LegisPL-BR - Dataset de projetos de leis brasileiros

Juan M. V. Marciano [1], Vinicius P. Machado[2], Arlino H. M. Araújo[2]

[1]Instituto Federal do Piauí (IFPI)

[2]Universidade Federal do Piauí (UFPI)

juan.morysson@ifpi.edu.br, vinicius@ufpi.edu.br, arlino@ufpi.edu.br

Abstract. *The area of legislative informatics has benefited from the use of structured data for political analysis. However, the dataset available in Brazil on legislative proposals is not well documented. This work presents the construction of a structured dataset with proposals of the type Bill (PL) from the Chamber of Deputies. The data were extracted from the official API and enriched with other information. As a result, an open and unified dataset is made available, with potential for legislative studies and applications in data science.*

Resumo. *A área de informática legislativa tem se beneficiado do uso de dados estruturados para análises políticas. Contudo, o conjunto de dados acessíveis no Brasil sobre proposições legislativas não são bem documentados. Este trabalho apresenta a construção de um dataset estruturado com proposições do tipo Projeto de Lei (PL) da Câmara dos Deputados. Os dados foram extraídos da API oficial e enriquecidos com outras informações. Como resultado, disponibiliza-se um conjunto de dados aberto e unificado, com potencial para estudos legislativos e aplicações em ciência de dados.*

1. Introdução

A crescente demanda social por maior transparência, participação cidadã e controle sobre a gestão dos recursos públicos tem impulsionado governos a disponibilizarem informações em formatos abertos, passíveis de apropriação e reuso pela sociedade [Barbalho 2018]. Outros autores citam a necessidade de transparência e eficiência em gestão publica, em específico nas compras de forma a assegurar a utilizaçõ adequada de recursos [Oliveira et al. 2024]. No Brasil, esse movimento foi institucionalizado pela Lei nº 12.527, de 18 de novembro de 2011, conhecida como Lei de Acesso à Informação (LAI), que assegura a qualquer cidadão o direito de solicitar e receber dados sob a guarda de órgãos públicos [Brasil 2011]. A publicação de Dados Abertos Governamentais (DAGs) é reconhecida como ferramenta estratégica para gerar valor social e econômico [Janssen 2011], fomentar a confiança nas instituições [Barbalho 2018], e melhorar a qualidade das análises e a formulação de políticas públicas.

O Poder Legislativo, como instância fundamental da democracia, produz uma vasta quantidade de dados, incluindo projetos de lei, autores e tramitações [Cavalcante et al. 2016]. A abertura dessas informações por meio de APIs públicas, como a oferecida pela Câmara dos Deputados, permite a análise sistemática do processo legislativo e viabiliza formas inovadoras de controle social e monitoramento cívico [Brandt 2018]. Entretanto, a simples disponibilização não é suficiente. O trabalho com grandes volumes de dados públicos frequentemente envolve desafios técnicos

como limpeza, padronização e interpretação, consumindo tempo e recursos substanciais [Fagundes and Ribeiro Junior 2020].

Adicionalmente, a qualidade dos dados publicados é um fator crítico para a efetividade das ações baseadas em evidências [Janssen et al. 2012]. Diversas barreiras técnicas, semânticas e institucionais dificultam o pleno aproveitamento dos DAGs, exigindo práticas que promovam interoperabilidade, documentação adequada e formatos legíveis [Breitman 2012]. Assim, torna-se essencial a criação de conjuntos de dados estruturados, documentados e prontos para uso, especialmente aqueles que se originam de fontes oficiais e confiáveis.

Nesse contexto, este artigo apresenta a construção e disponibilização de um conjunto de dados contendo Projetos de Lei (PL) da Câmara dos Deputados, abrangendo os anos de 2022 a 2024. O dataset inclui informações estruturadas como número, ano, autores, ementa e histórico da última tramitação. A metodologia aplicada contempla boas práticas de coleta automatizada via API pública, tratamento de dados textuais e enriquecimento com metadados legislativos. O repositório resultante é disponibilizado publicamente com o código-fonte completo, visando à reprodutibilidade, transparência e reutilização em projetos futuros. E, o conjunto de dados pode apoiar investigações sobre viés ideológico, priorização temática e padrões de autoria no processo legislativo, apoiada em tecnologias de aprendizado de máquina e processamento de linguagem natural.

O presente artigo está organizado da seguinte forma: a Seção 2 discute trabalhos relacionados com dados abertos legislativos e aplicação de PLN; a Seção 3 descreve o processo de aquisição e processamento dos dados; a Seção 5 detalha a disponibilização e os possíveis cenários de aplicação; a Seção 4 que faz uma análise descritiva e exploratória dos dados;e a Seção 6 apresenta as conclusões e direções para trabalhos futuros.

2. Trabalhos Relacionados

A publicação e o uso de Dados Abertos Governamentais (DAGs) têm sido amplamente discutidos na literatura, destacando-se a necessidade de modelos e *frameworks* padronizados para sua disponibilização, de forma a garantir unicidade e aderência a critérios internacionalmente reconhecidos [Dalenogare and Araújo 2019]. Embora a abertura de dados seja essencial, não garante por si só seu uso efetivo: frequentemente os dados não podem ser utilizados em seu formato original, sendo necessário submetê-los a etapas de avaliação, tratamento e padronização [Janssen et al. 2012].

Diversos modelos têm sido propostos. Por exemplo, o modelo SmartGov [Dzikrullah and Rinjani 2017] buscou integrar dados governamentais com tecnologias como *Big Data* e CKAN, mas não contemplou mecanismos de *feedback* sobre a qualidade. Ruijer et al. [Ruijer et al. 2017] apresentaram um modelo heurístico que relaciona DAGs a processos democráticos, ressaltando a importância da qualidade, porém sem tratá-la diretamente.

No contexto brasileiro, Brandt et al. [Brandt 2018] propuseram um modelo de dados legislativos da Câmara dos Deputados com base em *Linked Open Data* (LOD), utilizando RDF e vocabulários ontológicos. Embora relevante para representação semântica, o trabalho não abordou práticas consolidadas de metadados nem diretrizes de publicação. Rolim et. al. [Rolim et al. 2024] propuseram a criação de *dataset* semântico de pessoas

jurídicas, utilizando dados abertos da Receita Federal, sendo relevante por ter utilizados técnica de *Data Lakhouses*.

Complementando essas abordagens, iniciativas como o *Frictionless Data*, da *Open Knowledge Foundation*, buscam reduzir o "atrito"no uso de dados através de padrões leves como o *Data Package*, promovendo empacotamento, documentação e interoperabilidade [Fagundes and Ribeiro Junior 2020]. Outro marco relevante são os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*), inicialmente voltados a dados científicos, mas aplicáveis a qualquer tipo de dado estruturado [Wilkinson 2016].

Estudos recentes também têm explorado aplicações de técnicas de Processamento de Linguagem Natural (PLN) e Ciência de Dados sobre dados legislativos. Pesquisas incluem desde classificação de textos jurídicos [Mumcuoğlu 2021], extração de entidades [Bommarito et al. 2020], análise de tópicos [Chebolu et al. 2023] até análise de sentimento e identificação de viés argumentativo [Li et al. 2023]. Essas tarefas requerem dados bem estruturados e pré-processados, com atributos relevantes.

O presente trabalho alinha-se a essas iniciativas ao propor um conjunto de dados estruturado. Diferentemente de modelos generalistas, este estudo foca na criação e documentação de um *dataset* específico, aplicável a tarefas de PLN. A estrutura final inclui campos úteis para análises descritivas e preditivas, com destaque para classificação supervisionada, clusterização semântica e visualização exploratória. Ao disponibilizar o código-fonte, promove-se também a reprodutibilidade científica, um princípio fundamental na construção de ativos digitais confiáveis [Nay 2018].

3. Aquisição e Processamento dos Dados

O conjunto de dados descrito neste artigo foi construído a partir de informações legislativas disponibilizadas publicamente pela Câmara dos Deputados do Brasil, por meio da sua API oficial de Dados Abertos.[1] Foram coletadas proposições legislativas do tipo Projeto de Lei (PL) referentes aos anos de 2022, 2023 e 2024. Cada proposição inclui metadados, informações sobre autoria, tramitação e um resumo textual (ementa), consolidando informações úteis tanto para estudos sobre processos legislativos quanto para aplicações de Processamento de Linguagem Natural (PLN).

3.1. Extração e Agregação dos Dados

O processo de extração inicial foi automatizado com o uso de um *script python*, em torno de três funções principais: `obter_proposicoes()`, `obter_autores()` e `obter_tramitacao()`. As proposições foram consultadas por ano, em páginas de até 100 itens como forma de limitar o tamanho dos dados recebidos por vez. Para cada ano, o *script* chama o *endpoint* `/proposicoes` da API para coletar os metadados básicos das proposições do tipo Projeto de Lei (PL), com paginação controlada. Em seguida os dados extraídos foram armazenados em memória até completar cada ano.

Após a extração inicial das proposições por ano, cada item passou por duas etapas adicionais de agregação:

- **Autoria:** cada proposição foi submetida a uma requisição complementar para coletar os nomes e outros dados dos autores. Em casos de múltiplos autores, os nomes foram concatenados em uma única string separada por vírgulas.

[1] `https://dadosabertos.camara.leg.br/`

- **Tramitação:** uma terceira requisição foi realizada para recuperar o histórico completo de tramitação da proposição, sendo extraída a primeira e a última movimentação registrada, com algumas das seguintes informações: órgão responsável, descrição da ação e data.

A Figura 1 apresenta o fluxo geral de extração e enriquecimento dos dados, desde a chamada à API até a construção do *dataset* final.

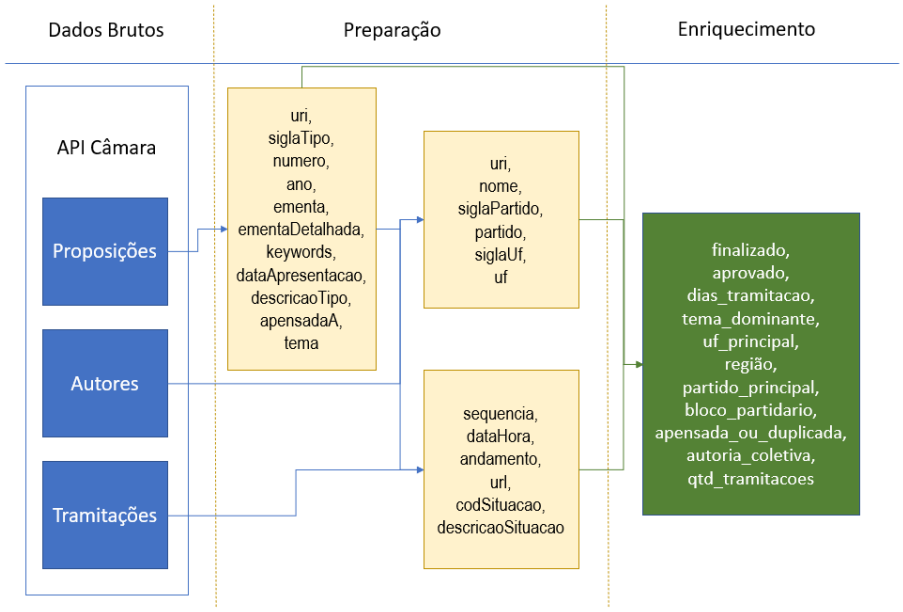

Figura 1. Fluxo geral de aquisição e enriquecimento dos dados legislativos

Todo o processo seguiu uma política de espaçamento entre as requisições (`time.sleep(0.5)`), a fim de respeitar a infraestrutura do servidor público da API.

3.2. Enriquecimento

Com a extração dos dados foi possível criar novas colunas a partir das colunas iniciais, da seguinte maneira:

- **finalizado:** indica se a proposição foi encerrada. Foi derivada da coluna `status atual`, com base em situações como ”Arquivada”, ”Rejeitada”, ”Transformada em norma jurídica”, entre outros, que são consideradas situações que finalizam o *status* da proposta, sem possibilidade de falsos positivos.
- **aprovado:** sinaliza se a proposição foi aprovada ou transformada em norma. Também derivada da coluna `status atual`, buscando situações como ”transformada em norma”, ”promulgado”, ”sanção”, entre outras, também sem possibilidade de falsos positivos.
- **dias_tramitacao:** calcula o tempo de tramitação da proposição, em dias. Obtida pela diferença entre as colunas `data da apresentação` e `data da última tramitação`.
- **tema_dominante:** categorização temática da proposição. Baseia-se na coluna `temas`, que é analisada com um mapeamento de palavras-chave para categorias amplas como ”Saúde”, ”Educação”, ”Meio Ambiente”, ”Economia”, ”Direitos

Humanos"e "Segurança Pública", convertendo os temas gerais em temas centrais. Aplicou-se um mapeamento simplificado com base em palavras-chave contidas nos temas originais, esse mapeamento foi gerado pelo ChatGPT, no algoritmo GPT-4o.

- **uf_principal:** extrai a unidade federativa (UF) do autor principal, a partir da *string* da coluna `autores`, onde a UF aparece abreviada (ex: "(PL-SP)"→ "SP").
- **região:** derivada da coluna `uf_principal`, classifica a proposição em uma das cinco regiões do Brasil (Norte, Nordeste, Centro-Oeste, Sudeste, Sul).
- **partido_principal:** identifica o partido do primeiro autor da proposição, extraído do campo `autores` a partir da sigla entre parênteses.
- **bloco_partidario:** classifica o partido principal em um dos grandes blocos políticos: "Esquerda", "Centrão"ou "Direita", conforme um dicionário predefinido de siglas partidárias e agregação sugerida por [Testa et al. 2024]
- **apensada_ou_duplicada:** *flag* booleano que verifica se a ementa da proposição aparece mais de uma vez no *dataset*.
- **autoria_coletiva:** indica se a proposição possui múltiplos autores, com base na presença de vírgulas na coluna `autores`.
- **qtd_tramitacoes:** estima a quantidade de tramitações da proposição. Caso a data de apresentação e a data da última tramitação sejam iguais, assume o valor 1; caso contrário, assume 2 como valor mínimo.

Ao término do processamento, os dados foram organizados em um `DataFrame` com vinte e seis colunas, sendo as quinze primeiras por coleta e as onze últimas por transformação.

3.3. Volume de Dados

Foram processados 3 anos de proposições legislativas, totalizando **12.334 proposições únicas**, distribuídas entre os anos de 2022 a 2024. A coleta foi realizada por meio da API da Câmara dos Deputados, com paginação automática e espaçamento entre requisições. O número de proposições varia de acordo com o ano, refletindo a dinâmica da produção legislativa brasileira. A Tabela 1 apresenta a distribuição exata por ano.

Tabela 1. Distribuição de proposições por ano

Ano	Quantidade de Proposições
2022	2.519
2023	5.438
2024	4.377

3.4. Dicionário de Dados

A Tabela 2, situada no Apêndice A, apresenta o dicionário de dados do conjunto de proposições legislativas coletadas. Cada registro representa uma proposição do tipo Projeto de Lei (PL) e está agregado com informações sobre autoria e tramitação

Todos os campos foram padronizados em tipos compatíveis com arquivos do tipo CSV e *Excel*, utilizando tipagens inteiras, textuais e datas. O objetivo deste dicionário é auxiliar pesquisadores, analistas e desenvolvedores na correta compreensão da estrutura

da base, promovendo transparência e reutilização adequada. Tornando fácil o carregamento em ferramentas de análise exploratória, bibliotecas de Processamento de Linguagem Natural (PLN), ou visualizadores interativos.

4. Análise Descritiva e Exploratória

Esta seção apresenta uma análise descritiva e exploratória da distribuição das proposições legislativas, tanto no recorte geográfico por Unidade Federativa (UF) quanto no recorte político por espectro ideológico.

4.1. Distribuição Geográfica das Proposições

A Figura 2 mostra, em sequência, a distribuição dos Projetos de Lei (PL) aprovados, rejeitados e submetidos por Unidade Federativa (UF). Destaca-se São Paulo (SP) como líder absoluto, com 3099 PLs submetidas e 40 aprovadas, seguido por Rio de Janeiro (RJ), Minas Gerais (MG) e Rio Grande do Sul (RS), este último com desempenho notável em aprovações (29), mesmo com menor volume de proposições em relação a RJ e MG.

Outros estados como Paraná (PR), Goiás (GO), Pernambuco (PE), Bahia (BA) e Ceará (CE) também apresentam participação expressiva, especialmente PE e CE, com 23 e 22 aprovações, respectivamente. Na outra ponta, estados como Amapá (AP), Acre (AC) e Roraima (RR) concentram os menores volumes. A análise revela que a distribuição das aprovações nem sempre acompanha o volume de submissões, indicando que alguns estados são proporcionalmente mais eficientes na conversão de proposições em normas jurídicas.

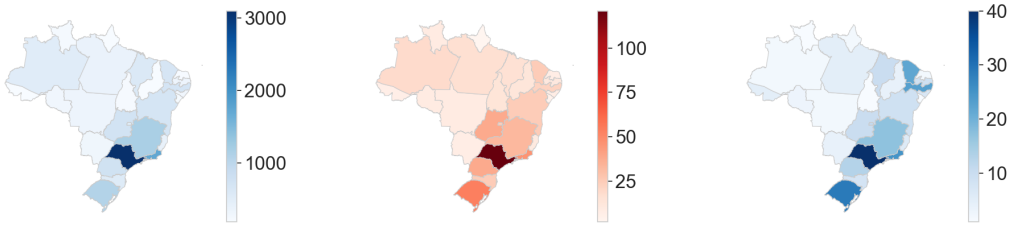

Figura 2. Distribuição de PLs submetidas (esquerda), rejeitadas (centro) e aprovadas (direita) por UF

4.2. Distribuição por Espectro Político

A Figura 3 mostra a distribuição das proposições legislativas por espectro político, classificadas em Esquerda, Centrão e Direita. No total de PLs submetidas, o Centrão lidera com 9910 participações em proposições, seguido pela Esquerda com 5107 e, bem atrás, pela Direita com apenas 1741 participações em proposições.

O mesmo padrão se observa nas aprovações, com o Centrão somando 133 participações em PLs aprovadas, a Esquerda 112 e a Direita apenas 26. Embora o Centrão concentre o maior volume, sua taxa de conversão não é muito superior à da Esquerda, que se mostra relativamente eficiente. Já a Direita apresenta um baixo desempenho tanto em volume quanto em aprovações no período analisado.

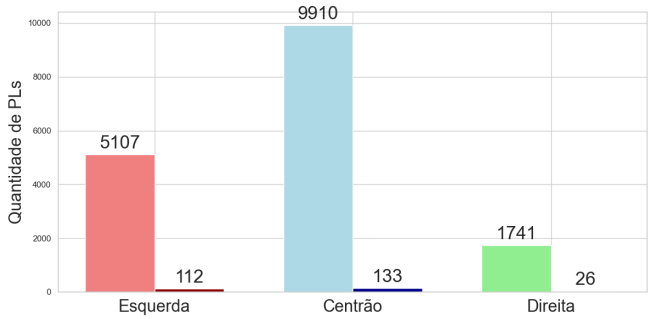

Figura 3. Distribuição de PLs submetidas (esquerda) e aprovadas (direita) por espectro político

4.3. Relação entre Espectro Político e Efetividade Legislativa

A Figura 4 e a Figura 5, apresentadas lado a lado, ilustram a relação entre espectro político e efetividade legislativa, considerando tanto o nível partidário quanto o agrupamento por bloco. Para o cálculo do índice de aprovação, utiliza-se a seguinte fórmula:

$$\text{Índice de Aprovação (\%)} = \frac{\text{Quantidade de PLs Aprovadas}}{\text{Quantidade de PLs Submetidas}} \times 100$$

Esse indicador reflete a proporção de proposições que cada partido ou bloco político consegue transformar em normas jurídicas. Observa-se que, proporcionalmente, o bloco da Esquerda apresenta o maior índice de aprovação, seguido pela Direita, enquanto o Centrão, apesar de concentrar o maior volume absoluto de proposições, apresenta uma taxa de conversão proporcionalmente inferior.

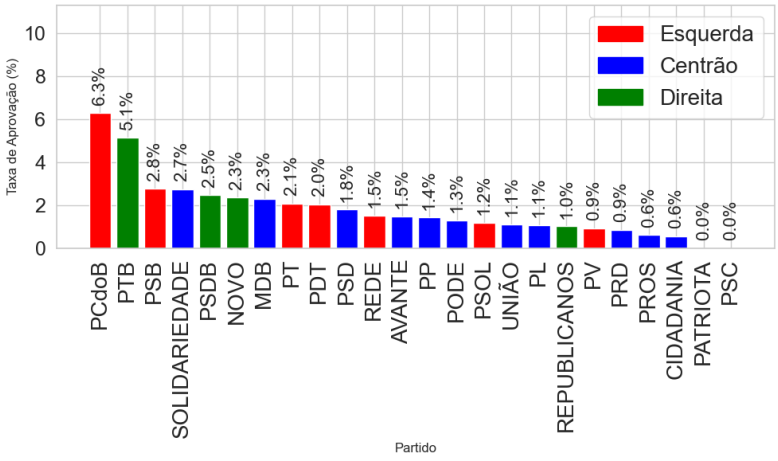

Figura 4. Índice de aprovação de PLs por partido

Esses resultados levantam questionamentos sobre os fatores que influenciam a eficiência legislativa. Estariam essas diferenças relacionadas ao tema ou tema dominante das proposições? À Unidade Federativa ou à região dos parlamentares? À autoria coletiva ou individual? Ou ainda, ao contexto político vivido no país durante o período

analisado? Este cenário sugere caminhos para investigações futuras, especialmente com o uso de técnicas de Inteligência Artificial e Aprendizado de Máquina, capazes de identificar padrões e variáveis com maior influência na efetividade legislativa.

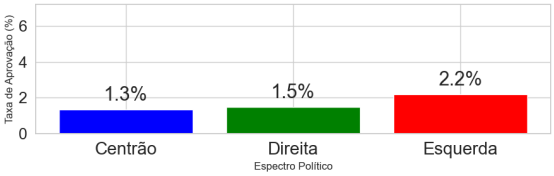

Figura 5. Índice de aprovação de PLs por espectro político

5. Disponibilização e Utilização

5.1. Disponibilização dos Dados e Código

O conjunto de dados resultante deste trabalho, bem como o código-fonte utilizado para sua obtenção e enriquecimento, estão disponibilizados publicamente em repositório do *GitHub* `https://github.com/juanmorysson/LegisPL-BR.git`. O *dataset* compreende proposições legislativas do tipo Projeto de Lei (PL) da Câmara dos Deputados entre os anos de 2022 e 2024, incluindo campos estruturados como autoria, tramitação e ementa textual.

O repositório inclui um *script* em *python* modularizado e comentado, permitindo sua reutilização e adaptação. Os usuários podem facilmente modificar a lista de anos ou o tipo de proposição a ser coletado, ampliando o escopo do *dataset* com novos dados obtidos diretamente da API pública. Além disso, o *script* oferece uma rotina completa de coleta paginada, enriquecimento dos registros com autores e informações de tramitação, normalização textual e exportação final em formato *Excel* (.xlsx), compatível com ferramentas de análise de dados.

5.2. Possíveis Cenários para Utilização

O conjunto de dados disponibilizado oferece uma ampla gama de possibilidades para aplicação em contextos de pesquisa, análise legislativa, ciência de dados e ensino. Como exemplo, o agrupamento (clusterização) de ementas com técnicas de Processamento de Linguagem Natural (PLN), a fim de identificar padrões temáticos entre proposições legislativas. A técnica de PLN pode ser aplicada no campo `ementa` coletado na primeira fase do processo, que contem uma descrição resumida do conteúdo da proposição. Com a extração atual existe um total de 12.334 textos, com 245.350 palavras, podendo ser ampliada para mais anos legislativos. A figura 6 demonstra a nuvem de palavras da base no campo `ementa`.

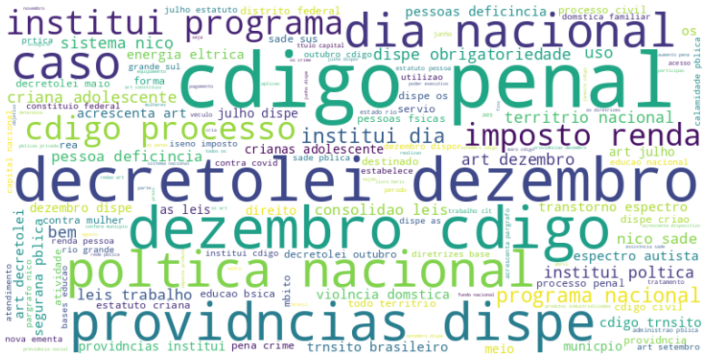

Figura 6. Nuvem de palavras da coluna `ementa`

Além disso, uma aplicação direta, pode ser a classificação supervisionada de proposições com base em atributos derivados, como o `tema_dominante`, a região geográfica do autor (`região`) ou a classificação ideológica (`bloco_partidario`). Com base em rótulos manuais ou históricos de tramitação, é possível treinar modelos para prever a aprovação de uma proposição (`aprovado`), sua tramitação acelerada ou sua associação a determinadas agendas temáticas (ex: saúde, segurança, economia). Técnicas de vetorização como TF-IDF e *embeddings* semânticos (como SBERT) podem ser combinadas com algoritmos como *Random Forest*, SVM ou redes neurais para esse fim. Em cenários com dados rotulados escassos, podem ser aplicadas estratégias de aprendizado fraco (*weak supervision*), supervisão distante (*distant supervision*) ou técnicas de autorotulagem com heurísticas baseadas nas colunas enriquecidas.

Outro cenário relevante pode ser a análise exploratória e estatística de padrões legislativos. As colunas `dias_tramitacao`, `qtd_tramitacoes` e `autoria_coletiva` permitem investigar correlações entre o tempo de tramitação e fatores como tema, autoria individual/coletiva ou filiação partidária. Análises por região ou bloco partidário também podem ser realizadas para identificar possíveis diferenças na pauta legislativa, no tempo de tramitação ou na taxa de aprovação.

O conjunto também pode ser utilizado como base para estudos qualitativos conduzidos por especialistas em ciência política ou jornalismo investigativo. Por exemplo, a partir da coluna `ementa` e da autoria, analistas podem rotular manualmente o viés ideológico das proposições (progressista, conservador, corporativista, etc.), alimentando pesquisas sobre alinhamento político, estratégias legislativas e polarização no parlamento brasileiro. Esses rótulos podem posteriormente ser utilizados para treinar modelos de previsão de viés, estudo de mudança de agenda ou detecção de *outliers* legislativos.

Por fim, o *dataset* representa uma ferramenta didática valiosa para disciplinas de ciência de dados, PLN, jornalismo de dados ou ciência política, ao oferecer um exemplo real de conjunto de dados governamentais abertos, estruturados e enriquecidos, com potencial para exercícios práticos de coleta, transformação, visualização e modelagem preditiva.

6. Considerações Finais e Trabalhos Futuros

Este trabalho apresentou a construção de um conjunto de dados estruturado e enriquecido de proposições legislativas do tipo Projeto de Lei (PL) da Câmara dos Deputados, abrangendo os anos de 2022 a 2024. A partir da API oficial de Dados Abertos da Câmara, foi possível extrair metadados relevantes e, por meio de um processo sistemático de transformação, gerar variáveis analíticas adicionais capazes de apoiar investigações quantitativas e qualitativas sobre o processo legislativo brasileiro.

O pipeline desenvolvido compreende etapas de coleta, normalização, enriquecimento semântico e categorização política dos autores, resultando em um *dataset* tabular com alto grau de organização, reprodutibilidade e potencial de reutilização. A disponibilização pública tanto dos dados quanto dos *scripts* utilizados visa fomentar a transparência e oferecer uma base sólida para futuras análises empíricas em ciência política e ciência de dados.

Como perspectiva para trabalhos futuros, pretende-se explorar a aplicação de técnicas PLN (como TF-IDF e SBERT) em conjunto com agrupamento não supervisionado (como K-Means e UMAP), com o objetivo de identificar padrões recorrentes entre as proposições legislativas a partir de suas ementas. Em especial, busca-se investigar a existência de alinhamentos temáticos com os blocos ideológicos dos partidos autores, contribuindo para a compreensão do comportamento legislativo sob a ótica do conteúdo das propostas e de seu viés político.

Além disso, outras possibilidades incluem a expansão do recorte temporal do *dataset*, a inclusão de novos tipos de proposições e a incorporação de dados complementares, como histórico de votação, tramitações intermediárias detalhadas ou contexto socioeconômico dos temas tratados. Tais avanços podem enriquecer ainda mais a capacidade explicativa da base de dados e ampliar seu uso em análises comparativas e preditivas.

Apêndice A – Tabela Complementar

Tabela 2. Dicionário de Dados do Dataset de Proposições Legislativas

Campo	Tipo	Descrição
id	Inteiro	Identificador único da proposição na API da Câmara dos Deputados
tipo	Texto	Sigla do tipo da proposição (ex: PL, PEC, etc.)
numero	Inteiro	Número da proposição no respectivo ano
ano	Inteiro	Ano de apresentação da proposição
ementa	Texto	Descrição resumida do conteúdo da proposição
temas	Texto	Palavras-chave temáticas associadas à proposição
status atual	Texto	Situação atual da proposição, segundo a API oficial
autores	Texto	Lista dos autores e seus partidos/UFs, separados por vírgula
tipo de autor	Texto	Tipo de autor (ex: Parlamentar, Comissão, etc.)
partido_principal	Texto	Sigla do partido do autor principal da proposição
bloco_partidario	Texto	Classificação ideológica do partido principal (Esquerda, Centrão, Direita, Outros)
autoria_coletiva	Booleano	Indica se a proposição tem mais de um autor
uf_principal	Texto	Unidade Federativa (UF) do autor principal
região	Texto	Região brasileira correspondente à UF principal
órgão inicial	Texto	Órgão responsável pela primeira tramitação da proposição
descrição inicial	Texto	Ação registrada na primeira tramitação
data da apresentação	Data	Data da apresentação da proposição
último órgão	Texto	Nome do órgão responsável pela última tramitação
última tramitação	Texto	Descrição da última movimentação da proposição
data da última tramitação	Data	Data da última tramitação no formato aaaa-mm-dd
dias_tramitacao	Inteiro	Número de dias entre a apresentação e a última tramitação
finalizado	Booleano	Indica se a proposição foi encerrada (arquivada, rejeitada, etc.)
aprovado	Booleano	Indica se a proposição foi aprovada ou transformada em norma
tema_dominante	Texto	Tema principal inferido a partir das palavras-chave
apensada_ou_duplicada	Booleano	Indica se a ementa aparece mais de uma vez no dataset
qtd_tramitacoes	Inteiro	Estimativa da quantidade de tramitações diferentes registradas

Referências

Barbalho, F. A. (2018). A emergência do campo de políticas públicas de dados abertos governamentais no brasil. *Conhecer: Debate entre o Público e o Privado*, 8(20):118–137.

Bommarito, M. J., Katz, D. M., and Detterman, E. M. (2020). Lexnlp: Natural language processing and information extraction for legal and regulatory texts. In *Research Handbook on Big Data Law*. Edward Elgar Publishing.

Brandt, M. B. e. a. (2018). Modelo de dados abertos conectados para informação legislativa. *Informação & Sociedade: Estudos*, 28(2):149–161.

Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011. `https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm`. Diário Oficial da União, Brasília, DF, 18 nov. 2011. Acesso em: 2 jun. 2025.

Breitman, K. e. a. (2012). Open government data in brazil. *IEEE Intelligent Systems*, 27(3):45–49.

Cavalcante, G. V., Sousa, F. R. d., Vaz, R. C. R., and Araujo, C. H. G. (2016). Dados abertos legislativos: o parlamento e o cidadão. *Anais da VIII Jornada de Pesquisa e Extensão*.

Chebolu, S. U. S., Dernoncourt, F., Lipka, N., and Solorio, T. (2023). A review of datasets for aspect-based sentiment analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing (IJCNLP 2023) and the 3rd Conference of the Asia-Pacific Chapter of the ACL (Volume 1: Long Papers)*, pages 611–628, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Dalenogare, L. G. C. and Araújo, M. A. D. d. (2019). Abordagens teóricas em dados governamentais abertos. *Revista Gestão e Tecnologia*, 19(5):296–314.

Dzikrullah, F. and Rinjani, M. A. (2017). A framework design to develop integrated data system for smart e-government based on big data technology. *Bulletin of Social Informatics Theory and Application*, 1(2):41–51.

Fagundes, M. F. and Ribeiro Junior, D. I. (2020). Modelo baseado em frictionless data aplicado aos dados abertos governamentais. *Revista Digital de Biblioteconomia e Ciência da Informação*, 18:e020034.

Janssen, K. (2011). The influence of the psi directive on open government data. *Government Information Quarterly*, 28(4):446–456.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4):258–268.

Li, A., Hua, X., Liao, Y., and Bansal, M. (2023). Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 14252–14270, Singapore. Association for Computational Linguistics.

Mumcuoğlu, E. e. a. (2021). Natural language processing in law: Prediction of outcomes in the higher courts of turkey. *Information Processing & Management*, 58(6):102684.

Nay, J. (2018). Natural language processing and machine learning for law and policy texts. Technical Report 3438276, Social Science Research Network (SSRN). Acesso em: 2 jun. 2025.

Oliveira, G. P., Silva, M. O., Costa, L. G. L., Dutra, M. T., and Pappa, G. L. (2024). Icpset: Um conjunto de dados estruturados de itens de compras p'ublicas. In *Proceedings of the VI Dataset Showcase Workshop (DSW)*, pages 103–113, Florian'opolis, SC, Brazil. SBC.

Rolim, T. V., Ávila, C. V. S., Freitas, R., Mariano, R. G., and Vidal, V. M. P. (2024). Construção do dataset semântico de pessoas jurídicas. In *Proceedings of the VI Dataset Showcase Workshop (DSW)*, pages 41–52, Florianópolis, SC, Brazil. SBC.

Ruijer, E., Grimmelikhuijsen, S., and Meijer, A. (2017). Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quarterly*, 34(1):45–52.

Testa, G., Mesquita, L., and Bolognesi, B. (2024). Do fisiologismo ao centro do poder: as reformas eleitorais e o centrão 2.0. *Caderno CRH*, 37(100):e024003.

Wilkinson, M. D. e. a. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.