



**UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
IEG - INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

MARCUS VINÍCIUS CARDOSO RÊGO

**QUAIS SÃO AS PRIORIDADES AMBIENTAIS DOS CANDIDATOS A
PREFEITO NO BRASIL? UMA VISÃO GERAL DAS PROPOSTAS PARA
AS ELEIÇÕES DE 2024**

**SANTARÉM - PARÁ
2025**

MARCUS VINÍCIUS CARDOSO RÊGO

QUAIS SÃO AS PRIORIDADES AMBIENTAIS DOS CANDIDATOS A
PREFEITO NO BRASIL? UMA VISÃO GERAL DAS PROPOSTAS PARA
AS ELEIÇÕES DE 2024

Trabalho de Conclusão de Curso apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação, pela Universidade Federal do Oeste do Pará, no Instituto de Engenharia e Geociências.

Orientador: Prof. Dr. Fábio Manoel França Lobato

SANTARÉM - PARÁ
2025

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/Ufopa

- R343q Rego, Marcus Vinícius Cardoso
Quais são as prioridades ambientais dos candidatos a prefeito no Brasil? Uma visão geral das propostas para as eleições de 2024./ Marcus Vinícius Cardoso Rego. – Santarém, 2025.
64 p.: il.
Inclui bibliografias.
- Orientador: Fábio Manoel França Lobato.
Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências, Bacharelado em Ciência da Computação.
1. Ciência de dados. 2. Processamento de linguagem natural. 3. Eleições municipais. I. Lobato, Fábio Manoel França, *orient.* II. Título.

CDD: 23 ed. 004

MARCUS VINÍCIUS CARDOSO RÊGO

**QUAIS SÃO AS PRIORIDADES AMBIENTAIS DOS CANDIDATOS A
PREFEITO NO BRASIL? UMA VISÃO GERAL DAS PROPOSTAS PARA
AS ELEIÇÕES DE 2024**

Trabalho de Conclusão de Curso apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação, pela Universidade Federal do Oeste do Pará, no Instituto de Engenharia e Geociências.

Data de aprovação:

29/07/2025

Banca Examinadora:

Prof. Dr. Fábio Manoel França Lobato
Universidade Federal do Oeste do Pará - UFOPA

Prof. Dr. Efren Lopes de Souza
Universidade Federal do Oeste do Pará - UFOPA

Prof^ª. Ma. Ana Letícia Mafra Salla
Universidade Federal do ABC - UFABC



SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO DE CIÊNCIA DA COMPUTAÇÃO
FORMULÁRIO DE AVALIAÇÃO DE TCC II

Identificação:


Título do Trabalho: QUAIS SÃO AS PRIORIDADES AMBIENTAIS DOS CANDIDATOS A PREFEITO NO BRASIL? UMA VISÃO GERAL DAS PROPOSTAS PARA AS ELEIÇÕES DE 2024
Aluno(a): Marcus Vinícius Cardoso Rêgo
Orientador(a): Fábio Manoel França Lobato

Avaliação:

Examinador(a) 1: Fábio Manoel França Lobato	Nota: 9,2
Assinatura:	

Examinador(a) 2: Éfren Lopes de Souza	Nota: 9,5
Assinatura:	

Examinador(a) 3: Ana Leticia Mafra Salla	Nota: 9,0
Assinatura:	

 Documento assinado digitalmente
ANA LETICIA MAFRA SALLA
Data: 30/07/2025 17:30:36-0300
Verifique em <https://validar.iti.gov.br>


Parecer:

Resumo da Avaliação:

<input checked="" type="checkbox"/>	Aceitação incondicional
<input type="checkbox"/>	Aceitação condicionada a modificações (especificar no verso)
<input type="checkbox"/>	Recusado

Nota Final: 9,2

São Carlos, SP, 28 de julho de 2025.

 Documento assinado digitalmente
FABIO MANOEL FRANCA LOBATO
Data: 30/07/2025 14:06:01-0300
Verifique em <https://validar.iti.gov.br>

Presidente da Banca Examinadora



SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO DE CIÊNCIA DA COMPUTAÇÃO
FORMULÁRIO INDIVIDUAL AVALIAÇÃO DE TCC II

Identificação:

Título do Trabalho: QUAIS SÃO AS PRIORIDADES AMBIENTAIS DOS CANDIDATOS A PREFEITO NO BRASIL? UMA VISÃO GERAL DAS PROPOSTAS PARA AS ELEIÇÕES DE 2024
Nome da Aluno(a): Marcus Vinícius Cardoso Rêgo
Orientador(a): Fábio Manoel França Lobato
Avaliador(a): Fábio Manoel França Lobato

Critérios de Avaliação:

	REDAÇÃO		
	0,0	0,5	1,0
Clareza, concisão e precisão do texto	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Organização do trabalho	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação e qualidade do embasamento teórico	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação das referências bibliográficas	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Adequação e qualidade do modelo proposto	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Adequação da metodologia e avaliações	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação e qualidade da contribuição	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
APRESENTAÇÃO			
Respeito ao tempo e qualidade do material	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Clareza na exposição das ideias	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Domínio do tema	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Observações:

1. Quesitos de avaliação que puderem ser corrigidos ou melhorados não deverão ter conceito inferior a seis;



**SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO DE CIÊNCIA DA COMPUTAÇÃO**

2. As sugestões para correções e melhoramentos deverão ser descritas de forma clara, concisa e precisa, no próprio trabalho do aluno, e com resumo no verso desta folha;

3. Conforme o Art. 23, § 4º, no caso do produto de TCC ser um artigo publicado, será avaliada apenas a apresentação oral.

Sugestões de notas para a redação de trabalhos publicados

Qualis	Notas
A1	10.0
A2	9.5
B1	9.0
B2	8.5
B3	8.0
B4	7.5
B5	7.0

São Carlos, SP, 29 de julho de 2025.

Assinatura



SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO DE CIÊNCIA DA COMPUTAÇÃO
FORMULÁRIO INDIVIDUAL AVALIAÇÃO DE TCC II

Identificação:

Título do Trabalho: QUAIS SÃO AS PRIORIDADES AMBIENTAIS DOS CANDIDATOS A PREFEITO NO BRASIL? UMA VISÃO GERAL DAS PROPOSTAS PARA AS ELEIÇÕES DE 2024
Nome da Aluno(a): Marcus Vinícius Cardoso Rêgo
Orientador(a): Fábio Manoel França Lobato
Avaliador(a): Efren Lopes de Souza

Critérios de Avaliação:

	REDAÇÃO		
	0,0	0,5	1,0
Clareza, concisão e precisão do texto	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Organização do trabalho	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação e qualidade do embasamento teórico	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação das referências bibliográficas	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação e qualidade do modelo proposto	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação da metodologia e avaliações	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação e qualidade da contribuição	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
APRESENTAÇÃO			
Respeito ao tempo e qualidade do material	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Clareza na exposição das ideias	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Domínio do tema	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Observações:

4. Quesitos de avaliação que puderem ser corrigidos ou melhorados



**SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO DE CIÊNCIA DA COMPUTAÇÃO**

não deverão ter conceito inferior a seis;

5. As sugestões para correções e melhoramentos deverão ser descritas de forma clara, concisa e precisa, no próprio trabalho do aluno, e com resumo no verso desta folha;

6. Conforme o Art. 23, § 4º, no caso do produto de TCC ser um artigo publicado, será avaliada apenas a apresentação oral.

Sugestões de notas para a redação de trabalhos publicados

Qualis	Notas
A1	10.0
A2	9.5
B1	9.0
B2	8.5
B3	8.0
B4	7.5
B5	7.0



Documento assinado digitalmente
EFREN LOPES DE SOUZA
Data: 30/07/2025 14:52:09-0300
Verifique em <https://validar.iti.gov.br>

São Carlos, SP, 29 de julho de 2025.

Assinatura



SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO DE CIÊNCIA DA COMPUTAÇÃO
FORMULÁRIO INDIVIDUAL AVALIAÇÃO DE TCC II

Identificação:

Título do Trabalho: QUAIS SÃO AS PRIORIDADES AMBIENTAIS DOS CANDIDATOS A PREFEITO NO BRASIL? UMA VISÃO GERAL DAS PROPOSTAS PARA AS ELEIÇÕES DE 2024
Nome da Aluno(a): Marcus Vinícius Cardoso Rêgo
Orientador(a): Fábio Manoel França Lobato
Avaliadora: Ana Leticia Mafra Salla

Critérios de Avaliação:

	REDAÇÃO		
	0,0	0,5	1,0
Clareza, concisão e precisão do texto	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Organização do trabalho	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Adequação e qualidade do embasamento teórico	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Adequação das referências bibliográficas	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação e qualidade do modelo proposto	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Adequação da metodologia e avaliações	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Adequação e qualidade da contribuição	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
APRESENTAÇÃO			
Respeito ao tempo e qualidade do material	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Clareza na exposição das ideias	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Domínio do tema	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Observações:

7. Quesitos de avaliação que puderem ser corrigidos ou melhorados



**SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO DE CIÊNCIA DA COMPUTAÇÃO**

não deverão ter conceito inferior a seis;

8. As sugestões para correções e melhoramentos deverão ser descritas de forma clara, concisa e precisa, no próprio trabalho do aluno, e com resumo no verso desta folha;

9. Conforme o Art. 23, § 4º, no caso do produto de TCC ser um artigo publicado, será avaliada apenas a apresentação oral.

Sugestões de notas para a redação de trabalhos publicados

Qualis	Notas
A1	10.0
A2	9.5
B1	9.0
B2	8.5
B3	8.0
B4	7.5
B5	7.0

São Paulo, SP, 29 de julho de 2025.

Assinatura

*Dedico aos meus pais, amigos e professores
que me auxiliaram durante o meu
processo de formação acadêmica*

AGRADECIMENTOS

Agradeço, acima de tudo, a Deus por me proporcionar a oportunidade de poder estudar em uma universidade, além de todos os meios necessários para que eu focasse firmemente no meio acadêmico.

A minha família, principalmente aos meus pais e avós, por me oferecerem todo o incentivo e suporte emocional e financeiro necessários para que eu pudesse avançar durante esse percurso.

Ao meu orientador, Professor Fábio Manoel França Lobato, pela orientação necessária e precisa, paciência e confiança no desenvolvimento do meu trabalho que foi de grande importância para meu aprendizado e contribuiu para minha melhora como profissional.

Aos professores, pesquisadores e servidores do Instituto de Engenharia e Geociências da Universidade Federal do Oeste do Pará e de outros institutos que foram fundamentais para a construção da minha formação acadêmica e para a realização desta pesquisa na área de Ciência de Dados.

“O objetivo é transformar dados em informações e informações em insights.”

Carly Fiorina, ex-CEO da Hewlett-Packard.

RESUMO

Este estudo apresenta uma abordagem híbrida de ciência de dados para analisar o discurso político e os perfis dos candidatos no contexto das eleições municipais brasileiras de 2024. Ela integra metadados eleitorais estruturados com dados textuais não estruturados derivados de mais de 15.000 propostas de governo apresentadas por candidatos a prefeito. O pipeline metodológico combina modelos de linguagem de grande porte (LLMs) de engenharia imediata com expansão de léxico usando *Word Embeddings*, permitindo extrair e classificar palavras-chave ambientais com maior sensibilidade contextual. A análise revela padrões demográficos, partidários e regionais, realçando temas-chave como a sustentabilidade e as alterações climáticas. Além disso, os gráficos de coocorrência fornecem informações valiosas sobre as estruturas do discurso e o envolvimento regional com tópicos ambientais. Os resultados demonstram o potencial da combinação da análise de metadados com o processamento de linguagem natural para a análise de textos políticos em grande escala, contribuindo significativamente para as ciências sociais computacionais e os estudos eleitorais.

Palavras-chave: Ciência de Dados; Processamento de Linguagem Natural; Eleições Municipais; Sustentabilidade; Discurso Político; Meio Ambiente; Conservação, Mudanças Climáticas.

ABSTRACT

This study presents a hybrid data science approach to analyze political discourse and candidate profiles in the context of the 2024 Brazilian municipal elections. It integrates structured electoral metadata with unstructured textual data derived from over 15,000 government proposals submitted by mayoral candidates. The methodological pipeline combines prompt-engineered large language models (LLMs) with lexicon expansion using Word Embeddings, enabling extracting and classifying environmental keywords with greater contextual sensitivity. The analysis reveals demographic, partisan, and regional patterns, emphasizing key themes such as sustainability and climate change. Additionally, co-occurrence graphs provide valuable insights into discourse structures and regional engagement with environmental topics. The results demonstrate the potential of combining metadata analytics with natural language processing for large-scale political text analysis, contributing significantly to computational social science and electoral studies.

Keywords: Data Science; Natural Language Processing; Municipal Elections; Sustainability; Political Discourse; Environment; Conservation; Climate Change.

LISTA DE FIGURAS

Figura 1 – Pipeline de análise das propostas e metadados. Fonte: Elaborado pelo autor (2025).	9
Figura 2 – Pipeline da modelagem dos dados. Fonte: Elaborado pelo autor (2025).	12
Figura 3 – Instruções do Prompt. Fonte: Elaborado pelo autor (2025).	14
Figura 4 – Distribuição do tamanho dos textos das propostas. Fonte: Elaborado pelo autor (2025).	16
Figura 5 – Tamanho médio das propostas por região. Fonte: Elaborado pelo autor (2025).	17
Figura 6 – Tamanho médio das propostas por capital. Fonte: Elaborado pelo autor (2025).	17
Figura 7 – Tamanho médio das propostas por partido. Fonte: Elaborado pelo autor (2025).	18
Figura 8 – Distribuição do Número de municípios por Região. Fonte: Elaborado pelo autor (2025).	19
Figura 9 – Distribuição do Número de municípios por Estado. Fonte: Elaborado pelo autor (2025).	19
Figura 10 – Distribuição do Número de Candidatos por Região. Fonte: Elaborado pelo autor (2025).	20
Figura 11 – Distribuição do número de candidatos por Estado. Fonte: Elaborado pelo autor (2025).	20
Figura 12 – Distribuição do Número de Candidatos por Município. Fonte: Elaborado pelo autor (2025).	21
Figura 13 – Distribuição de faixa etária dos candidatos. Fonte: Elaborado pelo autor (2025).	22
Figura 14 – Distribuição do nível de escolaridade entre os candidatos. Fonte: Elaborado pelo autor (2025).	22
Figura 15 – Distribuição do gênero dos candidatos por região. Fonte: Elaborado pelo autor (2025).	23
Figura 16 – Municípios com maior número de candidatos. Fonte: Elaborado pelo autor (2025).	23
Figura 17 – Candidatos que tentaram reeleição. Fonte: Elaborado pelo autor (2025).	24
Figura 18 – Partidos com mais candidatos registrados. Fonte: Elaborado pelo autor (2025).	24

Figura 19 – Partidos com o maior número total de gastos de campanha. Fonte: Elaborado pelo autor (2025).	25
Figura 20 – Partidos com o maior número de candidatos eleitos. Fonte: Elaborado pelo autor (2025).	25
Figura 21 – Partidos mais frequentemente presentes em coligações. Fonte: Elaborado pelo autor (2025).	26
Figura 22 – Rede de coligações partidárias. Fonte: Elaborado pelo autor (2025).	27
Figura 23 – Rede de coocorrência de palavras-chave do léxico ambiental com maior filtragem. Fonte: Elaborado pelo autor (2025).	28
Figura 25 – Mapa temático de eixos ambientais predominantes por município no Brasil. Fonte: Elaborado pelo autor (2025).	30
Figura 26 – Top 5 eixos temáticos predominantes por região. Fonte: Elaborado pelo autor (2025).	31
Figura 27 – Distribuição espacial dos eixos ambientais predominantes na região Norte. Fonte: Elaborado pelo autor (2025).	32
Figura 28 – Distribuição espacial dos eixos ambientais predominantes na região Nordeste. Fonte: Elaborado pelo autor (2025).	33
Figura 29 – Distribuição espacial dos eixos ambientais predominantes na região Sul. Fonte: Elaborado pelo autor (2025).	34
Figura 30 – Distribuição espacial dos eixos ambientais predominantes na região Centro-Oeste. Fonte: Elaborado pelo autor (2025).	35
Figura 31 – Distribuição espacial dos eixos ambientais predominantes na região Sudeste. Fonte: Elaborado pelo autor (2025).	35
Figura 32 – Top 30 palavras-chave relacionadas ao meio Ambiente nas propostas das capitais usando LLM. Fonte: Elaborado pelo autor (2025).	36

LISTA DE QUADROS

Quadro 1 – Resumo dos trabalhos relacionados	
Fonte: Elaborado pelo autor (2025).	7
Quadro 2 – Descrição dos atributos do conjunto de dados mais relevantes ao estudo	10
Quadro 3 – Palavras associadas aos eixos temáticos	
Fonte: Elaborado pelo autor (2025).	29

LISTA DE ABREVIATURAS E SIGLAS

ACM SIGKDD	Association for Computing Machinery – Special Interest Group on Knowledge Discovery and Data Mining
CLD	Causal Loop Diagram (Diagrama de Laços Causais)
CRISP-DM	Cross Industry Standard Process for Data Mining (Processo Padrão Interindustrial de Mineração de Dados)
DS_CARGO	Descrição do cargo concorrido
DS_COMPOSICAO_COLIGACAO	Partidos que compõem a coligação
DS_GENERO	Descrição do gênero do candidato
DS_GRAU_INSTRUCAO	Descrição do grau de instrução do candidato
DST	Data Science Trajectories (Trajetórias de Ciência de Dados)
IBGE	Instituto Brasileiro de Geografia e Estatística
INMET	Instituto Nacional de Meteorologia
LLM	Large Language Model (Modelo de Linguagem de Grande Escala)
MDB	Movimento Democrático Brasileiro
NM_CANDIDATO	Nome do candidato
NM_PARTIDO	Nome do partido
NM_UE	Nome da Unidade Estadual
NOVO	Partido Novo
OCR	Optical Character Recognition (Reconhecimento Óptico de Caracteres)
ODS	Objetivos de Desenvolvimento Sustentável
PCB	Partido Comunista Brasileiro
PCO	Partido da Causa Operária
PDT	Partido Democrático Trabalhista
PL	Partido Liberal
PLN	Processamento de Linguagem Natural
PMB	Partido da Mulher Brasileira
PP	Progressistas
PRD	Partido Renovador Democrático
PSB	Partido Socialista Brasileiro
PSDB	Partido da Social Democracia Brasileira

PSOL	Partido Socialismo e Liberdade
PSD	Partido Social Democrático
PSTU	Partido Socialista dos Trabalhadores Unificado
SEEG	Sistema de Estimativas de Emissões de Gases de Efeito Estufa
SG_UE	Sigla da Unidade Estadual
SG_UF	Sigla da Unidade Federal
SQ_CANDIDATO	Número identificador do candidato
ST_REELEICAO	Status de reeleição do candidato
TSE	Tribunal Superior Eleitoral
UNIÃO	União Brasil

SUMÁRIO

1	INTRODUÇÃO	1
1.1	Objetivos	1
1.2	Organização do trabalho	2
2	TRABALHOS RELACIONADOS	3
2.1	Fundamentação Teórica	7
3	MATERIAIS E MÉTODOS	9
3.1	Aquisição dos dados	10
3.2	Fusão dos metadados	10
3.3	Filtragem e conversão	11
3.4	Análise	11
3.5	Validação	13
4	RESULTADOS	16
4.1	Distribuição geral	18
4.2	Atividades partidárias	24
4.3	Discurso ambiental político	28
5	CONSIDERAÇÕES FINAIS	37
5.1	Contribuições do estudo	39
5.2	Dificuldades encontradas	40
5.3	Trabalhos futuros	40
	REFERÊNCIAS	41

1 INTRODUÇÃO

O avanço da ciência de dados, particularmente das técnicas avançadas de análise de dados, está impulsionando transformações em vários domínios, incluindo o campo da ciência política, convertendo dados em instrumentos de influência política (Ruppert *et al.*, 2017). No Brasil, o processo de informatização nacional facilitou um maior acesso a dados históricos eleitorais extensos e detalhados a partir de 1945, fornecidos pelo Tribunal Superior Eleitoral (TSE), permitindo assim a pesquisa empírica em ciência de dados (Silva *et al.*, 2020). Esse desenvolvimento cria oportunidades para o emprego de técnicas de ciência de dados focadas na análise de metadados eleitorais e do conteúdo das propostas de governo apresentadas pelos candidatos a prefeito, ambos disponíveis publicamente através do TSE. Além disso, a aplicação de técnicas de Inteligência Artificial (IA) a grandes conjuntos de dados está se tornando cada vez mais importante para identificar padrões com precisão e gerar *insights*, apoiando assim a tomada de decisões oportunas e informadas (Stroparo *et al.*, 2024).

Nesse contexto de maior disponibilidade de dados públicos eleitorais, as eleições municipais representam um momento propício para análise de dados, especialmente para características dos candidatos, suas coligações, partidos e comunicação com o eleitorado para revelar tendências relevantes sobre representatividade, estratégias de poder e mobilização política nos municípios como já abordado em trabalhos que exploram dados do TSE ao nível municipal como os em Jacintho *et al.* (2021). A escolha do eleitor é fortemente influenciada por fatores como as propostas políticas apresentadas, os traços de personalidade dos candidatos e os sentimentos gerados durante a campanha, o que reforça a importância da comunicação estratégica e da construção de imagem no contexto das disputas municipais (Silva; Mainardes, 2019).

1.1 Objetivos

A opinião pública exerce influência direta tanto sobre os temas abordados quanto sobre as posições adotadas nos discursos políticos, evidenciando um padrão de responsividade dos representantes às preferências da maioria (Hager; Hilbig, 2020). A análise do conteúdo político permite identificar padrões, aspectos textuais e possíveis vieses temáticos que podem indicar a relevância que esses políticos pretendem dar a diferentes demandas sociais e políticas em seus futuros mandatos. Entre essas demandas, destaca-se as questões ambientais. Nesse sentido, partidos e políticos são umas das fontes de destaque de informações a respeito de questões como as mudanças climáticas (Ejaz *et al.*, 2022).

Apesar da disponibilidade maior de dados abertos, observa-se uma escassez de trabalhos na literatura que integrem a análise dos perfis demográficos e partidários dos candidatos com técnicas de ciência de dados aplicadas ao conteúdo de suas propostas de governo, especialmente no contexto das eleições municipais brasileiras. Essa lacuna é ainda mais crítica no contexto municipal, marcado por ampla diversidade regional, partidária e linguística das propostas que tornam a tarefa ainda mais desafiadora. Nesse cenário, torna-se promissora a utilização de abordagens computacionais que combinem o uso de metadados eleitorais com análises textuais avançadas, permitindo revelar padrões discursivos e estruturais de forma escalável.

À luz do contexto apresentado, o presente trabalho tem como objetivo uma análise dos dados dos candidatos à prefeitura nas eleições municipais brasileiras de 2024, abarcando aspectos como gênero e escolaridade dos perfis demográficos, uma caracterização dos perfis partidários para entender as preferências de registro de candidaturas, além de aspectos textuais das propostas de governo, como o volume textual médio por região, partido e capital e, investigar quais são as prioridades ambientais expressas nos discursos desses candidatos. Para aprofundar o entendimento dos conteúdos dessas propostas, são aplicadas técnicas de mineração de texto por meio de uma abordagem híbrida que combina um modelo de *Large Language Model* (LLM) com um método de expansão de léxico baseados em *Word Embeddings*, visando identificar com maior precisão e sensibilidade os principais eixos temáticos presentes nos discursos políticos. Modelos de linguagem de grande porte e pré-treinados, como o GPT, têm alcançado desempenho elevado em diversas tarefas de PLN (Min *et al.*, 2021).

1.2 Organização do trabalho

O restante desta monografia está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados mais relevantes, discutindo abordagens anteriores de ciência de dados aplicadas em um contexto político. A Seção 3 descreve os materiais e métodos utilizados, incluindo o pipeline híbrido que combina as abordagens já mencionadas e descritas anteriormente. A Seção 4 apresenta os resultados obtidos a partir da análise exploratória dos metadados dos candidatos, das estatísticas textuais das propostas de governo e da identificação do viés no discurso político. Por fim, a Seção 5 apresenta as considerações finais do estudo e aponta direções para trabalhos futuros.

2 TRABALHOS RELACIONADOS

Considerando a natureza transdisciplinar do estudo, envolvendo ciências da computação, ciências políticas e aspectos ambientais, apenas para citar alguns, há uma vasta gama de trabalhos relacionados. A seguir é apresentada uma revisão da literatura do tipo *ad hoc* (Ralph; Baltes, 2022), com enfoque no uso de ciência de dados na política, discutindo também trabalhos voltados para as técnicas utilizadas, mais especificamente, grandes modelos de língua.

No estudo realizado por Cernev e Fenner (2020), é aplicada uma abordagem baseada em sistemas para identificar e priorizar os Objetivos de Desenvolvimento Sustentável (ODS). Na abordagem, foi desenvolvido um *Causal Loop Diagram* (CLD), inicialmente criando subsistemas para os setores ambientais, sociais e econômicos. Esse método permitiu visualizar as relações causais de cada ODS, classificadas como de influência direta ou indireta. Dentre as relações analisadas, mudança climática, aumenta a necessidade de ações climáticas (ligadas ao ODS 13). Já as ameaças cibernéticas podem atrapalhar o avanço de tecnologias, prejudicando o desenvolvimento de cidades sustentáveis (ODS 11), trabalho decente e crescimento econômico (ODS 8), e o acesso à energia limpa e acessível (ODS 7), afetando negativamente o avanço de outros ODS com os quais eles estão conectados. Essa representação foi feita de acordo com os possíveis riscos globais com maior impacto segundo os estudos em Fekete *et al.* (2018).

Em Anwar *et al.* (2019), os autores propõem uma metodologia que investiga a presença da temática das mudanças climáticas nos manifestos eleitorais dos três principais partidos políticos do Paquistão durante as eleições de 2018. É realizada uma abordagem que combina análise qualitativa com quantitativa. Outrossim, na etapa qualitativa, foi conduzida uma análise da frequência desses termos em manifestos partidários que indicam o nível de seriedade tratada para esse problema pelos partidos. A análise revelou que palavras-chave relacionadas ao meio ambiente como “*climate change*”, “*environment*” e “*carbon*” foram as mais recorrentes nos partidos e evidenciam as principais preocupações político-ambientais do Paquistão. Esse trabalho reforça a relevância de identificar não apenas quais termos são utilizados, mas também com que frequência aparecem no discurso político voltado a questões ambientais para entender sua importância na agenda política. O presente trabalho visa expandir essa abordagem usando técnicas de mineração de texto para viabilizar identificar novas palavras-chave recorrentes no discurso político-ambiental de candidatos às eleições municipais brasileiras

Tratando-se de trabalhos mais voltados para dados, em Vasconcelos *et al.* (2022)

as pessoas autoras construíram o CandiDATA, um conjunto de dados aprimorado e padronizado com informações sobre as eleições brasileiras entre 1945 e 2020. O principal objetivo do trabalho era desenvolver um pipeline estruturado que automatizasse a coleta, padronização e enriquecimento dos dados eleitorais disponibilizados pelo Tribunal Superior Eleitoral (TSE), mitigando problemas como inconsistências de formato, lacunas semânticas e variações entre diferentes eleições e anos. A abordagem proposta inclui desde a raspagem automatizada de informações brutas, passando por etapas de limpeza e transformação, até a exportação de arquivos tabulares e JSON acompanhados de dicionários de dados. A base resultante foi construída com informações obtidas no portal do TSE e que foram categorizados com rótulos como “Candidatos”, “Resultados” e entre outros. Entre os desafios da construção dessa base destaca-se a correção de campos nulos, mal formatados ou com alguma inconsistência passaram por correções ou foram removidos enquanto outros novos campos foram adicionados. Dessa forma, essa base de dados foi aprimorada para preencher lacunas existentes em informações eleitorais de antes de 1996. Com a criação da base de dados do CandiDATA e um *dashboard* complementar com informações sobre candidatos, partidos políticos e informações estatísticas, a ferramenta se torna útil para dar assistência a pesquisadores e gerentes interessados em informações mais detalhadas sobre as eleições brasileiras. O presente trabalho propõe expandir essa abordagem, indo além dos metadados eleitorais estruturados ao realizar uma análise semântica dos discursos políticos dos candidatos em propostas de governo com enfoque em temas ambientais nas eleições municipais de 2024.

Em Dekoninck e Schmuck (2022), os autores exploraram o papel dos influenciadores digitais das chamadas *Social Network Sites* (SNSs) na mobilização de comportamentos políticos e ambientais de seus seguidores por meio de um estudo longitudinal de curto prazo realizado durante as eleições parlamentares nacionais de 2019 na Áustria. Adiante, os autores analisaram se a participação política online induzida por influenciadores consegue aumentar a participação política offline dos indivíduos com o passar do tempo. O trabalho usou uma abordagem com métodos quantitativos como análise de caminhos (*path analysis*) por meio do pacote *lavaan* em R, além de modelos autorregressivos que avaliaram efeitos temporais das interações online sobre comportamentos offline. Com um conjunto de dados obtidos por meio de duas enquetes online, inicialmente com 1.206 participantes e aplicadas em dois momentos com intervalo de seis semanas cada. Os autores observaram que tanto as intenções de comportamento pró-ambiental quanto a participação online conseguiram prever, ao longo do tempo, um aumento na participação política offline. O trabalho aponta que não houve análise do conteúdo do discurso político dos influenciadores.

Em relação à ciência de dados e um fluxo de análise, Jacintho *et al.* (2020) desenvolveram um pipeline de ciência de dados para reconhecer padrões eleitorais nas eleições

presidenciais brasileiras entre 1994 e 2018. Foram utilizados no estudo autocorrelações espaciais (*Moran's Index*) para identificar padrões espaciais e *clustering* hierárquico para identificação de padrões temporais. O uso dessas ferramentas permitiu um estudo mais aprofundado a respeito do comportamento eleitoral dos brasileiros com o passar dos anos. Os resultados da análise de autocorrelação espacial mostraram que cidades vizinhas tendem a ter comportamentos eleitorais semelhantes. Adiante, a análise do *clustering* hierárquico permitiu descobrir que cidades vizinhas mudam esse comportamento eleitoral similarmente. Os autores apresentam um pipeline mais flexível e direto para orientação nas etapas de coleta, pré-processamento e análise espacial e padrões temporais nas eleições brasileiras. Ademais, a facilidade de obtenção da base de dados por ser disponibilizada publicamente por meio do site do TSE e acesso aos códigos utilizados pelos autores permitem que a reprodutibilidade de seus resultados seja garantida e mais facilmente revisada e inspecionada pela comunidade. Seguindo o direcionamento georreferenciado descrito por Jacintho *et al.* (2020), este trabalho propõe um pipeline para análise eleitoral, realizando uma análise espacial da distribuição geográfica dos municípios com maiores recorrências do tema das mudanças climáticas em propostas de candidatos. Esta análise espacial, busca compreender se há *clusters* regionais de engajamento ambiental nas eleições municipais de 2024.

A respeito da aplicabilidade de LLMs na ciência política, em Li *et al.* (2024) é produzido um exame sobre como explorar o potencial dessa ferramenta para o estudo de sistemas políticos e instituições pela ciência política. É descrito que os LLMs são pré-treinados com grandes volumes de textos oriundos da internet, o que lhes proporciona uma capacidade de compreender e analisar padrões linguísticos complexos. Em razão disso, essa ferramenta está sendo utilizada em documentos legislativos e discursos políticos para processar e interpretar conjuntos extensos de conteúdo político e permitindo uma análise automatizada em larga escala desses dados. Os autores abordam a engenharia de prompt pra refinamento como uma boa alternativa para direcionar a LLM na sumarização de aspectos de textos legislativos de forma precisa e concisa, fornecendo tarefas claras e estrutura de resposta definida. O presente estudo visa expandir essa abordagem para o contexto das eleições municipais brasileiras de 2024 ao analisar *corpus* de propostas por meio de LLM a fim de identificar discursos relacionados à causa ambiental.

Nos estudos realizados em Linegar *et al.* (2023), os autores buscam introduzir cientistas políticos aos LLMs, apresentando suas possibilidades de aplicação, os desafios técnicos e éticos envolvidos, e orientando o uso dessa ferramenta pela comunidade científica. Dentre os benefícios, o emprego de LLMs é capaz de flexibilizar os esforços de anotações manuais a respeito de conteúdos políticos e identificar elementos como polaridade política, sentimentos e discurso de ódio, assim como a simplificação do processo de extração de informação no discurso político para análise de discurso por cientistas sociais. Os LLMs

possuem vantagens na extração de informações em comparação com outros métodos por sua maior capacidade de compreender contexto. Assim como nesses estudos, este trabalho aplica uma abordagem híbrida de expansão de léxico e LLM com engenharia de prompt para o conteúdo textual de propostas políticas, garantindo que a análise dos discursos eleitorais seja feita levando em consideração a busca por documentos com foco nos temas ambientais.

No estudo desenvolvido em Muchlinski *et al.* (2021), os autores apresentam uma nova abordagem para estimar mais precisamente a conexão entre eleições e violência, utilizando uma rede neural convolucional para identificar manifestações de violência política diretamente a partir de textos não estruturados. Os tweets coletados foram rotulados manualmente e convertidos em vetores numéricos reais, chamados de *Word Embeddings*, utilizando o modelo *Word2Vec*. Essa técnica permite agrupar palavras que apareçam em contextos semelhantes, como “violência” e “agressão”. Os resultados evidenciaram que essa técnica obteve melhor performance do que outras abordagens tradicionais de *bag-of-words* para análise textual, por captarem o contexto linguístico das palavras e melhorarem a identificação e classificação de eventos violentos. O presente estudo adota uma abordagem que também envolve a busca de palavras-chave, aplicando técnicas de expansão de léxico com o uso de *Word Embeddings* do tipo *Word2Vec* para identificar termos ambientais relevantes em propostas políticas não estruturadas. O objetivo é ampliar as possibilidades de análise do conteúdo político-ambiental de forma mais contextualizada e eficaz.

Quadro 1 – Resumo dos trabalhos relacionados
Fonte: Elaborado pelo autor (2025).

Referência	Objetivo	Resultados	Comparativo com o Presente Trabalho
(Cernev; FENNER, 2020)	Priorizar ODS com abordagem de sistemas.	ODS 13 (Ação Climática) como chave para reduzir riscos globais.	Inspira foco em riscos climáticos e ODS no contexto das eleições.
(Anwar <i>et al.</i> , 2019)	Analisar discurso climático nos manifestos do Paquistão.	Alta recorrência de termos ambientais nos manifestos.	Método expandido para contexto brasileiro com mineração de texto.
(Vasconcelos <i>et al.</i> , 2022)	Criar base CandiDATA para dados eleitorais (1945–2020).	Dados eleitorais limpos, padronizados e acessíveis.	O presente trabalho aplica análise textual nas propostas, além dos metadados.
(Dekoninck; Schmuck, 2022)	Analisar influência de influenciadores na mobilização política.	Influenciadores ampliam participação política e ambiental.	Foco nos discursos dos candidatos pelo seu alto poder de influência eleitoral.
(Jacintho <i>et al.</i> , 2021)	Analisar padrões eleitorais espaciais e temporais no Brasil.	Cidades vizinhas têm comportamento eleitoral semelhante.	Aplica análise geoespacial para padrões regionais no discurso ambiental.
(Li <i>et al.</i> , 2024)	Aplicar LLMs na análise de documentos legislativos.	LLMs processam textos políticos com alta eficiência.	Adota LLMs para análise de propostas eleitorais focando em eixos temáticos.
(Linegar <i>et al.</i> , 2023)	Introduzir LLMs na análise de discurso político.	LLMs facilitam análise de sentimento e polarização.	A abordagem é usada para extrair temas de discurso nas eleições municipais.
(Muchlinski <i>et al.</i> , 2021)	Detectar violência política via redes neurais e embeddings.	Word2Vec supera modelos tradicionais na análise contextual.	A abordagem inspira uso de Word2Vec para expandir léxico nas propostas.
(Fekete <i>et al.</i> , 2018)	Mapear riscos globais (ambientais, econômicos e tecnológicos).	Riscos ambientais lideram o ranking. Cresce preocupação com eventos extremos e perda de biodiversidade.	Fundamenta o foco ambiental do trabalho, conectando riscos globais com demandas locais como saneamento.

2.1 Fundamentação Teórica

Os *Word Embeddings* consistem em representações vetoriais densas capazes de mapear palavras para um espaço numérico multidimensional, no qual palavras de significado semelhante ficam próximas entre si. Além disso, ferramentas como *Word2Vec*, *GloVe*

e *FastText* são amplamente utilizadas por identificar automaticamente os contextos e similaridades semânticas, potencializando tarefas de expansão de vocabulário e análise de coocorrência de termos. Trabalhos como o de Muchlinski *et al.* (2021) validam o uso dessas representações para identificar precisamente, termos relevantes em textos políticos não estruturados.

Em relação aos *Large Language Models* (LLMs), como GPT-3 e GPT-4o, revolucionaram a forma de abordar tarefas de processamento de linguagem natural em larga escala. Esses modelos, pré-treinados com grandes volumes de dados textuais, são capazes de realizar desde a análise e sumarização até a extração automática de temas em textos longos e complexos. Conforme apontado por Li *et al.* (2024) e Linegar *et al.* (2023), o uso de LLMs tem permitido, inclusive na ciência política, automatizar tarefas como a identificação de tópicos, extração de informações e análise de discurso, gerando bons resultados em geral.

A expansão de léxico refere-se ao processo de ampliação do conjunto de palavras e termos relevantes para determinado domínio temático, podendo ser feita de forma manual, automática ou híbrida para ampliar o alcance na busca por temáticas emergentes em grandes volumes de textos políticos. A aplicação dessa abordagem se mostrou especialmente útil diante da variedade e heterogeneidade dos textos analisados.

Em síntese, a escolha por técnicas como *Word Embeddings*, LLMs e expansão de léxico se justifica por sua aplicação e eficiência nos resultados em trabalhos na literatura, mas também por seu potencial de lidar com o grande volume e diversidade de dados textuais das propostas de governo. Dessa forma, o trabalho assegura maior sensibilidade na detecção dos principais eixos ambientais discutidos no contexto eleitoral brasileiro.

3 MATERIAIS E MÉTODOS

Este estudo realiza uma análise do conteúdo ambiental presente nas propostas de governo dos candidatos às eleições municipais de 2024, aliada a uma investigação dos metadados associados a esses candidatos. A estrutura metodológica foi inspirada no modelo *Data Science Trajectories* (DST) proposto por Martinez-Plumed *et al.* (2021), que incorpora atividades exploratórias ao método *Cross-Industry Standard Process for Data Mining* (CRISP-DM) (Wirth; Hipp, 2000). Essa abordagem foi empregada devido a necessidade de etapas exploratórias cruciais para o desenvolvimento das análises subsequentes.

O framework experimental foi desenhado com o objetivo de permitir organização, clareza e replicabilidade. Para tal, seguindo bons princípios de ciência de dados, estruturou-se um fluxo de execução conforme apresentado na Figura 1.

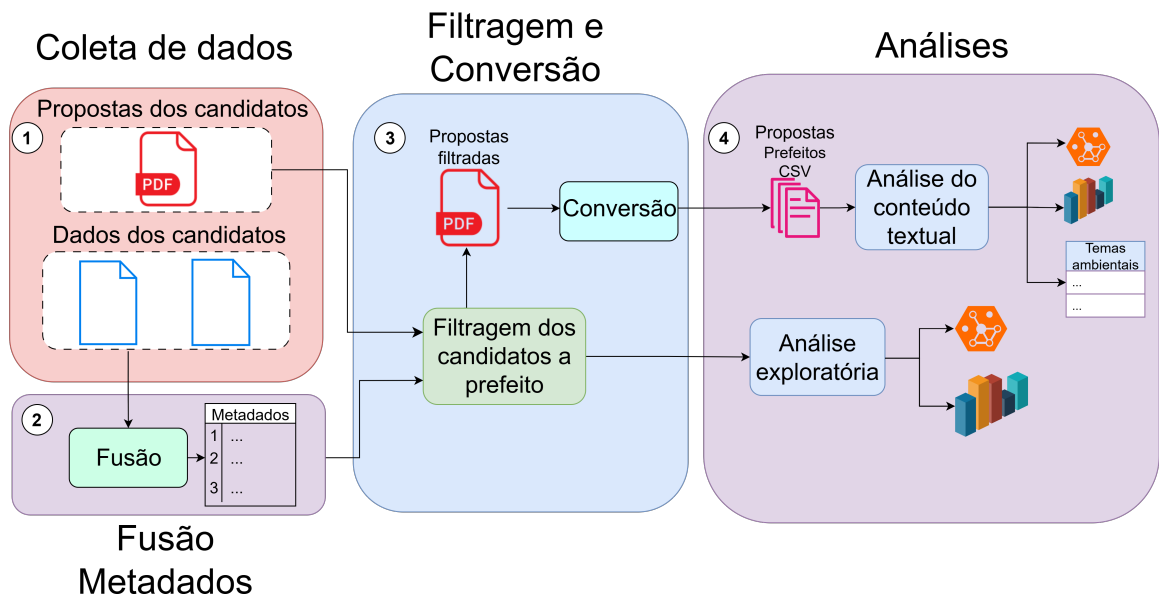


Figura 1 – Pipeline de análise das propostas e metadados.

Fonte: Elaborado pelo autor (2025).

Na fase inicial, foram determinados quais eram os objetivos e delimitada a problemática a ser explorada. Em razão disso, foi feita uma revisão na literatura do tipo *Ad Hoc* (Ralph; Baltes, 2022), apresentada na seção anterior. Esta revisão teve como objetivos uma melhor compreensão da relevância da temática do trabalho, identificar ferramentas e técnicas já aplicadas anteriormente em estudos de ciência de dados, além de revelar possíveis lacunas na literatura. As atividades foram agrupadas nas seguintes etapas: (I) Coleta de Dados, (II) Fusão de Metadados, (III) Filtragem e Conversão e (IV) Análises; as quais encontram-se detalhadas nas subseções a seguir.

3.1 Aquisição dos dados

Este trabalho utilizou dados abertos disponibilizados pelo governo federal via TSE ¹, coletados manualmente por meio de links de download fornecidos na página. Dessa forma, foram coletadas as informações dos candidatos às eleições municipais brasileiras de 2024 em formato tabulares. Esses registros encontravam-se dispostos em dois arquivos, um com as informações gerais dos candidatos e outro com informações complementares. As propostas dos candidatos estavam disponibilizadas no formato PDF. No total, foram coletados mais de 463.622 registros de candidatos aos cargos municipais de prefeito, vice-prefeito e vereador. Importante pontuar que o TSE não disponibilizava arquivos em separado para prefeitos, vice-prefeitos e vereadores, por isto este número de registros. As propostas de governo somaram 15.786.

3.2 Fusão dos metadados

Esta etapa consistiu na fusão dos arquivos de informações principais e complementares dos candidatos. A tarefa teve como objetivo centralizar os valores relevantes da coleta em um único conjunto, permitindo análises mais abrangentes e consistentes sobre os perfis dos candidatos. A Tabela ?? apresenta os principais atributos do conjunto de dados fusionado, que foram selecionados com base no escopo do presente trabalho.

Quadro 2 – Descrição dos atributos do conjunto de dados mais relevantes ao estudo

Atributo	Descrição
<i>SG_UF</i>	Sigla da Unidade Federal
<i>SG_UE</i>	Sigla da Unidade Estadual
<i>NM_UE</i>	Nome da Unidade Estadual
<i>DS_CARGO</i>	Descrição do cargo concorrido
<i>SQ_CANDIDATO</i>	Número identificador do candidato
<i>NM_CANDIDATO</i>	Nome do candidato
<i>TP_AGREMIACAO</i>	Tipo de agremiação do candidato
<i>NM_PARTIDO</i>	Nome do partido
<i>NM_COLIGACAO</i>	Nome da coligação
<i>DS_COMPOSICAO_COLIGACAO</i>	Partidos compostos na coligação
<i>DS_GENERO</i>	Descrição do gênero do candidato
<i>DS_GRAU_INSTRUCAO</i>	Descrição do grau de instrução do candidato
<i>ST_REELEICAO</i>	Status de reeleição do candidato

Fonte: Elaborado pelo autor (2025).

A Tabela ?? apresenta um total de 12 atributos, incluindo localização do candidato, cargo, partido, coligação, gênero, instrução/escolaridade e reeleição. Com os registros

¹<<https://dadosabertos.tse.jus.br/dataset/candidatos-2024>>

coletados e consolidados, foram identificados algumas inconsistências que necessitavam de filtragem, tal como descrito a seguir.

3.3 Filtragem e conversão

Durante o fusionalamento dos dados percebeu-se que eles apresentavam ruídos, como existência de duplicatas do número identificador dos candidatos que conseguiram avançar para um segundo turno, totalizando 204 registros. Ademais, em relação às propostas de governo, foi descoberto que alguns candidatos não tinham nenhuma proposta associada e algumas propostas não possuíam candidatos correspondentes e nem dados suficientes para identificar todas as suas informações corretamente pelo conteúdo textual. Para evitar interferência nos resultados das análises, tanto os registros duplicados quanto as propostas sem correspondência foram removidos da base de dados. Posteriormente, houve uma fase de filtragem dos candidatos em relação ao cargo, mantendo apenas os que estavam concorrendo ao cargo de prefeito e, com os metadados resultantes dessa filtragem, foram selecionadas as propostas que apresentavam um vínculo válido com candidatos, associando ambas as informações por meio do identificador único de cada registro de candidato.

A etapa posterior consistiu na conversão dos arquivos de propostas, originalmente em formato PDF, para o formato tabular, utilizando a técnica de Reconhecimento Óptico de Caracteres, mais conhecido por *Optical Character Recognition* (OCR). Esse método permitiu extrair o conteúdo textual presente em propostas com apenas imagens, convertendo-as para um formato legível e manipulável por computadores, sem comprometer a integridade do documento original. A partir disso, foi possível aplicar procedimentos de pré-processamento no texto das propostas, como a remoção de acentos e a normalização de capitalização das palavras no pré-processamento para evitar interferências nos resultados posteriores da análise textual.

3.4 Análise

Para fins de melhor visualização dos elementos da etapa de análises proposta na figura 1, foi criado um pipeline detalhando a criação do léxico e do prompt para LLM para a modelagem dos dados conforme apresentado na figura 2.

Na etapa de análises, os metadados dos candidatos foram explorados para entender a distribuição dos atributos descritos na Tabela ???. Representações gráficas desses dados auxiliaram em identificar características gerais dos perfis dos candidatos. Em paralelo, foram construídos grafos de coocorrência entre partidos com métricas de grau médio e modularidade utilizando a ferramenta de criação de redes *Gephi*, conforme descrito na abordagem metodológica por Couto *et al.* (2024). Esses grafos revelaram como estão

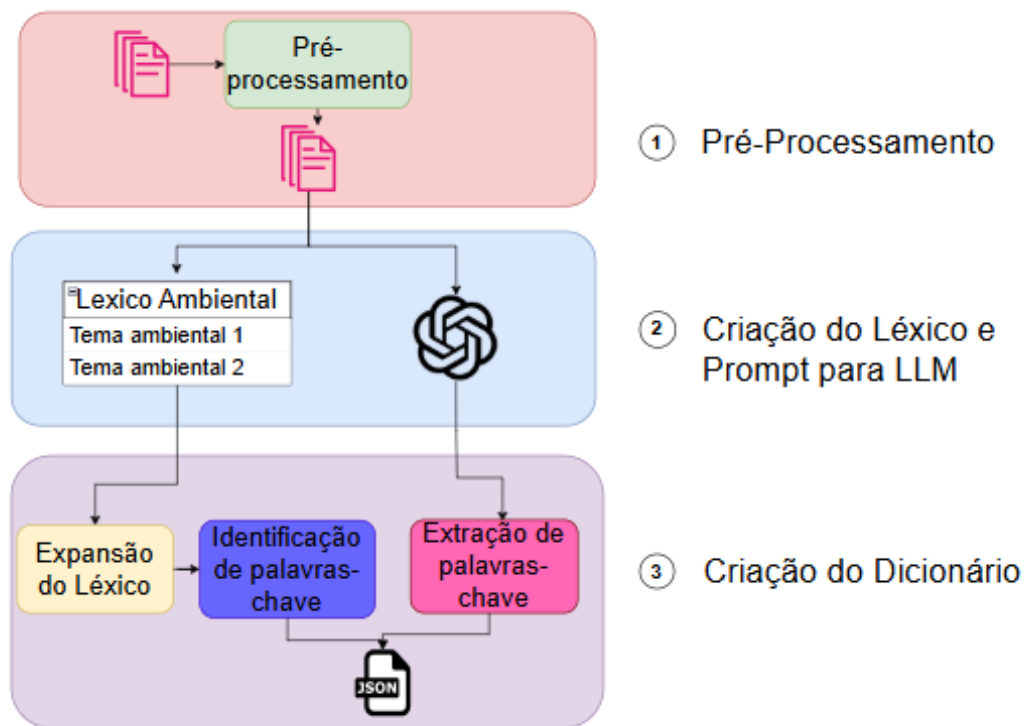


Figura 2 – Pipeline da modelagem dos dados.

Fonte: Elaborado pelo autor (2025).

definidos os diferentes sub agrupamentos políticos em um contexto mais atual. Considerando as tendências de análise de conteúdos textuais e o estado da arte em modelos de língua de grande escala, o presente estudo adotou uma abordagem híbrida visando maior confiabilidade, combinando o poder das LLMs com uma abordagem validada por especialista do domínio, baseada em Léxico.

Do ponto de vista técnico, a abordagem híbrida utilizada combinou a aplicação de engenharia de prompt com expansão de léxico baseada na técnica de *Word Embeddings*, essas abordagens foram exploradas de acordo com o ilustrado na figura 2. A utilização da engenharia de prompt com uma LLM foi respaldada pela sua capacidade de realizar interpretações contextuais e extrair termos relevantes conforme é explorado em outros trabalhos na literatura. Por outro lado, a expansão do léxico via *Word Embeddings*, mais especificamente, utilizando o *Word2Vec*, foi empregada com o objetivo de capturar relações semânticas entre palavras e identificar termos pertinentes de diferentes eixos ambientais que não estavam no conjunto original, considerando variações linguísticas. Esse método consiste em representar palavras como vetores em um espaço multidimensional, de forma que palavras com significados semelhantes estejam próximas entre si. Essa estratégia permitiu aumentar a cobertura da análise sem comprometer a precisão, tornando o processo mais sensível a nuances textuais e menos dependente de listas fixas de palavras. Convém destacar que todo o processo contou com especialistas do domínio em contato com

especialistas de análise, monitorando entrada e saída a fim de verificar a sua acuracidade e confiabilidade, em uma abordagem conhecida por *human-in-the-loop* (Zhang *et al.*, 2019).


No primeiro momento, as pessoas autoras construíram um conjunto de palavras que foram pré-selecionadas manualmente para formar um léxico ligado aos termos ambientais. Exemplos de termos incluem: “mudanças climáticas” e “desastres climáticos”, que seriam utilizados posteriormente para a etapa de expansão. Em razão das variações linguísticas presentes na língua portuguesa, fez-se necessária a expansão desse léxico. A segunda abordagem envolveu a aplicação da LLM para extração de palavras-chave em todas as propostas de candidatos à prefeitura de todos as capitais dos estados brasileiros já em formato tabular, isso é feito com o auxílio da técnica de engenharia de prompt para guiar a IA generativa com instruções e tarefas definidas para a identificação do item ambiental aplicado ao contexto do texto e, posteriormente, a sua devida extração definida na etapa. O trabalho adotou como LLM o ChatGPT-4o, considerando amplo uso, boa aceitação na comunidade de aprendizado de máquina e desempenho superior aos demais modelos proprietários. O prompt adotado é descrito na figura 3 com instruções detalhadas e especificadas do método de extração das palavras, além de informações adicionais de candidatos para fins de identificação posterior. Após isso, segue a resposta obtida em retorno descrita na Figura 3, confirmando a compreensão da tarefa pelo modelo por meio de uma exemplificação da estrutura de resposta.

Com as informações obtidas nestas abordagens, os resultados da etapa de análises foram validados com o auxílio dos conhecimentos de cientistas de dados e políticos especialistas de domínio, conforme descrito na seção de validação a seguir.

3.5 Validação

A etapa de validação dos dados, métodos e resultados foi conduzida de forma contínua e colaborativa, com foco na consistência e coerência com cada etapa do pipeline da Figura 1. Inicialmente, os resultados preliminares foram apresentados às pessoas autoras com *expertise* em ciência de dados para revisão técnica das etapas de filtragem, pré-processamento, expansão de léxico e extração de palavras-chave via LLM. Essa prática garantiu o alinhamento das decisões metodológicas com as boas práticas da área e a detecção antecipada de falhas ou inconsistências nos dados processados.

Além disso, foram realizadas apresentações periódicas à pessoa autora especialista de domínio nas áreas de ciência política. Esses encontros permitiram validar os termos extraídos e as visualizações espaciais obtidas, assegurando que os eixos temáticos identificados nos discursos políticos correspondiam aos debates contemporâneos sobre mudanças climáticas e meio ambiente. Durante esse processo, a especialista forneceu sugestões de




Solicito apenas a estrutura de um dicionário Python chamado candidato que contenha as seguintes informações extraídas de propostas de governo enviadas como arquivos:

- Nome do candidato: O nome completo do candidato.
- Partido: O partido político ao qual o candidato está afiliado.
- Estado: A sigla do estado de origem do candidato (ex.: SP, RJ).
- Município: O município correspondente, caso aplicável.
- Palavras-chave relacionadas ao meio ambiente: Uma lista de termos relevantes identificados no texto (ex.: sustentabilidade, energia renovável, desmatamento).
- Descrição: Um texto descritivo indicando se a proposta é favorável ou desfavorável ao meio ambiente, com base na análise do conteúdo.

As propostas de governo serão fornecidas como arquivos CSV separados para análise aqui no chatgpt. Os textos nesses arquivos devem ser processados no chatgpt para:

- Identificar as palavras-chave relacionadas ao meio ambiente.
- Estruturar as informações no formato especificado acima.

Após analisar o texto, retorne como output a estrutura do dicionário python mencionada anteriormente já com a adesão das informações obtidas do texto em suas variáveis correspondentes



Claro! Abaixo está a **estrutura de um dicionário Python chamado candidato**, conforme solicitado, já adaptada para armazenar as informações extraídas de uma proposta de governo:

```
candidato = {
    "nome": "Fulano de Tal",
    "partido": "XYZ",
    "estado": "SP",
    "municipio": "São Paulo",
    "palavras_chave_meio_ambiente": [
        "sustentabilidade",
        "energia renovável"
    ],
    "descricao": "A proposta apresenta ações concretas para a preservação ambiental, indicando compromisso com o desenvolvimento sustentável e políticas de mitigação climática."
}
```

-Os valores são exemplos simulados — eles devem ser **substituídos dinamicamente** a partir da leitura e análise dos arquivos CSV.

-A chave "palavras_chave_meio_ambiente" deve conter uma **lista de termos ambientais identificados no texto**.

-A chave "descricao" deve conter um **resumo avaliativo** baseado no conteúdo da proposta.

Quando quiser, envie um ou mais arquivos CSV com as propostas que deseja analisar, e eu extraio as informações automaticamente. Deseja enviar um agora?

Figura 3 – Instruções do Prompt.

Fonte: Elaborado pelo autor (2025).

palavras para construção do léxico inicial como “desastre climático” e “mitigação”, novas visualizações gráficas para identificação de possíveis clusters de temas e ranking de tópicos ambientais predominantes, além do confronto com a literatura cinza, incluindo dados públicos da plataforma Sistema de Estimativas de Emissões e Remoções de Gases de Efeito Estufa (SEEG)¹ sobre emissões de gases de efeito estufa por município e níveis de precipitação pluvial estaduais fornecidos pelo Instituto Nacional de Meteorologia² (INMET) durante o período de 2024.

Subsequentemente, utilizou-se um arquivo de malhas municipais auxiliar contendo dados geográficos dos municípios brasileiros fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE)³. O conteúdo foi integrado com informações de ocorrências de eixos ambientais nos discursos dos candidatos em suas respectivas cidades de candidatura. Após a remoção de valores nulos e duplicatas, possíveis distorções no mapeamento dos dados municipais foram minimizados. A partir disso, o arquivo serviu como referência para validar

¹ <<https://plataforma.seeg.eco.br/>>

² <<https://portal.inmet.gov.br/dadoshistoricos>>

³ <<https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html>>

a análise espacial com visualização dos eixos predominantes de cada município.

4 RESULTADOS

Os resultados foram apresentados conforme os principais achados obtidos a partir da análise dos dados referentes às eleições municipais brasileiras de 2024. Os resultados abrangem aspectos como o perfil dos candidatos, distribuição geográfica, características demográficas, atividades partidárias e a análise textual das propostas de governo dos candidatos, com enfoque nos diferentes eixos temáticos definidos.

Em primeiro momento, são exploradas características gerais das propostas, como seu tamanho médio por município, partido e região, permitindo inferir possíveis padrões sobre o grau de detalhamento dos programas de governo em diferentes grupos partidários e regiões geográficas. Em seguida, são analisados os perfis dos candidatos, considerando aspectos como idade, gênero, escolaridade e tentativa de reeleição, além da distribuição dos candidatos por estado e município. Na sequência, é detalhada na seção, a atuação dos partidos políticos, evidenciando aqueles com maior número de candidaturas, maiores gastos de campanha, maior número de prefeitos eleitos e maior presença em coligações.

A Figura 4 apresenta a distribuição da frequência do número de palavras nas propostas dos candidatos à prefeitura nas eleições municipais de 2024

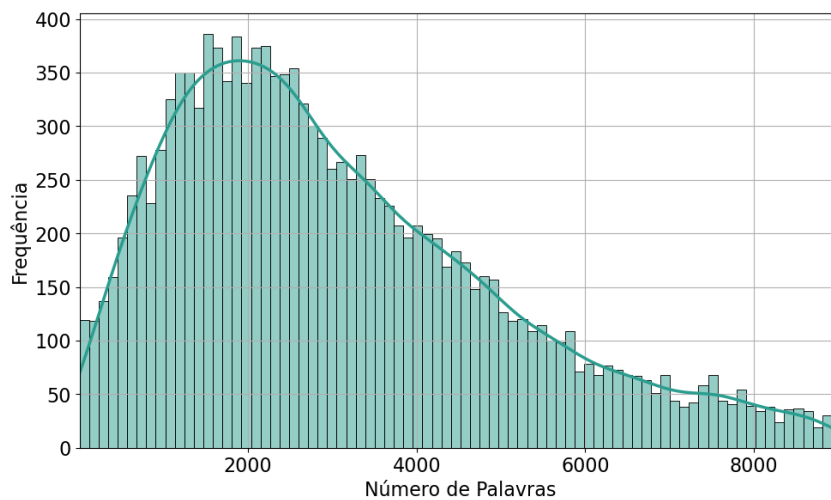


Figura 4 – Distribuição do tamanho dos textos das propostas.
Fonte: Elaborado pelo autor (2025).

A distribuição é assimétrica à direita, com um pico de frequência em torno de 2.000 palavras. Evidenciam-se poucos textos ultrapassando 6.000 palavras. A cauda longa sugere que há propostas excepcionalmente extensas, embora sejam casos menos frequentes. É possível observar a preferência da maioria dos candidatos por propostas com textos mais concisos

A figura 5 disponibiliza o tamanho médio das propostas por região.

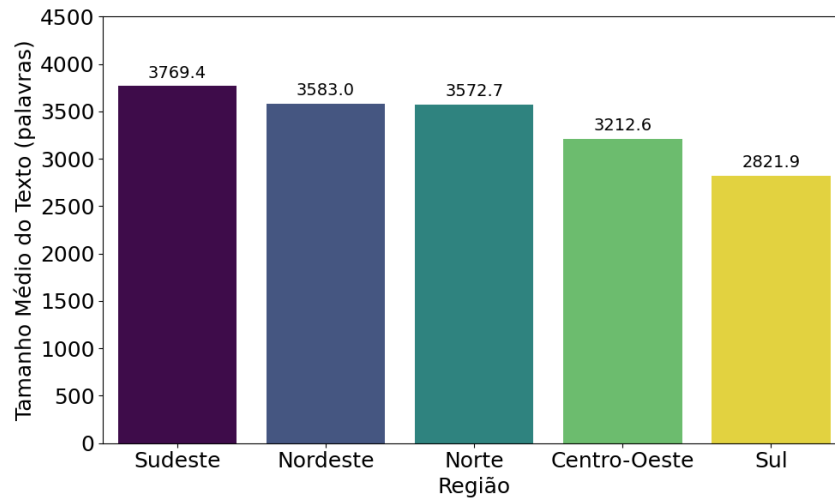


Figura 5 – Tamanho médio das propostas por região.
Fonte: Elaborado pelo autor (2025).

A região Sudeste lidera com a maior média, seguida pelo Nordeste e Norte. O Centro-Oeste, com uma média de 3.212 palavras por texto, e o Sul, com 2821, apresentam os menores valores médios. Esses resultados estão em consonância com os dados de propostas por capitais apresentados na figura

A Figura 6 mostra o tamanho médio dos textos por capital.

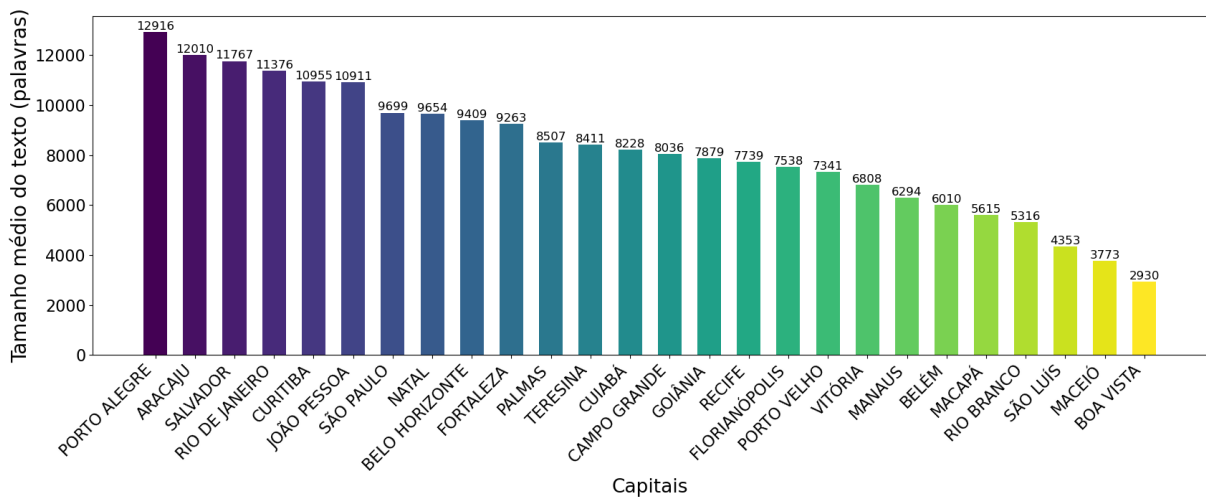


Figura 6 – Tamanho médio das propostas por capital.
Fonte: Elaborado pelo autor (2025).

Conforme mostrado na figura 6, Porto Alegre se destaca com a maior média de tamanho médio por capital do Brasil com aproximadamente 12.916 palavras, seguida por Salvador e Aracaju 12010 e 11767 palavras respectivamente. Em contraste, cidades como Boa Vista, Belém e Rio Branco possuem uma das menores médias com 5316, 6010 e 2930 palavras respectivamente. Esses resultados expõem uma predominância de textos

mais extensos em regiões do nordeste e sudeste do país em relação à região norte, com as propostas mais curtas.

Em relação a essas médias de propostas por partidos políticos, a figura 7 detalha o tamanho médio das propostas por partido.

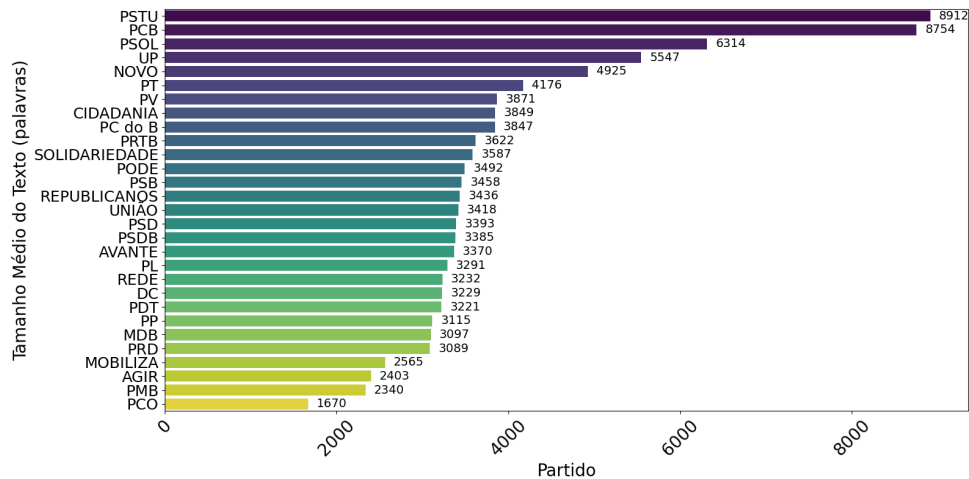


Figura 7 – Tamanho médio das propostas por partido.

Fonte: Elaborado pelo autor (2025).

Os partidos de esquerda, como o Partido Socialista dos Trabalhadores Unificado (PSTU) (8.912 palavras), o Partido Comunista Brasileiro (PCB) (8.754) e o Partido Socialismo e Liberdade (PSOL) (6.314), lideram com os textos mais longos, enquanto partidos como o Partido da Causa Operária (PCO) (949) e o Partido da Mulher Brasileira (PMB) (1.670) apresentam propostas significativamente mais curtas. Essa disparidade pode refletir o grau de detalhamento programático dos partidos, sugerindo que legendas com orientação ideológica mais marcada tendem a apresentar propostas mais extensas.

4.1 Distribuição geral

A Figura 8 apresenta a distribuição do número de municípios por região do Brasil, conforme registrado na base de dados das eleições municipais de 2024.

Observa-se que as regiões Nordeste e Sudeste concentram o maior número de municípios do país, com 1.725 e 1.659 municípios, respectivamente. O Sul aparece em seguida, com 1.178 municípios, enquanto o Centro-Oeste (466) e o Norte (448) apresentam as menores quantidades. Esse padrão evidencia a heterogeneidade na distribuição territorial dos municípios brasileiros, resultado de fatores históricos, econômicos e demográficos que moldaram a divisão administrativa do país. As regiões Nordeste e Sudeste, além de possuírem as maiores populações, são também as mais fragmentadas administrativamente, o que reflete na expressiva quantidade de municípios.

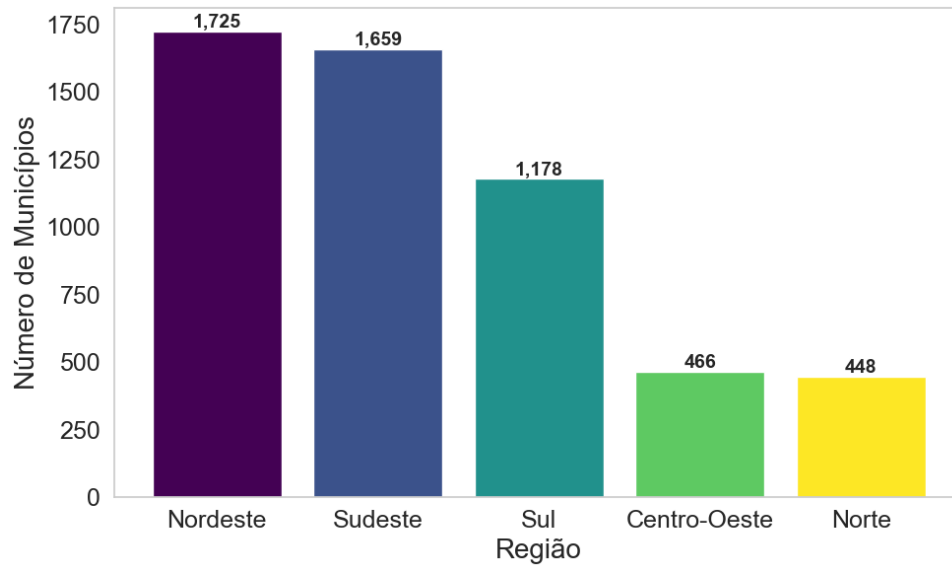


Figura 8 – Distribuição do Número de municípios por Região.
Fonte: Elaborado pelo autor (2025).

O Centro-Oeste e o Norte, por outro lado, possuem extensões territoriais muito amplas, mas contam com menor densidade municipal, indicando grandes áreas com baixa divisão administrativa. Esse cenário reforça o papel do Nordeste e do Sudeste como principais polos administrativos do país em termos de quantidade de municípios, enquanto Norte e Centro-Oeste se destacam por sua menor fragmentação territorial.

A Figura ?? exibe a distribuição do número de municípios por unidade federativa (UF) no Brasil, conforme dados das eleições municipais de 2024.

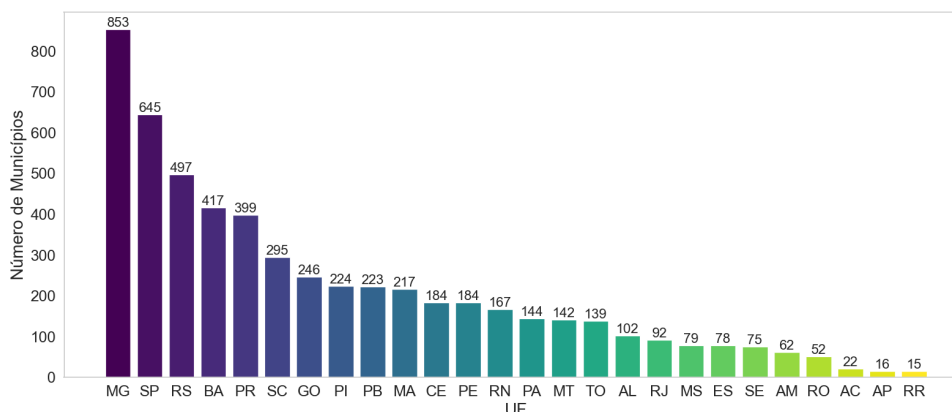


Figura 9 – Distribuição do Número de municípios por Estado.
Fonte: Elaborado pelo autor (2025).

Destaca-se que Minas Gerais é, de longe, o estado com maior número de municípios, totalizando 853, seguido por São Paulo (645) e Rio Grande do Sul (497). Esses estados, além de apresentarem grande população e relevância econômica, também se caracterizam por uma intensa fragmentação administrativa, enquanto estados do Norte e Centro-Oeste possuem menor fragmentação. Essas diferenças impactam diretamente na dinâmica política,

na gestão pública e no processo eleitoral, influenciando o número de candidatos e disputas municipais.

Na figura 10, A região Sudeste predomina com 5,111 candidatos, representando a maior parcela de registros. Isso acontece possivelmente pela sua densidade populacional em relação a outras regiões como Norte e Centro-Oeste que, embora dominem grande parcela da extensão territorial brasileira, não representam nem 20% da população.

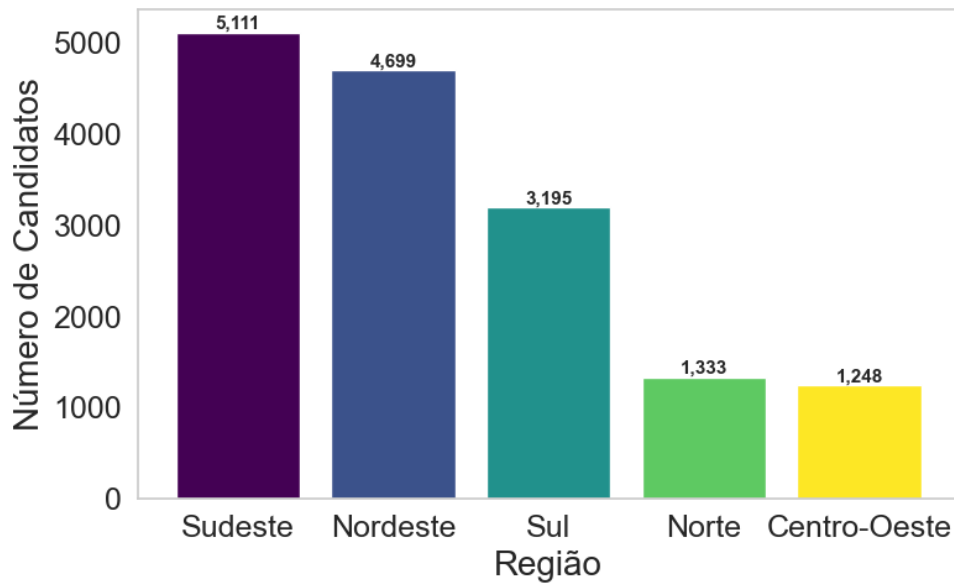


Figura 10 – Distribuição do Número de Candidatos por Região.
Fonte: Elaborado pelo autor (2025).

A Figura 11 apresenta a distribuição total dos candidatos ao cargo de prefeito em todo o Brasil separados por Estados.

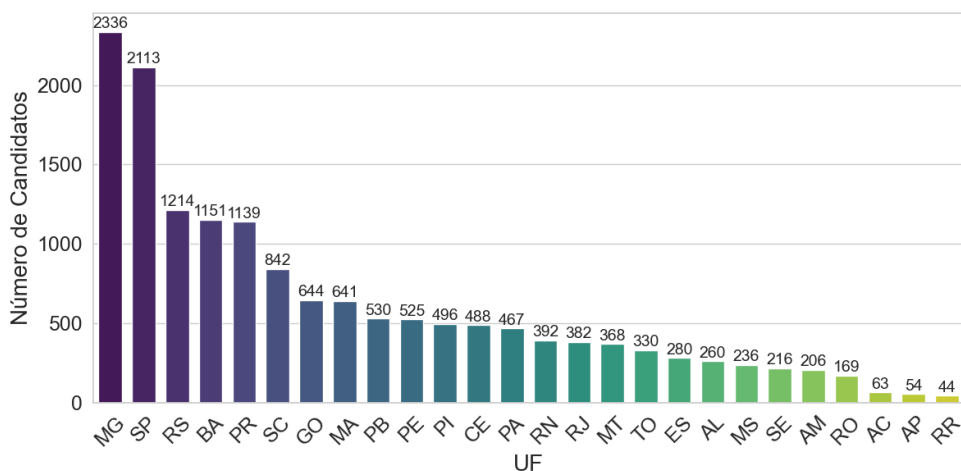


Figura 11 – Distribuição do número de candidatos por Estado.
Fonte: Elaborado pelo autor (2025).

Na figura, é possível revelar uma forte concentração em regiões mais populosas e com maior número de municípios, como Minas Gerais e São Paulo no Sudeste e Bahia no

Nordeste. Essa disparidade está associada tanto à dimensão territorial quanto à divisão administrativa dos estados, que impactam diretamente o volume de disputas municipais. Em contraste, estados da Região Norte, como Amazonas e Pará, apresentam uma participação significativamente menor, refletindo sua menor densidade populacional, embora sejam os maiores estados do Brasil.

A Figura 12 apresenta a distribuição do número de candidatos a prefeito registrados por município nas eleições municipais brasileiras de 2024.

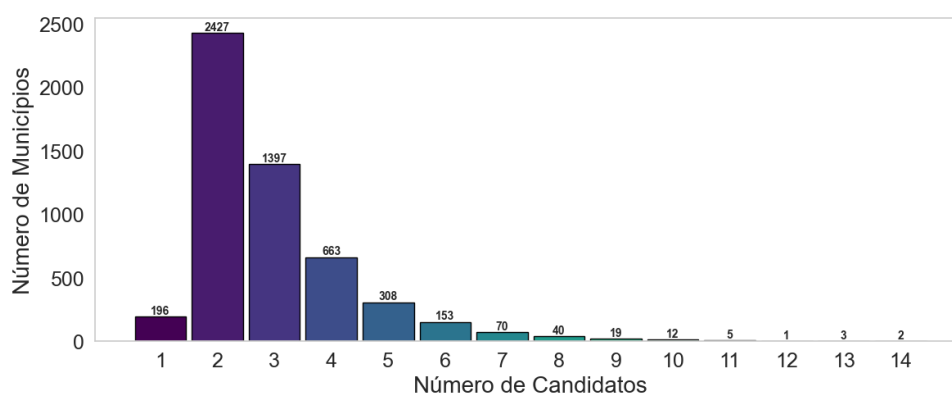


Figura 12 – Distribuição do Número de Candidatos por Município.
Fonte: Elaborado pelo autor (2025).

Observa-se que a distribuição é bastante desnivelada, com a grande maioria dos municípios registrando apenas dois ou três candidatos à prefeitura. O pico de frequência ocorre para municípios com dois candidatos (2.427 casos), seguido por três candidatos (1.397 casos) e quatro candidatos (663 casos). Esse padrão evidencia uma baixa competitividade na maioria dos municípios brasileiros e uma possível baixa pluralização política.

A Figura 13 apresenta o número de candidatos ao cargo de prefeito separados por faixas etárias.

Evidencia-se uma maior concentração de candidatos nas faixas etárias entre 40 e 59 anos. As faixas de 45–49 e 50–54 anos são as mais ocorrentes, com 2.459 e 2.355 candidatos, respectivamente. Em corroboração com isso, os estudos realizados por Eshima e Smith (2022) destacam uma preferência dos eleitores por candidatos na faixa dos 30 aos 55 anos, com uma queda estatisticamente significativa daqueles a partir da faixa entre 55 a 65 anos. Em razão disso, pode-se inferir que a distribuição etária dos candidatos está refletindo as preferências do eleitorado por perfis etários mais moderados.

A Figura 14 mostra que candidatos com maior nível de escolaridade, com destaque para os candidatos com curso superior, somam mais de 9147 registros, representando predominantemente 56.80% das candidaturas. Em seguida, candidatos com um diploma do ensino secundário representam 24.01%. As demais categorias com escolaridade menor têm

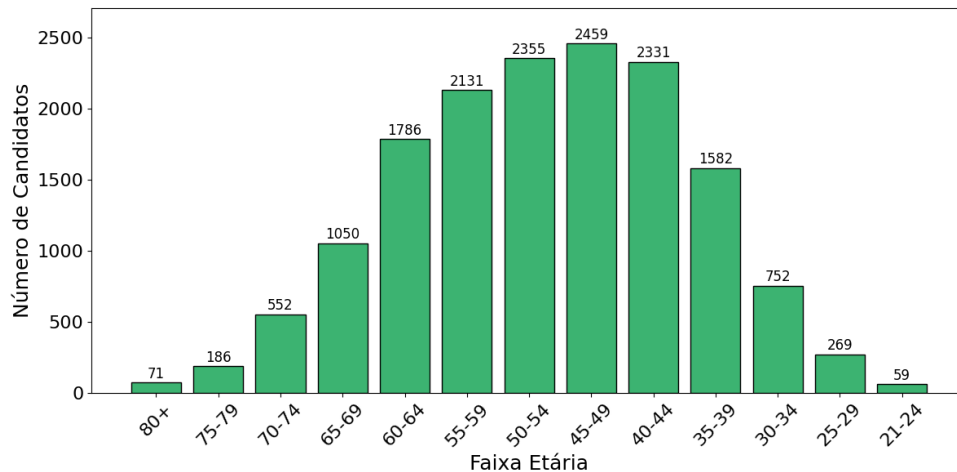


Figura 13 – Distribuição de faixa etária dos candidatos.
Fonte: Elaborado pelo autor (2025).

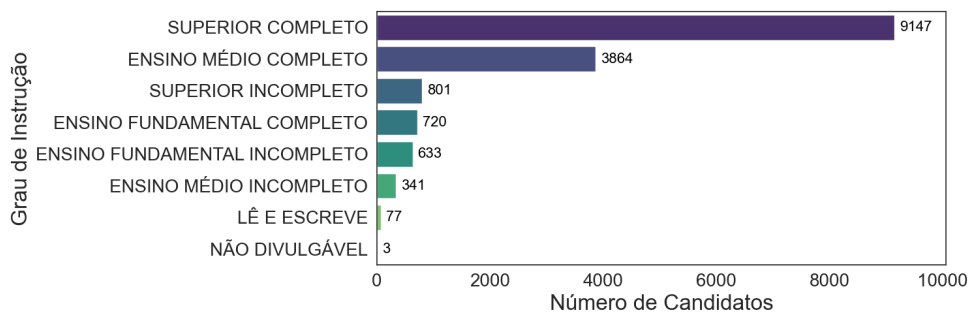


Figura 14 – Distribuição do nível de escolaridade entre os candidatos.
Fonte: Elaborado pelo autor (2025).

participação inferior a 10%. É possível inferir uma preferência em registrar candidatos com maior grau de instrução por serem considerados mais aptos a concorrer a cargos políticos importantes como o de prefeito.

A Figura 15 apresenta a distribuição de gênero dos candidatos em todas as regiões demográficas do Brasil. Pode-se inferir os candidatos do gênero masculino como os predominantes em todas as regiões do Brasil. Essa desigualdade de gênero é bastante expressiva ao destacar que as mulheres representam menos de 20% do total de candidatos em todas as regiões analisadas. Essa sub-representação feminina é recorrente, embora haja exigência legal de 30% de candidaturas femininas, o que, conforme estudo realizado em Ciocari e Mottinha (2021), não tem garantido maior presença efetiva das mulheres nos cargos eletivos.

A Figura 16 exhibe os municípios com maior número de candidaturas registradas para as eleições municipais. Pode-se destacar que Belém (PA) e Valença (BA) lideram o ranking, com 14 candidatos cada, seguidos por Campo Grande (MS) e São Domingos, com 13 candidaturas. Adiante, o restante da lista inclui cidades de diferentes portes e regiões do

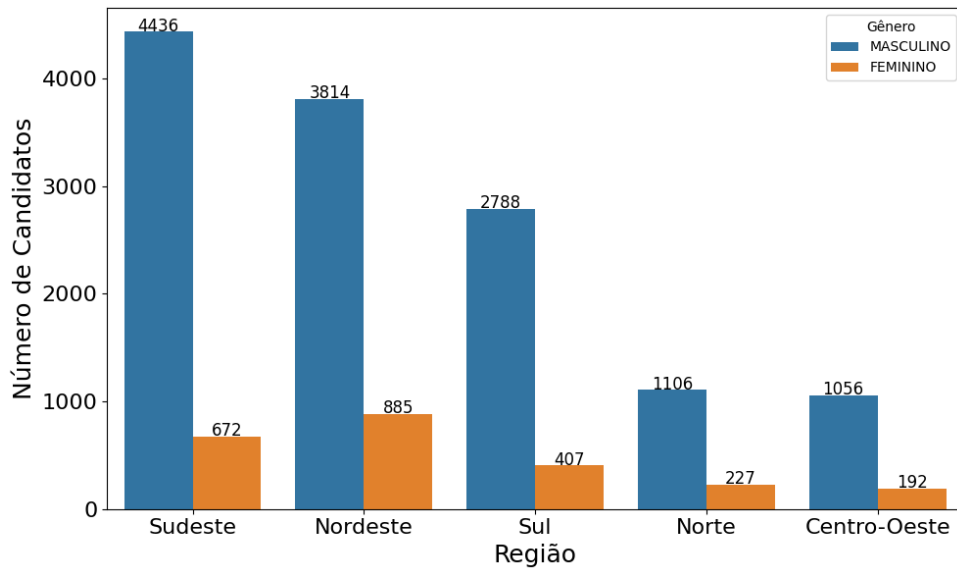


Figura 15 – Distribuição do gênero dos candidatos por região.
Fonte: Elaborado pelo autor (2025).

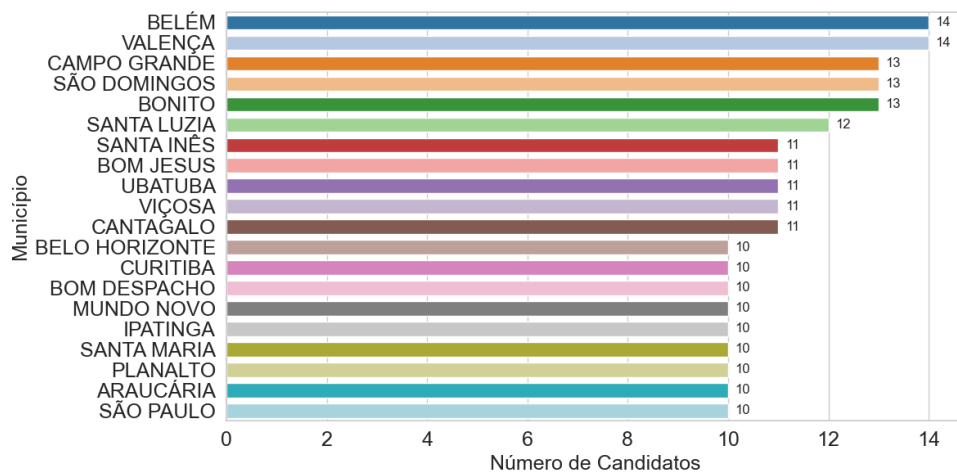


Figura 16 – Municípios com maior número de candidatos.
Fonte: Elaborado pelo autor (2025).

país, indicando que a competitividade eleitoral não se restringe a grandes centros urbanos, mas também está presente em municípios de menores escalas.

A Figura 17 apresenta um número expressivo de 3.039 candidatos que tentaram a reeleição para o cargo de prefeito, em relação ao total de 5.569 aos cargos de prefeito disputados. Evidencia-se que uma parcela significativa dos prefeitos em exercício opta por buscar a continuidade de seu mandato, o que pode refletir tanto a confiança na própria gestão quanto a percepção de vantagem competitiva em relação aos novos concorrentes.

Observa-se a taxa de sucesso na tentativa de reeleição de cerca de 81%. Dos candidatos que tentaram se manter no cargo, 2.460 obtiveram êxito, representando uma taxa de sucesso alta. Desse total, 2.455 prefeitos foram eleitos já no primeiro turno,

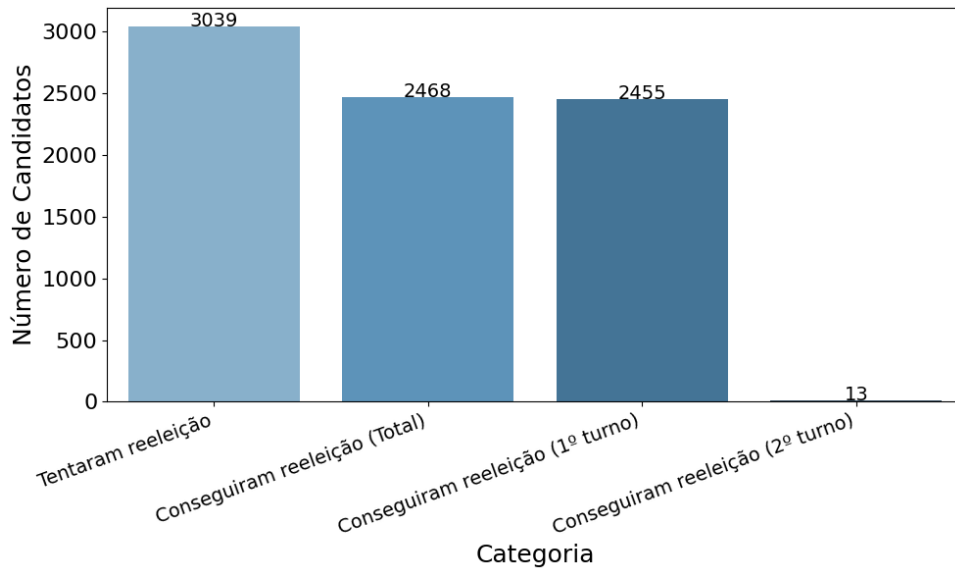


Figura 17 – Candidatos que tentaram reeleição.

Fonte: Elaborado pelo autor (2025).

enquanto apenas 13 tiveram que recorrer ao segundo turno para garantir sua permanência no cargo. Este dado evidencia que a alta taxa de reeleição tende a se concretizar nas tentativas iniciais de eleição, possivelmente pela maior visibilidade e consolidação política do candidato durante o mandato.

4.2 Atividades partidárias

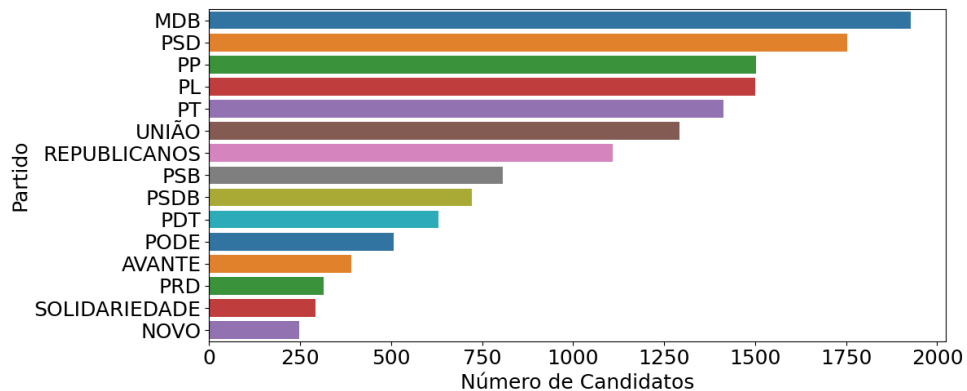


Figura 18 – Partidos com mais candidatos registrados.

Fonte: Elaborado pelo autor (2025).

O Movimento Democrático Brasileiro (MDB) lidera com folga, seguido por Partido Social Democrático (PSD), Progressistas (PP) e Partido Liberal (PL), que também apresentam forte presença nas disputas locais. Outros partidos tradicionais como o Partido dos Trabalhadores (PT), Partido da Social Democracia Brasileira (PSDB), Partido Democrático Trabalhista (PDT) e Partido Socialista Brasileiro (PSB) mantêm participação relevante. A distribuição reflete a fragmentação partidária do cenário político brasileiro. Observa-se ainda que partidos de menor expressão nacional, como o Partido Novo (NOVO)

e o Partido Renovador Democrático (PRD), apresentam registros mais modestos, o que pode estar associado a menor estrutura organizacional.

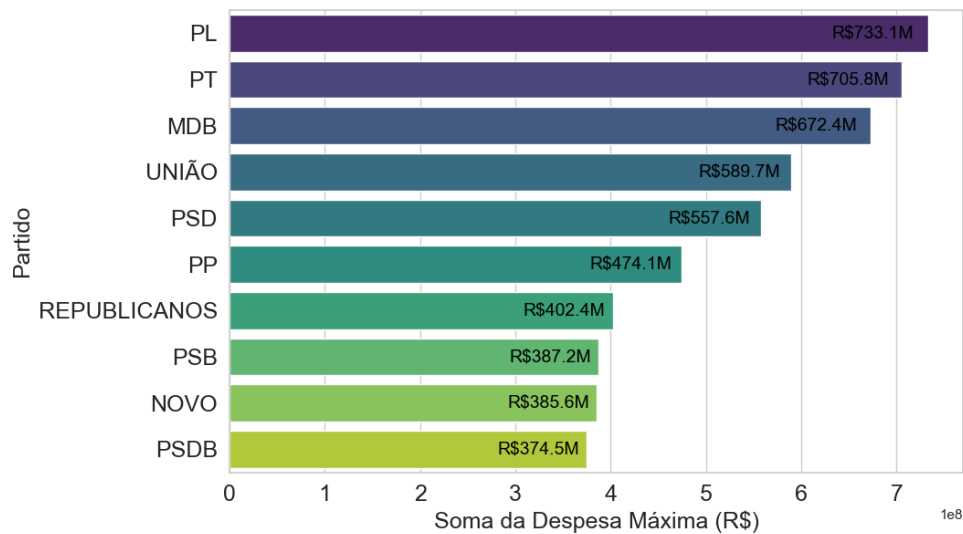


Figura 19 – Partidos com o maior número total de gastos de campanha.
Fonte: Elaborado pelo autor (2025).

Conforme apresentado na Figura 19, os partidos com maior soma da despesa máxima de campanha nas eleições municipais de 2024 foram o PL, com R\$733,1 milhões, seguido de perto por PT (R\$705,8 milhões) e MDB (R\$672,4 milhões). Esses valores indicam a alta capacidade de mobilização de recursos desses partidos, refletindo sua estrutura consolidada e abrangência nacional. Na sequência aparecem partidos como UNIÃO, PSD e PP, todos com mais de R\$400 milhões em despesas máximas autorizadas, evidenciando também forte presença nas disputas.

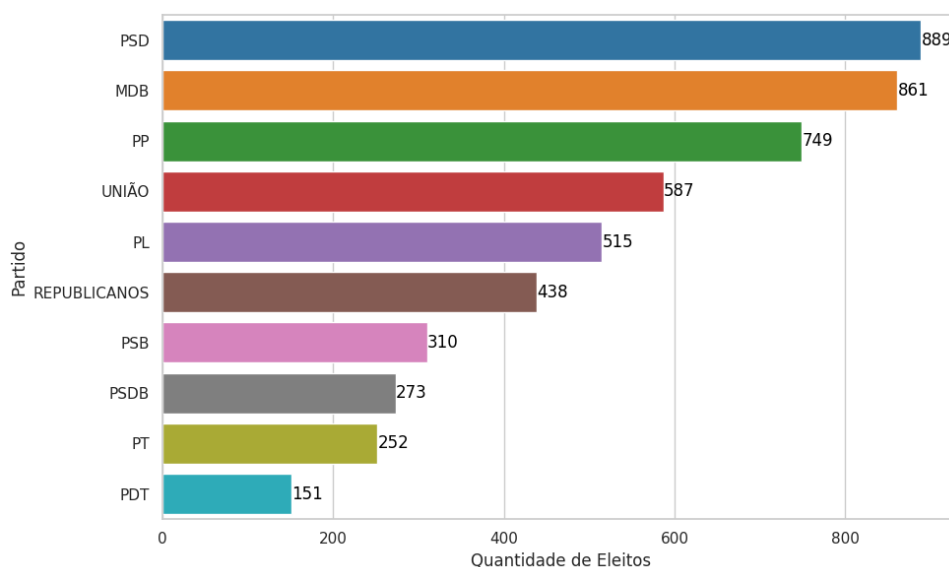


Figura 20 – Partidos com o maior número de candidatos eleitos.
Fonte: Elaborado pelo autor (2025).

Por meio da análise da Figura 20, observa-se que o PSD lidera com 889 prefeitos eleitos, seguido pelo MDB (861) e PP (749). Em contrapartida, partidos como PSB, PSDB, PT e PDT aparecem com quantidades mais modestas de eleitos. Pode-se inferir desses dados, ao comparar com a Figura 19, a predominância de partidos com candidatos com mais gastos de campanha entre os mais eleitos, conforme endossado nos estudos em Campos-Valdés *et al.* (2021).

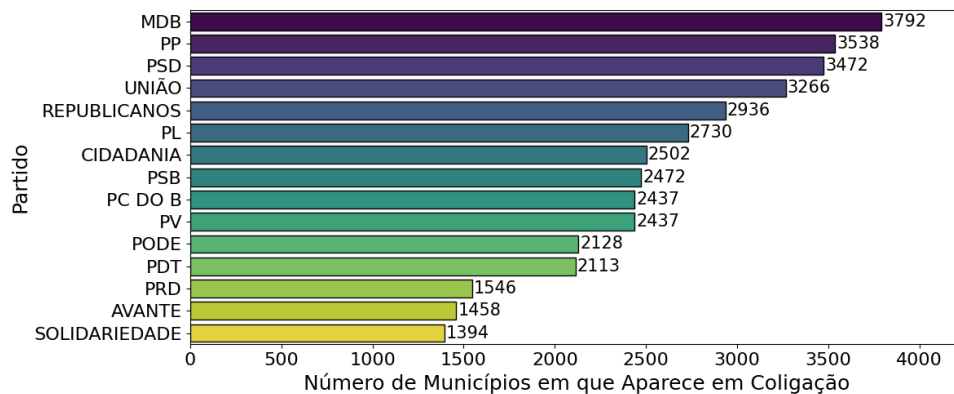


Figura 21 – Partidos mais frequentemente presentes em coligações.
Fonte: Elaborado pelo autor (2025).

A Figura 21 apresenta os partidos que mais participaram de coligações nas eleições municipais de 2024. O MDB aparece como o mais presente, integrando coligações em 3.792 municípios, seguido por PP (3.538), PSD (3.472) e UNIÃO (3.266). Em um comparativo com a Figura 20, nota-se uma correlação evidente entre a presença nas coligações e o número de candidatos eleitos, sugerindo que a formação de coligações tem um papel estratégico importante para o sucesso nas urnas.

A figura 22 apresenta um grafo das redes de coligações partidárias das eleições municipais brasileiras de 2024.

Na imagem do grafo, cada nó representa um partido político, enquanto as arestas representam as coligações vinculadas. O nome do nó proporcional ao tamanho do seu agrupamento, a espessura e a quantidade das conexões dessas arestas, refletem a intensidade das relações, e as cores distinguem agrupamentos com maior proximidade na rede.

Ao observar a estrutura do grafo, nota-se que determinados partidos ocupam posições de maior destaque em razão de seus agrupamentos robustos na rede. Entre eles, destacam-se MDB, PP, PSD, UNIÃO e PL, que não apenas apresentam mais conexões, mas também desempenham papel estratégico na articulação entre diferentes grupos políticos. Isso evidencia uma elevada capacidade desses partidos em formar alianças, maximizando sua competitividade eleitoral e aumentando probabilidade de terem candidatos eleitos.

O MDB, partido tradicional e com grande influência histórica no Brasil, visualizado

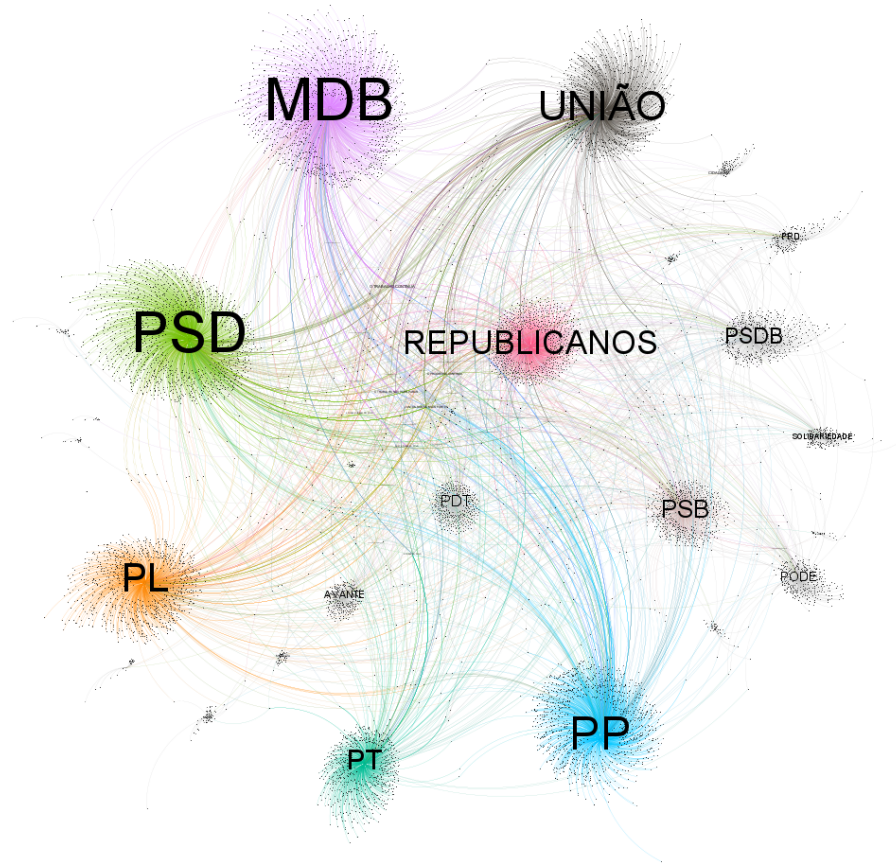


Figura 22 – Rede de coligações partidárias.
Fonte: Elaborado pelo autor (2025).

na região superior esquerda com coloração roxa, aparece como um dos principais articuladores, possuindo conexões densas tanto internas quanto externas. Da mesma maneira, o PP, identificado na parte inferior direita com coloração azul, também demonstra um elevado grau de conectividade, especialmente em articulações com partidos como PT e PSD. O bloco representado pelo PSD (em verde) mostra-se igualmente relevante, articulando-se fortemente tanto com partidos do seu entorno quanto com outras siglas centrais. O UNIÃO (em cinza-escuro), posicionado na parte superior, se destaca pela sua extensa ramificação, atuando como uma ponte entre diversos agrupamentos. Além dos partidos majoritários, observa-se que agremiações como Republicanos, PSDB, PL e PSB possuem papéis mais secundários na rede, com menor centralidade e menos conexões robustas. Entretanto, eles são pontos de articulação em nichos específicos.

Ao analisar o grafo da Figura 22, observa-se que partidos com maior centralidade e conectividade na rede de coligações, como MDB, PP, PSD, UNIÃO e PL, são em sua maioria aqueles que mais elegeram candidatos, conforme evidenciado também na Figura 20.

4.3 Discurso ambiental político

A Figura 23 apresenta um grafo bipartido com uma rede de coocorrência entre termos da expansão do léxico ambiental extraídos das propostas de governo dos candidatos à prefeitura.

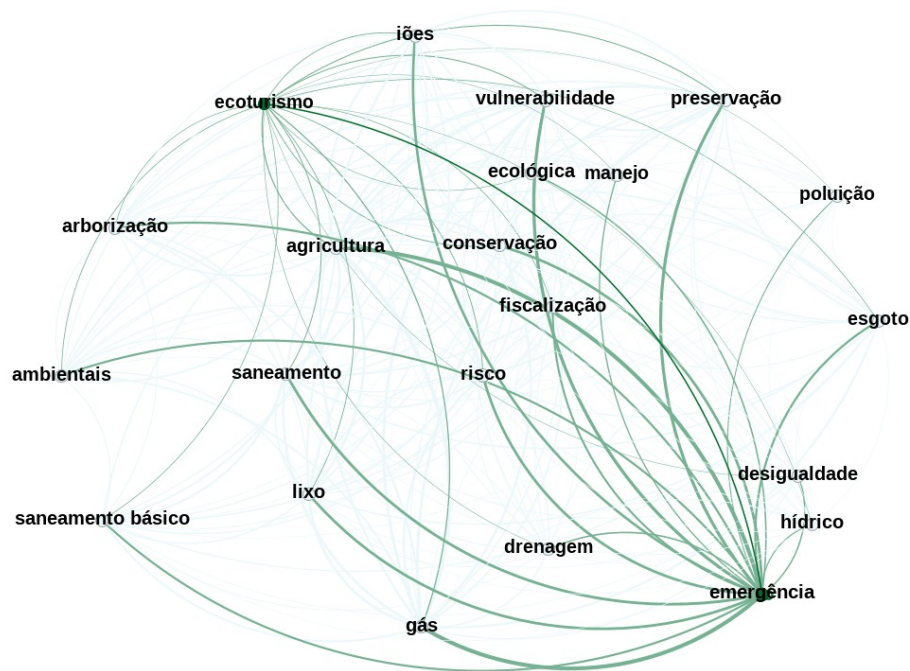


Figura 23 – Rede de coocorrência de palavras-chave do léxico ambiental com maior filtragem.

Fonte: Elaborado pelo autor (2025).

Cada nó da rede representa uma palavra-chave associada ao discurso temático. A conexão entre esses nós e a grossura das arestas representam o número de vezes que pares de palavras estavam presentes no mesmo discurso ambiental. Após aplicação de filtragem para destacar as conexões mais frequentes, observou-se que termos como “conservação”, “fiscalização”, “drenagem”, “emergência” e “desigualdade hídrica” ocupam posições de destaque na rede, indicando alta frequência de coocorrência e possível papel de importância no discurso político.

Outros termos, como “recuperação”, “manejo”, “ecossistema”, “saneamento”, “esgoto” e “preservação”, compõem subgrupos temáticos que se interconectam de maneira a formar um panorama abrangente das pautas presentes nas propostas. No entanto, a

densidade das conexões varia, evidenciando certos temas com maior protagonismo nos programas, enquanto outros surgem de forma periférica.

Portanto, a análise da rede de coocorrência revelou que o discurso ambiental nas propostas analisadas se organiza em torno de três eixos principais: I diz respeito à gestão dos recursos hídricos e ao enfrentamento de emergências urbanas; II envolve ações de conservação, manejo e fiscalização ambiental; III aborda o reconhecimento das desigualdades socioambientais, com ênfase naquelas relacionadas ao acesso à água e aos serviços urbanos de saneamento.

Quadro 3 – Palavras associadas aos eixos temáticos

Fonte: Elaborado pelo autor (2025).

Eixo Temático	Palavras associadas
Clima	mudanças climáticas, crise climática, emergência climática, aquecimento global, acidificação dos oceanos, variabilidade climática, mitigação, mitigação climática, adaptação climática, resiliência climática
Conservação e Biodiversidade	conservação, preservação, biodiversidade, ecologia, ecossistema, ecossistemas, fiscalização ambiental, unidades de conservação, manejo, reflorestamento, restauração florestal, silvicultura, arborização urbana, ecoturismo, impacto antrópico, atividade antropogênica
Desmatamento e Solo	desmatamento, desflorestamento, degradação do solo, desertificação, erosão, assoreamento, garimpo, uso do solo, recuperação de áreas degradadas
Água e Enchentes	enchentes, alagamentos, inundações, enxurradas, cheias, estiagem, seca, chuvas intensas, recursos hídricos, água potável, drenagem urbana, irrigação, dessalinização, abastecimento de água
Saneamento e Resíduos	saneamento, saneamento básico, esgotamento sanitário, esgoto, resíduos sólidos, coleta seletiva, reciclagem, logística reversa, aterro sanitário, lixo, compostagem, poluição, poluente, contaminação, resíduos perigosos, substâncias tóxicas
Desastres Naturais	desastres naturais, desastre climático, deslizamentos, desabamentos, escorregamentos, derrocadas, queimadas, incêndios florestais, defesa civil, plano de contingência, mapeamento de risco
Energia e Emissões	combustão, combustíveis, gás natural, emissões, gases de efeito estufa, carbono, dióxido de carbono (CO ₂), inventário de emissões, neutralidade de carbono, transição energética, energia renovável, energia solar, energia eólica, biomassa, biogás, eficiência energética, segurança energética
Compostos Químicos e Técnicos	amônia, bário, nitrogênio, dióxidos, óxidos, íons, secagem, evaporação, filtração, incineração, extração, adsorção, tratamento
Infraestrutura e Vulnerabilidade	congestionamento, precariedade, vulnerabilidade socioambiental, vulnerabilidades, risco, riscos, emergência, descarte irregular, coleta, custo-efetivo, microclima, ilha de calor, infraestrutura verde, permeabilidade do solo

A Tabela 3 apresenta o vocabulário expandido resultante da etapa de expansão lexical descrita na seção de metodologia. Inicialmente, o conjunto de termos obtido por meio dessa abordagem era significativamente maior; no entanto, os termos mais relevantes foram filtrados para serem associados a um dos eixos temáticos políticos definidos por palavras com maior proximidade textual e semântica. Esses nomes de eixos temáticos foram montados a partir de pesquisas realizadas tanto na literatura científica tradicional quanto na literatura cinza por meio das plataformas SEEG e INMET. Além disso, sugestões de nomes por especialistas de domínio e inteligência artificial generativa também contribuíram

para definição dos nomes.

A Figura 32 apresenta as palavras-chave mais frequentemente mencionadas nas propostas de governo dos candidatos das capitais brasileiras, com base na extração automatizada feita com LLM.

O termo “meio ambiente” aparece como o mais recorrente, seguido de perto por “sustentabilidade” e “conservação”, indicando uma ênfase geral nos compromissos mais comuns das pautas ambientais. Outros temas com destaque incluem “reciclagem”, “poluição”, “mudanças climáticas” e “preservação ambiental”, sugerindo uma preocupação com gestão de resíduos e impactos do aquecimento global. Os eixos temáticos são representados na figura 25, apresentando um mapa temático dos eixos ambientais predominantes nas propostas dos candidatos à prefeitura nas eleições municipais de 2024 delimitadas por municípios.

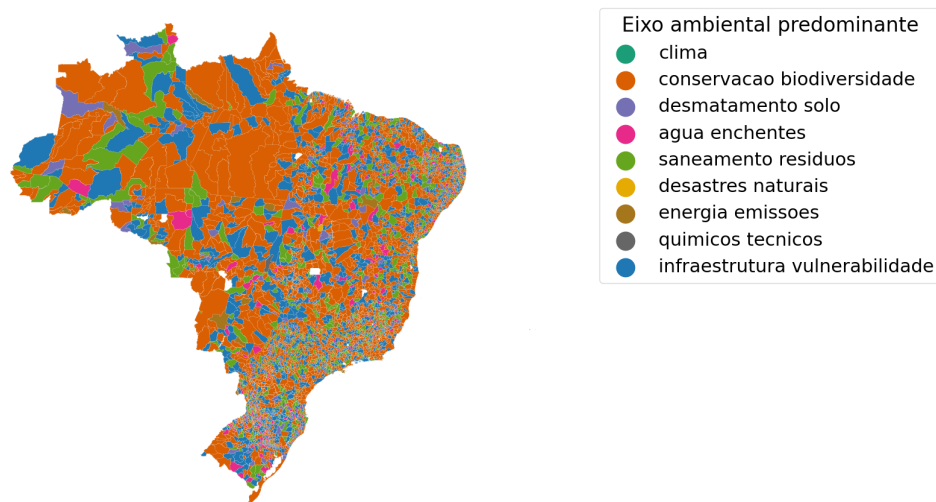


Figura 25 – Mapa temático de eixos ambientais predominantes por município no Brasil. Fonte: Elaborado pelo autor (2025).

Cada município é representado por uma cor associada ao eixo temático com maior destaque entre as propostas da localidade. Observa-se o eixo de conservação e biodiversidade como majoritário na maioria do território nacional, especialmente nas regiões Norte, Sudeste e Sul, evidenciando uma preocupação recorrente com a preservação dos ecossistemas e recursos naturais.

Além disso, temas como infraestrutura e vulnerabilidade e saneamento e resíduos também aparecem com frequência, refletindo demandas locais por melhorias urbanas e serviços básicos. Eixos como clima e energia e emissões surgem com menor abrangência territorial, indicando um foco mais localizado ou menos recorrente nessas agendas.

A Figura 26 sintetiza, em forma de diversos gráfico de barras, os cinco eixos ambientais mais mencionados por região.

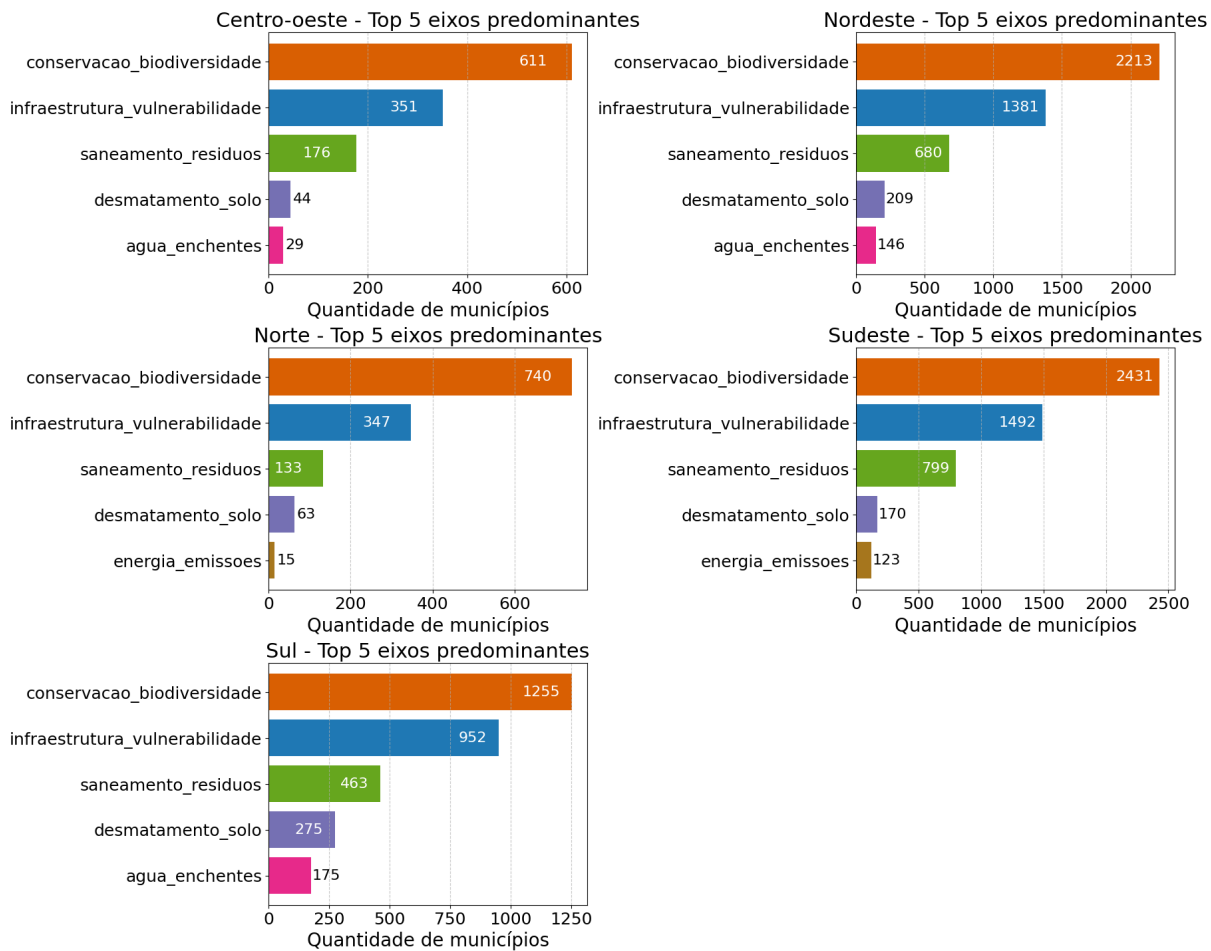


Figura 26 – Top 5 eixos temáticos predominantes por região.
Fonte: Elaborado pelo autor (2025).

Vale destacar que, em todas as regiões, “saneamento e resíduos” e “conservação e biodiversidade” aparecem com frequência, indicando preocupações estruturais comuns no país. No entanto, há variações importantes: o eixo “infraestrutura e vulnerabilidade” ganha destaque no Nordeste e Sul, enquanto “energia e emissões” aparece com mais frequência no Centro-Oeste e Norte. Além disso, o eixo “água enchentes” é bastante recorrente na região Sul do Brasil, provavelmente em razão das grandes enchentes ocorridas no estado do Rio Grande do Sul em 2024.

A Figura 27 mostra a diversidade temática ambiental na região Norte. Observa-se uma predominância dos eixos “conservação e biodiversidade” e “saneamento e resíduos”, especialmente em municípios dos estados do Amazonas, Pará e Rondônia.

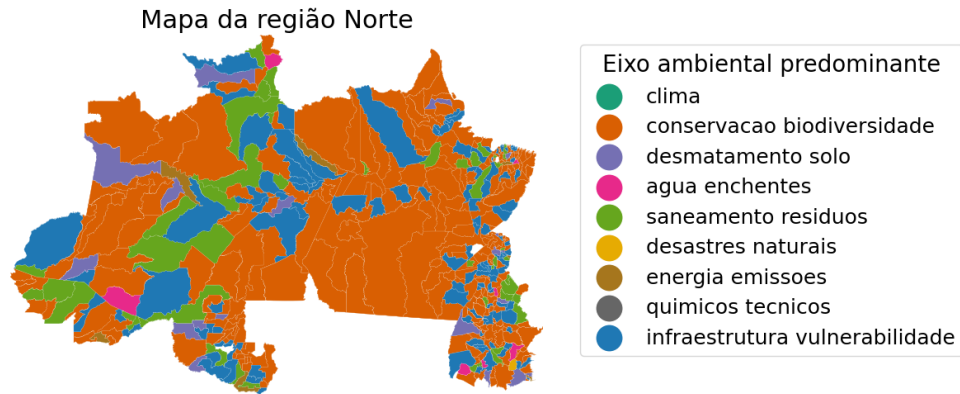


Figura 27 – Distribuição espacial dos eixos ambientais predominantes na região Norte.
Fonte: Elaborado pelo autor (2025).

Contudo, eixos como “infraestrutura e vulnerabilidade” também aparecem com frequência em áreas mais urbanizadas ou afetadas por eventos extremos, refletindo diferentes realidades socioambientais na região.

A Figura 28 revela uma forte presença do eixo “saneamento e resíduos” como o mais recorrente nos municípios nordestinos. Este padrão é evidente em estados como Bahia, Pernambuco e Ceará. Além disso, o eixo “infraestrutura e vulnerabilidade” também se destaca em várias localidades, o que pode estar relacionado a deficiências estruturais e aos riscos associados a eventos climáticos extremos. A distribuição é bastante heterogênea, com outros temas como “conservação e biodiversidade” e “água e enchentes” aparecendo pontualmente.

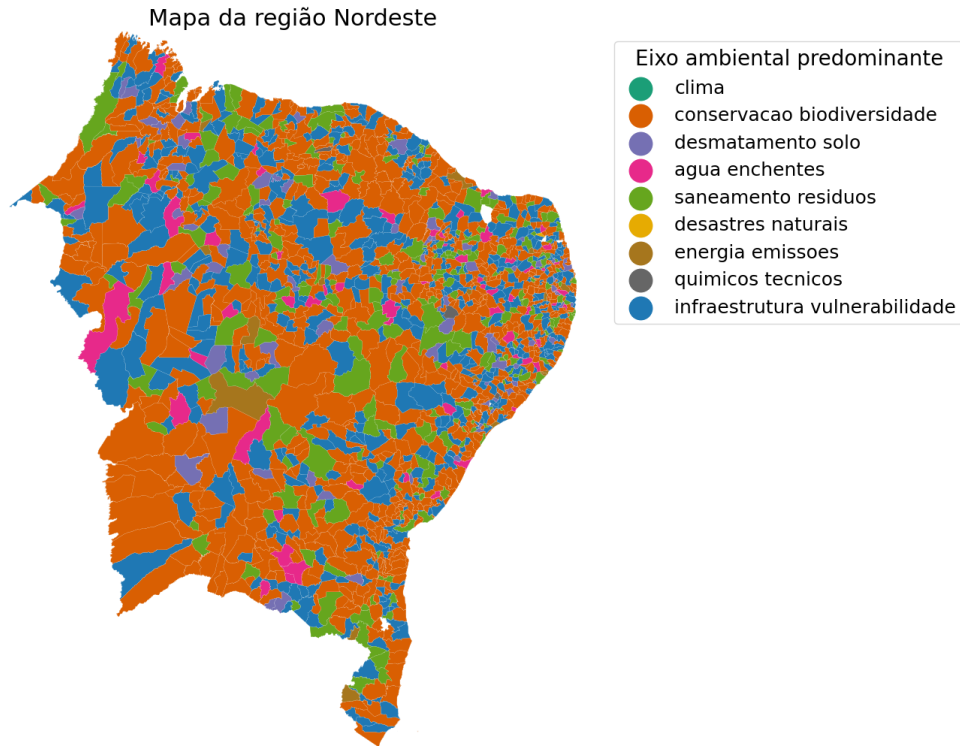


Figura 28 – Distribuição espacial dos eixos ambientais predominantes na região Nordeste.
Fonte: Elaborado pelo autor (2025).

Na região Sul, como apresentado na Figura 29, os eixos “saneamento e resíduos”, “conservação e biodiversidade” e “infraestrutura vulnerabilidade” predominam de forma bastante equilibrada. Municípios do interior do Rio Grande do Sul e Santa Catarina priorizam frequentemente a questão da biodiversidade, enquanto áreas mais densamente povoadas mostram maior preocupação com o saneamento. Ademais, a diversidade de eixos sugere que diferentes aspectos ambientais têm relevância nos planejamentos municipais da região. Um ponto a ser ressaltado é a concentração de “água enchentes” como eixo predominante em muitas localidades da região central do estado do Rio Grande do Sul em áreas com precipitação pluvial acima da média em razão das fortes enchentes ocorridas no final de abril de 2024, bem próximo do período das eleições municipais.

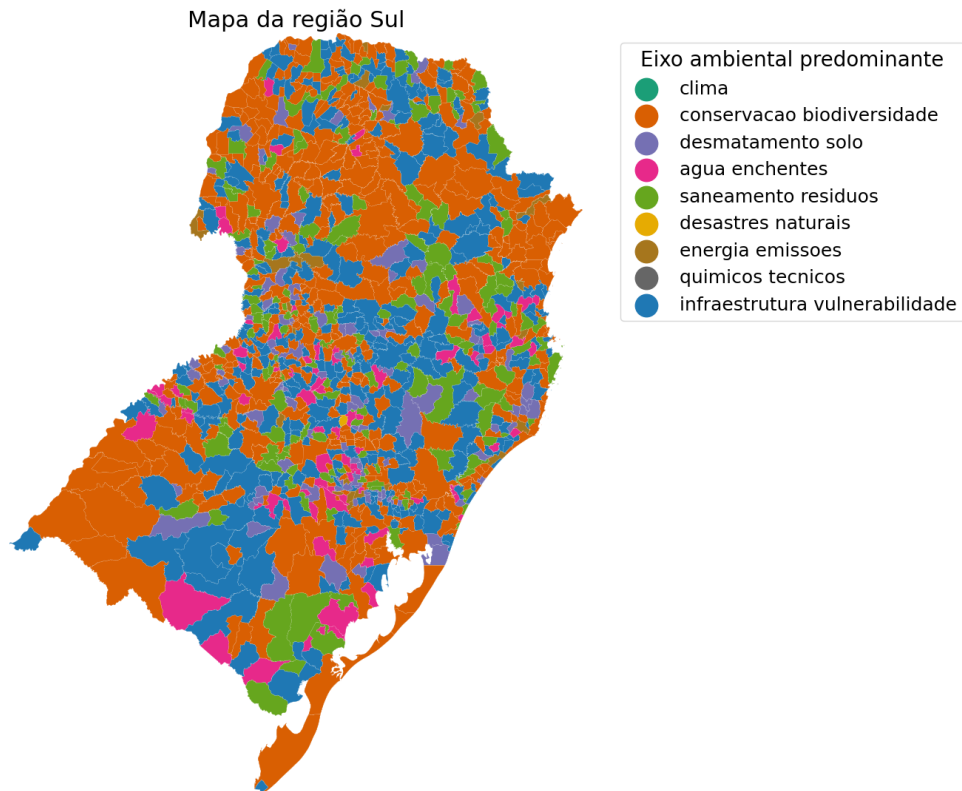


Figura 29 – Distribuição espacial dos eixos ambientais predominantes na região Sul.
Fonte: Elaborado pelo autor (2025).

Conforme ilustrado na Figura 30, os municípios do Centro-Oeste apresentam uma forte ênfase nos eixos “conservação e biodiversidade” e “saneamento e resíduos”. Essa prevalência pode estar relacionada à presença do bioma Cerrado e ao crescimento urbano em cidades médias. Alguns municípios de Mato Grosso e Mato Grosso do Sul também destacam eixos como “energia e emissões” e “clima”, apontando para discussões emergentes sobre impactos da agropecuária e mudanças climáticas.

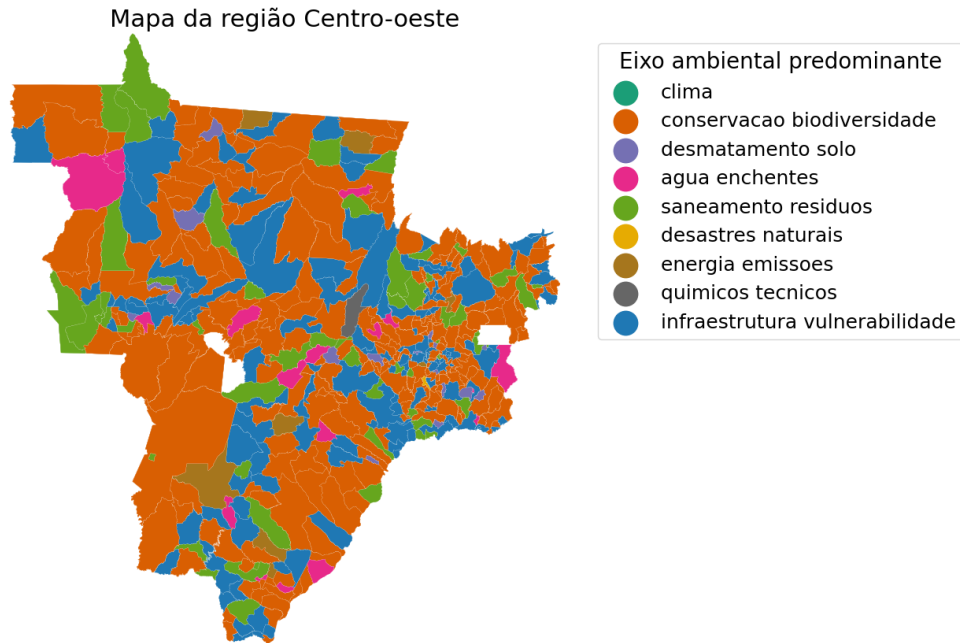


Figura 30 – Distribuição espacial dos eixos ambientais predominantes na região Centro-Oeste.

Fonte: Elaborado pelo autor (2025).

Já na Figura 31, mostra-se que o Sudeste é a região com maior heterogeneidade temática. O eixo “conservação e biodiversidade” aparece com destaque em estados como Minas Gerais e São Paulo, enquanto “saneamento e resíduos” predomina em cidades de médio porte. Municípios com infraestrutura mais densa, como os do Rio de Janeiro, também apresentam foco em “infraestrutura e vulnerabilidade”. Essa diversidade reflete o contraste entre áreas urbanas e rurais, além da complexidade socioambiental da região.



Figura 31 – Distribuição espacial dos eixos ambientais predominantes na região Sudeste.

Fonte: Elaborado pelo autor (2025).

A Figura 4.25 apresenta um ranking das 30 palavras-chave mais frequentes associadas ao tema ambiental nas propostas de governo das capitais.

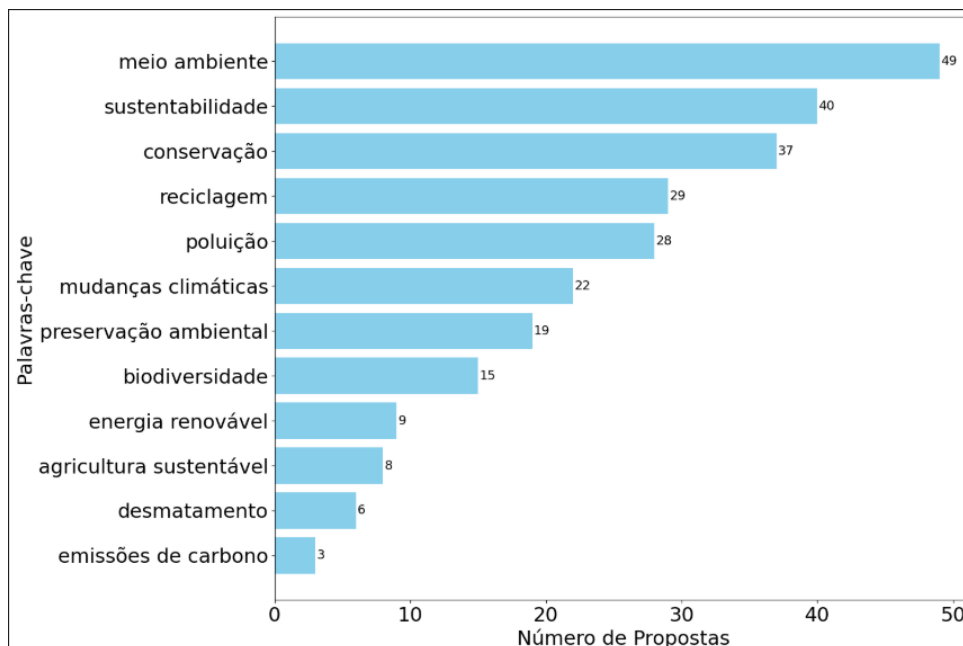


Figura 32 – Top 30 palavras-chave relacionadas ao meio Ambiente nas propostas das capitais usando LLM.

Fonte: Elaborado pelo autor (2025).

A análise demonstra que o termo “meio ambiente” lidera com ampla vantagem, com citações em 49 propostas, revelando a agenda ambiental, mesmo que de forma mais genérica, possuindo presença considerável no discurso político. Na sequência, termos como “sustentabilidade” (40 ocorrências) e “conservação” (37 ocorrências) aparecem com destaque. Além disso, a presença relevante de termos operacionais e práticos como “reciclagem” (29 ocorrências), “poluição” (20 ocorrências) e “mudanças climáticas” (22 ocorrências) indica que uma parte das propostas aborda problemas ambientais urbanos específicos, como gestão de resíduos sólidos, controle da poluição e enfrentamento das alterações climáticas. Vale ressaltar que, “mudanças climáticas” figurar com 22 menções sinaliza uma crescente incorporação desse tema na política local, refletindo um avanço na compreensão da relevância do tema mesmo em delimitações municipais.

De maneira geral, os resultados mostram a pauta ambiental como presente nas campanhas com uma concentração maior em propostas voltadas à gestão de resíduos, conservação de áreas e mitigação da poluição urbana.

5 CONSIDERAÇÕES FINAIS

A disponibilidade de dados eleitorais abertos tem facilitado análises mais abrangentes e detalhadas de aspectos como perfis de candidatos, filiações partidárias e propostas. No entanto, ainda existe uma lacuna significativa na literatura quando se trata de integrar uma análise dos aspectos mencionados por meio de ferramentas de ciência de dados, visando maior escalabilidade e automatização para os grandes volumes textuais municipalmente. Nesse contexto, o presente trabalho propôs uma abordagem híbrida para analisar o discurso ambiental nas propostas de governo dos candidatos a prefeito nas eleições municipais brasileiras de 2024 por ser uma época em que há maior pluralidade partidária, regional e linguística nas propostas. Isso é atingido integrando técnicas de mineração de texto, expansão lexical via *Word Embeddings* e o uso de um LLM.

O presente trabalho realizou uma análise dos dados de candidatos e compreender quais focos temáticos das propostas de governo dos candidatos às eleições municipais brasileiras de 2024. A metodologia adotada baseou-se no modelo *CRISP-DM*, adaptado com elementos do *framework Data Science Trajectories* (DST), permitindo uma abordagem mais flexível e exploratória. Primeiramente, foram coletados dados abertos baixados manualmente a partir de links disponibilizados pelo Tribunal Superior Eleitoral (TSE). Esse conjunto incluiu informações de dados tabulares de mais de 463 mil candidatos aos cargos de prefeito, vice-prefeito e vereador, além de aproximadamente 15.786 propostas de governo, disponibilizadas em arquivos PDF.

Após a coleta, os metadados foram organizados por meio da fusão de dois arquivos principais, contendo dados gerais e complementares dos candidatos. Foram selecionados os atributos mais relevantes para as análises destacados na tabela ?? buscando um panorama detalhado dos perfis dos candidatos. Após o fusionamento, os dados passaram por um processo de filtragem e limpeza. Foram removidas duplicatas, como aquelas relacionadas a candidatos que participaram de segundo turno, e registros sem correspondência entre propostas e candidatos. Também foi realizado um filtro para considerar apenas os candidatos ao cargo de prefeito. As propostas em PDF foram convertidas para texto utilizando a técnica de Reconhecimento Óptico de Caracteres (OCR), permitindo a análise textual automatizada.

A abordagem híbrida proposta para a análise do conteúdo ambiental presente nas propostas de governo destes candidatos, integra técnicas de mineração de texto, expansão de léxico usando a abordagem de *Word Embeddings* e o uso de uma LLM aberta se mostrou eficaz para identificar com maior precisão e sensibilidade os principais eixos temáticos

presentes nos discursos políticos. Essa metodologia auxiliou na superação de desafios encontrados ao analisar textos políticos em português. Por fim, na etapa de análise, os metadados foram explorados por meio de gráficos e visualizações para identificar padrões e características gerais dos candidatos. Além disso, foram construídos grafos de coocorrência entre partidos usando a ferramenta Gephi, com métricas como grau médio e modularidade, para identificar agrupamentos políticos relevantes.

Os resultados da análise das eleições municipais de 2024 melhoraram a compreensão sobre o perfil dos candidatos, as estratégias partidárias e o conteúdo temático das propostas de governo. Observou-se que a maioria dos textos apresentados pelos candidatos é relativamente concisa, com concentração em torno de 2.000 palavras, com exceção de alguns casos que poderiam chegar até 8000 palavras. Ainda que propostas extensas sejam menos frequentes, partidos com forte orientação ideológica à esquerda, como PSTU, PCB e PSOL, apresentaram documentos mais longos, evidenciando possíveis propostas mais complexas e mais bem elaboradas. O perfil dos candidatos evidenciou uma predominância de indivíduos entre 40 e 59 anos, predominantemente do sexo masculino, com elevado grau de escolaridade, especialmente ensino superior. Os resultados mostraram que as mulheres representam menos de 20% dos candidatos efetivamente registrados, elucidando uma persistente desigualdade de gênero na política. Outro ponto relevante é o alto número de prefeitos tentando a reeleição, o que evidencia uma cultura de permanência no poder e, em muitos casos, sucesso nas urnas.

A análise das coligações entre partidos demonstra algumas estratégias eleitorais significativas, com partidos como MDB, PP, PSD e UNIÃO demonstrando uma maior formação de alianças políticas, o que se correlaciona diretamente com o sucesso em eleger seus membros. Os dados sobre financiamento de campanha, com destaque para PL, PT e MDB como os partidos com maior soma de despesa máxima autorizada, indicam o impacto dos recursos financeiros como elemento chave das disputas eleitorais municipais. A relevância de estudos que analisam dados eleitorais ultrapassa o interesse acadêmico, de forma que oferece insumos para melhor compreensão do comportamento político eleitoral dos brasileiros. Ao transformar grandes *corpus* em conhecimento estruturado, a ciência de dados possibilita uma tomada de decisão estratégica e baseada em dados tanto para as entidades que formulam as políticas públicas quanto para os eleitores.

Em relação ao discurso nas propostas, os dados enfatizam palavras como “meio ambiente”, “sustentabilidade” e “conservação” como uma das mais recorrentes. A análise léxica por coocorrência mostrou que palavras como “drenagem”, “emergência” e “desigualdade hídrica” também se destacam, refletindo preocupações com infraestrutura e riscos climáticos. A delimitação das temáticas em destaque delimitadas por município revelou eixos como “conservação e biodiversidade” como o mais frequente na maioria do

território nacional, especialmente nas regiões Norte, Sudeste e Sul, eixos como “saneamento e resíduos” e “infraestrutura e vulnerabilidade” são os mais recorrentes em seguida. Um destaque para a região sul com alta recorrência da temática de “água enchentes” pelo fato das eleições terem ocorrido após as grandes enchentes que abalaram os municípios da região, mais acentuadamente no estado do Rio Grande do Sul e tornou-se pauta de discussão frequente nas propostas governamentais.

5.1 Contribuições do estudo

O presente trabalho contribui ao disponibilizar um pipeline replicável para análise de dados eleitorais no contexto das eleições municipais brasileiras. Além disso, serão disponibilizados os dados processados, os prompts utilizados e as etapas do pipeline validadas por especialistas de domínio, de modo a garantir transparência e reprodutibilidade para futuras pesquisas acadêmicas ou aplicações práticas.

Do ponto de vista científico, este trabalho amplia os estudos no campo das ciências sociais com uso de ferramentas computacionais e dos estudos eleitorais, ao integrar análises textuais e metadados eleitorais para revelar padrões regionais, partidários e discursivos ligados à diferentes pautas de discurso político. Trata-se de uma aplicação pouco explorada na literatura brasileira, especialmente em eleições municipais. O conhecimento extraído permite identificar os eixos temáticos predominantes na agenda ambiental dos candidatos, fornecendo dados empíricos que podem subsidiar análises políticas, jornalísticas ou acadêmicas.

Além disso, como produto científico do presente estudo, foi elaborado o artigo intitulado “*What are the environmental priorities of mayoral candidates in Brazil? An overview of proposals for the 2024 elections*”, aceito para apresentação como *COMMUNICATION PAPER* no FedCSIS 2025 (20th Conference on Computer Science and Intelligence Systems), evento internacional indexado em bases de dados científicas como *Scopus* e *Web of Science*.

Sob o aspecto social, os resultados têm potencial impacto na sociedade, pois oferecem subsídios para cidadãos, organizações civis e jornalistas avaliarem as propostas políticas de forma mais informada, especialmente no tema ambiental. As informações levantadas podem auxiliar o eleitorado na identificação de candidatos ou partidos mais alinhados às pautas temáticas consideradas mais relevantes, bem como orientar políticas públicas e estratégias políticas de candidatos que buscam ampliar seu compromisso com questões como sustentabilidade e mudanças climáticas.

5.2 Dificuldades encontradas

Durante o percorrer do trabalho, a dificuldade inicial foi compreender quais dados relevantes ao objeto de pesquisa seriam coletados. Em seguida, embora os dados abertos governamentais estivessem estruturados e com poucas lacunas, tive dificuldades para remover valores nulos e registros duplicados, além de lidar com inconsistências no número de candidatos registrados nos *datasets*, que não correspondiam aos números oficiais por erros de fusão de arquivos realizados, o que ocorreu pela minha pouca *expertise* no tratamento dos dados para a etapa de filtragem.

Além disso, outro obstáculo significativo ocorreu pelas limitações computacionais da minha máquina. O grande volume de dados envolvido fez com que muitas operações, especialmente as voltadas ao pré-processamento e análise textual, demandassem tempo de execução elevado. Em alguns casos, podendo levar horas para serem concluídas, o que impactou na produtividade e o andamento do projeto.

5.3 Trabalhos futuros

Como direções para trabalhos futuros, propõe-se:

- Aplicar LLMs mais avançadas e multilinguísticas para melhorar a identificação de temas específicos nos discursos.
- Estender o uso do dicionário de palavras-chave para novos ciclos eleitorais e outros cargos políticos (como deputados e vereadores).

REFERÊNCIAS

- Anwar, M. A.; ZHOU, R.; SAJJAD, A.; ASMI, F. Climate change communication as political agenda and voters' behavior. **Environmental Science and Pollution Research**, Springer, v. 26, p. 29946–29961, 2019.
- CAMPOS-VALDÉS, C.; ÁLVAREZ-MIRANDA, E.; QUIROGA, M. M.; PEREIRA, J.; DURAN, F. L. The impact of candidates' profile and campaign decisions in electoral results: A data analytics approach. **Mathematics**, MDPI, v. 9, n. 8, p. 902, 2021.
- Cernev, T.; FENNER, R. The importance of achieving foundational sustainable development goals in reducing global risk. **Futures**, v. 115, 2020. ISSN 00163287.
- CIOCCARI, D.; MOTTINHA, R. **ELEIÇÕES 2020: QUAL É O PERFIL DE COR/RAÇA E GÊNERO DOS CANDIDATOS E DOS ELEITOS?** [S.l.], 2021.
- COUTO, M.; CARMO, F.; JR, A. J.; MARCACINI, R.; LOBATO, F. Characterization of co-authorship networks of cnpq productivity fellows: an approach based on data science. In: **Anais do XII Symposium on Knowledge Discovery, Mining and Learning**. Porto Alegre, RS, Brasil: SBC, 2024. p. 113–120. ISSN 2763-8944.
- Dekoninck, H.; Schmuck, D. The mobilizing power of influencers for pro-environmental behavior intentions and political participation. **Environmental Communication**, v. 16, p. 458–472, 5 2022. ISSN 1752-4032.
- Ejaz, W.; MUKHERJEE, M.; FLETCHER, R.; NIELSEN, R. How we follow climate change: Climate news use and attitudes in eight countries. **Oxford Climate Journalism Network**, Reuters Institute for the Study of Journalism, 2022.
- ESHIMA, S.; SMITH, D. M. Just a number? voter evaluations of age in candidate-choice experiments. **The Journal of Politics**, The University of Chicago Press Chicago, IL, v. 84, n. 3, p. 1856–1861, 2022.
- Fekete, B. M.; REVENGA, C.; TODD, M. The global risks report 2018 13th edition, [geneva: World economic forum, 2018] http://www3.weforum.org/docs/wef_grr18_report.pdf. **Agricultural Economics**, JSTOR, v. 31, n. 1, p. 50–67, 2018.
- Hager, A.; Hilbig, H. Does public opinion affect political speech? **American Journal of Political Science**, Wiley Online Library, v. 64, n. 4, p. 921–937, 2020.
- JACINTHO, L. H. M.; SILVA, T. P. d.; PARMEZAN, A. R. S.; BATISTA, G. E. d. A. P. A. Brazilian presidential elections: Analysing voting patterns in time and space using a simple data science pipeline. In: **Proceedings**. [S.l.: s.n.], 2020.
- Jacinto, L. H. M.; SILVA, T. P. da; PARMEZAN, A. R. S.; BATISTA, G. E. A. P. A. Analysing spatio-temporal voting patterns in brazilian elections through a simple data science pipeline. **Journal of Information and Data Management**, v. 12, 8 2021. ISSN 2178-7107.

JACINTHO, L. H. M.; SILVA, T. P. da; PARMEZAN, A. R. S.; BATISTA, G. E. A. P. A. Analyzing spatio-temporal voting patterns in brazilian elections through a simple data science pipeline. **Journal of Information and Data Management**, v. 12, n. 1, p. 31–47, Aug 2021.

Li, L.; LI, J.; CHEN, C.; GUI, F.; YANG, H.; YU, C.; WANG, Z.; CAI, J.; ZHOU, J. A.; SHEN, B. *et al.* Political-llm: Large language models in political science. **arXiv preprint arXiv:2412.06864**, 2024.

Linegar, M.; KOCIELNIK, R.; ALVAREZ, R. M. Large language models and political science. **Frontiers in Political Science**, Frontiers Media SA, v. 5, p. 1257092, 2023.

MARTINEZ-PLUMED, F.; CONTRERAS-OCHANDO, L.; FERRI, C.; HERNANDEZ-ORALLO, J.; KULL, M.; LACHICHE, N.; RAMIREZ-QUINTANA, M. J.; FLACH, P. Crisp-dm twenty years later: From data mining processes to data science trajectories. **IEEE Transactions on Knowledge and Data Engineering**, v. 33, 2021. ISSN 15582191.

Min, B.; ROSS, H.; SULEM, E.; VEYSEH, A. P. B.; NGUYEN, T. H.; SAINZ, O.; AGIRRE, E.; HEINZ, I.; ROTH, D. Recent advances in natural language processing via large pre-trained language models: A survey. **ACM Computing Surveys**, v. 56, n. 4, p. 1–40, 2021.

Muchlinski, D.; YANG, X.; BIRCH, S.; MACDONALD, C.; OUNIS, I. We need to go deeper: measuring electoral violence using convolutional neural networks and social media. **Political Science Research and Methods**, v. 9, n. 1, p. 122–139, 2021.

Ralph, P.; Baltés, S. Paving the way for mature secondary research: the seven types of literature review. In: **Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering**. [S.l.: s.n.], 2022. p. 1632–1636.

Ruppert, E.; ISIN, E.; BIGO, D. Data politics. **Big Data Society**, v. 4, 2017.

Silva, M. B. da; AZEVEDO, R. F. d. B. C.; ARAÚJO, D. O. de; FRANÇA, F. P.; PEREIRA, M. da S. A publicidade de dados abertos pelo tribunal superior eleitoral (tse): o caso do repositório de dados eleitorais. **Ciência da Informação**, v. 49, n. 3, 2020.

Silva Érika R. T.; Mainardes, E. W. O comportamento do eleitor brasileiro na escolha do candidato. **Revista Gestão e Desenvolvimento**, Associação Pro-Ensino Superior em Novo Hamburgo, v. 16, p. 201, 2 2019. ISSN 1807-5436.

Stroparo, T. R.; ARAÚJO, J. H. K. de; BORTOLOTTI, M. A.; JUNIOR, O. da S. L. Inteligência artificial na gestão de custos: Avanços, desafios e oportunidades. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, Revista Ibero-Americana de Humanidades, Ciências e Educação, v. 10, p. 1446–1456, 6 2024.

Vasconcelos, F. F.; TAVARES, J. V. S.; OLIVEIRA, M. G. S.; COUTINHO, F. J.; CLARINDO, J. P. Candidata: an enhanced dataset for data analysis of elections in brazil from 1945 to 2020. **Journal of Information and Data Management**, v. 13, 8 2022. ISSN 2178-7107.

Wirth, R.; Hipp, J. Crisp-dm: Towards a standard process model for data mining. In: MANCHESTER. **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. [S.l.], 2000. v. 1, p. 29–39.

Zhang, S.; HE, L.; DRAGUT, E.; VUCETIC, S. How to invest my time: Lessons from human-in-the-loop entity extraction. In: **Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining**. [S.l.: s.n.], 2019. p. 2305–2313.