SPEAR: Um Método para Descoberta de Padrões Psicolinguísticos e Identificação de Discursos Presidenciais Destoantes

Isabele Figueiredo[1], Matheus Amendoeira[1], Kele Belloze[1], Eduardo Bezerra[1], Eduardo Ogasawara[1], Gustavo Guedes[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

{isabele.figueiredo, matheus.amendoeira}@aluno.cefet-rj.br,
{kele.belloze, eduardo.bezerra}@cefet-rj.br,
eogasawara@ieee.org, gustavo.guedes@cefet-rj.br

Resumo. *A análise computacional de textos tem ampliado a investigação de como escolhas linguísticas revelam traços de comunicação e posicionamentos políticos. No entanto, ainda há lacunas na identificação de padrões psicolinguísticos em mandatos presidenciais brasileiros e na detecção de discursos destoantes. Para enfrentar essa lacuna, este trabalho propõe o SPEAR, um método que combina extração de dimensões do LIWC, remoção de outliers via LOF e agrupamento com K-Means, seguido da reintrodução dos discursos mais destoantes. Os resultados indicam melhora consistente da qualidade dos agrupamentos em 11 mandatos analisados.*

Abstract. *Computational text analysis has expanded the investigation of how linguistic choices reveal communication traits and political positions. However, gaps remain in the identification of psycholinguistic patterns across complete Brazilian presidential terms and in the detection of deviant speeches. To address this gap, this paper proposes SPEAR, a method that combines LIWC dimension extraction, LOF-based outlier removal, and K-Means clustering, followed by the reinsertion of the most deviant speeches. The results indicate a consistent improvement in clustering quality across the 11 analyzed terms.*

1. Introdução

A análise textual tem se consolidado como um recurso relevante nos estudos políticos, pois permite compreender tanto as dinâmicas de circulação de mensagens políticas nas redes sociais [Gadelha et al., 2023; Nascimento et al., 2025] quanto aquilo que os discursos de representantes políticos revelam sobre os próprios atores [Ahmadian et al., 2017] e sobre os temas em debate [Silva-Muller and Sposito, 2024]. O crescimento do número de artigos publicados sobre análise de discursos políticos nos últimos anos evidencia esse interesse crescente [Randour et al., 2020].

Esse crescimento decorre, em parte, do desenvolvimento de técnicas computacionais, como Processamento de Linguagem Natural, algoritmos para análises quantitativas, estatística de texto e aprendizado de máquina, que automatizam essas análises e as tornam mais escaláveis e sistemáticas [Boyd, 2017]. Dentre os métodos computacionais de análise textual, o *Linguistic Inquiry and Word Count (LIWC)* destaca-se por quantificar, em diferentes dimensões (*e.g.*, pronomes impessoais e verbos auxiliares), a frequência relativa de uso das palavras em um texto [Tausczik and Pennebaker, 2010].

Na literatura, observa-se uma lacuna na análise de discursos presidenciais brasileiros com foco em períodos delimitados, como mandatos, e nas características que emergem ao longo desses intervalos [Randour et al., 2020]. Em particular, ainda são escassas as abordagens que permitam caracterizar, dentro de um mesmo mandato, padrões psicolinguísticos predominantes e identificar, de forma sistemática, discursos que deles se afastam sem comprometer a consistência da análise. Neste trabalho, entende-se por padrão psicolinguístico a recorrência de combinações semelhantes entre as dimensões do LIWC em um conjunto de discursos. Para enfrentar essa lacuna, este trabalho propõe o método *Speech Psycholinguistic Extraction and Anomaly Recognition (SPEAR)*.

O método *SPEAR* parte da extração, em cada discurso, das dimensões do LIWC, que neste trabalho funcionam como representação psicolinguística dos textos. A partir dessa representação, identifica-se a estrutura local dos dados com o *Local Outlier Factor (LOF)*, removendo temporariamente observações anômalas antes do agrupamento por K-Means. Após a formação dos grupos, esses discursos são reinseridos para análise como discursos destoantes. O método é aplicado a um conjunto de dados composto por discursos presidenciais brasileiros proferidos entre 1985 e 2020, organizados por mandato.

As principais contribuições do trabalho são o desenvolvimento do método *SPEAR*, sua avaliação em 11 mandatos presidenciais brasileiros com base na métrica de silhueta e a realização de uma análise qualitativa no segundo mandato de Lula. Nesse contexto, o trabalho busca verificar se a abordagem melhora a qualidade dos agrupamentos e favorece a interpretação dos padrões predominantes e dos discursos que deles se afastam, preservando esses discursos como objeto analítico em vez de tratá-los apenas como ruído.

Além desta introdução, a Seção 2 contextualiza o problema investigado. A Seção 3 apresenta os trabalhos relacionados. A Seção 4 descreve o método *SPEAR*. Por fim, as Seções 5 e 6 apresentam, respectivamente, os resultados obtidos e as conclusões.

2. Fundamentação Teórica

Esta seção reúne os principais fundamentos do trabalho. O LIWC é empregado para a extração de dimensões psicolinguísticas, a detecção de *outliers* é utilizada para separar observações anômalas antes do agrupamento, e o *Bidirectional Encoder Representations from Transformers Topic (BERTopic)* é adotado para a interpretação temática dos grupos.

2.1. Linguistic Inquiry and Word Count (LIWC)

O *Linguistic Inquiry and Word Count (LIWC)* é um programa de análise textual que, com base em um dicionário, realiza a contagem de palavras correspondentes a cada categoria, gera sua frequência e atribui uma porcentagem de uso de cada categoria em relação ao total de palavras [Tausczik and Pennebaker, 2010]. Esse software fornece informações tanto sobre o uso de diferentes tipos de palavras quanto sobre as emoções associadas aos textos. Ele é formado por dimensões linguísticas, como pronomes e artigos, e por dimensões psicológicas, como tristeza e raiva, definidas em um dicionário que contém uma lista extensa de palavras e construções verbais categorizadas [Carvalho et al., 2019].

Além dessas dimensões básicas, existem outras quatro dimensões compostas: autenticidade, pensamento analítico, influência e tom emocional. Essas dimensões fornecem métricas mais complexas, formadas pela combinação e pela normalização das dimensões

básicas. Cada dimensão composta é derivada por um algoritmo proprietário, não divulgado, que combina e analisa os valores de frequência das dimensões básicas. O resultado gera uma pontuação padronizada em percentil, que varia de 1 a 99 e indica presença alta ou baixa da dimensão analisada no texto [Tausczik and Pennebaker, 2010]. Nesse estudo, utiliza-se o *Brazilian Portuguese LIWC dictionary* (BP-LIWC2015), versão mais recente do dicionário LIWC em português brasileiro, que apresenta mais categorias e maior correlação com a versão em inglês do que a versão anterior [Carvalho et al., 2019, 2024].

2.2. Detecção de Outliers

Em contextos de agrupamento, a presença de *outliers* pode distorcer a formação dos grupos e gerar agrupamentos compostos inteiramente por observações atípicas [Aggarwal, 2016]. A detecção prévia dessas observações contribui para grupos mais coesos e representativos. O *Local Outlier Factor (LOF)* [Breuniq et al., 2000] é um método não supervisionado que avalia a densidade local da vizinhança de cada ponto, identificando anomalias em relação ao contexto local sem pressupor distribuição dos dados, o que o diferencia de métodos baseados em distribuições globais ou limiares fixos de distância.

Por operar de forma não paramétrica, o LOF adapta-se a espaços de alta dimensionalidade com distribuições irregulares. O algoritmo detecta com eficácia *outliers* locais, isto é, pontos anômalos mesmo em regiões densas do espaço. Sua principal limitação é o custo computacional elevado em conjuntos grandes, já que o cálculo de densidade local exige consulta à vizinhança de cada ponto.

2.3. BERTopic

O *Bidirectional Encoder Representations from Transformers Topic (BERTopic)* [Grootendorst, 2022] identifica tópicos com modelos de linguagem pré-treinados, preservando o contexto semântico dos documentos e gerando tópicos densos e interpretáveis. Seu *pipeline* opera em três etapas: os documentos são convertidos em *embeddings* via *Transformers*; a dimensionalidade é reduzida com UMAP [McInnes et al., 2020]; e os documentos são agrupados com HDBSCAN [Malzer and Baum, 2020], que identifica grupos de densidade variável e classifica como *outliers* documentos sem correlação semântica suficiente.

Para representar cada tópico, o BERTopic utiliza a métrica C-TF/IDF, que trata todos os documentos de um grupo como um único documento combinado c. A pontuação de importância $W_{x,c}$ de uma palavra x no tópico c é dada pela Equação 1:

$$W_{x,c} = ||tf_{x,c}|| \times \log\left(1 + \frac{A}{f_x}\right) \tag{1}$$

Onde $tf_{x,c}$ é a frequência de x no agrupamento c, f_x é a frequência total de x em todos os grupos e A é a média do número de palavras por grupo. Assim, destacam-se termos frequentes em um tópico específico e raros nos demais.

3. Trabalhos Relacionados

Esta seção revisa trabalhos relacionados à análise de discursos políticos. Os estudos foram identificados por meio da estrutura *Population, Exposure, Comparator, Outcome (PECO)* [Morgan et al., 2018], aplicada ao *Elicit* [Bernard et al., 2025]. Nesse processo, a população correspondeu a discursos políticos, a exposição a métodos computacionais de análise

psicolinguística, o comparador a diferentes atores, períodos ou abordagens, e o desfecho à identificação de padrões e desvios discursivos. A busca foi guiada pela seguinte pergunta: *How have computational methods of psycholinguistic analysis, such as LIWC, been used to identify patterns and deviations in political speeches?*

Os trabalhos que empregam o LIWC concentram-se na identificação de traços estilísticos, psicológicos e de identidade política em discursos de lideranças específicas. Nesse eixo, observam-se estudos voltados à comparação entre líderes e candidatos, com ênfase em diferenças de estilo de liderança, autenticidade, pensamento analítico e assertividade [Körner et al., 2022; Kangas, 2014; Jordan et al., 2019; Okuno et al., 2018]. Também há investigações sobre tendências de longo prazo na linguagem de lideranças políticas e sobre mudanças no *messaging* político ao longo do tempo [Jordan et al., 2019; Stromer-Galley et al., 2021]. Em conjunto, esses trabalhos mostram que o LIWC é útil para capturar variações psicolinguísticas relevantes em discursos políticos, mas tendem a privilegiar comparações entre atores ou momentos específicos, em vez de caracterizar padrões predominantes dentro de um mesmo mandato.

Além do LIWC, métodos computacionais baseados em aprendizado de máquina e modelagem quantitativa têm sido utilizados para analisar regularidades, mudanças temporais e posicionamentos ideológicos em discursos políticos. Nesse grupo, há estudos que estimam proporções ideológicas em discursos eleitorais [Sim et al., 2013], investigam mudanças de linguagem e mensagem ao longo do tempo [Stromer-Galley et al., 2021] e analisam, com apoio de aprendizado de máquina, como diferentes fatores moldam a construção discursiva de temas políticos [Silva-Muller and Sposito, 2024]. Embora esses trabalhos avancem na identificação de padrões e variações discursivas, seu foco recai mais sobre classificação, comparação temporal ou interpretação temática, e não sobre a descoberta de padrões psicolinguísticos e identificação de discursos destoantes.

No cenário brasileiro, embora menos explorado, também há estudos relevantes. Parte dessa literatura compara traços de autenticidade, estilo e identidade política entre discursos presidenciais brasileiros e estadunidenses [Sposito, 2025]; outra parte investiga diferenças discursivas entre presidentes brasileiros em temas específicos, como política externa [Vilela and Neiva, 2011]. Esses trabalhos confirmam a relevância do contexto brasileiro e mostram que o discurso presidencial é sensível ao contexto histórico e político, mas permanecem centrados em comparações entre atores, temas ou recortes específicos.

De forma geral, os trabalhos apresentados evidenciam a crescente utilização de ferramentas computacionais, como o LIWC e algoritmos de aprendizado de máquina, para examinar discursos políticos. No entanto, permanece menos explorada a combinação entre descoberta de padrões psicolinguísticos predominantes em discursos presidenciais e identificação de discursos destoantes preservados para análise posterior. Em contraste, este trabalho propõe uma contribuição metodológica centrada nessa articulação, ao combinar extração de dimensões psicolinguísticas, detecção de anomalias, agrupamento e interpretação qualitativa dos grupos formados.

4. Metodologia

Esta seção apresenta a metodologia do SPEAR, voltada à identificação de padrões psicolinguísticos predominantes em discursos presidenciais e à detecção de discursos destoantes. A Figura 1 resume o fluxo metodológico.

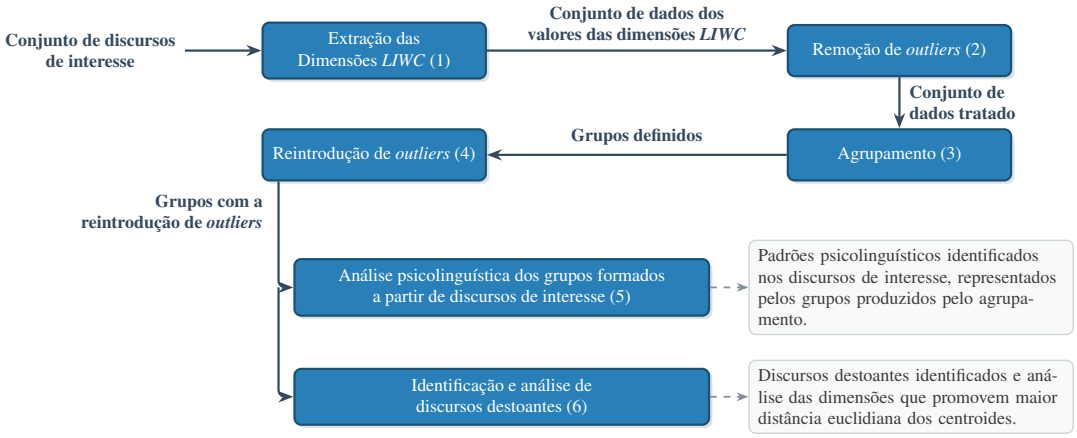


Figura 1. Diagrama de atividades da metodologia do trabalho.

Dado um conjunto C_d de discursos presidenciais, a primeira etapa do método consiste em aplicar o LIWC a cada discurso $D_i \in C_d$, produzindo o conjunto C_l, no qual os textos passam a ser representados por vetores de 77 dimensões psicolinguísticas. A escolha do LIWC justifica-se por sua capacidade de operacionalizar aspectos afetivos, cognitivos e sociais da linguagem relevantes para a caracterização de padrões psicolinguísticos. Diferentemente de abordagens como *Term Frequency/Inverse Document Frequency (TF/IDF)*, que verificam apenas a presença e a frequência de termos, o LIWC captura essas dimensões de forma mais aderente ao objetivo deste trabalho. Cabe destacar que Pennebaker et al. [2015] recomendam não aplicar pré-processamento de linguagem natural antes dessa extração, pois palavras funcionais, como pronomes e artigos, frequentemente removidas nesse processo, são relevantes para o LIWC. As dimensões são então normalizadas via *z-score* [Han et al., 2011] dentro de cada mandato, de modo a preservar a comparação intra-mandato e garantir comparabilidade nas etapas subsequentes, que utilizam distância euclidiana [Jain, 2010]. A partir dessa etapa, apenas as entradas de C_l são utilizadas, e o conjunto C_d deixa de ser necessário.

Na segunda etapa, observações anômalas são removidas de C_l antes do agrupamento. Discursos inconsistentes ou significativamente diferentes dos demais podem distorcer a formação dos grupos. Um efeito típico é a criação de grupos compostos exclusivamente por *outliers*, o que prejudica a representatividade dos padrões identificados. O agrupamento não supervisionado é então aplicado sobre as observações restantes: a ausência de rótulos prévios para os padrões psicolinguísticos torna essa abordagem adequada, pois permite a descoberta automática de estruturas latentes nos dados. Neste trabalho, o algoritmo de agrupamento adotado é o K-Means [MacQueen, 1967], escolhido por sua aderência a vetores numéricos contínuos, pela interpretação direta via centroides e por sua compatibilidade com a etapa posterior de reintrodução dos discursos removidos com base em distância ao centroide[1]. A hipótese operacional subjacente é que a remoção prévia de observações anômalas pode melhorar a coesão e a separação dos grupos sem eliminar definitivamente esses discursos. Seja G_j o grupo j produzido pelo agrupamento

[1]Métodos como DBSCAN, HDBSCAN, agrupamento hierárquico ou GMM poderiam produzir partições distintas, especialmente em estruturas não esféricas, mas fogem ao escopo comparativo deste trabalho e permanecem como alternativa para estudos futuros.

e μ_j o seu centroide. A distância entre uma observação $x_i \in C_l$ e o centroide de um grupo é dada por $d(x_i, \mu_j) = \|x_i - \mu_j\|^2$. Após o agrupamento, cada *outlier* removido é reintroduzido no grupo cujo centroide minimiza essa distância, isto é, $g(x_i) = \min_j d(x_i, \mu_j)$.

Essa etapa permite que todos os discursos do mandato recebam uma identidade de grupo e possam ser analisados individualmente como discursos destoantes. Assume-se, ainda, que a reintrodução posterior desses casos busca preservar sua utilidade interpretativa sem comprometer a estrutura dos agrupamentos obtidos.

A análise final divide-se entre os padrões formados, em que cada grupo é interpretado pelas dimensões LIWC de maior contribuição e pelas palavras mais frequentes. Primeiramente, ocorre a remoção de *stopwords* apenas nesse ponto, o que permite caracterizar o perfil psicolinguístico predominante de cada padrão. Em seguida, para a análise dos discursos destoantes, se utiliza a distância euclidiana ao centroide do grupo para identificar e ranquear as entradas mais afastadas. Seja $\bar{d}_j$ a média das distâncias ao centroide no grupo G_j e σ_j o desvio padrão correspondente, um discurso reatribuído ao grupo G_j é mantido para análise de destoamento quando sua distância satisfaz $d(x_i, \mu_j) > \bar{d}_j + 3\sigma_j$.

Dessa forma, o LOF atua como mecanismo de detecção local de anomalias, enquanto o filtro de 3σ restringe a análise qualitativa aos casos mais destoantes em relação à estrutura interna de cada grupo. Esse filtro é utilizado como critério heurístico-estatístico para priorização de extremos, e não como evidência de aderência estrita a um modelo probabilístico específico para as distâncias observadas. Além disso, a magnitude dessa distância em desvios padrão permite comparar a dispersão entre grupos: grupos com maior dispersão média indicam menor coesão interna nos padrões de comunicação. Para fins de análise, os discursos destoantes de cada grupo são ranqueados em ordem decrescente de $d(x_i, \mu_j)$, priorizando-se os casos mais afastados do centroide esperado.

5. Avaliação Experimental

Esta seção apresenta a avaliação experimental do método proposto. Inicialmente, são descritos o conjunto de dados e a configuração experimental. Em seguida, são apresentados os resultados quantitativos obtidos nos 11 mandatos presidenciais analisados. Por fim, é desenvolvido um estudo de caso no segundo mandato de Lula, com foco na interpretação qualitativa dos grupos formados e dos discursos destoantes.

5.1. Conjunto de Dados e Configuração Experimental

O conjunto de dados foi obtido no Harvard Dataverse [Cezar, 2022] e contém 5.888 discursos presidenciais brasileiros, do mandato de José Sarney até parte do mandato de Jair Bolsonaro. Após a remoção de 16 discursos em outros idiomas via *langdetect* [Mimino and Nakatani, 2021], restaram 5.872 entradas válidas, distribuídas conforme a Tabela 1.

O mesmo protocolo metodológico foi aplicado aos 11 mandatos avaliados, variando apenas os parâmetros definidos automaticamente em função dos dados, como o número de vizinhos do LOF e a quantidade de grupos k. A avaliação foi organizada em três frentes. A primeira corresponde à vetorização de cada discurso em 81 dimensões utilizando o BP-LIWC2015 (77 dimensões básicas, além de autenticidade, pensamento analítico, influência e tom emocional). A segunda apresenta uma análise quantitativa comparando a abordagem regular (explicada a seguir) e o SPEAR em todos os mandatos, com base na quantidade de grupos, nos valores de silhueta e na detecção de *outliers*. A

Tabela 1. Discursos por mandato

Mandato	Presidente	Período	Total	Outros idiomas	Válidos
Sarney	José Sarney	1985–1990	581	0	581
Collor	Fernando Collor	1990–1992	150	0	150
Itamar	Itamar Franco	1992–1994	56	0	56
FHC I	Fernando H. Cardoso	1995–1998	781	0	781
FHC II	Fernando H. Cardoso	1999–2002	670	0	670
Lula I	Luiz Inácio Lula da Silva	2003–2006	953	0	953
Lula II	Luiz Inácio Lula da Silva	2007–2010	1110	1	1109
Dilma I	Dilma Rousseff	2011–2014	709	14	695
Dilma II	Dilma Rousseff	2015–2016	237	0	237
Temer	Michel Temer	2016–2018	380	1	379
Bolsonaro (parc.)	Jair Bolsonaro	2019–2022	246	0	246
Total	–	–	5.888	16	5.872

terceira consiste em um estudo de caso no segundo mandato de Lula, escolhido por concentrar o maior volume de discursos válidos (1.109), o que favorece a análise qualitativa.

O experimento foi executado para comparar duas abordagens: (i) a abordagem regular, em que o K-Means é aplicado diretamente ao conjunto de dados, sem tratamento prévio de *outliers*; e (ii) a metodologia SPEAR, que prevê a remoção dos *outliers* antes do agrupamento e sua posterior reintrodução para análise específica. Para a detecção de *outliers*, adotou-se o LOF, cujo parâmetro de número de vizinhos foi otimizado no intervalo de 5 a 50, selecionando-se o valor que maximizou a silhueta [Shahapure and Nicholas, 2020]. A silhueta foi utilizada como principal métrica quantitativa de avaliação, por refletir simultaneamente a coesão interna e a separação entre os grupos [Shahapure and Nicholas, 2020], enquanto a análise qualitativa foi empregada como validação complementar da interpretabilidade dos padrões identificados e dos discursos destoantes. O valor de k do K-Means foi definido pelo método *Kneedle* [Satopää et al., 2011].

A Tabela 2 apresenta, por mandato, as métricas de agrupamento e detecção de *outliers* comparando a abordagem regular com a metodologia SPEAR. O SPEAR aumenta a silhueta em todos os mandatos, com exceção de Itamar, no qual o valor se manteve igual. Esse resultado sugere que o tratamento prévio de *outliers* tende a melhorar a qualidade do agrupamento na maior parte dos mandatos, ainda que com magnitudes distintas. O caso Bolsonaro é o mais expressivo: mesmo com redução de 6 para 4 grupos, a silhueta aumentou de 0.062 para 0.179, sugerindo forte influência de *outliers* no agrupamento regular. Dois mandatos apresentaram aumento no número de grupos, FHC II e Dilma I, evidenciando que a abordagem regular pode subestimar a quantidade de padrões psicolinguísticos tanto por excesso quanto por falta. Em conjunto, esses resultados reforçam o ganho principal da proposta em termos de qualidade de agrupamento, enquanto a interpretabilidade dos grupos aparece como consequência desse ganho.

A mesma Tabela 2 também resume, por mandato, a quantidade de *outliers* identificados pelo LOF e o subconjunto filtrado por 3σ. Os valores mostram que o LOF atua como etapa de detecção sensível de candidatos a anomalia, enquanto o filtro de 3σ restringe a análise aos casos mais extremos. No segundo mandato de Lula, por exemplo, o LOF marcou 91 discursos, reduzidos a 23 pelo filtro de 3σ, o que ilustra a diferença entre uma triagem ampla e uma seleção mais conservadora para análise qualitativa. O mandato

Tabela 2. Métricas por mandato: agrupamento e *outliers*

Mandato	**Discursos**	**Viz.**	**LOF**	**K Reg.**	***K SPEAR***	3σ **Reg.**	3σ **SPEAR**	**Silh. Reg.**	***Silh. SPEAR***
Sarney	581	25	57	7	5	8	9	0.037	**0.069**
Collor	150	30	9	5	4	2	2	0.069	**0.075**
Itamar	56	35	0	6	6	0	0	0.130	0.130
FHC I	781	10	66	4	4	16	17	0.095	**0.122**
FHC II	670	35	55	3	4	13	14	0.086	**0.089**
Lula I	953	45	70	4	4	17	20	0.062	**0.068**
Lula II	1109	10	91	5	4	23	23	0.101	**0.122**
Dilma I	695	5	83	4	5	16	16	0.077	**0.087**
Dilma II	237	5	23	5	5	5	6	0.049	**0.084**
Temer	379	5	41	4	4	7	8	0.097	**0.106**
Bolsonaro	246	15	18	6	4	2	3	0.062	**0.179**

de Itamar é um caso limite: com apenas 56 discursos, o LOF não identificou nenhum *outlier*, o que sugere que a escassez de pontos dificulta a formação de núcleos de densidade local contrastantes o suficiente para classificar uma observação como anômala e ajuda a explicar por que a silhueta não variou nesse mandato.

5.2. Estudo de Caso - Segundo Mandato de Lula

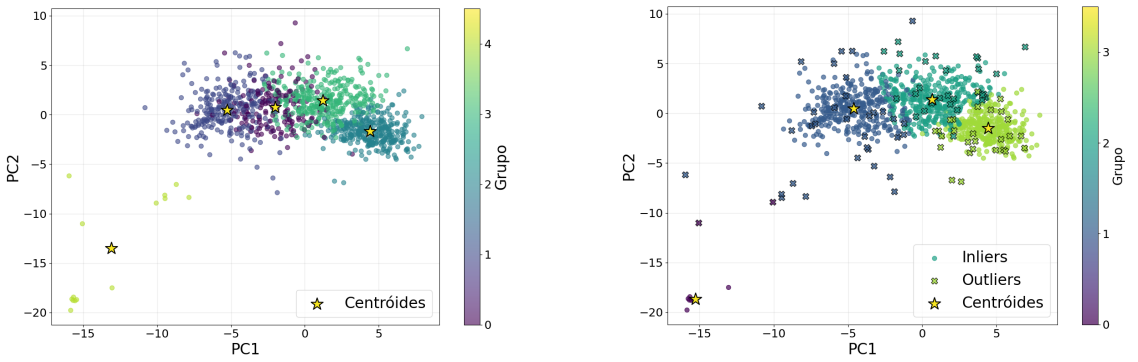
Esta seção apresenta o estudo de caso do segundo mandato de Lula, por ser o mandato com mais discursos no conjunto de dados. Para a visualização das análises, foram utilizadas nuvens de palavras, BERTopic e gráfico radar.

5.2.1. Grupos Obtidos pelo Algoritmo de Agrupamento

A análise dos agrupamentos produzidos com e sem a metodologia SPEAR ilustra a relevância do método proposto. A Figura 2(a) exibe a formação de grupos da abordagem regular (cinco grupos). O grupo roxo aparenta constituir uma fronteira entre os grupos azul e verde, com pontos muito próximos dos respectivos centroides. A Figura 2(b) apresenta os grupos com a metodologia SPEAR (quatro grupos). Essa redução, com realocação dos pontos, resulta em melhor separação e maior homogeneidade entre os grupos azul e verde, o que se traduz em maior silhueta, conforme apresentado na Subseção 5.1. Para fins de visualização bidimensional, os vetores psicolinguísticos foram projetados em duas dimensões por PCA, de modo que a figura sintetiza a estrutura relativa dos grupos sem substituir a análise no espaço original de 81 dimensões.

5.2.2. Análise Qualitativa do Padrão Psicolinguístico

A partir do agrupamento gerado, cada grupo concentra um padrão psicolinguístico. O grupo 2, por reunir a maior quantidade de discursos, 390, foi escolhido entre os quatro grupos para a análise qualitativa. Para fins de visualização, somente nesta etapa interpretativa foi realizado pré-processamento do texto; a extração das dimensões do LIWC descrita na metodologia utilizou os discursos sem esse tratamento. Foi gerada uma nuvem de palavras a partir de texto pré-processado, com remoção de *stopwords* e símbolos de pontuação. Além disso, com base no princípio de Pareto [Wilkinson, 2006], foram removidas 20% das palavras mais comuns, como critério operacional para reduzir termos excessivamente frequentes e destacar vocabulário mais informativo na análise interna do

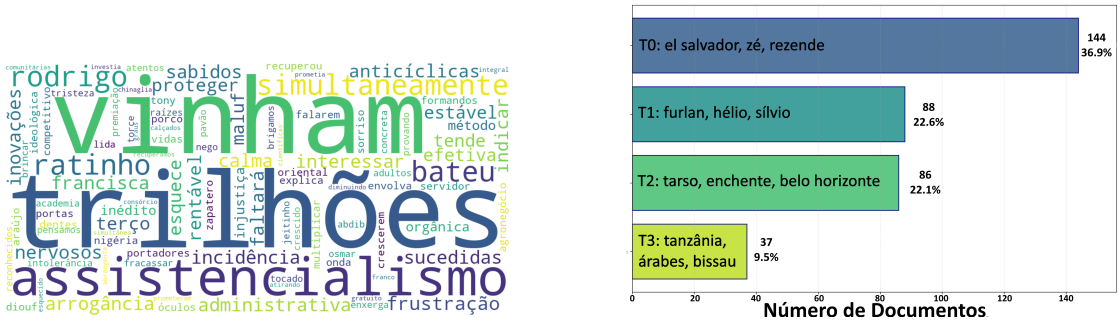


(a) (b)

Figura 2. Agrupamentos do segundo mandato de Lula: (a) antes da remoção de *outliers*; (b) após a remoção com LOF.

grupo. Para a modelagem de tópicos, utilizou-se o BERTopic, preservando mais contexto semântico para a identificação dos tópicos mais relevantes do grupo.

A Figura 3(a) apresenta a nuvem de palavras do grupo 2, enquanto a Figura 3(b) apresenta os quatro tópicos definidos a partir de palavras recorrentes nos discursos desse grupo. As palavras mais destacadas na nuvem, como *assistencialismo*, *trilhões* e *administrativa*, estão relacionadas às dimensões *work* e *number* e indicam um padrão voltado para ações governamentais e questões financeiras. A nuvem também apresenta a palavra *anticíclicas*, associada a estratégias de estabilização da economia em momentos de crise [de Brito Gadelha, 2011], como a crise financeira internacional de 2008.



(a) (b)

Figura 3. Análise visual do grupo 2 no segundo mandato de Lula: (a) nuvem de palavras; (b) tópicos e quantidade de documentos.

Nos tópicos gerados, o tópico zero (T0), que reúne a maioria dos discursos, menciona El Salvador, o que condiz com o contexto do mandato, no qual Lula estreitava as relações diplomáticas entre o Brasil e esse país. Observam-se discursos recorrentes sobre visitas, acordos para isenção de visto e acordos comerciais. O tópico um (T1) apresenta os nomes Furlan (Luiz Fernando Furlan), ministro do Desenvolvimento, Comércio e Indústria, e Hélio (Hélio Calixto da Costa), ministro das Comunicações. Ambos eram ministros naquele período e aparecem em menções possivelmente correlacionadas com projetos financeiros e investimentos, como as palavras detectadas na nuvem. O tópico dois (T2) também merece destaque, pois corresponde às fortes chuvas que causaram enchentes em Belo Horizonte, provocando danos em residências e vias públicas [Lucas et al., 2015].

Por fim, o tópico três (T3) mostra menções a relações internacionais estabelecidas por questões comerciais com a Tanzânia e a Guiné-Bissau.

Em conjunto, a nuvem de palavras e o BERTopic indicam que o grupo 2 concentra discursos associados à economia e a ações governamentais. Esse resultado sugere que o padrão psicolinguístico do grupo não corresponde a um único tema específico, mas a uma forma recorrente de tratar agendas de gestão, investimento público e articulação institucional. Assim, o SPEAR não substitui a análise temática, mas ajuda a delimitar subconjuntos discursivos nos quais certos temas passam a ser examinados em associação com uma forma relativamente estável de comunicação.

5.2.3. Análise Qualitativa do Discurso Mais Destoante

Os discursos com maior distância euclidiana ao centroide são considerados destoantes quando são identificados como *outliers* pelo LOF e filtrados pelos três desvios padrão. Para esses discursos, analisam-se as 15 dimensões mais relevantes do LIWC, escaladas por normalização min-max entre 0 e 1 para construção do gráfico radar.

O discurso do grupo 2 é o pronunciamento de José Alencar, então vice-presidente, na abertura do 56° Encontro da Frente Nacional de Prefeitos, em Fortaleza, em 30 de novembro de 2009. Como Alencar discursou em nome de Lula, o conteúdo permaneceu próximo ao esperado para o grupo, mas as características do orador contribuíram para o caráter destoante. O discurso destaca desenvolvimento econômico, investimentos, ajuda financeira aos municípios e programas nacionais. Assim, ele não destoa do tema esperado para o grupo 2, mas da forma de comunicação. A Figura 4 compara seus valores com a média esperada do grupo.

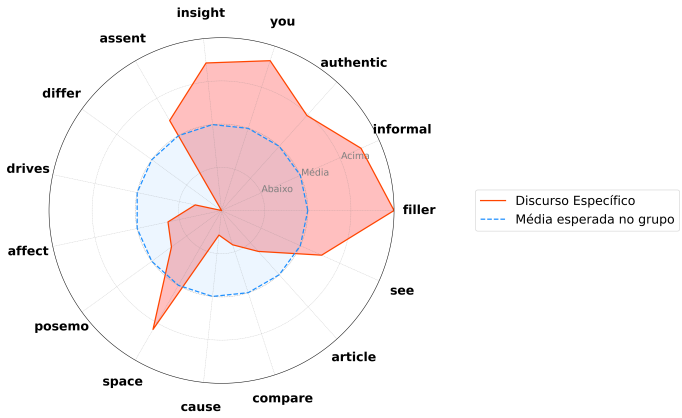


Figura 4. Dimensões LIWC do discurso mais destoante do grupo 2.

Pelo gráfico, observam-se dimensões com valores acima e abaixo da média esperada do grupo, o que também pode ser evidenciado no texto. O aumento de *space* está relacionado ao uso de palavras associadas a espaço, como município, estado e país. As dimensões *informal* e *filler* geralmente não são esperadas em discursos formais. Essas dimensões podem ter se destacado pelo uso de uma introdução mais informal e de palavras de preenchimento de fala, como *então*, evidentes nos trechos iniciais: "*Então* contei para eles lá que os antigos, em Minas..." e "E, *então*, sempre ouvia isso com o maior respeito".

A dimensão *insight* apresenta teor reflexivo no texto, o que corrobora o propósito do discurso de refletir sobre as realizações do governo que também contemplaram os municípios. O uso de palavras como *preferir* e *saber* também contribui para esse aumento e é observado em alguns momentos, como "Todos *sabem* que o presidente Lula está fora do Brasil [...] *sabendo* que eu estava vindo aqui" e "Eu, normalmente, *prefiro* falar sem papel na mão, mas são tantas as realizações do governo".

Embora, em sua maior parte, o texto siga um tom informativo sobre realizações e valores monetários movimentados, a dimensão *authentic* também ficou um pouco acima da média. O uso de verbos e pronomes na primeira pessoa, principalmente na introdução, pode estar relacionado a esse aumento. Isso pode ser observado logo no primeiro trecho: "Vocês já imaginaram como *eu estou*, não é? *Eu acabei* de receber, às 2 horas, 3 horas, o título de Cidadão Honorário de Fortaleza".

5.3. Discussão

A Subseção 5.1 mostrou que a abordagem regular obteve valores de silhueta inferiores ou iguais aos da metodologia proposta em todos os mandatos, conforme a Tabela 2. A evidência quantitativa transversal em 11 mandatos sustenta o principal ganho empírico do método: melhorar a qualidade dos agrupamentos após o tratamento de observações anômalas, sem removê-las definitivamente. No estudo de caso de Lula II, essa diferença aparece com a redução no número de grupos e maior delimitação de quatro padrões psicolinguísticos. A análise qualitativa com nuvem de palavras e BERTopic detalha como esse ganho quantitativo se traduz em padrões mais interpretáveis. Ao mesmo tempo, os valores absolutos de silhueta recomendam moderação interpretativa: o método melhora a estrutura observada, mas não elimina a sobreposição esperada entre estilos discursivos em um corpus político amplo e heterogêneo.

Outra análise relevante refere-se à identificação de *outliers*, também apresentada na Tabela 2. A metodologia SPEAR identificou, em alguns mandatos, até três discursos destoantes a mais no filtro de 3σ em comparação com a abordagem regular. Em conjunto, esses resultados mostram que a abordagem permite tratar discursos destoantes também como objeto de análise, e não apenas como ruído. No caso analisado, o discurso destoante foi proferido pelo vice-presidente José Alencar em nome de Lula e, embora não destoe da temática principal do grupo, a mudança de orador parece alterar a forma esperada de comunicação. O gráfico radar mostra em quais dimensões psicolinguísticas do LIWC essa variação ocorre. Assim, a evidência qualitativa complementa a evidência quantitativa ao mostrar como o método pode apoiar a interpretação dos grupos e dos discursos que deles se afastam. O caso de Itamar, por sua vez, sinaliza uma limitação importante: em mandatos com poucos discursos, a baixa densidade amostral pode reduzir a sensibilidade do LOF e limitar o ganho observável do agrupamento.

Esse resultado também contraria uma intuição simplificadora sobre os componentes do método quando considerados isoladamente. Poder-se-ia supor que apenas remover *outliers* antes do agrupamento já seria suficiente para melhorar os resultados ou, alternativamente, que bastaria ranquear discursos pela distância ao centroide após o K-Means. No entanto, a evidência quantitativa da Tabela 2, em conjunto com a análise qualitativa do estudo de caso, sugere que nenhuma dessas estratégias isoladas reproduz o ganho do SPEAR. A remoção isolada favoreceria a coesão dos grupos, mas impediria a reintrodução e o ranqueamento dos discursos mais destoantes; já a análise apenas por distância

ao centroide, sem a detecção local prévia com LOF, reduziria a capacidade de distinguir observações anômalas da estrutura principal dos grupos. Assim, o método mostra que o ganho não decorre de uma etapa isolada, mas da articulação entre remoção inicial, reintrodução e ranqueamento.

6. Conclusão

Este trabalho apresenta o SPEAR, método para descoberta de padrões psicolinguísticos e identificação de discursos destoantes em corpora presidenciais. A metodologia combina extração de dimensões do LIWC, remoção de *outliers* via LOF, agrupamento K-Means e reintrodução dos pontos removidos ao grupo mais próximo. Os resultados quantitativos sobre 11 mandatos presidenciais brasileiros mostram aumento da silhueta na maior parte dos mandatos e estabilidade no mandato de Itamar em relação à abordagem regular. Merece destaque o mandato de Bolsonaro, que apresenta valores de silhueta de 0.062 com a abordagem tradicional e 0.179 com o SPEAR, sustentando a contribuição principal do método em termos de qualidade de agrupamento.

Em conjunto, esses resultados indicam que o SPEAR melhora a qualidade dos agrupamentos e oferece apoio interpretativo para examinar padrões formados e discursos que deles se afastam. No estudo de caso do segundo mandato de Lula, o método identifica quatro padrões predominantes; a análise qualitativa do maior grupo revela o uso elevado de termos econômicos, corroborada pelo BERTopic. O discurso mais destoante, proferido pelo vice-presidente José Alencar em substituição a Lula, não destoa pelo tema, mas pela forma e pela estrutura do orador, sugerindo, nesse caso, que o método pode captar variações estilísticas além do conteúdo. Assim, o trabalho oferece uma resposta metodológica à lacuna identificada na literatura ao combinar descoberta de padrões intra-mandato e preservação de discursos destoantes.

Esses achados devem ser interpretados à luz de algumas limitações, entre as quais se destacam a dependência de transcritos oficiais, que podem conter ruídos, e o próprio uso do LIWC como representação baseada em dicionário, o que privilegia categorias previamente definidas e pode deixar de captar nuances contextuais, semânticas ou retóricas mais finas. Soma-se a isso a restrição do BP-LIWC2015 a textos em português, o que exige a remoção de discursos em outros idiomas e delimita o escopo de generalização deste estudo ao contexto linguístico analisado. Como extensões naturais, destacam-se comparações entre mandatos de um mesmo presidente ou entre países que utilizem dicionários LIWC adaptados, análise temporal intra-mandato e aprofundamento da análise de destoantes por meio do cruzamento entre BERTopic, dimensões LIWC e eventos históricos. Em seu escopo atual, o SPEAR oferece uma alternativa metodológica para analisar discursos presidenciais sem reduzir discursos destoantes a mero ruído.

Agradecimentos

Os autores agradecem às agências CAPES, CNPq e FAPERJ pelo apoio parcial à pesquisa.

Uso de IA

Ferramentas de IA generativa são utilizadas apenas para revisão linguística, com foco em clareza textual. Os autores revisaram todas as sugestões e permanecem integralmente responsáveis pelo conteúdo do artigo.

Referências

Aggarwal, C. C. (2016). *Outlier Analysis*. Springer International Publishing.

Ahmadian, S., Azarshahi, S., and Paulhus, D. L. (2017). Explaining Donald Trump via communication style: Grandiosity, informality, and dynamism. *Personality and Individual Differences*, 107:49 – 53.

Bernard, N., Sagawa, Y., Bier, N., Lihoreau, T., Pazart, L. H., and Tannou, T. (2025). Using artificial intelligence for systematic review: the example of elicit. *BMC Medical Research Methodology*, 25.

Boyd, R. L. (2017). Psychological Text Analysis in the Digital Humanities. In Hai-Jew, S., editor, *Data Analytics in Digital Humanities*, pages 161–189. Springer International Publishing, Cham.

Breuniq, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: Identifying density-based local outliers. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 29:93 – 104.

Carvalho, F., Junior, F. P., Ogasawara, E., Ferrari, L., and Guedes, G. (2024). Evaluation of the Brazilian Portuguese version of linguistic inquiry and word count 2015 (BP-LIWC2015). *Language Resources and Evaluation*, 58:203 – 222.

Carvalho, F., Rodrigues, R., Santos, G., Cruz, P., Ferrari, L., and Guedes, G. (2019). Avaliação da versão em português do LIWC Lexicon 2015 com análise de sentimentos em redes sociais. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 24–34, Porto Alegre, RS, Brasil. SBC.

Cezar, R. F. (2022). Brazilian Presidential Speeches from 1985 to July 2020.

de Brito Gadelha, S. R. (2011). Countercyclical fiscal policy, international financial crisis and economic growth in Brazil. *Revista de Economia Politica*, 31:794 – 812.

Gadelha, T., Monteiro, J. M., Machado, J., Claudino, I., Santos, R., Galick, L., and Santos, C. (2023). Ativismo da extrema direita brasileira no WhatsApp: O que mudou das eleições de 2018 para 2022? In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 336–341, Porto Alegre, RS, Brasil. SBC.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, United States, 3 edition.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31:651 – 666.

Jordan, K. N., Sterling, J., Pennebaker, J. W., and Boyd, R. L. (2019). Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences of the United States of America*, 116:3476 – 3481.

Kangas, S. E. (2014). What can software tell us about political candidates? A critical analysis of a computerized method for political discourse. *Journal of Language and Politics*, 13:77 – 97.

Körner, R., Overbeck, J. R., Körner, E., and Schütz, A. (2022). How the Linguistic Styles of Donald Trump and Joe Biden Reflect Different Forms of Power. *Journal of Language and Social Psychology*, 41:631 – 658.

Lucas, T. P. B., Augusto, P., Reis, S. d., and Rocha, S. C. (2015). Impactos hidrometeóricos em Belo Horizonte-MG. *Revista Brasileira de Climatologia*, 16.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pages 281–298. University of California press.

Malzer, C. and Baum, M. (2020). A Hybrid Approach To Hierarchical Density-based Cluster Selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 223–228. IEEE.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Mimino, M. and Nakatani, S. (2021). langdetect: Port of Google's language-detection library to Python.

Morgan, R. L., Whaley, P., Thayer, K. A., and Schünemann, H. J. (2018). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environment International*, 121:1027 – 1031.

Nascimento, F., Monteiro, J. M., and Machado, J. (2025). Dinâmicas de Grupos de WhatsApp da Extrema Direita no Brasil: Uma Análise Comparativa Pré e Pós-Eleição de 2022. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 563–575, Porto Alegre, RS, Brasil. SBC.

Okuno, H. Y., Carvalho, F., Guedes, G. P., and Torres, M. (2018). At analysis - Toward a psycholinguistic method to analyze video textual information. In *CEUR Workshop Proceedings*, volume 2170, pages 73 – 79.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Technical report, University of Texas at Austin, Austin, TX.

Randour, F., Perrez, J., and Reuchamps, M. (2020). Twenty years of research on political discourse: A systematic review and directions for future research. *Discourse and Society*, 31:428 – 443.

Satopää, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a "kneedle"in a haystack: Detecting knee points in system behavior. In *Proceedings - International Conference on Distributed Computing Systems*, pages 166 – 171.

Shahapure, K. R. and Nicholas, C. (2020). Cluster quality analysis using silhouette score. In G, W., Z, Z., V.S, T., G, W., M, V., and L, C., editors, *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, pages 747 – 748. Institute of Electrical and Electronics Engineers Inc.

Silva-Muller, L. and Sposito, H. (2024). Which Amazon Problem? Problem-constructions and Transnationalism in Brazilian Presidential Discourse since 1985. *Environmental Politics*, 33:398 – 421.

Sim, Y., Acree, B. D. L., Gross, J. H., and Smith, N. A. (2013). Measuring Ideological Proportions in Political Speeches. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, Seattle, Washington, USA. Association for Computational Linguistics.

Sposito, H. (2025). Radiating Truthiness: Authenticity Performances in Politics in Brazil and the United States. *Political Studies*, 73:770 – 794.

Stromer-Galley, J., Rossini, P., Hemsley, J., Bolden, S. E., and McKernan, B. (2021). Political Messaging Over Time: A Comparison of US Presidential Candidate Facebook Posts and Tweets in 2016 and 2020. *Social Media and Society*, 7.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24 – 54.

Vilela, E. and Neiva, P. (2011). Temas e regiões nas políticas externas de Lula e Fernando Henrique: comparação do discurso dos dois presidentes. *Revista Brasileira de Política Internacional*, 54:70–96.

Wilkinson, L. (2006). Statistical computing and graphics: Revising the Pareto chart. *American Statistician*, 60:332 – 334.