja distância pudesse ser avaliada [Jain et al., 1999]).

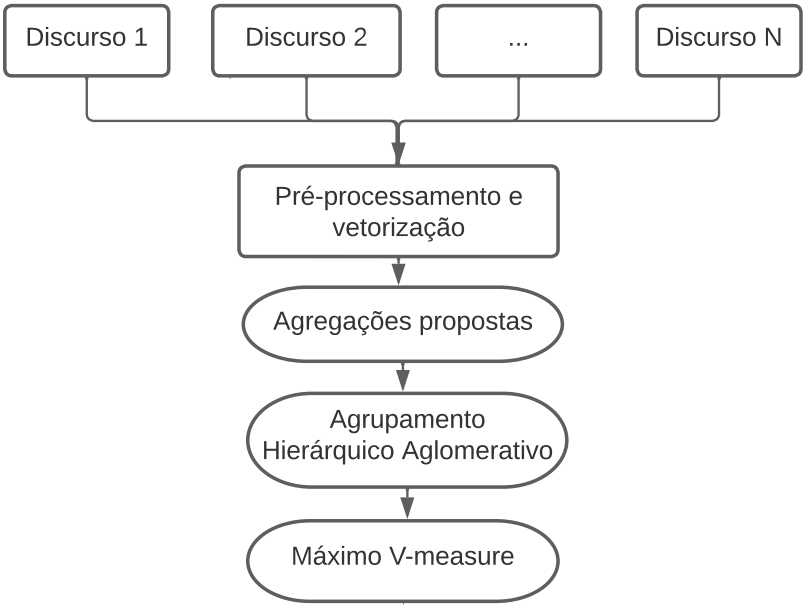

Figura IV.5: Fluxograma de obtenção da medida de avaliação.

No fluxograma apresentado na Figura IV.5, o procedimento para obtenção da medida de avaliação recebe o texto dos discursos em sua forma integral, em seguida realiza o pré-processamento e vetorização que consiste em aplicar os algoritmos de tokenização, *steming* e remoção de *stop words* e transformar o texto em uma matriz TF-IDF. Logo após realizar algumas das agregações discutidas na Seção IV.3, obtém-se como resultado um grafo completo entre cada parlamentar com os pesos das arestas representando a semelhança dos discursos. Por último, é aplicado o algoritmo de Agrupamento Hierárquico Aglomerativo a fim de agrupar os parlamentares de acordo com a similaridade dos seus discursos utilizando os três tipos de *linkage* mostrados na Seção II.7. Ao final do processo o *linkage* que fornece o melhor valor de *V-measure* é mantido.

Capítulo V Avaliação Experimental

Este Capítulo apresenta os experimentos realizados com os dados e métodos abordados no Capítulo IV. Os resultados dos experimentos que serão apresentados neste Capítulo foram obtidos utilizando os 8.378 discursos referentes a 627 deputados federais ativos durante a 54^a legislatura da Câmara, que compreende os anos entre 2011 e 2015. Cada parlamentar foi tratado como membro do partido ao qual ele pertencia quando foi eleito, desconsiderando eventuais mudanças de partido durante o mandato. Todos os experimentos foram realizados utilizando a linguagem *Python* na plataforma *Google Colaboratory* de forma gratuita.

V.1 Avaliação das Medidas de Distância Baseadas no TF-IDF

É indispensável avaliar e comparar o desempenho dos diferentes tipos de medida de distância entre parlamentares, e seu impacto na avaliação da coesão partidária e na diversidade intra-partidária de discursos. O TF-IDF faz uso de alguns hiperparâmetros, sendo sensível aos valores assumidos por eles. Os hiperparâmetros avaliados neste estudo foram os que regulam a conversão dos discursos para um formato vetorial, refletindo diretamente no agrupamento dos parlamentares.

Utilizando as combinações de hiperparâmetros para vetorização baseada na estatística TF-IDF e no algoritmo de agrupamento hierárquico aglomerativo, diferentes versões do método para aferir a distância foram usados para comparação. A fim de obter os resultados de todas as configurações possíveis em um conjunto de valores especificado para cada hiperparâmetro, foi realizada uma busca em grade (*Grid Search*) [Pedregosa et al., 2011], com 120 configurações de valores considerando, além do *linkage*, as seguintes dimensões, para o método TF-IDF:

- **max_df** (MD): Ao construir o vocabulário, ignora-se os termos que tem uma frequência de documento superior ao limiar fornecido. Os valores considerados para este hiperparâmetro foram: $2^0, 2^{-1}, \cdots, 2^{-5}$;
- **max_features** (MF): Constrói um vocabulário selecionando não mais que este número de atributos, ordenados por suas frequências em todo o corpus. Os valores considerados para este hiperparâmetro foram: $10^1, 10^2, \cdots, 10^5$;
- **sublinear_tf** (ST): Substitui *tf (term frequency)* por $1 + log(tf)$ no cálculo do TF-IDF. Os

valores considerados para este hiperparâmetro foram: Sim e Não;

- **use_idf** (UI): Ativa o uso do *inverse-document-frequency* (idf) no processo de vetorização, o qual também pode ser realizado considerando apenas o *tf*. Os valores considerados para este hiperparâmetro foram: Sim e Não.

A Tabela V.1 apresenta os melhores resultados dentre todas as configurações de hiperparâmetros avaliados. O *V-Measure* foi usado para determinar essa classificação, o qual agrega *i) homogeneidade* e *ii)* completude. No contexto deste experimento, a homogeneidade representa a minimização da diversidade partidária dentro de cada grupo de parlamentares inferido com base nas semelhanças entre seus discursos. A completude representa a preservação da unidade partidária no agrupamento obtido, de forma que a fragmentação de parlamentares originalmente relacionados, seja evitada. A maioria dos resultados apresentados utilizam a abordagem de agregação das distâncias aqui proposta (Seção IV.3). A configuração que obteve melhor resultado foi a que utilizou a média das distâncias; apenas uma das sete melhores é baseada na abordagem *User-as-Document*.

Abordagem	**V-measure**	**Linkage**	**Max_df**	**Max_features**	**ST**	**UI**
Média	0,2264	average	2^{-4}	10^5	Não	Sim
UaD	0,2156	average	2^{-4}	10^3	Não	Sim
MédiaMax	0,2015	complete	2^{-3}	10^5	Sim	Sim
AGMin	0,1950	complete	2^{-4}	10^4	Sim	Não
MédiaMin	0,1659	average	2^{-4}	10^4	Sim	Sim
Máximo	0,1635	complete	2^0	10^5	Sim	Sim
Mínimo	0,1598	complete	2^{-4}	10^2	Sim	Não

Tabela V.1: Melhores configurações conforme *V-measure*.

Observando os hiperparâmetros, abordagens e métodos de agregação individualmente, é possível constatar que a opção $max\_df = 2^{-4}$ em geral tem o melhor desempenho dentre os valores avaliados, o que pode ser interpretado como um indicativo de que há muitos termos relativamente frequentes e pouco discriminativos no corpus analisado. O parâmetro *max_features* para os métodos que calculam a distância entre cada par de deputados obtém melhores resultados com a quantidade máxima de atributos igual ou maior a dez mil, enquanto a abordagem *User-as-Document* obteve seu melhor resultado utilizando no máximo 10^3 atributos. Em geral, os resultados utilizando ambos os parâmetros ST ($sublinear\_tf$) e UI (use_idf) obtiveram melhores resultados quando fornecido *True* como argumento.

Para finalizar o comparativo das medidas de distância entre os parlamentares, a Figura V.1 apresenta a matriz de correlação das medidas consideradas neste estudo. Tal matriz está expressando um mapa, colorido como um mapa de calor para uma melhor visualização. Neste trabalho utilizou-se o τ de Kendall como estatística de correlação, que é uma medida de correlação não-paramétrica,

a qual é capaz capturar até mesmo relacionamentos não-lineares entre variáveis aleatórias, tornando a análise mais flexível e robusta. Segundo a matriz de correlação obtida, é evidente a separação da medida UaD das demais, estando a medida Média posicionada como uma intermediária entre esta primeira e as cinco restantes. Tal organização é consistente com a definição de cada uma destas medidas. Entretanto, chama atenção, o fato notável de que a correlação mais expressiva ocorre entre as alternativas AGMin e MédiaMax. Essas duas medidas possuem princípios fundamentais que não compartilham qualquer semelhança aparente. Essa análise sugere uma relação não convencional entre AGMin e MédiaMax, indicando a complexidade das interações subjacentes entre as variáveis analisadas.

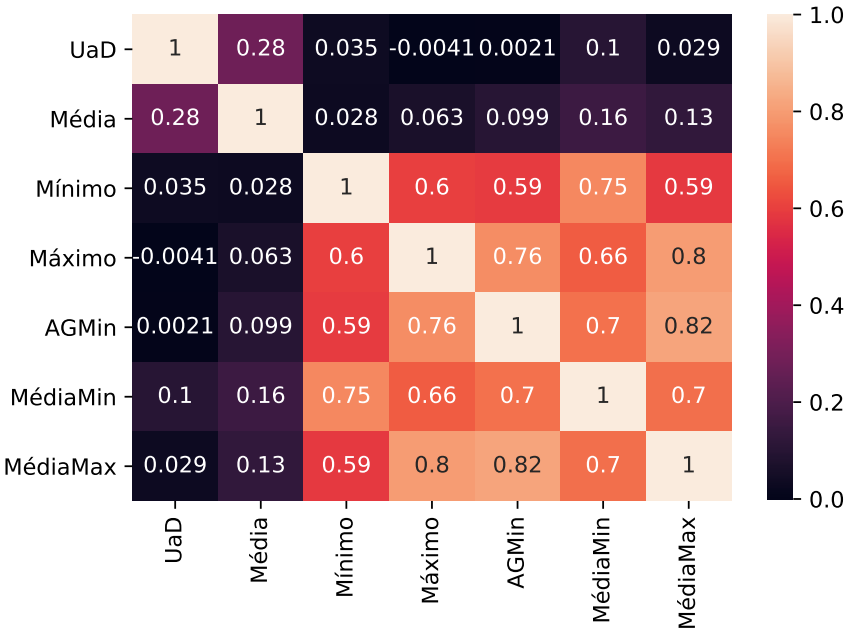

Figura V.1: Matriz de correlação entre as medidas usadas para estimar a similaridade entre os parlamentares.

V.2 Avaliação das Medidas de Distância Baseadas no USE

Na presente seção, abordaremos a condução de experimentos fundamentados na avaliação de medidas de distância baseadas no USE. Como delineado anteriormente, o USE representa uma ferramenta poderosa na análise semântica de sentenças, empregando uma abordagem de *embedding* para capturar representações semânticas significativas.

Com o propósito de avaliar a eficácia das medidas de distância oriundas do USE, foram feitas avaliações comparativas de frases distintas e antagônicas, quantificando a similaridade semântica entre elas. Essa abordagem não apenas evidencia a capacidade do USE em capturar nuances semânticas, mas também fornecerá uma base sólida para a validação das medidas de distância em análise.

A sentença identificada no conjunto de dados foi a seguinte: "*O PSD orienta 'sim', Sr. Presidente, pelo motivo que já manifestamos: o aumento da oferta de crédito e a oferta de crédito às micro e pequenas empresas.*". Duas proposições foram formuladas mediante a consideração da sentença previamente selecionada; estas representam uma orientação semântica oposta à sentença de origem.

Na primeira proposição formulada manualmente: "*O PSD orienta 'não', Sr. Presidente, pelo motivo que já manifestamos: a diminuição da oferta de crédito e a rejeição de crédito às micro e pequenas empresas.*"; aferiu-se uma medida de similaridade de 0,91, o que denota uma significativa correspondência semântica entre ambas as sentenças, não obstante sua natureza antagônica.

A segunda proposição foi formulada com auxílio do Modelo Largo de Linguagem (MLL) [OpenAI, 2022]: "*O PSD orienta 'não', Sr. Presidente, com base na preocupação de que um aumento descontrolado da oferta de crédito pode levar a riscos financeiros e instabilidade econômica, prejudicando a sustentabilidade a longo prazo.*". Nela foi realizado o cálculo de uma medida de similaridade de 0,36, evidenciando uma menor afinidade semântica. Tal divergência é atribuível à disparidade na exposição temática, aliada à distinção na extensão argumentativa do voto, embora os temas tratados guardem semelhança.

V.3 Análise da Distribuição dos Nós nos Clusters

Nesta seção, os resultados obtidos a partir da análise da distribuição dos parlamentares em *clusters* são apresentados. O objetivo primário desta etapa consistiu em avaliar o desempenho dos métodos de agrupamento partindo das duas técnicas utilizadas, USE e TF-IDF, na categorização dos parlamentares com base em seus perfis partidários.

V.3.1 Distribuição de Parlamentares Utilizando o Método TF-IDF

As Figuras V.2 e V.3 ilustram a matriz resultante da aplicação do método TF-IDF para a clusterização dos parlamentares, em números absolutos e proporcionais a cada partido, respectivamente. Cada célula da matriz representa a relação entre o partido original de um parlamentar (eixo X) e o partido ao qual foi atribuído no processo de clusterização (eixo Y). A matriz revela uma notável concentração de parlamentares no *cluster* 1, obtida utilizando a melhor combinação de parâmetros mostrada na Tabela V.1, que utiliza a abordagem Média e dimensionalidade do *embedding* 10^5 . Esta análise sugere que o método TF-IDF tendeu a agrupar um grande número de parlamentares sob um único *cluster*, indicando uma maior homogeneidade nos perfis políticos considerados.

Tal concentração na clusterização possivelmente decorre da limitação do método TF-IDF em compreender semanticamente os discursos, uma vez que este se baseia exclusivamente nas palavras utilizadas nos discursos. Dado o contexto comum em que todos os parlamentares se expressam,

termos políticos comuns e slogans podem estar contribuindo para dificultar uma distribuição de parlamentares mais equilibrada entre os *clusters*

Esta concentração em um *cluster* específico pode ter implicações relevantes para a análise política, sugerindo possíveis áreas de convergência ou polarização entre os partidos. A alta incidência de parlamentares em um único *cluster* também pode indicar a existência de temas ou agendas comuns entre os partidos considerados.

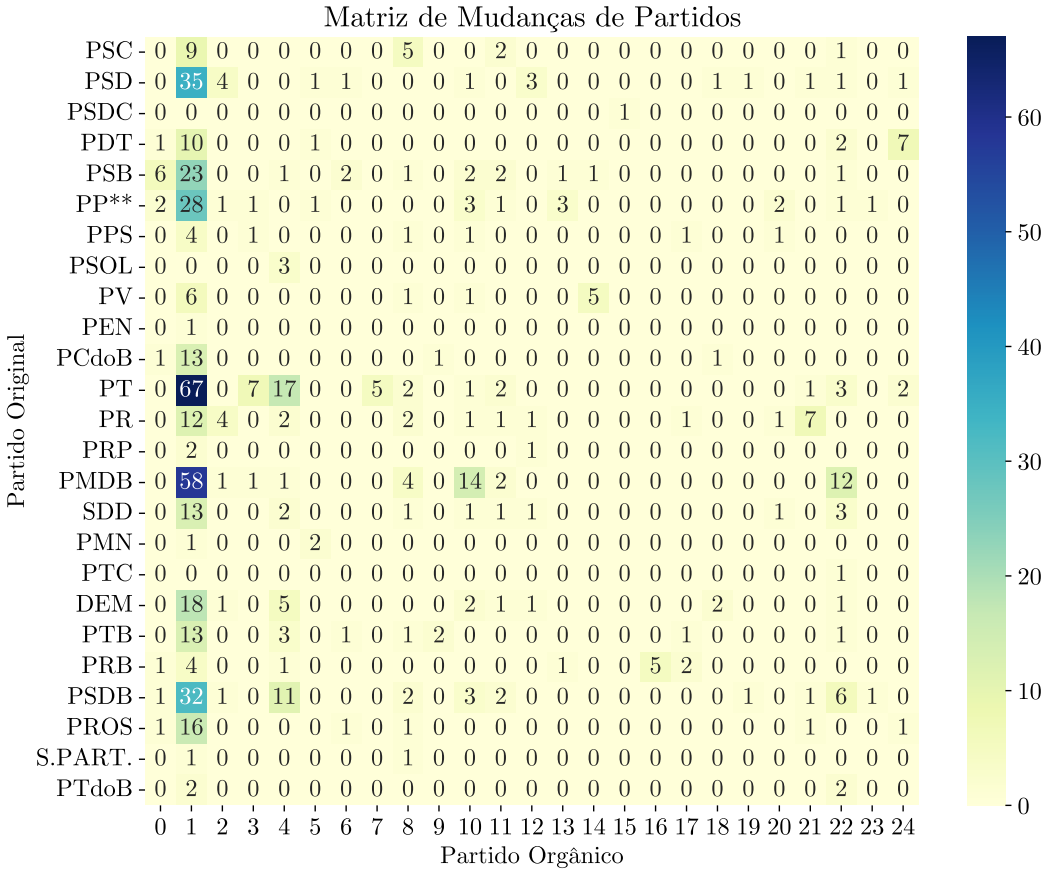

Figura V.2: Matriz de Mudança de Partido utilizando o método TF-IDF em números absolutos.

V.3.2 Distribuição de Parlamentares Utilizando o Método USE

Contrastando com o método TF-IDF, a aplicação do método USE, utilizando a agregação Média, embedding de dimensionalidade 512 e o algoritmo *Multilingual universal sentence encoder for semantic retrieval* [Yang et al., 2019], proposta neste trabalho, resultou em uma distribuição notavelmente equitativa dos parlamentares dentre os partidos orgânicos, como pode ser observado nas Figuras V.4 e V.5, que apresentam a matriz de confusão de mudança de partidos em números absolutos e proporcionais, respectivamente. Destaca-se que, para este método, observou-se uma

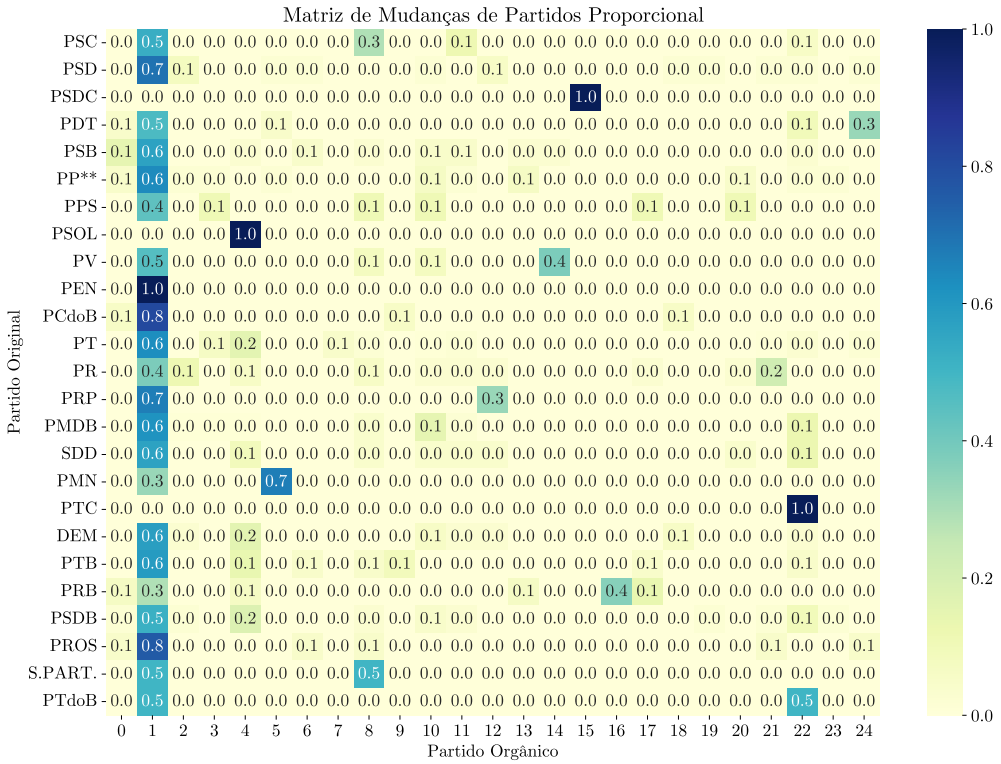

Figura V.3: Matriz de Mudança de Partido utilizando o método TF-IDF em números proporcionais.

propensão para que os parlamentares fossem dispersos de forma homogênea entre os diferentes *clusters*, evidenciando uma distribuição mais equilibrada dos representantes políticos.

Este resultado indica que o método USE proporcionou uma abordagem eficaz para a categorização dos parlamentares, minimizando a tendência de concentração em *clusters* específicos e refletindo, assim, a representatividade dos partidos originais.

A análise das matrizes resultantes fornece informações sobre o desempenho dos métodos de agrupamento utilizados. Enquanto o método USE demonstrou uma distribuição mais equitativa dos parlamentares entre os clusters, o método TF-IDF revelou uma tendência para a formação de um cluster dominante, ocasionando um *v-measure* alto e uma clusterização concentrada resultante do método TF-IDF . A partir da premissa de que os partidos originais estão corretos, parte-se do pressuposto de que agrupar discursos de indivíduos pertencentes ao mesmo partido em um único *cluster* representa a abordagem correta. Os resultados do USE sinalizam que o agrupamento produzido por esse método mostra-se mais coerente que o agrupamento dos partidos originais, explicitando a diferença entre partido original e partido baseado nos discursos.

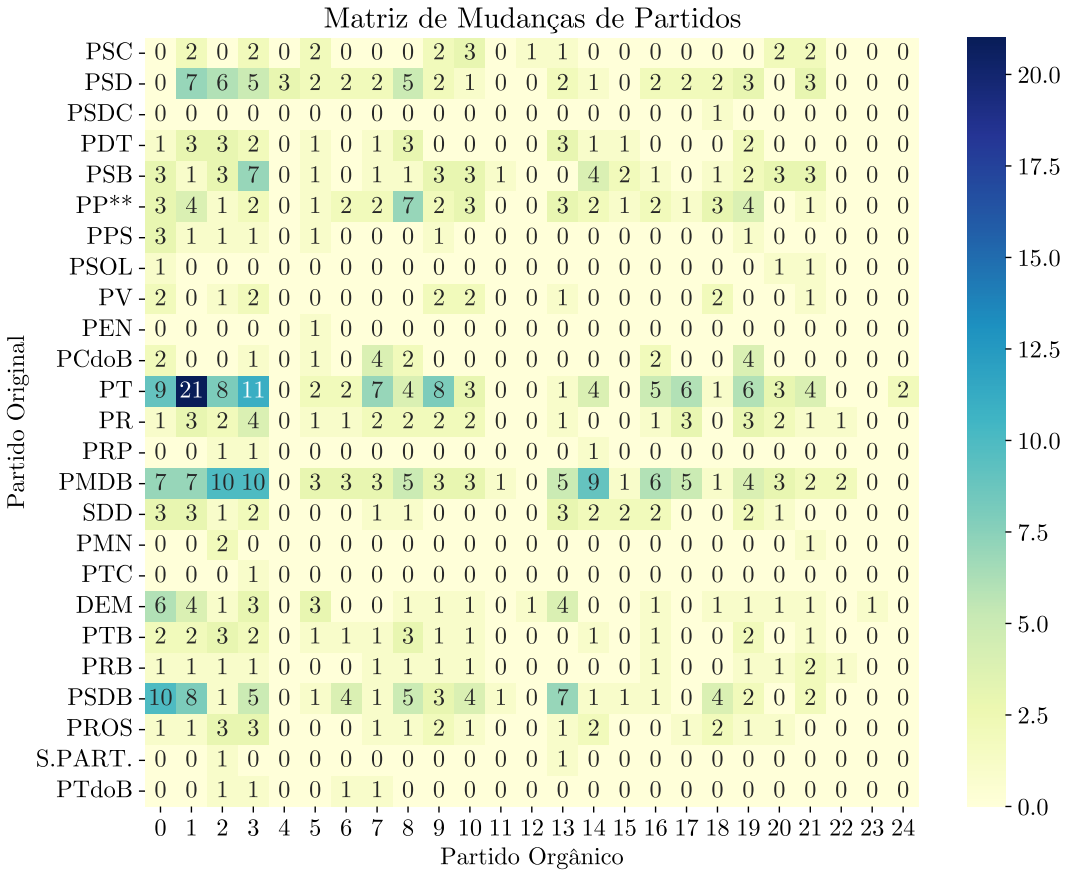

Figura V.4: Matriz de Mudança de Partido utilizando o método USE em número absoluto de parlamentares.

V.3.3 Análise da Média das Distâncias entre os Partidos

A avaliação e comparação de técnicas de representação de dados são cruciais para a obtenção de informações precisas e significativas em contextos analíticos. Neste capítulo é realizada a uma análise detalhada da eficácia de duas abordagens distintas na quantificação de similaridades semânticas entre discursos parlamentares, a saber, o esquema TF-IDF e o USE.

A Figura V.6 apresenta uma representação gráfica da média de distâncias utilizando o TF-IDF na forma de matriz de confusão, na qual os eixos X e Y denotam os diferentes partidos sob análise. Cada célula da matriz reflete a média das distâncias calculadas entre os discursos proferidos por cada par de parlamentares pertencentes aos respectivos partidos. Esta métrica de distância é fundamentada na comparação sintática dos discursos, uma vez que se baseia na dissimilaridade entre os termos e suas ponderações através do esquema TF-IDF. Observa-se que as médias de distância, notavelmente, convergem para valores próximos de 0,99 e 0,97, indicando uma tendência de similaridade acentuada entre os discursos analisados.

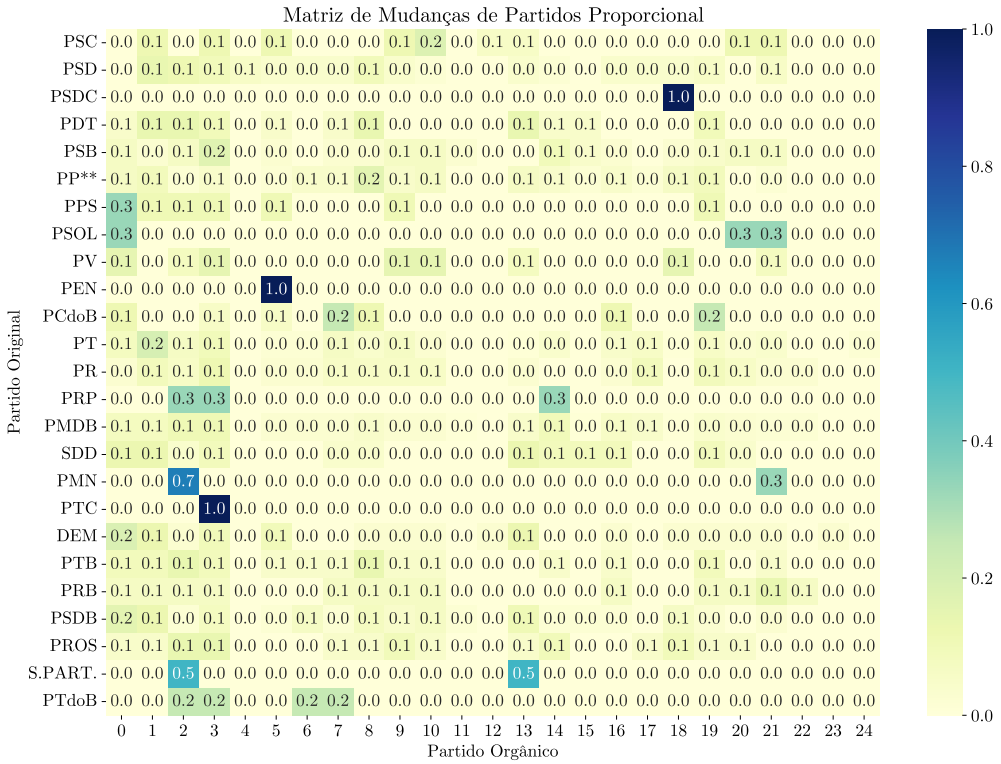

Figura V.5: Matriz de Mudança de Partido utilizando o método USE em número proporcional de parlamentares.

Tal resultado aponta para uma limitação significativa no desempenho do *TF-IDF* no que se refere à discriminação efetiva das nuances semânticas nos discursos parlamentares. A alta uniformidade nas médias de distância, sem distinção substancial entre os diferentes partidos, sugere que o *TF-IDF* tende a atribuir valores de distância relativamente elevados para todos os pares de discursos. Esta análise revela uma necessidade crítica de explorar abordagens alternativas para a quantificação das diferenças semânticas entre discursos parlamentares, a fim de obter uma representação mais precisa e discriminativa das relações interpartidárias.

Já a Figura V.7, oferece uma análise comparativa da representação das distâncias interpartidárias utilizando o USE. Assim como na Figura V.6, os eixos X e Y denotam os diferentes partidos, com cada célula da matriz representando a média das distâncias entre os discursos proferidos por todos os parlamentares por pares de partidos. Os resultados obtidos com a aplicação do USE são notavelmente distintos em relação à abordagem baseada em TF-IDF. As médias das distâncias variam entre 0.34 e 0.44, indicando uma dispersão significativamente maior nas medidas de dissimilaridade entre os discursos analisados. Este resultado demonstra a eficácia aprimorada do USE em destacar as diferenças semânticas subjacentes aos discursos proferidos por parlamentares de diferentes partidos.

A variação mais ampla nas médias de distância indica uma capacidade mais robusta do USE em discriminar com precisão as nuances semânticas nos discursos, resultando em uma representação mais fidedigna das relações interpartidárias. A significativa redução nas médias de distância, quando comparadas à abordagem baseada em TF-IDF, evidencia uma maior sensibilidade do USE à diversidade lexical e semântica presente nos discursos parlamentares.

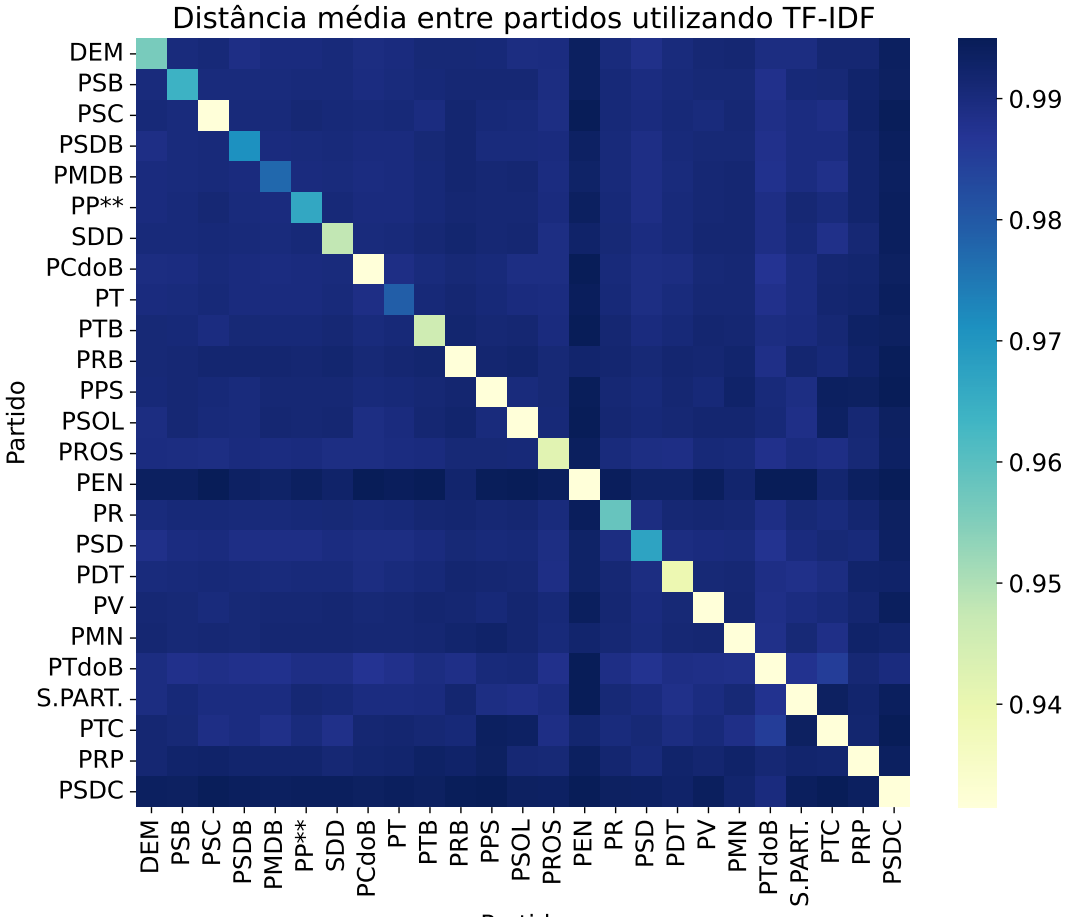

Figura V.6: Matriz de distância média entre partidos utilizando TF-IDF.

V.4 Análise da Coesão Partidária

Nesta seção são exibidos nas figuras V.8 e V.9 os resultados da análise da coesão partidária baseada na medida de distância parlamentar em sua já detalhada configuração ótima (Média) utilizando TF-IDF e USE, respectivamente. A ordem dos partidos está disposta da esquerda para a direita do maior para o menor coeficiente de agrupamento do subgrafo induzido pelos nós relativos aos parlamentares de cada partido. O valor do coeficiente de agrupamento é exibido na legenda no eixo Y, juntamente com a sigla do partido. O coeficiente de agrupamento avalia o grau com que os nós de um grafo tendem a coligar-se, considerando a distância entre os nós de um mesmo

grupo, com grande aplicação em grafos densos [Hagberg et al., 2008]. Partidos cujos parlamentares tem discursos com pequenas distâncias entre si, terão coeficientes de agrupamento maiores, e de forma análoga, coeficientes de agrupamento menores indicam maiores distâncias entre os discursos dos parlamentares.

As barras vermelhas representam o tamanho do partido em quantidade de parlamentares, enquanto as barras azuis caracterizam a entropia presente na distribuição dos parlamentares de um partido em grupos. Um valor de entropia zero significa que todos os membros de um partido foram agrupados em um mesmo cluster, e à medida que os membros de um partido foram dispersados em mais grupos o valor da entropia cresce.

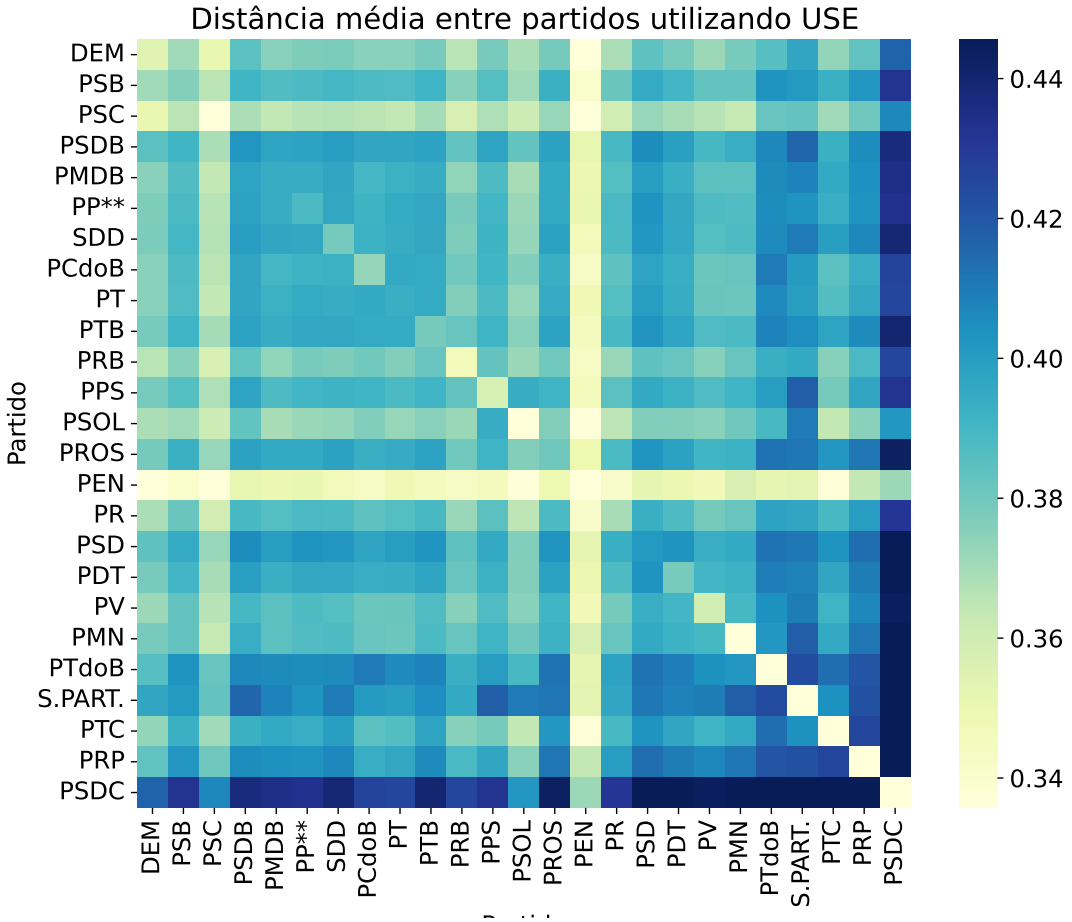

Figura V.7: Matriz de distância média entre partidos utilizando USE.

V.4.1 Análise da Coesão Partidária com TF-IDF

A partir da Figura V.8 é possível analisar a relação entre o tamanho do partido, o coeficiente de agrupamento e a entropia alcançados utilizando o grafo de distâncias entre os parlamentares e o agrupamento. É esperado que partidos com maiores quantidades de parlamentares tenham entropia maior e coeficiente de agrupamento menores, resultando em uma menor coesão partidária,

pois existe uma chance maior de divergências entre os seus parlamentares, devido ao tamanho do partido a que pertencem. É possível observar que há casos onde esta expectativa se confirma, como no segundo partido com melhor coeficiente de agrupamento, o Partido Socialismo e Liberdade (PSOL), cuja quantidade de parlamentares é pequena e sua entropia e coeficiente de agrupamento também são pequenos. Outros exemplos de partidos que também seguem esta expectativa são o Partido dos Trabalhadores (PT), o Partido Socialista Brasileiro (PSB) e o Progressistas (PP), onde a quantidade de parlamentares é grande, e a entropia também é alta e o coeficiente de agrupamento é menor. Há também partidos que não correspondem a este comportamento, como o PRP e o Partido Trabalhista do Brasil (PTdoB), que tem um número bastante reduzido de parlamentares e uma entropia relativamente alta e o coeficiente de agrupamento baixo, indicando discursos pouco similares e uma grande distância entre os discursos dos parlamentares de um mesmo partido.

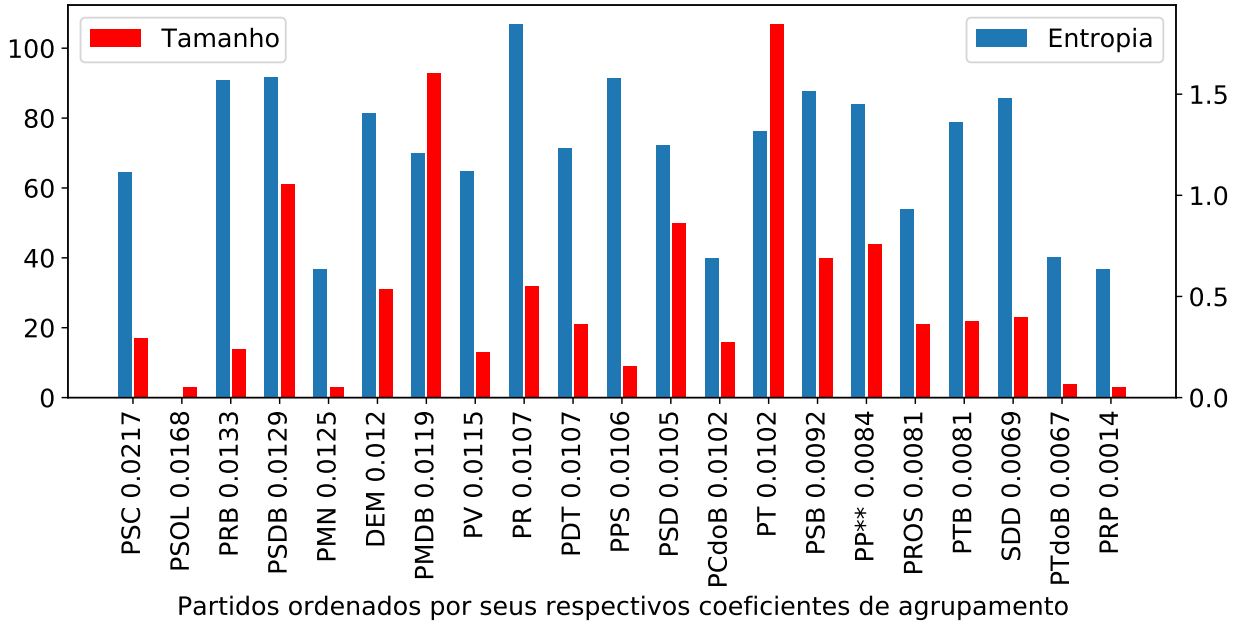

Figura V.8: Avaliação da similaridade entre parlamentares utilizando o método TF-IDF.

O Partido do Movimento Democrático Brasileiro (PMDB) e o Partido da Social Democracia Brasileira (PSDB), partidos com grandes quantidades de parlamentares, são exemplos de partidos com um valor relativamente alto de coeficiente de agrupamento, e com um valor alto de entropia, indicando distâncias pequenas entre os parlamentares de um único partido, e também uma dispersão dos parlamentares de um partido em outros partidos. Isto significa que o processo de agrupamento difundiu os parlamentares desses partidos em diversos outros grupos, apesar da semelhança entre os discursos dos seus parlamentares em média. Isto pode ser visto como uma evidência da existência de alas coesas e distantes entre si nestes partidos.

V.5 Análise da Coesão Partidária com USE

A Figura V.9 apresenta os resultados da análise da coesão partidária fundamentada na aplicação do algoritmo USE. É possível observar uma entropia e um coeficiente de agrupamento médio maior que a encontrada na Figura V.8 . A hipótese inicial postulava que o coeficiente de agrupamento e a entropia seriam métricas adequadas para avaliar a coesão dos partidos políticos, partindo do fato que os partidos originais seriam os corretos. No caso do TF-IDF, observou-se inicialmente uma baixa entropia, o que sugere uma forte coesão entre os membros de um mesmo partido. No entanto, análises realizadas na Seção V.3 revelam que essa baixa entropia pode ser atribuída à tendência do TF-IDF de agrupar a maioria dos parlamentares em um único nó (cluster), implicando em uma interpretação distorcida da coesão partidária real. Em contrapartida, ao empregar o USE, mesmo observando que a entropia dos partidos tenha apresentado valores mais elevados, os resultados foram substancialmente mais informativos e representativos. O USE demonstrou uma capacidade mais equitativa de clusterização, levando em consideração de maneira mais acurada a distância semântica entre os discursos dos parlamentares. Esta característica foi determinante para uma representação mais confiável da verdadeira coesão partidária.

É possível observar também na Figura V.9 que a entropia é alta para todos os partidos, mesmo considerando partidos pequenos ou com grande coeficiente de agrupamento. Este fenômeno indica que apesar da proximidade *intra-clusters*, parlamentares de um mesmo partido próximos entre si são agrupados em *clusters* diferentes, pois apesar da pequena distância *intra-cluster*, a distância *inter-cluster* tende a ser menor.

V.6 Análise de Clusterização para Identificação do Número Ideal de Partidos

A fim de refinar a compreensão da dinâmica partidária no contexto em estudo, foram realizados procedimentos de clusterização sobre o conjunto de dados contendo os discursos dos 21 partidos políticos inicialmente considerados. A clusterização é uma técnica estatística amplamente empregada para agrupar elementos similares em subconjuntos distintos, permitindo assim uma visão mais clara das tendências e padrões subjacentes.

Para determinar o número ótimo de *clusters* que melhor representam a estrutura subjacente dos dados, empregou-se a métrica de silhueta. A silhueta mede a coesão *intra-cluster* e a separação *inter-cluster*, proporcionando uma avaliação objetiva da validade dos agrupamentos. A partir de uma sequência de experimentos, variando o número de *clusters* de 2 a 19, obteve-se o gráfico de cotovelo apresentado na Figura V.10.

A Figura V.10 ilustra a relação entre o número de *clusters* e a pontuação de silhueta e *V-Measure* correspondente. Este comportamento revela um cenário peculiar com base na silhueta,

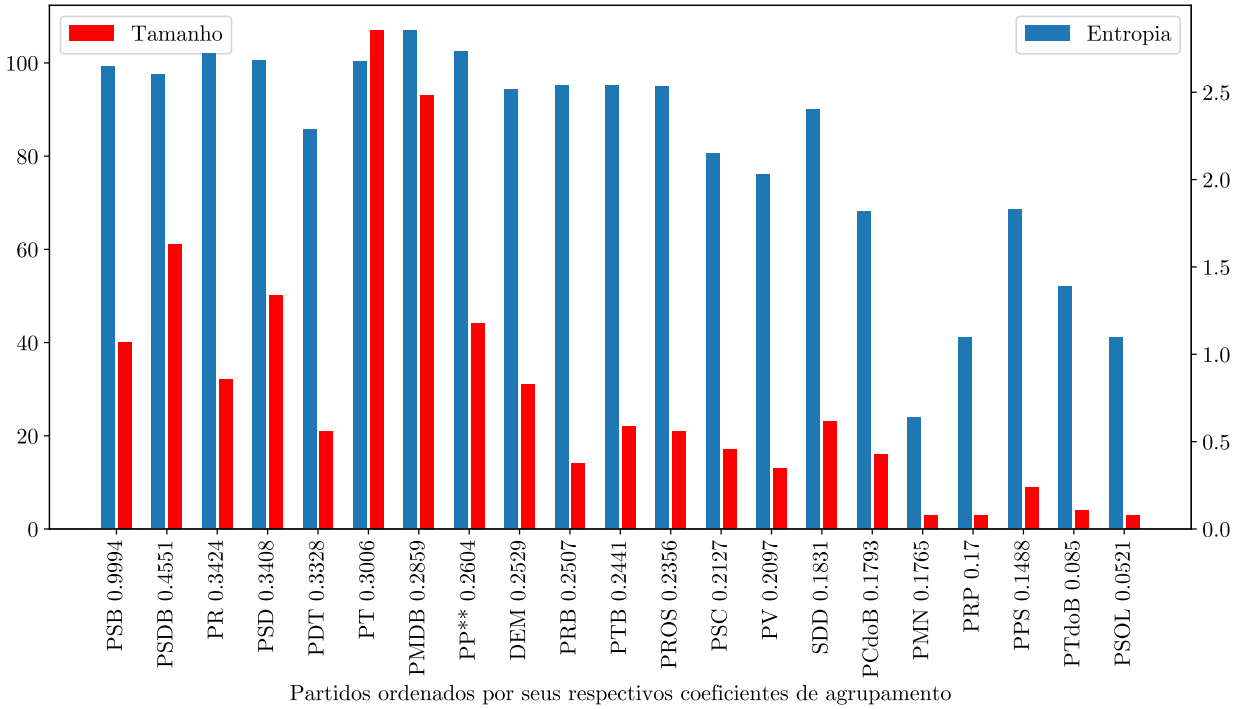

Figura V.9: Avaliação da similaridade entre parlamentares utilizando o método USE.

indicando que o aumento do número de *clusters* além de um determinado ponto pode resultar em agrupamentos menos representativos e coesos . Este fenômeno é crucial para a determinação da quantidade ideal de *clusters* que melhor reflete a estrutura subjacente dos dados de discursos. Ao examinar o gráfico do índice *V-Measure*, evidencia-se a correlação positiva entre o aumento deste índice e a ampliação simultânea da completude e homogeneidade do *cluster*. Este fenômeno sugere que o incremento no número de *clusters* até um ponto específico pode resultar na formação de agrupamentos mais representativos e coesos.

Neste contexto, observando gráfico apresentado na Figura V.10 o ponto não está claramente definido, e a interpretação do número ótimo de *clusters* torna-se uma consideração mais complexa. Este resultado sublinha a necessidade de uma abordagem cuidadosa na definição do número de *clusters*, visando evitar a divisão excessiva dos parlamentares, criando assim uma quantidade maior de partidos, com o intuito de preservar a representatividade dos *clusters* identificados.

Dessa maneira, a análise de clusterização aliada à métrica de silhueta proporcionou uma perspectiva valiosa sobre a estrutura dos dados de discursos. A compreensão da relação entre a pontuação de silhueta e o número de *clusters* é um passo crítico na interpretação dos padrões discursivos emergentes no conjunto de dados em questão.

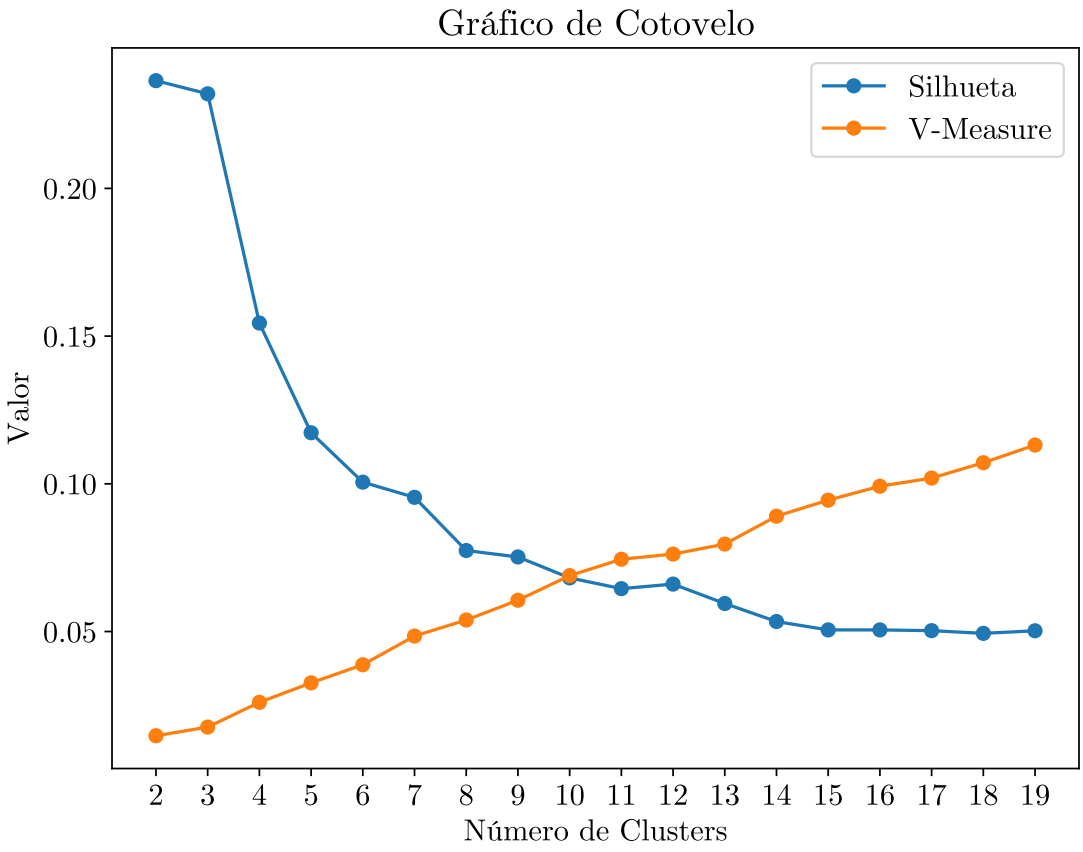

Figura V.10: Gráfico de cotovelo variando a quantidade de clusters utilizando o método USE.

Capítulo VI Conclusões

Analisar as relações entre deputados e partidos é uma atividade interessante para a democracia, pois fornece dados sobre os deputados e possíveis conflitos de interesse. Neste trabalho foi analisada a relação e a coesão partidária dos parlamentares da Câmara dos Deputados ativos durante a 54^a legislatura, que compreende os anos entre 2011 a 2015, de acordo com seus discursos sob duas metodologias, o USE e o TF-IDF, bem como uma análise sobre a quantidade ideal de partidos segundo agrupamentos por discursos. Para a modelagem das relações entre os deputados foram utilizados os conceitos de redes complexas, e para a análise da coesão partidária foi utilizado agrupamento hierárquico aglomerativo, uma abordagem de aprendizado de máquina não-supervisionada.

A metodologia proposta avalia a afinidade entre qualquer par de parlamentares com base nos conjuntos de discursos de cada um destes. Para isso foi considerado um grafo bipartido completo para cada par de deputados cujos nós seriam seus discursos, divididos em duas partições, cada uma representando um dos parlamentares do par analisado no momento. As arestas destes grafos bipartidos são ponderadas, sendo os valores dos pesos referente à similaridade textual dos respectivos discursos, calculados utilizando USE ou TF-IDF, representados pelos vértices localizados em suas extremidades. A afinidade entre dois parlamentares pode então ser calculada pela agregação destes pesos conforme uma das seis alternativas de agregação apresentadas na Seção IV.3 deste trabalho e consideradas para tal tarefa.

Os resultados obtidos com a metodologia proposta utilizando TF-IDF foram superiores à medida de referência previamente estabelecida na literatura. Neste ínterim, o método de agregação que obteve melhores resultados foi aquele relativo à média das similaridades de cada par de discursos de quaisquer dois parlamentares, aqui chamado apenas de Média. Em contrapartida o USE demonstrou uma capacidade mais equitativa de clusterização, levando em consideração de maneira mais acurada a distância semântica entre os discursos dos parlamentares, minimizando a tendência de concentração em *clusters* específicos e refletindo, assim, a representatividade dos partidos originais. Esta característica foi determinante para uma representação mais confiável da verdadeira coesão partidária, utilizando também o método aqui chamado Média.

A pesquisa em questão abre portas para futuras investigações e aprimoramentos na área. Uma área promissora para trabalhos futuros seria a comparação direta entre o TF-IDF e o *Universal Sentence Encoder*, utilizando tamanhos de *embedding* uniformes. Essa abordagem permitiria uma

análise mais precisa e justa do desempenho relativo dos modelos, uma vez que a dimensão dos *embeddings* pode influenciar significativamente os resultados.

Referências

Abercrombie, G. and Batista-Navarro, R. T. 'aye'or 'no'? speech-level sentiment analysis of hansard uk parliamentary debate transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Aires, V., da Silva, A., Nakamura, F., and Nakamura, E. An evaluation of structural characteristics of networks to identify media bias in news portals. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 225–232, 2020.

Alajmi, A., Saad, E., and Darwish, R. Toward an arabic stop-words list generation. *International Journal of Computer Applications*, 46(8):8–13, 2012.

Alanis, A. Y., Hernandez-Vargas, E. A., Ramirez, N. F., and Rios-Rivera, D. Neural Control for Epidemic Model of Covid-19 with a Complex Network Approach. *IEEE Latin America Transactions*, 19(6):866–873, 2021.

Balahur, A., Kozareva, Z., and Montoyo, A. Determining the polarity and source of opinions expressed in political debates. In *Computational Linguistics and Intelligent Text Processing: 10th International Conference, CICLing 2009, Mexico City, Mexico, March 1-7, 2009. Proceedings 10*, pages 468–480. Springer, 2009.

Barabási, A.-L. and Pósfai, M. *Network science*. Cambridge University Press, Cambridge, United Kingdom, 2016.

Beel, J., Gipp, B., Langer, S., and Breitinger, C. Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.*, 17(4):305–338, 2016.

Biessmann, F. Automating political bias prediction. *arXiv preprint arXiv:1608.02195*, 2016.

Bird, S. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

Bondy, J. A. and Murty, U. S. R. *Graph Theory with Applications*. Macmillan Education UK, 1976.

Brito, A. C. M., Silva, F. N., and Amancio, D. R. A complex network approach to political analysis: Application to the brazilian chamber of deputies. *Plos one*, 15(3):e0229928, 2020.

Bursztyn, V. S., Nunes, M. G., and Figueiredo, D. R. How brazilian congressmen connect: homophily and cohesion in voting and donation networks. *Journal of Complex Networks*, 8(1):cnaa006, 2020.

Caetano, J. A., Lima, H. S., Santos, M. F., and Marques-Neto, H. T. Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 american presidential election. *Journal of internet services and applications*, 9(1):1–15, 2018.

Camacho-Collados, J. and Pilehvar, M. T. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780*, 2017.

Cardoso, D. O., Lima, W. P., Silva, G. G., and Assis, L. S. An approach for probabilistic modeling and reasoning of voting networks. In *International Conference on Computational Science*, pages 90–104. Springer, 2023.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Chen, W., Zhang, X., Wang, T., Yang, B., and Li, Y. Opinion-aware knowledge graph for political ideology detection. In *IJCAI*, volume 17, pages 3647–3653, 2017.

Cherepnalkoski, D., Karpf, A., Mozetič, I., and Grčar, M. Cohesion and coalition formation in the european parliament: Roll-call votes and twitter activities. *PloS one*, 11(11):e0166586, 2016.

Cossu, J.-V., Labatut, V., and Dugué, N. A review of features for the discrimination of twitter users: Application to the prediction of offline influence. *Social Network Analysis and Mining*, 6(1):25, 2016.

Cristiani, A., Lieira, D., and Camargo, H. A sentiment analysis of brazilian elections tweets. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 153–160. SBC, 2020.

Dal Maso, C., Pompa, G., Puliga, M., Riotta, G., and Chessa, A. Voting behavior, coalitions and government strength through a complex network analysis. *PloS one*, 9(12):e116046, 2014.

Dal Maso, C., Pompa, G., Puliga, M., Riotta, G., and Chessa, A. Voting behavior, coalitions and government strength through a complex network analysis. *PLOS ONE*, 9(12):1–13, 2015.

Dallmann, A., Lemmerich, F., Zoller, D., and Hotho, A. Media bias in german online newspapers. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 133–137, 2015.

Dezembro, D. G. *Uma medida de similaridade híbrida para correspondência aproximada de múltiplos padrões*. PhD thesis, Universidade de São Paulo, 2019.

Dias, M., Braz, P., Bezerra, E., and Goldschmidt, R. Contextual Information Based Community Detection in Attributed Heterogeneous Networks. *IEEE Latin America Transactions*, 17(02):236–244, 2019.

Dolamic, L. and Savoy, J. When stopword lists make the difference. *Journal of the American Society for Information Science and Technology*, 61(1):200–203, 2010.

dos Deputados, C. Dados abertos da câmara dos deputados. `https://dadosabertos.camara.leg.br/swagger/api.html`, 2021.

dos Santos, J. C. and Siqueira, S. W. M. Classificação de opinião no twitter em português utilizando o multilingual universal sentence encoder para apoiar pesquisas sobre filter bubble. In *Anais do XV Simpósio Brasileiro de Sistemas Colaborativos*, pages 68–73. SBC, 2019.

Elejalde, E., Ferres, L., and Herder, E. The nature of real and perceived bias in chilean media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 95–104, 2017.

Fernandes, M. S. Tenho dito: uma aplicação para análise de discursos parlamentares utilizando técnicas de processamento de linguagem natural, 2017.

Gerrish, S. M. and Blei, D. M. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 2011.

Gomes Ferreira, C. H., de Sousa Matos, B., and Almeira, J. M. Analyzing dynamic ideological communities in congressional voting networks. In *Social Informatics: 10th International Conference, SocInfo 2018, St. Petersburg, Russia, September 25-28, 2018, Proceedings, Part I 10*, pages 257–273. Springer, 2018.

Greene, D. and Cross, J. P. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, 25(1):77–94, 2017.

Hagberg, A., Swart, P., and S Chult, D. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Halberstam, Y. and Knight, B. Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of public economics*, 143:73–88, 2016.

Indurkhya, N. and Damerau, F. J. *Handbook of natural language processing*. Chapman and Hall/CRC, 2010.

Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60(5):493–502, 2004.

Kirkland, J. H. and Gross, J. H. Measurement and theory in legislative networks: the evolving topology of congressional collaboration. *Social Networks*, 36, 2014.

Kummervold, P. E., De la Rosa, J., Wetjen, F., and Brygfjeld, S. A. Operationalizing a national digital library: The case for a norwegian transformer model. *arXiv preprint arXiv:2104.09617*, 2021.

Lee, S. H., Magallanes, J. M., and Porter, M. A. Time-dependent community structure in legislation cosponsorship networks in the congress of the republic of peru. *Journal of Complex Networks*, 5(1):127–144, 2016.

Lima, W. P., Marques, L. C., Assis, L. S., and Cardoso, D. O. An analysis of political parties cohesion based on congressional speeches. In *International Conference on Computational Science*, pages 105–119. Springer, 2023.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

Lovins, J. B. Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, 11(1-2):22–31, 1968a.

Lovins, J. B. Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, 11(1-2):22–31, 1968b.

Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to information retrieval*. Cambridge University Press, 2008.

Menini, S. and Tonelli, S. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2461–2470, 2016.

Metz, J., Calvo, R., Seno, E. R., Romero, R. A. F., Liang, Z., et al. Redes complexas: conceitos e aplicações., 2007.

Nay, J. J. Predicting and understanding law-making with word vectors and an ensemble model. *PloS one*, 12(5):e0176999, 2017.

Neal, Z. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Soc. Networks*, 39:84–97, 2014.

Onnela, J.-P., Saramäki, J., Kertész, J., and Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, 2005.

OpenAI. Gpt-3.5. `https://www.openai.com`, 2022.

Pastor-Galindo, J., Zago, M., Nespoli, P., Bernal, S. L., Celdrán, A. H., Pérez, M. G., Ruipérez-Valiente, J. A., Pérez, G. M., and Mármol, F. G. Spotting political social bots in twitter: A use case of the 2019 spanish general election. *IEEE Transactions on Network and Service Management*, 17(4):2156–2170, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rao, A. and Spasojevic, N. Actionable and political text classification using word embeddings and lstm. *arXiv preprint arXiv:1607.02501*, 2016.

Rosenberg, A. and Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

Saif, H., Fernández, M., He, Y., and Alani, H. On stopwords, filtering and data sparsity for sentiment analysis of twitter, 2014.

Saligrama, A. Knowbias: Detecting political polarity in long text content. *arXiv preprint arXiv:1909.12230*, 2019.

Saramäki, J., Kivelä, M., Onnela, J.-P., Kaski, K., and Kertesz, J. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007.

Schwarz, a., Traber, D., and Benoit, K. Estimating intra-party preferences: comparing speeches to votes. *Political Science Research and Methods*, 5(2):379–396, 2017.

Silva, N. F. d., Silva, M. C. R., Pereira, F. S., Tarrega, J. P. M., Beinotti, J. V. P., Fonseca, M., Andrade, F. E. d., and de Carvalho, A. C. d. L. Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 104–120. Springer, 2021.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 178–185, 2010.

Van Dijk, T. A. Political discourse and ideology. *Doxa Comunicación. Revista interdisciplinar de estudios de comunicación y ciencias sociales*, (1):207–225, 2003.

Van Dijk, T. A. Discourse and ideology. *Discourse studies: A multidisciplinary introduction*, pages 379–407, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

Wives, L. K. Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de"clustering", 1999.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*, 2019.

Zhitomirsky-Geffet, M., David, E., Koppel, M., and Uzan, H. Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites. *Online Information Review*, 2016.