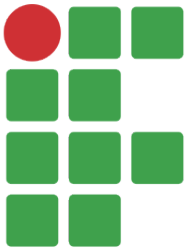
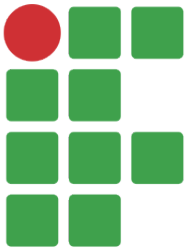


Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Coordenação do Curso Superior Bacharelado em Engenharia de Computação

Democracia descomplicada: utilizando Processamento de Linguagem Natural para classificar propostas de parlamentares na Câmara dos Deputados.

Arthur Maurício Thomaz Soares

Orientador: Prof. Marcelo José Siqueira Coutinho de Almeida, DSc.



Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Coordenação do Curso Superior de Tecnologia em Telemática

Democracia descomplicada: Utilizando Processamento de Linguagem Natural para para classificar propostas de parlamentares na Câmara dos Deputados.

Arthur Maurício Thomaz Soares

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação, do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - *Campus* Campina Grande, em cumprimento às exigências parciais para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Marcelo José Siqueira Coutinho de Almeida, DSc.

Campina Grande, setembro de 2024

S776d Soares, Arthur Maurício Thomaz
Democracia descomplicada: utilizando processamento de linguagem natural para classificar propostas de parlamentares na Câmara dos Deputados. / Arthur Maurício Thomaz Soares. - Campina Grande, 2024.
62 f. : il.

Trabalho de Conclusão de Curso (Curso Superior de Bacharelado em Engenharia de Computação) - Instituto Federal da Paraíba, 2024.
Orientador: Prof. Dr. Marcelo José Siqueira Coutinho de Almeida.

1. Transparência governamental 2. Processamento de linguagem natural 3. Aprendizagem de máquina I. Almeida, Marcelo José Siqueira Coutinho de II.Título.

CDU 004.8:351

Arthur Maurício Thomaz Soares

Prof. Marcelo José Siqueira Coutinho de Almeida

Profa. Iana Daya Cavalcante Facundo Passos
Membro da Banca

Prof. Henrique do Nascimento Cunha
Membro da Banca

Campina Grande, Paraíba, Brasil
Setembro/2024

"Tudo vale a pena quando a alma não é pequena."
(Fernando Pessoa)

Agradecimentos

Ao concluir este trabalho agradeço primeiramente a Deus, por ter me dado a oportunidade de seguir nesta jornada com saúde e coragem. Agradeço à toda minha família em nome dos meus pais, Antônio e Mércia, dos meus irmãos Pedro e Vinícius e da minha companheira Ana Paula por todo o carinho, amor e paciência nos dias difíceis.

Agradeço aos colegas de turma (que se tornaram valiosos amigos) que me foram presenteados nesta instituição em nome de Daniel Lemos, Jhonatan Andrade, Beatriz Araújo, Hércules Silva, Bianca Rangel, Isaac Antônio e Matheus Alves e a todos os amigos de caminhada que carrego no coração em nome de Jonatas Duarte, Hugo Paulino, João Paulo Dantas, Sara Ayme e Luiz Eduardo pois com eles a vida se torna mais leve e divertida. Dentre eles um especial agradecimento a Mateus Lisboa pelas orientações e *feedbacks* fundamentais a construção deste trabalho.

Agradeço a todos os professores que fizeram parte desta jornada em nome do professor Marcelo José Siqueira, meu orientador, professora Iana Daya e professor Henrique Cunha pela confiança, conselhos e ensinamentos.

Resumo

Em uma sociedade democrática, a transparência das ações governamentais é fundamental. No contexto do poder legislativo brasileiro, é crucial que a sociedade civil possa acompanhar de forma clara e objetiva as decisões tomadas pelos parlamentares na Câmara dos Deputados. Atualmente, a complexidade e a dificuldade de acesso às informações legislativas dificultam o acompanhamento da sociedade civil, e esta dificuldade demanda soluções que aproximem as pessoas das atividades legislativas do Brasil, como por exemplo o portal ”Ranking dos Políticos” (<politicos.org.br>). O presente trabalho visa contribuir neste aspecto através da criação de uma base de dados relacional, contendo as proposições de deputados federais e seus partidos, além da categorização das proposições em tópicos. Para realizar essa categorização foram utilizadas técnicas de Processamento de Linguagem Natural, Aprendizagem de Máquina e Visão Computacional para classificar automaticamente as proposições em tópicos relevantes. É esperado que a base de dados resultante auxilie no desenvolvimento de aplicações que democratizem a informação sobre as atividades legislativas promovendo maior transparência acerca do que é proposto da Câmara dos Deputados do Brasil.
Palavras-chave: Processamento de Linguagem Natural, Aprendizado de Máquina, Câmara dos Deputados, Proposições, Democracia, Visão Computacional

Abstract

In a democratic society, transparency in government actions is fundamental. In the context of the Brazilian legislative branch, it is crucial that civil society can clearly and objectively monitor the decisions made by parliamentarians in the Chamber of Deputies. Currently, the complexity and difficulty in accessing legislative information hinder civil society's ability to keep track, and this difficulty demands solutions that bring people closer to Brazil's legislative activities, such as the "Ranking dos Políticos"portal (politicos.org.br). This work aims to contribute to this aspect by creating a relational database containing the proposals of federal deputies and clustering theese proposal into topics. To achieve this goal, Natural Language Processing, Machine Learning, and Computer Vision techniques were used to automatically classify the proposals into relevant topics. It is expected that the resulting database will aid in the development of applications that democratize information about legislative activities, promoting greater transparency regarding what is proposed in the Brazilian Chamber of Deputies.
Keywords: Natural Language Processing, Machine Learning, Chamber of Deputies, Legislative proposals, Democracy, Computer Vision

Sumário

Lista de Siglas e Abreviaturas

IA	Inteligência Artificial
API	Application Programming Interface (Interface de Programação de Aplicação)
EMC	Emenda à Constituição
HTTP	HyperText Transfer Protocol (Protocolo de Transferência de Hiper Texto)
LLMs	Large Language Models (Modelo de linguagem geral)
OCR	Reconhecimento Ótico de Caracteres
PDF	Portable Document Format (Formato de documento portátil)
CSV	Comma-Separated Value (Valores separados por vírgula)
PEC	Proposta de Emenda à Constituição
PL	Projeto de Lei
PLN	Processamento de Linguagem Natural
TF-IDF	Frequência do Termo – Frequência Inversa dos Documentos
TICs	Tecnologias da Informação e Comunicação
URL	Uniform Resource Locator (Localizador Uniforme de Recursos)

Lista de Figuras

Lista de Símbolos

$idf(t)$ Inverso da frequência de termos nos documentos

$tf(t, d)$ Frequência do termo no documento

$TFIDF(t, d)$ Frequência do termo–inverso da frequência nos documentos

Capítulo 1

Introdução

1.1 Contextualização

Desde 1889 vigora no Brasil a democracia, forma de governo utilizada ao redor do mundo que tem origem associada à cidade de Atenas, capital da Grécia. Os valores democráticos desempenham um papel de grande importância na sociedade moderna, influenciando desde questões básicas tais como aquelas que definem nosso modelo educacional até outras mais complexas como a relação entre as nações. Uma das principais características desse modelo político é que a população participa da escolha dos seus representantes por meio de eleições diretas.

As Tecnologias da Informação e Comunicação (TICs) estão promovendo uma profunda transformação nos comportamentos individuais e nas interações sociais. Essa mudança se reflete também na democracia, que passa por significativas alterações em seus processos internos devido à influência das TICs. Conforme BATISTA C.; VIANA (2006), as TICs geram um novo panorama de relações e comportamentos sociais, exemplificado pela disseminação de notícias falsas e o uso de *deepfakes* durante as eleições (MULHOLLAND C.; DE OLIVEIRA, 2021).

De acordo com a ENAP (Escola Nacional de Administração Pública, 2022) 51% dos 718 gerentes públicos latino-americanos pesquisados reconhecem ter um déficit severo ou muito severo em suas habilidades de análise de dados. O próprio Governo Federal reconhece tal fato e por isso em 2020 foram lançadas pelo Governo Federal as 7 competências transversais, e dentre elas foi citada a seguinte competência: "resolução de problemas com base em dados"[2].

Para CASTELLS (1999) uma nova sociedade surge quando há uma transformação estrutural nas relações de produção, de poder e de experiência. A chegada das TICs também podem ser consideradas como tais mudanças, gerando um novo arranjo social a partir da maior facilidade de disseminação de informações. Porém, estas mudanças muitas vezes podem não atingir todas as camadas da população da mesma forma. Podemos constatar isso no trabalho de MAX (2012), que realizou uma pesquisa de perfis de acesso ao portal da Câmara dos Deputados[1]. Essa pesquisa mostra que 71% dos usuários deste portal são homens, 67,5% possuem ensino superior completo e mais da metade dos acessos a esse portal estão distribuídos no Distrito Federal e em outros estados

[1] <http://camara.leg.br>

do Sul e Sudeste.

ROVER (2006) chama essa nova democracia de “Democracia Aberta”, “Democracia Eletrônica” ou “Democracia Digital”. Nesta nova Democracia Digital são necessárias iniciativas que agreguem, simplifiquem e espalhem essas informações com uma maior capilaridade atingindo toda a pluralidade do Brasil, visando sanar as desigualdades no acesso de informações referentes aos processos que ocorrem nesta nova democracia, como a iniciativa do portal politicos.org[2] que desde 2011 avalia senadores e deputados federais em exercício, classificando-os de acordo com os critérios de combate aos privilégios, aos desperdícios e à corrupção no poder público.

1.2 Objetivos

1.2.1 Objetivo geral

Analisar, organizar e classificar os dados da Câmara dos Deputados do Brasil para habilitar novas soluções que aumentem a visibilidade das atividades dos deputados.

1.2.2 Objetivos específicos

- Disponibilizar uma base de dados relacional com as informações de deputados, partidos e proposições de fácil entendimento e pronta para ser utilizada em projetos.
- Categorizar proposições a partir do seu conteúdo em tópicos.

1.3 Organização do trabalho

Este trabalho está subdividido em três capítulos:

1.3.1 Fundamentação teórica

O capítulo de fundamentação teórica aborda os conceitos e ferramentas que sustentam o desenvolvimento do trabalho. Ele detalha as etapas do processo de mineração de dados, explica o método utilizado para organizar o projeto e também explora as proposições legislativas. Em suma, ele fornece o embasamento teórico essencial para a compreensão das etapas e tecnologias empregadas no desenvolvimento da pesquisa.

Este capitulo também tem uma seção reservada aos trabalhos relacionados, listando pesquisas anteriores que utilizam técnicas de PLN (Processamento de Linguagem Natural) e aprendizado de máquina para analisar dados legislativos, como discursos e proposições. Esses estudos destacam o potencial dessas ferramentas para promover a transparência, a participação cidadã e a compreensão do processo legislativo no Brasil.

[2] <http://politicos.org>

1.3.2 Metodologia e Desenvolvimento

O capítulo de metodologia e desenvolvimento descreve detalhadamente o processo de coleta, pré-processamento e análise dos dados, bem como as ferramentas e tecnologias utilizadas para construir o modelo de aprendizado de máquina. Ele aborda o que foi feito desde a obtenção dos dados da *API* (*Application Programming Interface*) da Câmara dos Deputados até a implementação do modelo e sua avaliação, oferecendo uma visão completa das etapas e decisões tomadas durante o desenvolvimento do trabalho.

1.3.3 Considerações finais e Trabalhos Futuros

O capítulo de considerações finais e trabalhos futuros recapitula as principais descobertas da pesquisa, avalia o alcance dos objetivos iniciais e propõe direções para pesquisas futuras. Ele destaca o potencial do modelo desenvolvido para o desenvolvimento de aplicações que aprimorem o acesso à informação referente aos processos da Câmara dos Deputados.

Capítulo 2

Fundamentação Teórica

Foi feita uma pesquisa bibliográfica sobre dados abertos e atividades da Câmara dos Deputados, análise do funcionamento de proposições e legislação, estudo das *APIs* da Câmara[1] para coleta de dados sobre partidos, deputados e proposições. O pré-processamento dos dados incluiu técnicas de *multi-threading*, tratamento de dados inconsistentes e extração de texto de arquivos PDF utilizando visão computacional e reconhecimento ótico de caracteres (OCR). O capítulo também aborda a definição de modelos conceituais para o banco de dados PostgreSQL[2] e o uso de Docker[3] para facilitar o desenvolvimento e a reprodução do ambiente do projeto. Em resumo, o capítulo detalha o processo de pesquisa, coleta, tratamento e organização dos dados, ferramentas e tecnologias utilizadas para o desenvolvimento do projeto e por fim a pesquisa bibliográfica listando trabalhos relacionados a área de dados públicos.

2.1 Proposições

Segundo o Portal da Câmara dos Deputados[4], proposições são os diversos tipos de documentos da Câmara dos Deputados do Brasil que são submetidos à consideração, discussão e votação dos deputados federais no âmbito da instituição legislativa brasileira. Elas podem abranger uma ampla gama de assuntos, indo desde projetos de lei até emendas constitucionais, passando por propostas de emenda à Constituição (PECs), medidas provisórias, requerimentos, indicações e diversos outros tipos de documentos.

As proposições são objetos cruciais do poder legislativo que é exercido hoje no Brasil. Elas passam por um processo legislativo complexo, que envolve discussão, análise em comissões especializadas e votação em plenário, antes de serem encaminhadas para apreciação do Senado Federal e, eventualmente, sanção pelo Presidente da República.

O Portal da Câmara dos Deputados[5] disponibiliza todas as proposições que foram produzidas

[1] <http://camara.leg.br>
[2] <https://www.postgresql.org>
[3] <https://www.docker.com>
[4] <www.camara.leg.br/>
[5] <www.camara.leg.br/>

na câmara. No momento deste trabalho foram encontrados 76 tipos diferentes de proposições legislativas, disponíveis no ANEXO A.

Com o intuito de diminuir o escopo, foram escolhidos os seguintes tipos de proposições:

1. **PL (Projeto de Lei)**: Proposta legislativa apresentada por um parlamentar ou comissão com o objetivo de criar, alterar ou revogar uma lei ordinária.

2. **EMC (Emenda à Constituição)**: Proposta legislativa que visa a alteração da Constituição Federal. Esse tipo de proposição requer um processo legislativo mais rigoroso e exige um maior quórum para aprovação.

3. **RDF (Redação Final)**: Proposição que consiste na versão final do texto de uma proposição legislativa, após ser aprovada em todas as etapas do processo legislativo. É a versão que será enviada para sanção ou promulgação.

4. **SBT (Substitutivo)**: Um Substitutivo é um tipo de proposição que consiste em uma nova redação para um projeto de lei, elaborada por um parlamentar ou comissão, substituindo integralmente o texto original do projeto.

Esses tipos são relacionados a assuntos mais importantes da Câmara, como mudanças na constituição e criação de leis. A motivação para a escolha deles foi criar um foco em proposições pouco subjetivas e com intenções mais abrangentes, removendo por exemplo proposições relacionadas a licitações, indicações e convocações para reuniões ordinárias da Câmara.

2.2 Mineração de dados

A mineração de dados é definida por BERRY Michael J. A.; LINOFF (1997) como a exploração e a análise, por meio automático ou semiautomático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativas. Ela pode ser usada para resolver uma variedade de problemas que envolvem grandes quantidades de dados nas áreas de negócio, ciência e governo. Algumas aplicações relevantes são:

- Analisar dados de clientes para encontrar padrões de consumo
- Detecção de fraudes
- Identificação de padrões a fim de melhorar diagnósticos médicos
- Reduzir a dimensionalidade grandes quantidades de dados para encontrar padrões e gerar informação a partir disto

As informações da Câmara dos Deputados são disponibilizadas por meio de uma API de acesso aos recursos HTTP e por arquivos.

2.2.1 API HTTP

Uma *API HTTP* (HyperText Transfer Protocol), ou Interface de Programação de Aplicações baseada em Protocolo de Transferência de Hipertexto, é um conjunto de padrões e protocolos que permitem a comunicação entre diferentes softwares por meio da internet utilizando o protocolo *HTTP*. Segundo a RFC 9110 (IETF, 2022) o protocolo *HTTP* por sua vez é a base da comunicação de dados para a internet, permitindo a transferência de recursos, como documentos *HTML* (Hypertext Markup Language), *JSON* (JavaScript Object Notation) e *XML* (Extensible Markup Language) entre clientes e servidores.

As *APIs* servem como especificações de como componentes de software devem interagir uns com os outros, permitindo o acesso e a manipulação de recursos e dados de uma aplicação ou serviço de forma padronizada e segura.

Essa solicitação é feita através de uma *URL* (Uniform Resource Locator), que especifica o recurso desejado e inclui informações como o método *HTTP* a ser utilizado (ex: GET para obter dados, POST para enviar dados) e cabeçalhos *HTTP* que fornecem metadados sobre a solicitação ou o cliente. O servidor processa a solicitação e responde com um código de status (por exemplo, 200 corresponde a OK, 404 a *Not Found* e assim por diante) e, se aplicável, o recurso solicitado no corpo da resposta, também acompanhado de cabeçalhos *HTTP* que podem conter informações sobre o corpo da resposta ou do servidor.

Utilizando as *APIs* disponibilizadas pela câmara é possível coletar informações sobre entidades específicas (por exemplo, as informações de um deputado em específico) ou listar entidades (por exemplo, listar todas as proposições relacionadas a um deputado).

2.3 Pré-processamento dos dados

TAN P. N.; STEINBACH (2014) explica que o pré-processamento de dados deve ser utilizado para deixar os dados mais propícios para a mineração de dados. O pré-processamento de dados é uma área muito ampla que consiste em diversas estratégias e técnicas que estão interligadas para atingir o objetivo citado acima.

A preparação adequada dos dados de treinamento, incluindo a normalização e a remoção de características indesejadas, é essencial para evitar o enviesamento do modelo e garantir um processamento preciso. Neste trabalho, foram aplicadas etapas específicas de pré-processamento para otimizar os dados adquiridos, assegurando a qualidade e a confiabilidade dos resultados obtidos.

2.3.1 Visão computacional

A visão computacional é um ramo da ciência da computação que capacita sistemas computacionais a interpretar e compreender o conteúdo visual presente em imagens (PRINCE, 2012).

Através da aplicação de técnicas de processamento de imagens, reconhecimento de padrões e aprendizado de máquina, a visão computacional possibilita que computadores extraiam informa-

ções significativas de representações gráficas, identificando objetos, reconhecendo faces, interpretando gestos e executando diversas tarefas relacionadas à percepção visual.

Neste estudo, a biblioteca OpenCV[6] (*Open Source Computer Vision Library*) foi empregada para a definição de quadrantes em arquivos PDF de ementas produzidas pelos deputados na Câmara dos Deputados. Essa biblioteca livre e multiplataforma, compatível com diversas linguagens de programação, permitiu a identificação de regiões nos documentos que contêm textos mais relevantes sobre as propostas apresentadas, contribuindo para a análise e compreensão do conteúdo legislativo.

2.3.2 Reconhecimento Ótico de Caracteres (OCR)

De acordo com NAGY G.; NARTKER (2000) o Reconhecimento Ótico de Caracteres (OCR) desempenha um papel fundamental na conversão de texto presente em imagens ou mapas de bits em um formato legível por sistemas computacionais. No presente trabalho, o OCR foi utilizado em conjunto com a Visão Computacional para extrair o texto relevante dos arquivos PDF de ementas legislativas. Após a identificação das regiões de interesse nos documentos através da visão computacional, o OCR possibilitou a transformação do texto presente nessas áreas em um formato adequado para análise e processamento computacional.

A ferramenta *Tesseract*[7], um software livre de OCR originalmente desenvolvido pela Hewlett-Packard e posteriormente mantido pelo Google, foi empregada para implementar o OCR neste estudo.

2.3.3 Remoção de ruídos do texto

A etapa de pré-processamento textual é crucial para otimizar a qualidade dos dados e melhorar o desempenho dos modelos de análise de texto. Neste trabalho, foram aplicadas as seguintes técnicas de pré-processamento:

- **StopWords(ou palavras vazias)**: Este conceito foi descrito inicialmente por LUHN (1960) e pode ser definido como a remoção de palavras comuns, como conectivos (e, nem, também, etc.), de algum idioma especifico que geralmente não possuem valor semântico relevante para o processamento de linguagem natural.

- **TF-IDF (frequência do termo–inverso da frequência nos documentos)**: Trata-se de uma medida estatística que avalia a importância de um termo em um determinado documento, considerando tanto a frequência do termo no documento quanto sua frequência em todo o conjunto de documentos analisados. Esse método surgiu a partir da definição de idf: "A especificidade de um termo pode ser quantificada como função inversa do número de documentos em que ocorre."(JONES, 1972). O valor TF-IDF de uma palavra aumenta proporci-

[6] <https://opencv.org>
[7] <https://github.com/tesseract-ocr/tesseract>

onalmente ao número de vezes que ela aparece em um documento, mas é ponderado pela frequência da palavra em todo o *corpus*.

A fórmula que representa o TF-IDF é:

$$\text{TFIDF}(t, d) = tf(t, d) * idf(t) \tag{2.1}$$

Onde: $tf(t, d)$ é a frequência do termo t no documento d. $idf(t)$ é o inverso da frequência do termo t nos documentos, calculado como:

$$idf(t) = \log \frac{\text{Número de documentos}}{\text{Número de documentos onde o termo t aparece}} \tag{2.2}$$

Essas técnicas de pré-processamento contribuem para a redução de ruídos nos dados textuais, realçando as informações relevantes e aprimorando a eficácia das análises subsequentes.

2.4 Algoritmos de aprendizagem de máquina não supervisionados

Para MULLER e GUIDO (MULLER A.; GUIDO, 2016) a aprendizagem não supervisionada abrange todos os tipos de aprendizagem de máquina onde o resultado não é conhecido e o algoritmo utilizado não tem nenhum tutor auxiliando no aprendizado. Na aprendizagem não supervisionada o algoritmo tem apenas os dados de entrada é esperado que ele gere algum conhecimento na sua saída. Podemos dividir a aprendizagem não supervisionada em dois tipos gerais:

- Transformações de um conjunto de dados
- Clusterização de um conjunto de dados

Neste trabalho foi utilizado o algoritmo de aprendizagem de máquina KMeans.

2.4.1 Representação de texto

As técnicas de representação de texto são formas de transformarmos conjuntos de caracteres que seres humanos conseguem compreender em representações numéricas que, de alguma forma, mantenham o significado que o texto contém e possam ser processadas por um computador.

2.4.2 Processamento de linguagem natural

O Processamento de Linguagem Natural (*NLP*) busca capacitar computadores a entender, interpretar e gerar linguagem humana de forma significativa. É um campo interdisciplinar que combina conhecimentos de linguística, ciência da computação e inteligência artificial, com o objetivo

de criar sistemas capazes de interagir com a linguagem humana de maneira natural e eficiente (JURAFSKY D.; MARTIN, 2024).

Dentre os desafios do PLN estão: compreensão de língua natural, fazer com que computadores extraiam sentido de linguagem humana ou natural e geração de língua natural.

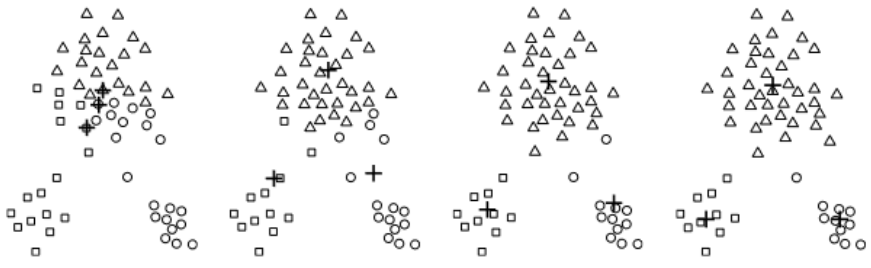
Neste trabalho foi utilizado o modelo pré-treinado de representação textual *fasttext*[8]. O *fasttext* é uma ferramenta de código aberto para processamento de linguagem natural e representação de texto em geral que se destaca por disponibilizar representações de texto para qualquer linguagem. Ela transforma texto em vetores contínuos processáveis por uma máquina que poderão ser utilizados posteriormente em análises voltadas à semântica dos textos em um idioma específico. A habilidade de capturar elementos semânticos de textos vem do seu pré-treinamento utilizando a Wikipédia[9] e a iniciativa *Common Crawl*[10].

2.4.3 KMeans

O Modelo de aprendizagem não supervisionada KMeans define um protótipo em termos de um centroide, que geralmente é a média de um grupo de pontos, e é tipicamente aplicado a objetos em um espaço contínuo n-dimensional (onde n é a dimensionalidade da entrada do modelo). Segundo TAN P. N.; STEINBACH (2014) a técnica de agrupamento KMeans consiste em:

- Escolher K centroides iniciais, onde K é um parâmetro especificado pelo usuário, ou seja, o número de clusters desejados. Cada ponto é então atribuído ao centroide mais próximo, e cada coleção de pontos atribuidos a um centroide é um **cluster**.
- O centroide de cada cluster é então atualizado com base nos pontos atribuídos ao cluster.
- A etapa anterior é repetida até que nenhum ponto mude de cluster, ou equivalentemente, até que os centroides permaneçam os mesmos.

Figura 2.1: *Exemplo de escolha de três centroides na execução de um algoritmo de KMeans em um plano de duas dimensões.*



Fonte: TAN P. N.; STEINBACH (2014)

[8] <https://fasttext.cc/>
[9] <https://pt.wikipedia.org>
[10] <https://commoncrawl.org/>

Como pode ser visto na Figura 2.1, os centroides são recalculados em cada iteração do algoritmo até uma escolha ótima (ou até atingir uma limite de parada).

2.4.4 Modelos de LLM

Os *Large Language Models* (LLMs), ou Grandes Modelos de Linguagem, representam uma classe de modelos de aprendizagem de máquina com capacidade de processar e gerar texto em linguagem natural. VASWANI (2017) propõe uma arquitetura de redes neurais chamada Transformer, baseada apenas em mecanismos de atenção, dispensando recorrência e convoluções inteiramente. Esta arquitetura forneceu a base para a criação de modelos poderosos de LLM devido principalmente a sua adaptabilidade, escalabilidade, paralelismo e eficiência.

Treinados com vastos conjuntos de dados textuais, esses modelos aprendem a reconhecer padrões, estruturas gramaticais e até mesmo nuances semânticas da linguagem, permitindo-lhes realizar uma variedade de tarefas, desde a tradução automática até a geração de conteúdo criativo. Como afirma RARFORD (2019), "Os LLMs demonstram uma notável capacidade de generalização, aplicando o conhecimento adquirido durante o treinamento a uma ampla gama de tarefas linguísticas, mesmo sem treinamento específico para elas". Essa flexibilidade e capacidade de aprendizado contínuo tornam os LLMs ferramentas poderosas em diversas áreas, desde a pesquisa científica até o desenvolvimento de chatbots e assistentes virtuais.

2.5 Persistência dos dados

2.5.1 PostgreSQL

O PostgreSQL[11] é um sistema de gerenciamento de banco de dados relacional de código aberto, robusto e versátil amplamente utilizado no mercado de tecnologia. Ele dispõe de diversas funções que auxiliam na consulta e inserção dos dados, além de ser otimizado para grandes cargas de escrita. Todos os dados adquiridos nas etapas anteriores foram armazenados em um banco PostgreSQL.

2.5.2 Docker

O Docker[12] é uma ferramenta para criação de contêineres que permite encapsular a configuração de componentes de software necessários para o desenvolvimento de uma solução, como linguagens de programação, compiladores, bancos de dados, sistemas operacionais específicos dentre outros.

[11] <https://www.postgresql.org>
[12] <https://www.docker.com>

2.6 Kanban

A metodologia *Kanban*, originária da Toyota e adaptada para o desenvolvimento de software, foi utilizada neste trabalho como ferramenta de gestão visual para o acompanhamento do projeto. Segundo RIBEIRO R. A.; FARINA (2023), o Kanban permite a organização das tarefas em um quadro, facilitando a visualização do progresso e a identificação de gargalos. Além disso, de acordo com MARODIN G.; FRANK (2018), o Kanban pode auxiliar na identificação de desperdícios e na melhoria contínua do processo, aspectos cruciais para o sucesso do projeto de engenharia proposto neste trabalho.

2.7 Trabalhos relacionados

Nesta seção é exposta a pesquisa bibliográfica sobre trabalhos que utilizam dos dados públicos (da Câmara, Governo Federal, Prefeituras e etc.) para gerar conhecimento e, principalmente, treinar modelos de aprendizagem de máquina que consigam extrair informações relevantes sobre trabalhos feitos por deputados (tanto no Brasil quanto no exterior). Estes trabalhos serviram de inspiração para o desenvolvimento deste e estão no decorrer do texto.

2.7.1 FERNANDES (2017)

O trabalho tem como objetivo extrair o perfil temático dos deputados federais, através do processamento dos textos obtidos de seus discursos e proposições, bem como desenvolver uma aplicação web para que os resultados dessa pesquisa sejam apresentados de forma lúdica e amigável (FERNANDES, 2017). O autor processa e classifica os textos dos sumários dos discursos da câmara dos deputados e disponibilizar uma interface web lúdica e amigável onde é possível consultar os assuntos dos discursos proferidos no Pequeno Expediente, que é a primeira parte da sessão ordinária do Plenário, e das proposições do tipo “Projetos de Lei”.

Figura 2.2: *Captura de tela do site do projeto "Tenho dito".*

Fonte: FERNANDES (2017)

Foi feito o uso dos sumários dos discursos e também do tesauro da Câmara dos Deputados. O tesauro é uma espécie de biblioteca de classificações disponibilizada pela câmara. O autor realizou um agrupamento dessas classificações em temas e macro-temas.

Para a classificação dos sumários nos temas citados anteriormente foi utilizado o algoritmo *NaiveBayesClassifier* disponível na biblioteca scikit-learn. Ele é um método de classificação supervisionada baseado no teorema de Bayes, que calcula a probabilidade de uma determinada amostra pertencer a uma classe específica com base nas características observadas. Ele assume independência condicional entre as características, ou seja, cada característica contribui de forma independente para a probabilidade de pertencer a uma classe. Durante o treinamento, o algoritmo calcula as probabilidades a priori de cada classe e as probabilidades condicionais das características para cada classe. Para cada documento são atribuídas probabilidades do mesmo se encaixar em um determinado tema, e o tema com maior probabilidade é o **tema principal** daquele documento. Ao tornar transparente os temas debatidos e propostos pelos legisladores, FERNANDES (2017) contribui para o fortalecimento da democracia no Brasil, promovendo uma maior compreensão das atividades parlamentares e facilitando o acompanhamento e a fiscalização do trabalho dos representantes eleitos pela sociedade.

Por fim, o autor levanta a ideia de analisar mais diretamente os textos das proposições da Câmara, que no momento do em que o trabalho foi realizado ainda não eram disponibilizados de forma direta.

2.7.2 MOLINARI (2020)

O trabalho de MOLINARI (2020) se propõe a analisar os padrões ideológicos e a polarização política na Câmara dos Deputados do Brasil, entre a 52ª e a 55ª legislatura, através da análise de 317.980 discursos parlamentares. Ao utilizar técnicas de processamento de linguagem natural

e aprendizado de máquina, o estudo busca identificar os discursos característicos de diferentes espectros ideológicos e quantificar a evolução da polarização política no período. A pesquisa contribui para a democracia aberta ao trazer transparência para o debate político, permitindo que a sociedade compreenda melhor as posições e estratégias dos seus representantes, promovendo assim uma maior participação e fiscalização cidadã.

A classificação dos discursos é realizada em duas etapas principais. Primeiramente, os partidos são classificados em esquerda, centro e direita com base no trabalho de ZUCCO C.; POWER (2019), e essa classificação é atribuída aos discursos de seus membros. Em seguida é o TF-IDF (*Term Frequency - Inverse Document Frequency*) é aplicada para criar vetores de vocabulário para cada grupo ideológico, destacando as palavras mais relevantes para cada um. Adicionalmente, um modelo de *Naive Bayes Classifier* é treinado para prever a ideologia dos discursos, e sua acurácia é utilizada como uma medida de polarização.

Os resultados do trabalho indicam um aumento da polarização política na Câmara dos Deputados[13] entre a 52ª e a 55ª legislatura, com uma inflexão na 54ª. A análise dos vocabulários evidencia uma divergência crescente entre esquerda e direita, especialmente a partir da 54ª legislatura, e uma aproximação entre centro e direita no mesmo período. A acurácia do modelo de classificação também corrobora esse aumento da polarização, com um salto significativo a partir da 54ª legislatura. O estudo conclui que a polarização se intensificou no período analisado, impulsionada por pautas como o *impeachment* da presidente Dilma Rousseff, e destaca a importância de pesquisas futuras que investiguem as causas e consequências desse fenômeno para a democracia brasileira.

2.7.3 MAX (2012)

A dissertação de MAX (2012) teve como objetivo principal analisar a interação dos usuários com o Portal da Câmara dos Deputados e como as ferramentas oferecidas por este atendem às necessidades de participação política dos cidadãos. Para tanto, a pesquisa utilizou duas metodologias: análise das estatísticas de acesso ao Portal e aplicação de um web survey para coletar a opinião dos usuários.

Um dos principais achados da dissertação é que a democracia eletrônica proporcionada pelo Portal da Câmara dos Deputados tem servido mais aos profissionais da política do que aos cidadãos comuns. No entanto, o autor ressalta que o cidadão comum interessado em participar da vida política tem procurado as ferramentas online disponibilizadas. Essa conclusão aponta claramente um vácuo na produção de informações compiladas e de fácil acesso acerca dos trâmites da Câmara dos Deputados, criando a necessidade de iniciativas que facilitem o acesso de tais informações ao cidadão comum.

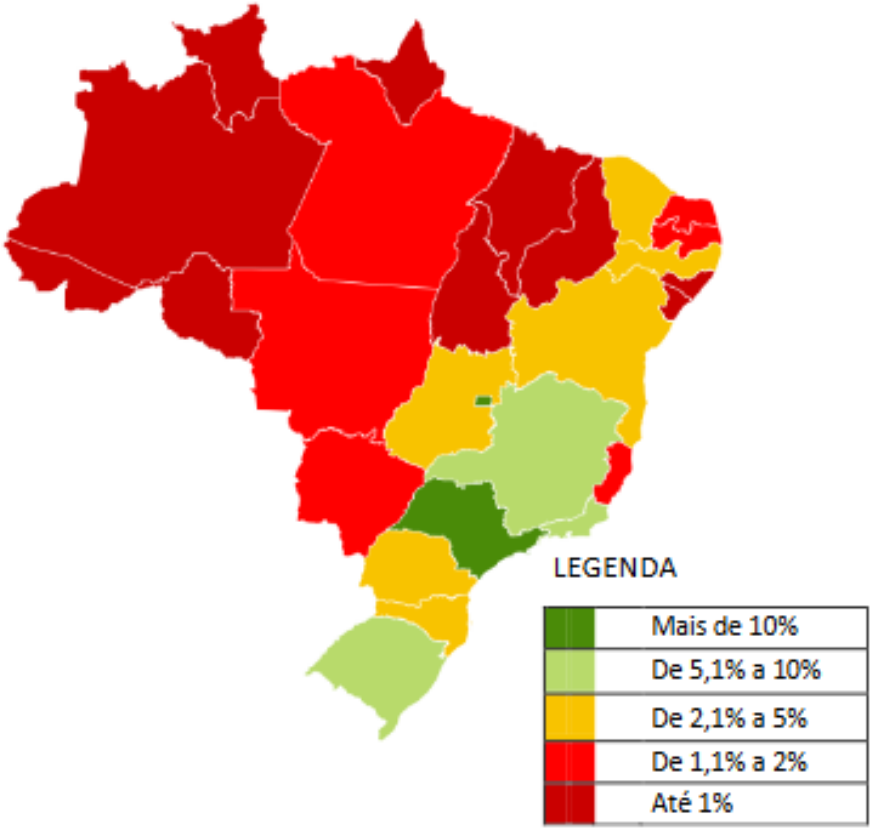
Outros resultados importantes apontados por MAX (2012) são:

- Perfil dos usuários: A maioria dos usuários do Portal acessa para fins de cidadania, são servidores públicos, possuem ensino superior e têm entre 41 e 65 anos.

[13] <https://www.camara.leg.br>

- Localização dos usuários: Sul e Sudeste, portanto, representam mais de 50% dos acessos ao Portal. O Distrito Federal tem a menor taxa de novas visitas (20,2%) e de rejeição (37%), o que reflete basicamente o acesso primordialmente profissional e contínuo ao Portal, principalmente feito pelas burocracias federais e dos profissionais da área da política que se encontram em Brasília.
- Avaliação do Portal: Os usuários que acessam o Portal para fins profissionais tendem a avaliá-lo melhor do que aqueles que o acessam para fins de cidadania. As ferramentas de interação por e-mail são as mais bem avaliadas, enquanto as ferramentas de interação em tempo real, como bate-papos e o Portal e-Democracia, são as piores avaliadas.
- Portal e-Democracia: A maioria dos usuários reconhece a importância do Portal e-Democracia, mas a participação nas discussões ainda é baixa. Os principais motivos para a não participação são a falta de tempo e o desconhecimento sobre o funcionamento do site.
- Sugestões dos usuários: Os usuários pedem por mais interatividade com os parlamentares, melhorias na linguagem e na navegabilidade do Portal, mais informações sobre as atividades parlamentares e a implementação de mecanismos de democracia direta.

Figura 2.3: *Mapa de acessos ao Portal da Câmara por UF*



Fonte: MAX (2012)

MAX (2012) conclui que, para que o Portal da Câmara dos Deputados possa atender às necessidades de participação política dos cidadãos comuns, é necessário facilitar a linguagem, melhorar a interatividade entre os atores políticos e os cidadãos. No contexto atual podemos clareamento utilizar de tecnologias de Inteligência Artificial e Aprendizagem de máquina para melhorar os pontos trazidos pela autora.

2.7.4 PETERSON (2018)

O trabalho de PETERSON (2018) propõe um método inovador para medir a polarização política em sistemas parlamentares, utilizando a acurácia de algoritmos de aprendizado de máquina para classificar discursos parlamentares por partido. Ao quantificar a distinção entre os discursos de diferentes partidos, os autores oferecem uma medida da polarização política ao longo do tempo. Essa abordagem contribui para a democracia aberta ao fornecer uma ferramenta para monitorar e analisar o nível de polarização no debate político, promovendo a transparência e o entendimento público sobre o posicionamento dos partidos e seus representantes.

A classificação dos discursos é realizada através de algoritmos de aprendizado supervisionado, que são treinados em um conjunto de discursos previamente rotulados com a identificação partidária do autor. Os algoritmos aprendem a identificar padrões linguísticos e temáticos que distinguem os discursos de diferentes partidos. Uma vez treinados, os modelos são capazes de prever a filiação partidária de novos discursos com base em seu conteúdo. A acurácia desses modelos, ou seja, a proporção de classificações corretas, é então utilizada como uma medida da polarização: quanto maior a acurácia, mais fácil é distinguir os discursos dos partidos, indicando um maior grau de polarização. Neste trabalho foi utilizada a estratégia de TF-IDF para capturar os termos mais importantes dos documentos e, assim como o trabalho citado anteriormente, o algoritmo de *NaiveBayesClassifier* como aprendizagem supervisionada.

Os autores aplicaram seu método a um conjunto de discursos da Câmara dos Comuns britânica abrangendo 78 anos. Seus resultados indicam que a polarização política no Reino Unido variou ao longo do tempo, com um período de relativa baixa polarização durante o "consenso do pós-guerra", seguido por um aumento significativo durante o governo Thatcher e um declínio subsequente. Esses achados, consistentes com análises históricas qualitativas e quantitativas, demonstram a validade do método proposto e seu potencial para gerar conhecimento sobre a evolução da polarização política em sistemas parlamentares.

Capítulo 3

Metodologia e Desenvolvimento

Neste capítulo é detalhado como é feito o uso de dados e estratégias de Mineração de Dados e Aprendizagem de Máquina com foco nas proposições disponibilizadas pelo Portal da Câmara dos Deputados do Brasil[1].

Através do portal da Câmara dos Deputados, a comunidade tem acesso facilitado a uma vasta gama de dados, desde gastos do cartão corporativo de cada deputado até pautas legislativas e discursos parlamentares. Em um contexto onde a transparência e a análise de dados desempenham um papel fundamental na governança democrática, essa disponibilidade de informações oferece oportunidades significativas para pesquisas e análises diversas.

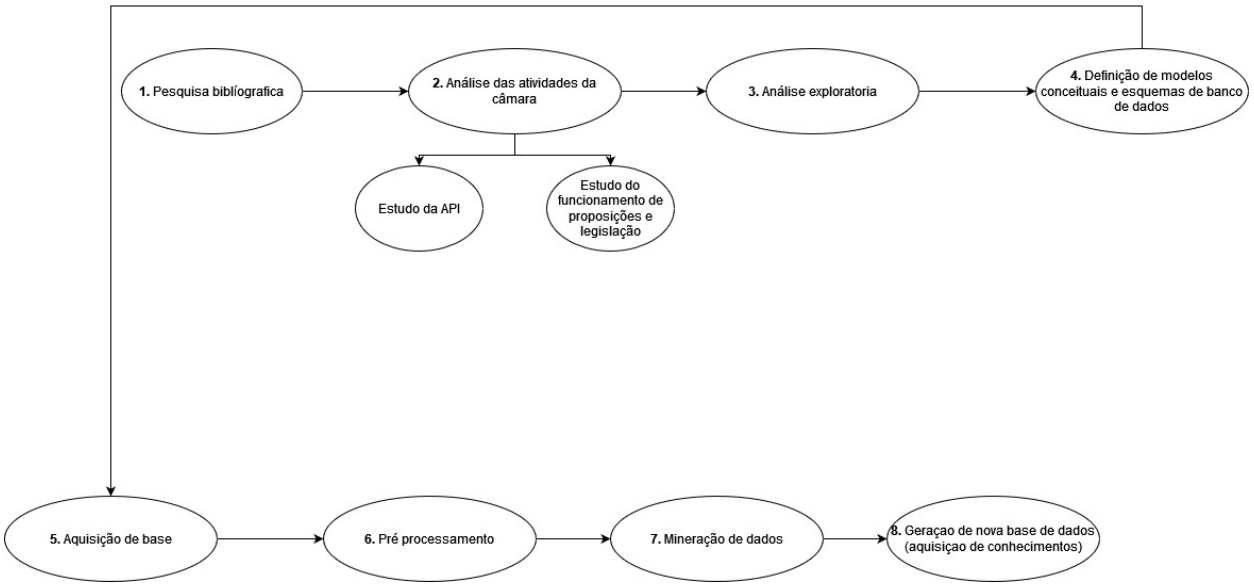
Utilizando aprendizagem não supervisionada será disponibilizada uma base de dados relacional, normalizada e sucinta. Para alimentar essa base (além das informações devidamente tratadas e organizadas que são disponibilizados pela Câmara dos Deputados) serão inferidos *clusters* para as proposições, que foram inferidas utilizando modelos de aprendizagem não supervisionada, visão computacional, reconhecimento ótico de caracteres (OCR) para leitura de arquivos PDF e algoritmos de clusterização. Para essas inferências foram utilizadas as informações disponibilizadas pela próprio Câmara (arquivos PDF das proposições e palavras chave de cada proposição) e o *fastText*[2] uma ferramenta de código aberto que transforma texto em representações vetoriais, possibilitando o uso de modelos de aprendizado de máquina. Além disso, todos os passos para a aquisição dos dados, tratamento, pré-processamento, enriquecimento e montagem da base relacional serão descritos, possibilitando o uso dessa ferramenta para outros períodos de funcionamento da Câmara.

3.1 Organização do trabalho

Inicialmente foi desenvolvido um cronograma por etapas que serão detalhadas no capítulo de Metodologia e desenvolvimento (Figura 4.1).

[1] <http://camara.leg.br>
[2] <https://fasttext.cc>

Figura 3.1: *Cronograma visual de etapas do desenvolvimento deste trabalho.*



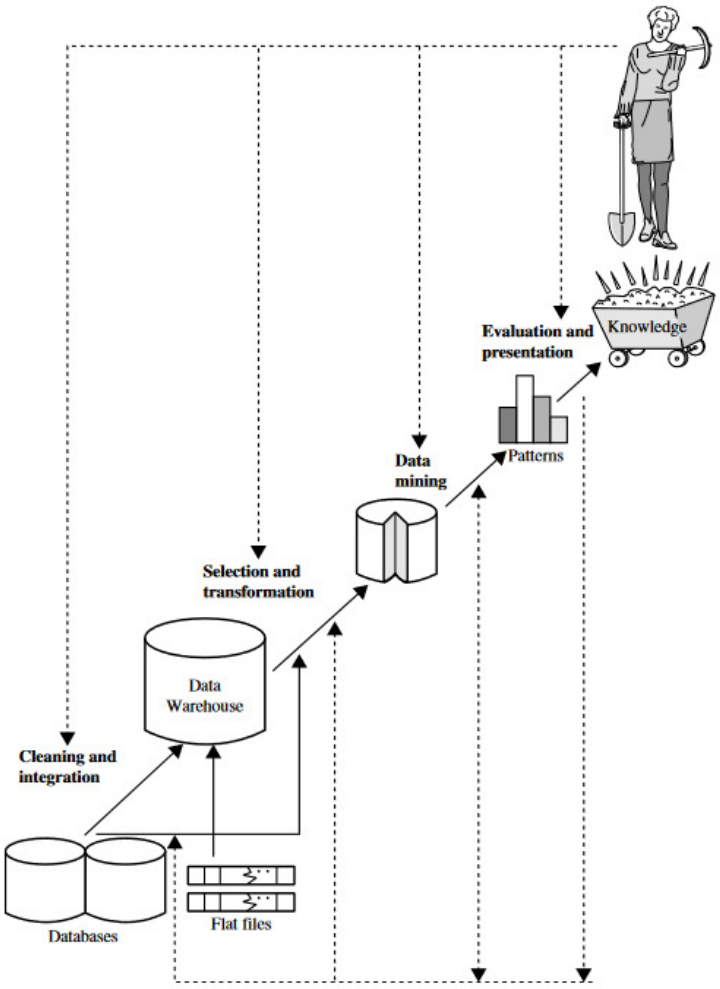
Fonte: Elaborado pelo autor

A primeira etapa é referente à pesquisa bibliográfica responsável por encontrar referências necessárias para o embasamento teórico, já as demais são partes do processo de criação de um modelo de mineração de dados, que foram transformadas em etapas de um projeto KanBan.

HAN J.; KAMBER (2012) sugerem cinco etapas principais o desenvolvimento de trabalhos de mineração de dados:

- Entendimento do problema: Compreender os objetivos do projeto e os requisitos necessários para serem trabalhados.
- Entendimento dos dados: Coletar, explorar e entender os dados disponíveis, atestando sua qualidade e relevância para o problema.
- Preparação dos dados: Limpar, transformar e integrar os dados para torná-los adequados para a modelagem.
- Modelagem: Selecionar e aplicar algoritmos de mineração de dados para descobrir padrões e construir modelos preditivos ou descritivos.
- Avaliação: Avaliar a qualidade e a utilidade dos modelos gerados em relação aos objetivos de negócio.

Figura 3.2: *Ilustração das etapas de trabalhos de mineração de dados.*



Fonte: HAN J.; KAMBER (2012)

As primeiras três etapas se baseiam na aquisição e **pré-processamento dos dados**. É nessas etapas onde os dados são consultados, limpos, normalizados e transformados de acordo com a necessidade do problema.

Nas duas últimas etapas são utilizados esses dados para **construção de modelos de aprendizado e/ou análise de padrões** (em sua maioria estatística).

Essas etapas não são independentes, isso quer dizer que na etapa de **modelagem**, por exemplo, pode ser necessário que algum dado seja revisado e **preparado** novamente. Assim como a **avaliação** do artefato gerado pelo trabalho pode resultar em alterações na **modelagem** ou no **entendimento dos dados**.

É importante salientar que essas etapas não são uma regra e podem variar de acordo com trabalhos ou autores, mas em geral todos os trabalhos de mineração de dados seguem um padrão semelhante ao citado.

Foram feitos acompanhamentos com o orientador e um desenvolvimento incremental de todos os entregáveis de cada etapa.

3.2 Pesquisa bibliográfica

A primeira etapa foi uma profunda pesquisa bibliográfica de trabalhos relacionados a utilização dos dados abertos da Câmara dos Deputados[3] em análise de dados, produção de relatórios sobre gastos e trabalhos, confecção de modelos de aprendizagem de máquina e também trabalhos relacionados a dados políticos no exterior, onde os trabalhos mais relevantes foram explorados na seção de "Trabalhos Relacionados". Nesta pesquisa foi possível constatar que o uso de dados públicos se mostra muito importante na nossa nação, pois temos dados em abundância e pouca visibilidade e extração dessas informações para a sociedade civil.

Foram encontrados diversos trabalhos de grande qualidade sobre democracia digital, análise de sentimento de textos da rede social X[4] (antigo Twitter) sobre política, polarização de deputados regentes do Reino Unido dentre outros.

Esses trabalhos ajudaram a refinar este projeto, oferecendo ideias sobre implementação, dados, métricas e, crucialmente, identificando problemas sociais que esta pesquisa pode solucionar. Todos os textos analisados convergem na necessidade de democratizar informações para uma sociedade democrática e fortalecer a representatividade.

A pesquisa bibliográfica se estendeu por quase todo o desenvolvimento deste trabalho, porém com mais ênfase nos primeiros meses do desenvolvimento.

3.3 Análise das atividades da câmara

3.3.1 Estudo do funcionamento de proposições e da legislação

Foi feita uma análise de como funcionam as atividades da Câmara dos Deputados do Brasil e quais são os significados de termos utilizados nessas atividades. O principal método de pesquisa dessas atividades foi o Portal da Câmara dos Deputados, onde é possível ver postagens das atividades que estão ocorrendo na Câmara além de filtrar propostas legislativas, bancadas de deputados, temas abordados na Câmara dentre outras informações.

Nessa análise das atividades da câmara o assunto escolhido para este trabalho foram as proposições. Segundo o Regimento Interno da Câmara dos Deputados, proposição é toda matéria sujeita à deliberação da Câmara. Apesar dessa ampla definição, os tipos de proposição considerados principais, visto que originam as normas descritas no art. 59 da Constituição Federal, são: Propostas de Emenda à Constituição (PEC), Projetos de Lei Complementar (PLP), Projetos de Lei Ordinária (PL), Projetos de Decreto Legislativo (PDC), Projetos de Resolução (PRC) e Medidas Provisó-

[3] <www.camara.leg.br/>
[4] <www.x.com/>

rias (MPV). Há ainda mais tipos de proposição apreciados pela Câmara, tais como: pareceres, emendas, propostas de fiscalização de controle, indicações, etc.

De acordo com as informações disponibilizadas pela Câmara dos Deputados no portal Clique Regimento[5], esta é a explicação de cada um desses tipos de proposição de acordo com o Portal da Câmara dos Deputados:

- Projeto de Lei (PL):
  - O que é: Um PL é a proposta inicial de uma nova lei ou alteração de uma lei existente. Ele pode ser apresentado por deputados, senadores, comissões da Câmara, o Presidente da República ou cidadãos (através de iniciativa popular).
  - Objetivo: Criar, modificar ou revogar leis, abordando uma ampla gama de temas, desde questões sociais e econômicas até políticas públicas e regulamentações setoriais.
  - Tramitação: Passa por várias etapas, incluindo análise em comissões, votação em plenário e, se aprovado, segue para o Senado e, em alguns casos, para sanção presidencial.
- Emenda Constitucional (EMC):
  - O que é: Uma EMC propõe alterar a Constituição Federal, o documento mais importante do sistema jurídico brasileiro. Devido à sua relevância, a aprovação de uma EMC exige um processo mais rigoroso.
  - Objetivo: Modificar dispositivos constitucionais, adaptando-os às necessidades da sociedade ou corrigindo eventuais lacunas ou contradições.
  - Tramitação: Requer aprovação em dois turnos de votação em ambas as Casas do Congresso (Câmara e Senado), com um quórum qualificado de três quintos dos membros.
- Requerimento de Direito de Fiscalização e Controle (RDF):
  - O que é: Um RDF é um instrumento utilizado pelos deputados para exercer sua função de fiscalizar o Poder Executivo e controlar a aplicação de recursos públicos.
  - Objetivo: Solicitar informações, convocar autoridades para prestar esclarecimentos, realizar audiências públicas e investigar denúncias de irregularidades.
  - Tramitação: Depende do tipo de requerimento e do objeto da fiscalização, podendo envolver a análise em comissões e a votação em plenário.
- Substitutivo (SBT):
  - O que é: Um SBT é uma proposta alternativa apresentada durante a tramitação de um projeto de lei, substituindo integral ou parcialmente o texto original.
  - Objetivo: Aperfeiçoar o projeto original, incorporar sugestões de outros parlamentares ou apresentar uma visão diferente sobre o tema.

[5] <https://educacaoadistancia.camara.leg.br/clique_regimento/>

  - Tramitação: Se aprovado, o substitutivo passa a ser o texto principal em discussão, seguindo as demais etapas do processo legislativo.

Essa escolha foi feita para simplificar a análise dos textos, pois são documentos sobre projetos de lei, emendas na constituição, versões finais de textos que foram aprovados e substituições integrais de projetos de lei.

3.3.2 Estudo das APIs da Câmara dos Deputados

Sabendo os dados que seriam necessários para progredir com o trabalho, foi necessário buscar formas práticas de capturar esses dados. Para isso foi utilizada a API de Dados Abertos da Câmara dos Deputados[6], que disponibiliza uma API HTTP e uma série de arquivos em diversos formatos contendo os dados dos mais diversos trâmites da Câmara dos Deputados.

Para a confecção do banco de dados deste trabalho foram necessários os dados das seguintes entidades:

- Partido
- Deputado
- Proposições

Para as informações dos partidos políticos foi utilizado a API HTTP GET <https://dadosabertos.camara.leg.br/api/v2/partidos>. Essa API disponibiliza diversos filtros, mas no contexto desse trabalho foi utilizado apenas o filtro de idLegislatura. Esse filtro foi utilizado para trazer apenas os partidos que foram relevantes nas duas últimas legislaturas.

Legislaturas são os períodos de quatro anos em que os deputados federais exercem seus mandatos. Cada legislatura é dividida em quatro sessões legislativas anuais, durante as quais os deputados elaboram leis, fiscalizam o poder executivo e debatem questões de interesse nacional. As legislaturas são numeradas sequencialmente desde a primeira, que se iniciou em 1826.

Para as informações dos deputados foram utilizadas as APIs HTTP GET <https://dadosabertos.camara.leg.br/api/v2/deputados> e GET <https://dadosabertos.camara.leg.br/api/v2/deputados/ID>. A primeira API retorna a lista de todos os deputados das determinadas legislaturas, a partir do filtro idLegislatura. Ela não retornou todos os dados do deputado, e por isso foi preciso utilizar a segunda API de filtro individual para preencher essas informações faltantes.

Para as informações das proposições foram utilizadas as seguintes APIs HTTP:

- GET <https://dadosabertos.camara.leg.br/api/v2/proposicoes/ID/autores>
- GET <https://dadosabertos.camara.leg.br/api/v2/referencias/proposicoes/codSituacao>

[6] <https://dadosabertos.camara.leg.br/swagger/api.html>

Além dessas duas APIs foi utilizado o arquivo proposicoes-2023.csv, disponível na URL <https://dadosabertos.camara.leg.br/arquivos/proposicoes/csv/proposicoes-2023.csv>. Este arquivo contém informações sobre todas as proposições do ano de 2023. A primeira API HTTP listada acima retorna os autores de uma proposição e a segunda retorna a descrição dos códigos de situação das proposições. As situações possíveis para as proposições são:

- Aguardando Deliberação
- Aguardando Despacho do Presidente da Câmara dos Deputados (Chancela)
- Retirado pelo Autor
- Tramitando em Conjunto
- Aguardando Análise de Parecer
- Aguardando Vistas
- Pronta para Pauta
- Aguardando Apreciação pelo Senado Federal
- Aguardando Designação - Aguardando Devolução de Relator que deixou de ser Membro
- Aguardando Deliberação de Recurso
- Arquivada
- Aguardando Parecer
- Aguardando Despacho de Arquivamento
- Aguardando Definição Encaminhamento
- Aguardando Despacho do Presidente da Câmara dos Deputados (Análise)
- Aguardando Recurso
- Aguardando Designação de Relator
- Aguardando Encaminhamento
- Aguardando Apensação
- Aguardando Sanção
- Transformado em Norma Jurídica

Todas as requisições e tratamento de arquivos foram feitos utilizando a linguagem de programação Python.

3.4 Análise exploratória

Nesta etapa foram explorados os dados de deputados, partidos e proposições.

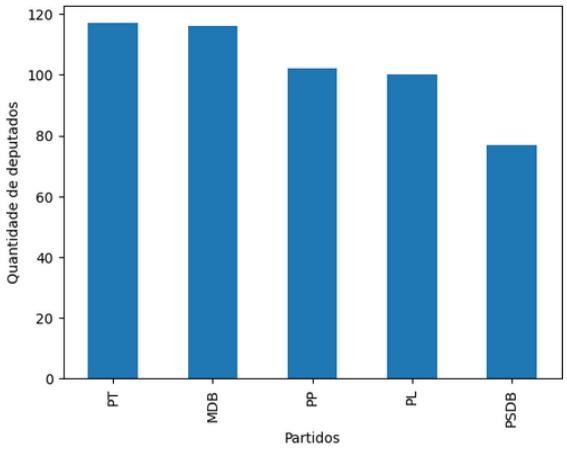
Como o intuito deste trabalho é disponibilizar uma base relacional de dados normalizada contendo os dados de proposições, deputados e partidos interligados de forma que seja simples construir soluções utilizando estes dados.

Vale ressaltar que essa base de dados contém as informações referentes ao ano de 2023.

3.4.1 Partidos e deputados

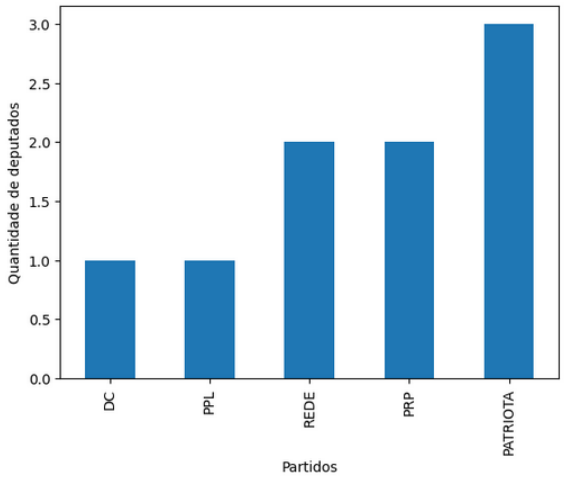
Seguem gráficos com os partidos com mais deputados e menos deputados no ano de 2023 (Figura 4.1 e Figura 4.2):

Figura 3.3: *Partidos com mais deputados no ano de 2023.*



Fonte: Elaborado pelo autor

Figura 3.4: *Partidos com menos deputados no ano de 2023.*



Fonte: Elaborado pelo autor

Em média cada partido tinha 34 deputados no ano de 2023. Podemos utilizar a estratégia de percentis para verificar a de deputados em cada partido. Percentis são medidas de posição que dividem um conjunto de dados ordenados em 100 partes iguais. Um percentil indica a porcentagem de dados que se encontram abaixo de um determinado valor. Por exemplo, o percentil 25 (ou primeiro quartil) representa o valor abaixo do qual se encontram 25% dos dados. Percentis são amplamente utilizados em estatística para analisar dados, permitindo comparar observações individuais com o conjunto de dados como um todo. Eles são especialmente úteis para identificar valores atípicos e entender a distribuição dos dados.

O percentil 75 da quantidade de deputados por partido é aproximadamente 55, e o percentil 25 é aproximadamente 9. Esses números mostram uma variação considerável na quantidade de deputados por partido.

3.4.2 Proposições

Analisando as proposições registradas foi constatado que cerca de 9 mil proposições (dentre as 66 mil consultadas no Portal da Câmara dos Deputados[7]) não contêm autores relacionados.

Nessa análise também foi constatado que apenas cerca de 4 mil das proposições consultadas contêm o campo *keywords* e apenas 212 contêm o campo *ementaDetalhada* preenchido.

Na análise dos dados disponibilizados pelo Portal da Câmara dos Deputados essas foram as *features* (dados a serem analisados nos modelos de aprendizagem de máquina) inicialmente escolhidos para serem trabalhadas, porém após constatar essas informações foi preciso buscar outras alternativas.

Nessa busca foi encontrado o campo *uriInteiroTeor*, que está preenchido para 59 mil dentre as 60 mil proposições consultadas. Esse campo contém um link para o arquivo PDF (Portable Document Format) para a proposição na íntegra. Os campos *uriInteiroTeor* juntamente com o campo *keywords* foram usados para desenvolver os modelos descritos posteriormente.

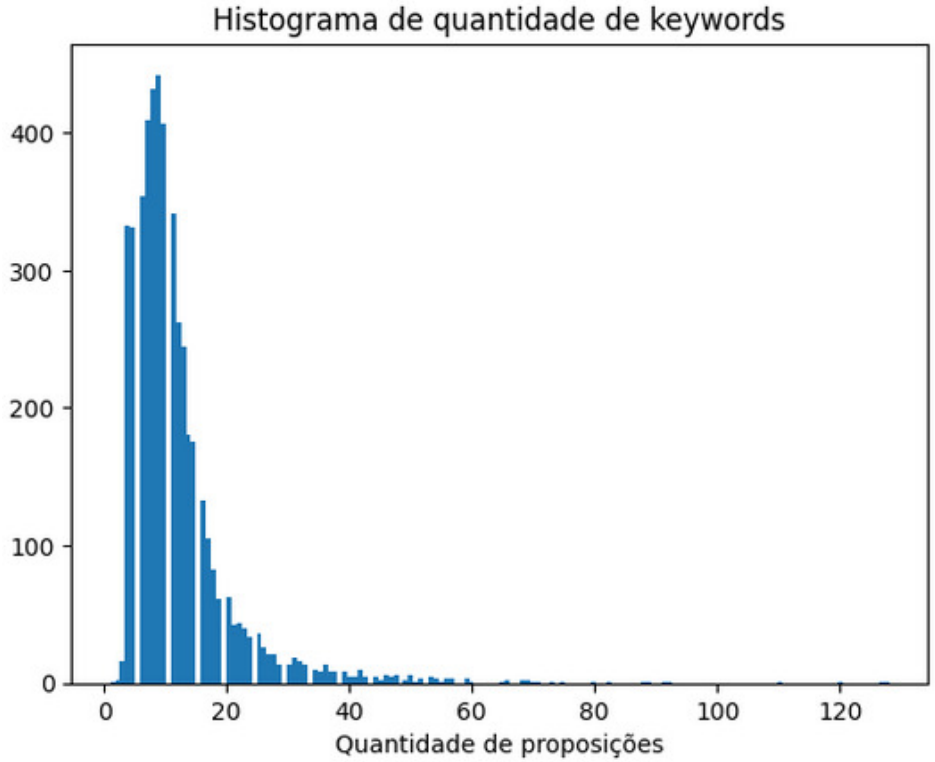
Em média, as proposições consultadas contém 12 termos no campo keywords.

O percentil 99 da quantidade de termos no campo *keywords* para as proposições consultadas é de 50, enquanto o percentil 1 é de 4.

Também é possível visualizar essa distribuição em um histograma (Figura 4.2):

[7]<www.camara.leg.br/>

Figura 3.5: *Histograma de quantidade de keywords.*



Fonte: Elaborado pelo autor

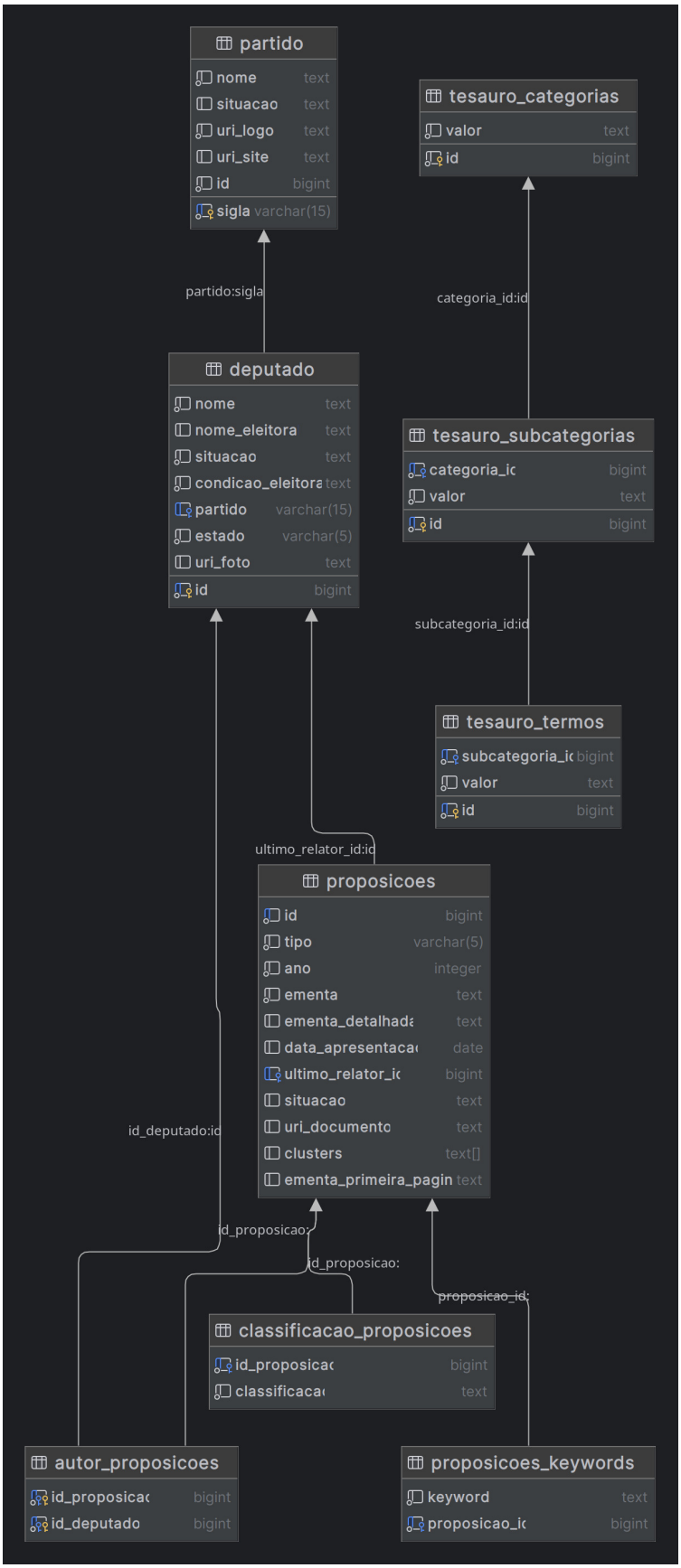
Nele podemos ver que boa parte das proposições têm entre quatro e vinte termos no campo keywords. Seguem alguns exemplos dos termos contidos neste campo:

- 'alteracao', 'lei do fundo nacional de seguranca publica', 'treinamento', 'qualificacao', 'agente de transito'
- 'alteracao', 'lei dos crimes ambientais', 'crime ambiental', 'aumento da pena', 'pesquisa', 'lavra mineral', 'extracao mineral', 'ausencia', 'autorizacao', 'permissao de lavra garimpeira _ pena', 'prazo em dobro', 'garimpo ilegal', 'terras indigenas', 'risco sanitario', 'perigo para a vida de outrem', 'dano ambiental', 'utilizacao', 'maquina', 'equipamento', 'arma de fogo'
- 'comissao de agricultura', 'pecuaria', 'abastecimento e desenvolvimento rural (capadr)', 'fiscalizacao', 'omissao', 'governo federal', 'invasao de propriedade', 'propriedade rural', 'movimento social', 'movimento dos trabalhadores rurais sem-terra (mst)'

3.5 Definição de modelos conceituais de banco de dados

Na etapa de definição de modelo conceitual foi feito o tratamento de todas as informações citadas anteriormente e a criação do modelo do banco de dados. Foi utilizado o banco de dados PostgreSQL para esse trabalho com o seguinte modelo entidade-relacionamento (Figura 4.3):

Figura 3.6: *Modelo de banco de dados desenvolvido no trabalho.*



Fonte: Elaborado pelo autor

As tabelas *tesauro_termos*, *tesauro_subcategorias* e *tesauro_categorias* não foram utilizadas. Nessa etapa todos os dados foram persistidos no banco de dados e após isso foi gerado um arquivo *SQL* da definição da base de dados, que é uma cópia das informações armazenadas no banco de dados. Junto com este arquivo foram disponibilizados arquivos de configuração do Docker que possibilitam executar o banco de dados e o código responsável por gerar os modelos e capturar os dados. Este banco de dados, junto com o modelo de análise de documentos apresentado anteriormente e os modelos de aprendizagem de máquina que serão utilizados posteriormente são as contribuições esperadas para este trabalho.

3.6 Aquisição de base

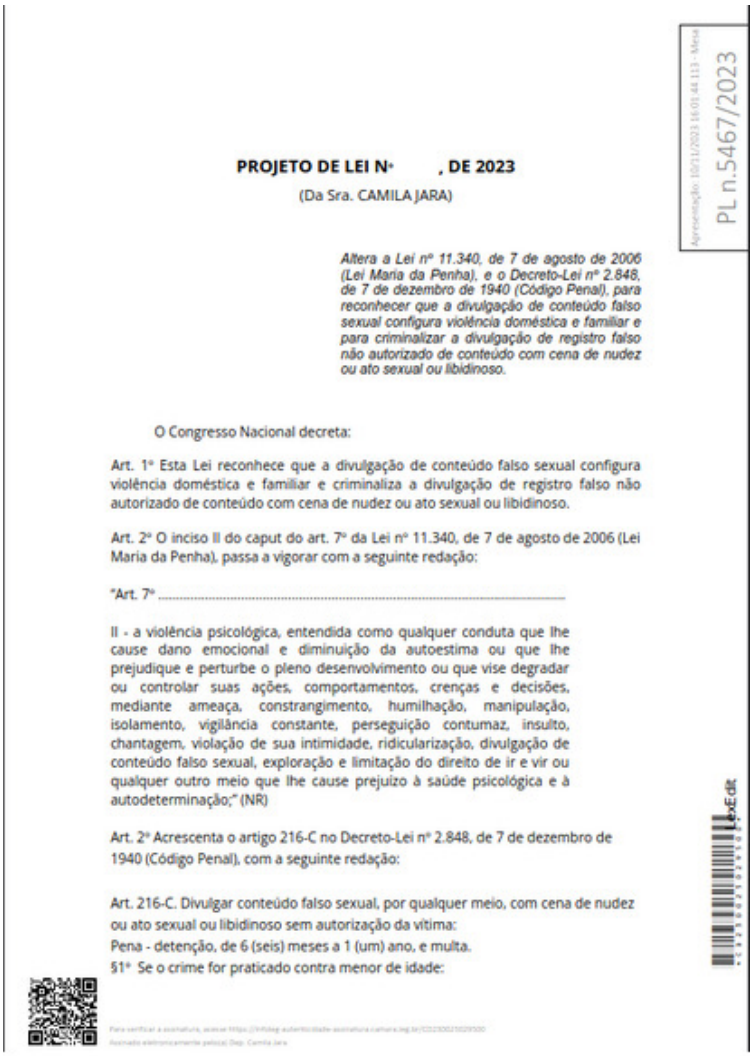
Para agilizar a aquisição de dados foram utilizadas estratégias para consultar as informações do Portal da Câmara dos Deputados de forma paralela. Essa abordagem permite a divisão das tarefas em *threads* concorrentes, que são executadas simultaneamente em múltiplos núcleos de processamento. Com isso, busca-se reduzir o tempo total de execução e aumentar a eficiência do sistema, aproveitando ao máximo os recursos computacionais disponíveis. A biblioteca de paralelismo nativa da linguagem de programação Python (multiprocessing) foi utilizada para gerenciar a criação, sincronização e comunicação entre as *threads*.

Foi utilizada a linguagem de programação Python e a biblioteca psycopg2, responsável por fazer a conexão com o PostgreSQL.

3.7 Pré processamento

Como mencionado na Análise exploratória, um dos dados que estavam inconsistentes na API de proposições era o campo *ementaDetalhada*, que não estava preenchido na maioria das proposições. Para preencher a informação do texto das proposições analisadas foi utilizado o campo *uri_documento*, que estava presente em todas as proposições analisadas. Esse campo contém a URL para um arquivo PDF como este: (Figura 4.4)

Figura 3.7: *Projeto de lei capturado pela API da Câmara dos Deputados.*



PL n.5467/2023

PROJETO DE LEI Nº , DE 2023

(Da Sra. CAMILA JARA)

Altera a Lei nº 11.340, de 7 de agosto de 2006 (Lei Maria da Penha), e o Decreto-Lei nº 2.848, de 7 de dezembro de 1940 (Código Penal), para reconhecer que a divulgação de conteúdo falso sexual configura violência doméstica e familiar e para criminalizar a divulgação de registro falso não autorizado de conteúdo com cena de nudez ou ato sexual ou libidinoso.

O Congresso Nacional decreta:

Art. 1º Esta Lei reconhece que a divulgação de conteúdo falso sexual configura violência doméstica e familiar e criminaliza a divulgação de registro falso não autorizado de conteúdo com cena de nudez ou ato sexual ou libidinoso.

Art. 2º O inciso II do caput do art. 7º da Lei nº 11.340, de 7 de agosto de 2006 (Lei Maria da Penha), passa a vigorar com a seguinte redação:

"Art. 7º .....

II - a violência psicológica, entendida como qualquer conduta que lhe cause dano emocional e diminuição da autoestima ou que lhe prejudique e perturbe o pleno desenvolvimento ou que vise degradar ou controlar suas ações, comportamentos, crenças e decisões, mediante ameaça, constrangimento, humilhação, manipulação, isolamento, vigilância constante, perseguição contumaz, insulto, chantagem, violação de sua intimidade, ridicularização, divulgação de conteúdo falso sexual, exploração e limitação do direito de ir e vir ou qualquer outro meio que lhe cause prejuízo à saúde psicológica e à autodeterminação;" (NR)

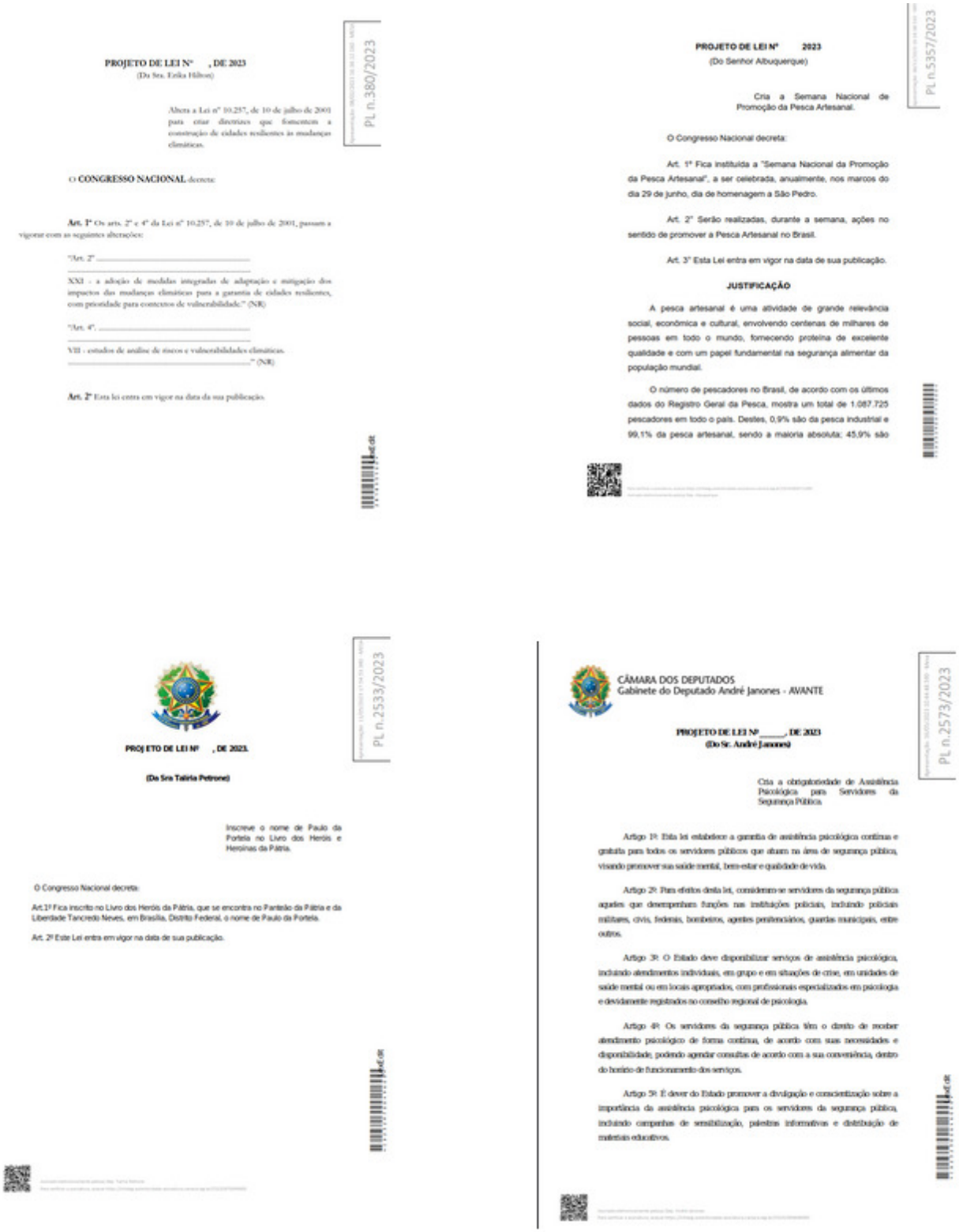
Art. 2º Acrescenta o artigo 216-C no Decreto-Lei nº 2.848, de 7 de dezembro de 1940 (Código Penal), com a seguinte redação:

Art. 216-C. Divulgar conteúdo falso sexual, por qualquer meio, com cena de nudez ou ato sexual ou libidinoso sem autorização da vítima:
Pena - detenção, de 6 (seis) meses a 1 (um) ano, e multa.
§1º Se o crime for praticado contra menor de idade:

Disponível em <https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2358932>. Acesso em 06 de agosto de 2024.

Analisando esses arquivos foi possível encontrar um padrão onde em todas as primeiras páginas contêm um resumo do documento: (Figura 4.5)

Figura 3.8: *Exemplo de primeiras páginas de projetos de lei capturados na API da Câmara dos Deputados.*



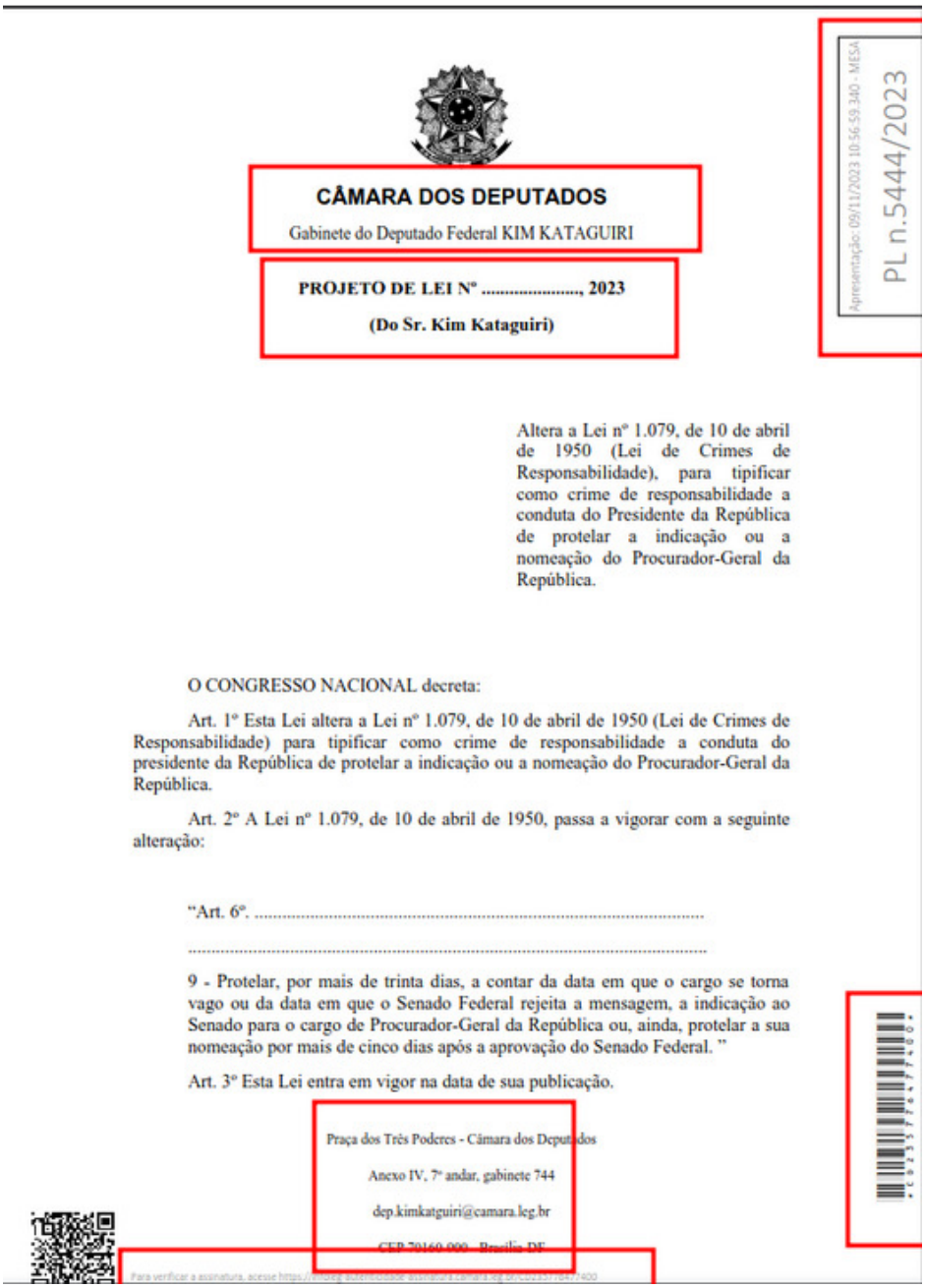
Disponível em <https://dadosabertos.camara.leg.br/swagger/api.html>. Acesso em 06 de agosto de 2024.

Tendo essa informação, foi decidido que o texto analisado neste trabalho será o conteúdo da primeira página de cada proposição dos tipos selecionados. Primeiro foi feito o download dos

arquivos PDF de todas as proposições analisadas (também de forma paralela utilizando multi-threading). Após isso foi utilizada a biblioteca *pdf2image* para transformar um arquivo PDF em vários arquivos de imagem em PNG para cada página do texto integral da proposição.

Com os arquivos de cada página disponíveis foi utilizada visão computacional para extrair o texto que descreve a proposição da primeira página. Houve uma tentativa de capturar todo o texto da primeira página da proposição sem nenhum tratamento, porém pela natureza do arquivo essa estratégia se mostrou imprecisa pois o documento contém muito texto repetido, códigos de barra, identificadores de processos e metadados do documento que não são úteis para a análise feita neste trabalho.

Figura 3.9: *Exemplo de Projeto de Lei capturado pela API da Câmara dos Deputados.*



CÂMARA DOS DEPUTADOS

Gabinete do Deputado Federal KIM KATAGUIRI

PROJETO DE LEI Nº ....., 2023

(Do Sr. Kim Kataguiri)

Apresentação: 09/11/2023 10:56:59.340 - MESA

PL n.5444/2023

Altera a Lei nº 1.079, de 10 de abril de 1950 (Lei de Crimes de Responsabilidade), para tipificar como crime de responsabilidade a conduta do Presidente da República de protelar a indicação ou a nomeação do Procurador-Geral da República.

O CONGRESSO NACIONAL decreta:

Art. 1º Esta Lei altera a Lei nº 1.079, de 10 de abril de 1950 (Lei de Crimes de Responsabilidade) para tipificar como crime de responsabilidade a conduta do presidente da República de protelar a indicação ou a nomeação do Procurador-Geral da República.

Art. 2º A Lei nº 1.079, de 10 de abril de 1950, passa a vigorar com a seguinte alteração:

"Art. 6º. .....

.....

9 - Protelar, por mais de trinta dias, a contar da data em que o cargo se torna vago ou da data em que o Senado Federal rejeita a mensagem, a indicação ao Senado para o cargo de Procurador-Geral da República ou, ainda, protelar a sua nomeação por mais de cinco dias após a aprovação do Senado Federal. "

Art. 3º Esta Lei entra em vigor na data de sua publicação.

Praça dos Três Poderes - Câmara dos Deputados

Anexo IV, 7º andar, gabinete 744

dep.kimkatguiri@camara.leg.br

CEP 70160-900 - Brasília-DF

Disponível em <https://dadosabertos.camara.leg.br/swagger/api.html>. Acesso em 06 de agosto de 2024.

Como pode ser visto na imagem acima (Figura 4.6), todas as áreas grifadas contêm informações irrelevantes para a análise do conteúdo daquela proposição.

Para extrair as informações apenas da área central do documento (onde existem informações mais úteis para nossa análise) foi utilizada a biblioteca *opencv-python*.

Visão computacional, segundo (BALLARD, 1982) pode ser definida como a construção de modelos computacionais explícitos e significativos de processos de percepção visual. Ela é um campo híbrido que interage com inteligência artificial, processamento digital de sinais e ótica, utilizado para lidar com a complexidade visual do mundo em sistemas computacionais. Primeiramente todo o arquivo foi re-colorido em escala de cinza, para facilitar o processamento do texto presente na imagem. Isso é uma prática comum em aplicações de visão computacional para análise de texto, pois imagens em escala de cinza possuem apenas um canal de cor (luminância), enquanto imagens coloridas possuem três (vermelho, verde e azul). Esta redução na quantidade de informações facilita o processamento da imagem pelo algoritmo, tornando-o mais rápido e eficiente, além de não resultar em perda de precisão na análise das imagens pois o nosso objetivo não está relacionado a distribuição cores das mesmas. A conversão para escala de cinza também pode ajudar a aumentar o contraste entre o texto e o fundo da imagem, tornando os caracteres mais nítidos e fáceis de identificar. Isso é especialmente útil em imagens com cores vibrantes ou fundos complexos, onde o texto pode se perder visualmente. Após isso foi aplicado um filtro gaussiano (na classe *GaussianBlur* da biblioteca *open-cv*) para reduzir possíveis ruídos das imagens analisadas. A suavização da imagem esse filtro é amplamente utilizada na etapa de pré-processamento para operações de visão computacional, como detecção de bordas, segmentação de imagem e reconhecimento de objetos (que serão os próximos passos dessa etapa). Com a imagem pré-processada foi utilizado o método *cv2.findContours* para encontrar áreas dentro da imagem que contêm informações. Segue o trecho de código que faz essa análise:

Figura 3.10: *Trecho de código do pré-processamento das imagens.*

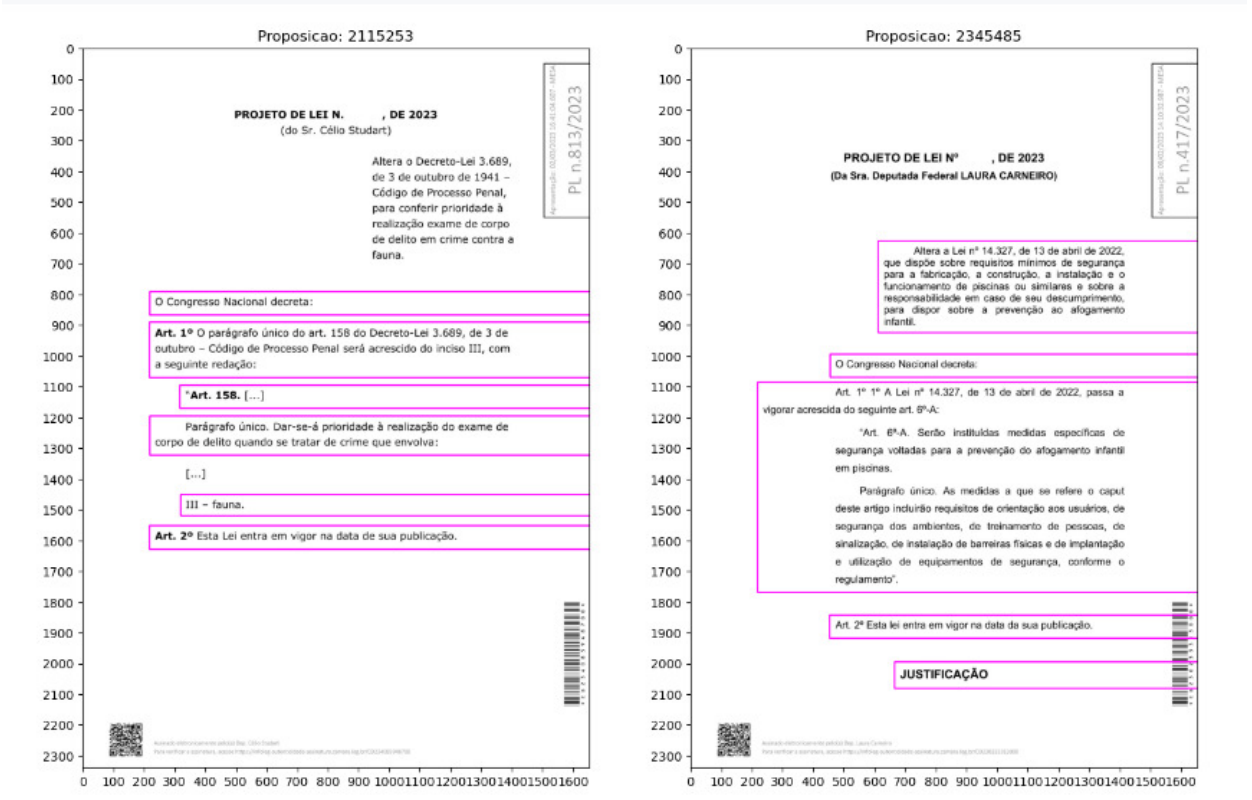
```
im = cv2.imread(image_path)

gray = cv2.cvtColor(im, cv2.COLOR_BGR2GRAY) blur = cv2.GaussianBlur(gray, (9, 9), 0)
thresh = cv2.adaptiveThreshold(blur, 255, cv2.ADAPTIVE_THRESH_GAUSSIAN_C, cv2.THRESH_BINARY_INV, 11,
30)

kernel = cv2.getStructuringElement(cv2.MORPH_RECT, (12, 12))
dilate = cv2.dilate(thresh, kernel, iterations=4)

cnts = cv2.findContours(dilate, cv2.RETR_LIST, cv2.CHAIN_APPROX_SIMPLE)
cnts = cnts[0] if len(cnts) == 2 else cnts[1]
```

Fonte: Elaborado pelo autor.

A sexta linha do código acima aplica a operação de dilatação na imagem binarizada thresh utilizando o kernel criado anteriormente. A dilatação expande os objetos brancos na imagem, tornando-os mais espessos. O parâmetro iterations=4 indica que a dilatação será repetida quatro vezes.

A sétima e oitava linhas de código utilizam a imagem dilatada para encontrar áreas retangulares que contêm informações destoantes dos arredores (no nosso caso, são áreas retangulares com caracteres). Segue exemplos de algumas proposições que passaram por esse algoritmo: (Figura 4.8)

Figura 3.11: *Exemplo de Projeto de Lei capturado pela API da Câmara dos Deputados processados.*



Fonte: Elaborado pelo autor.

Foi gasto em média cerca de quatro segundos para realizar esse processo para cada página analisada utilizando um computador pessoal com especificações medianas.

Com as áreas já delimitadas foi utilizado reconhecimento ótico de caracteres (OCR) para extrair as palavras das áreas desejadas da imagem. O reconhecimento ótico de caracteres é uma área que caminha junto da visão computacional responsável pela conversão de imagens de texto impresso ou manuscrito em cadeias de caracteres compreensíveis por uma máquina. Jain e Duin

(2004) definem o OCR como uma subárea da visão computacional que se concentra na extração de textos de imagens e documentos digitalizados.

Após a extração do texto da imagem foram removidas todas as palavras com pouco (ou nenhum) valor semântico, palavras essas conhecidas como stop words. Segundo RAJARAMAN e ULLMAN (RAJARAMAN A.; ULLMAN, 2012) stop words são palavras comuns em um idioma que, geralmente, carregam pouco significado relevante para a análise de dados textuais. Palavras como "o", "a", "e", "de", "que", entre outras, são exemplos típicos em português. A remoção delas é uma etapa fundamental de pré-processamento em tarefas de mineração de dados que envolvem análise de texto, como:

- Redução de dimensionalidade: Diminuir a quantidade de termos a serem processados
- Melhora da relevância: Permitir que algoritmos se concentrem em palavras com mais significância, melhorando os resultados.

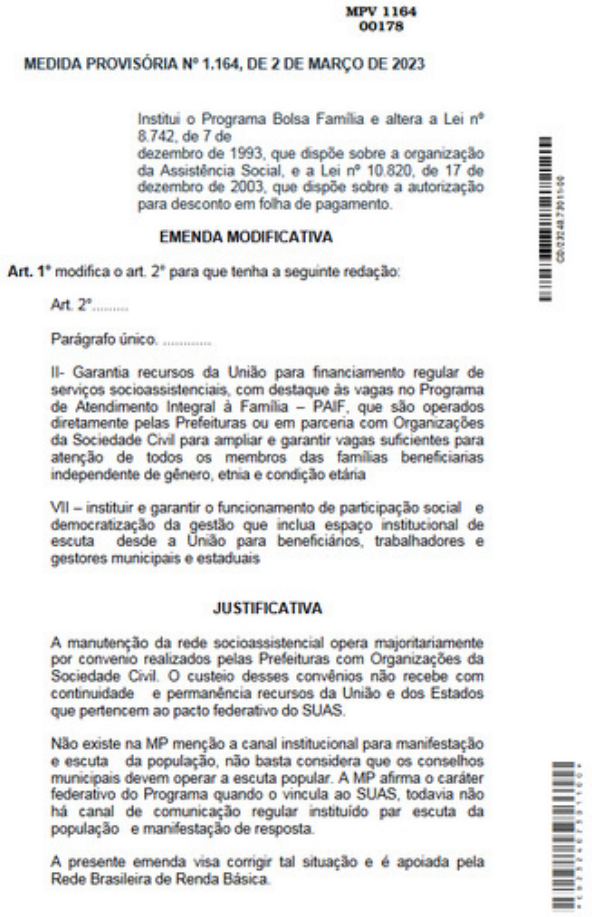
A lista de stop words varia de acordo com a língua analisada, que no contexto deste trabalho é o portugues. Foi utilizada a lista de stop words disponibilizada pela biblioteca nltk (Natural Language Toolkit), amplamente utilizada pela comunidade científica no momento do desenvolvimento deste trabalho. Segue uma lista das primeiras 50 stop words consideradas neste trabalho:

[a, à, ao, aos, aquela, aquelas, aquele, aqueles, aquilo, as, às, até, com, como, da, das, de, dela, delas, dele, deles, depois, do, dos, e, é, ela, elas, ele, eles, em, entre, era, eram, éramos, essa, essas, esse, esses, esta, está, estamos, estão, estar, estas, estava, estavam, estávamos, este, 'esteja']

Exemplo de stop words da biblioteca NLTK. Disponível em <https://www.nltk.org/>. Acesso em 06 de agosto de 2024.

Para que o processamento das 4000 proposições que continham uma referência ao arquivo fossem processadas apenas uma vez, todo o resultado da extração de texto foi salvo em uma planilha .CSV que foi utilizada nas outras etapas. Segue um exemplo da primeira página de uma proposição e o texto extraído pelo algoritmo (Figura 4.9):

Figura 3.12: *Exemplo de Projeto de Lei capturado pela API da Câmara dos Deputados processados.*



MPV 1164
00178

MEDIDA PROVISÓRIA Nº 1.164, DE 2 DE MARÇO DE 2023

Institui o Programa Bolsa Família e altera a Lei nº 8.742, de 7 de dezembro de 1993, que dispõe sobre a organização da Assistência Social, e a Lei nº 10.820, de 17 de dezembro de 2003, que dispõe sobre a autorização para desconto em folha de pagamento.

EMENDA MODIFICATIVA

Art. 1º modifica o art. 2º para que tenha a seguinte redação:

Art. 2º .....

Parágrafo único. .....

II- Garantia recursos da União para financiamento regular de serviços socioassistenciais, com destaque às vagas no Programa de Atendimento Integral à Família – PAIF, que são operados diretamente pelas Prefeituras ou em parceria com Organizações da Sociedade Civil para ampliar e garantir vagas suficientes para atenção de todos os membros das famílias beneficiarias independente de gênero, etnia e condição etária

VII – instituir e garantir o funcionamento de participação social e democratização da gestão que inclua espaço institucional de escuta desde a União para beneficiários, trabalhadores e gestores municipais e estaduais

JUSTIFICATIVA

A manutenção da rede socioassistencial opera majoritariamente por convenio realizados pelas Prefeituras com Organizações da Sociedade Civil. O custeio desses convênios não recebe com continuidade e permanência recursos da União e dos Estados que pertencem ao pacto federativo do SUAS.

Não existe na MP menção a canal institucional para manifestação e escuta da população, não basta considera que os conselhos municipais devem operar a escuta popular. A MP afirma o caráter federativo do Programa quando o vincula ao SUAS, todavia não há canal de comunicação regular instituído par escuta da população e manifestação de resposta.

A presente emenda visa corrigir tal situação e é apoiada pela Rede Brasileira de Renda Básica.

Disponível em <https://dadosabertos.camara.leg.br/swagger/api.html>. Acesso em 06 de agosto de 2024.

“A presente emenda visa corrigir tal situação apoiada Rede Brasileira Renda Básica Não existe MP menção canal institucional manifestação escuta população basta considera conselhos municipais devem operar escuta popular AÀ MP afirma caráter º federativo Programa vincula SUAS todavia canal comunicação regular instituído par escuta população manifestação resposta A manutenção rede socioassistencial opera majoritariamente convenio realizados Prefeituras Organizações Sociedade Civil O custeio desses convênios recebe continuidade permanência recursos União Estados pertencem pacto federativo SUAS JUSTIFICATIVA VII instituir garantir funcionamento participação social democratização gestão inclua espaço institucional escuta desde União beneficiários trabalhadores gestores municipais estaduais Garantia recursos União financiamento regular serviços socioassistenciais destaque vagas Programa Atendimento Integral Família PAIF operados diretamente Prefeituras parceria Organizações Sociedade Civil ampliar garantir vagas suficientes atenção todos membros famílias beneficiarias independente gênero etnia condição etária Parágrafo único Art Y Art º modifica art º seguinte redação É EMENDA MODIFICATIVA É ã”

3.8 Mineração de dados

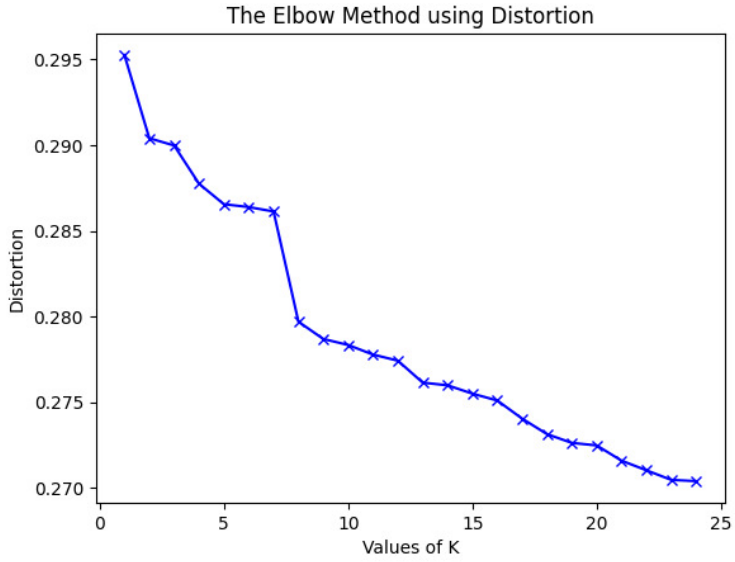
Com a base de dados populada e curada foi realizada a confecção de dois modelos de aprendizagem de máquina não supervisionada.

3.8.1 Agrupamento de keywords

Utilizando aprendizagem não supervisionada foi feito o agrupamento das keywords presentes na base de dados em *clusters*. Neste agrupamento as *features* foram todos os itens diferentes presentes no campo *keywords* de todas as proposições dos tipos 'PL', 'EMC', 'RDF', 'SBT', 'EMP' e 'PEC'. Para essas proposições foram obtidos 9743 termos distintos. Dentre estes termos foram removidos os 5% que mais apareciam e menos apareciam. Após isso foram removidas as *stopwords* desses termos e por fim todos eles foram vetorizados utilizando a biblioteca *fasttext*.

O modelo utilizado neste agrupamento foi o KMeans. Foram testados valores de K *clusters* entre 1 e 25 para este modelo, e para metrificar a qualidade destes agrupamentos foi utilizada a métrica de distorção de "Elbow" (Figura 4.10).

Figura 3.13: *Métrica de Elbow para distorção de clusters.*



Fonte: Elaborado pelo autor.

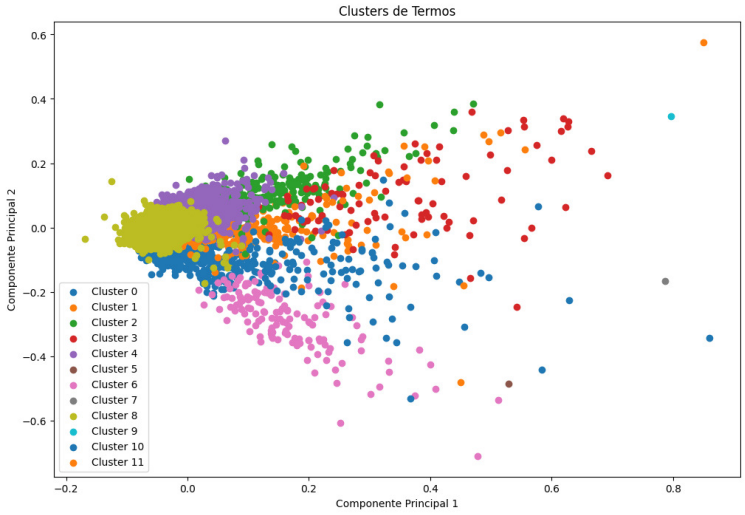
Esta métrica pode ser traduzida de forma direta como "método do ombro utilizando distorção". Ela é uma métrica utilizada para determinar a quantidade de *clusters* mais próxima de ótima para um conjunto de dados, e para isso buscamos um valor de K grande o bastante para que a distorção seja aceitável e que, para uma quantidade de *clusters* N > K a distorção varie tanto (evitando assim o *overfitting*). Este método pode ser considerado muito subjetivo para alguns casos de uso, por isso são necessárias validações além do resultado desta métrica. Após essa análise baseada na distorção

entre os *clusters*, a quantidade escolhida para gerar o modelo de aprendizagem não supervisionada foi 12.

Utilizando a estratégia de KMeans (com a classe KMeans da biblioteca *scipy*), sendo a entrada do modelo um conjunto de 6620 keywords disponibilizadas pela Câmara dos Deputados[8], foi feito o treinamento do modelo, e o mesmo foi exportado para um arquivo da extensão PICKLE (estratégia utilizada no Python para salvar objetos, evitando re-processamento desnecessário). É importante ressaltar que o dado de entrada do modelo (também chamado de feature) é um termo (keyword) vetorizado, que consiste em um vetor de 300 números naturais que representa aquela sentença obtido a partir da utilização do modelo pré-treinado de representação de texto.

Para facilitar a visualização dos resultados deste modelo, foi feita a redução de dimensionalidade desses vetores de 300 para dois, permitindo a visualização deste agrupamento em duas dimensões. Essa redução foi possível utilizando a estratégia de PCA (Principal component analysis ou Análise de Componentes Principais), que utiliza matrizes de covariância, cálculo de autovetores, ordenação de componentes e seleção de componentes para reduzir o conjunto de dados original em um sub-espaço com uma grandeza menor que a original, mantendo a maior parte da informação original. Feita esta redução foi confeccionado um gráfico colorido que representa visualmente a distribuição dos termos analisados nos seus devidos *clusters* (Figura 4.11):

Figura 3.14: *Distribuição de termos disponibilizados pela Câmara dos Deputados.*

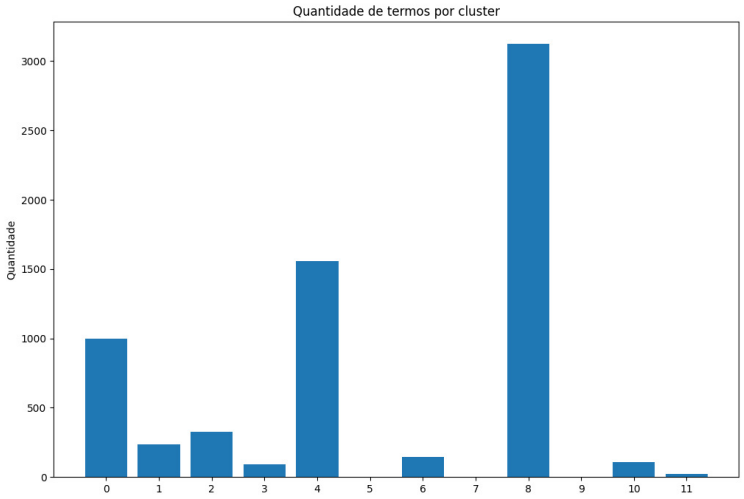


Fonte: Elaborado pelo autor.

Também foi confeccionado um gráfico com a quantidade de termos agrupados em cada cluster (Figura 4.12):

[8]<www.camara.leg.br/>

Figura 3.15: *Quantidade de termos termos disponibilizados pela Câmara dos Deputados em cada cluster.*



Fonte: Elaborado pelo autor.

Visualizando estes gráficos é perceptível que existe uma grande concentração de termos semelhantes no *cluster* 8 e que, mesmo os outros *clusters* mais populosos (0 e 4) estão graficamente próximos do *cluster* 8. Os outros *clusters* parecem "pulverizados" dentro do plano, dificultando que o algoritmo de agrupamento os relacione em grupos mais consistentes. Essas características podem ser melhor entendidas visualizando alguns exemplares de termos designados nos *clusters* gerados. Seguem 25 exemplos de termos agregados no *cluster* superpopulado (*cluster* número 8):

['lei fundo nacional seguranca publica', 'empresa brasileira correios telegrafos ', 'politica nacional longo prazo ' ,'planejamento estrategico','desenvolvimento nacional' 'lei defesa concorrencia ','infracao contra ordem economica' ,'anticompetitividade','protocolo nao nao' ,'selo nao nao - mulheres seguras','perseguicao obsessiva','alistamento militar _proibicao','certificado reservista','certificado dispensa incorporacao','radiodifusao sonora frequencia modulada','defesa direitos animais','atendimento veterinario','legislacao tributaria federal ','profissional educacao fisica','carreira auditoria receita federal brasil','programa transferencia renda','obrigacao tributaria' 'coproprietario','financiamento habitacional','reproducao humana assistida']

25 termos agregados em *cluster* superpopulado de keywords . Fonte: Elaborado pelo autor.

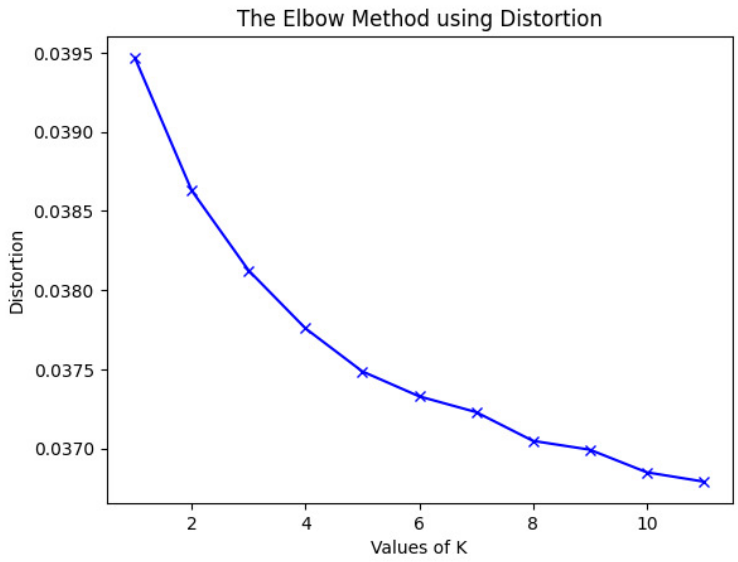
Salvo alguns termos como "defesa direitos animais" e "reprodução humana assistida" todos os outros contêm palavras relacionadas a documentos, legislação, processos legais, dentre outros. O *cluster* 0 também demonstra essa característica, com termos como "simplificação", "carga tributária", "natureza tributária" dentre outros.

Após a análise deste modelo foi necessário buscar outra alternativa para categorizar as proposições, tendo em vista que os resultados obtidos não foram muito satisfatórios.

3.8.2 Agrupamento de proposições

Utilizando da mesma estratégia listada acima, foi realizada uma tentativa de agrupar os textos retirados da primeira página das ementas das proposições analisadas. Assim como o modelo anterior, foi feita a transformação do texto limpo da primeira página da ementa da proposição em um vetor de números naturais utilizando o *fasttext* e foi realizado o agrupamento utilizando o algoritmo de K-Means. Foi utilizado o método de Elbow para definir a quantidade de *clusters* (Figura 4.13):

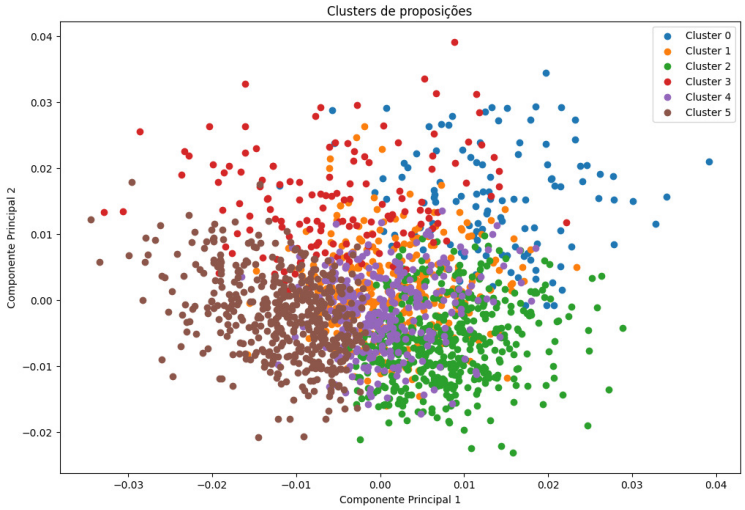
Figura 3.16: *Métrica de Elbow para distorção de clusters.*



Fonte: Elaborado pelo autor.

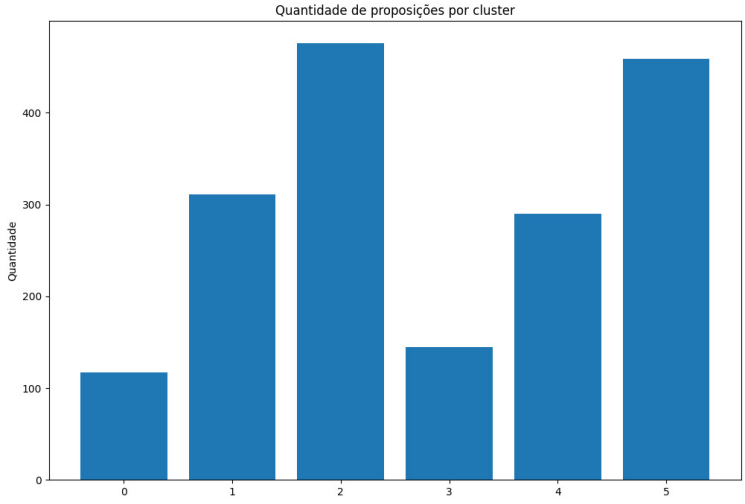
Podemos perceber uma progressão mais suave do que a do agrupamento de *keywords*, o que pode ser interpretado como uma variação mais gradual da distorção e maior facilidade do algoritmo de encontrar similaridades entre as features analisadas. Para este modelo a quantidade de *clusters* escolhida foi de 6. Também foi feita a análise visual da distribuição de proposições entre cada *cluster*(Figuras 4.14 e 4.15):

Figura 3.17: *Distribuição de proposições por cluster.*



Fonte: Elaborado pelo autor.

Figura 3.18: *Quantidade de proposições por cluster.*



Fonte: Elaborado pelo autor.

Esta análise demonstra claramente uma discrepância menor da quantidade de itens em cada subconjunto e, no gráfico de distribuição uma organização mais clara de cada grupo em setores do plano cartesiano (mesmo com algumas sobreposições).

3.8.3 Nomeação dos Clusters

Foi feita a nomeação dos *clusters* encontrados pelo algoritmo de K-Means. Essa nomeação é necessária para evidenciar o as características encontradas pelo algoritmo no conjunto de dados,

pois um número não significa nada quando confrontado com cada proposição analisada neste trabalho. A nomeação de *clusters* é uma etapa delicada da área de aprendizagem de máquina, pois se feita de forma errada pode não descrever o verdadeiro significado daquele agrupamento de entidades, além de estar passível a vieses inerentes do nomeador ou até mesmo descrever apenas um subconjunto do *cluster* nomeado. Para evitar tais falhas e vieses humanos, foi utilizado um modelo de LLM. Estes são modelos baseados em PLN treinados com gigantescas bases de dados capazes de processar textos e elaborar respostas. Neste trabalho foi utilizado o LLM Google Gemini[9] na sua versão Advanced. Os modelos de LLM tem como entrada uma estrutura chamada prompt, que é basicamente um texto que o modelo irá processar e responder da forma adequada. O prompt disponibilizado para o modelo foi:

> *nomeie cada conjunto de proposições a seguir:*
> *conjunto 1: . . .*
> *conjunto 2: . . .*
> *. . .*
> *conjunto X: . . .*
> *esses conjuntos são compostos por ementas de proposições da câmara dos deputados do brasil*
> *quero que o nome do conjunto descreva sobre que áreas da sociedade essas proposições estão relacionadas*
> *quero a resposta em um arquivo JSON onde a chave é o índice do* cluster *e o valor é o nome dado a ele*

Prompt utilizado para nomear os *clusters*. Elaborado pelo autor.

Nas linhas onde estão os conjuntos foram adicionadas 25 proposições aleatórias agrupadas em cada *cluster* agrupado anteriormente. As últimas linhas descrevem qual é a saída desejada, explicando que a resposta deve ser um JSON com os nomes dados a cada *cluster* a partir dos exemplos de proposições disponibilizadas. Esta foi a resposta dada pelo modelo:

> *{"0": "Segurança Pública e Proteção de Vulneráveis", "1": "Economia, Tributos e Direitos do Consumidor", "2": "Saúde, Educação e Cultura", "3": "Administração Pública e Território", "4": "Trabalho, Educação e Cultura", "5": "Saúde, Educação, Cultura e Proteção de Vulneráveis"}*

Resposta disponibilizada pelo Google Gemini. Elaborado pelo autor.

3.9 Aquisição de conhecimento

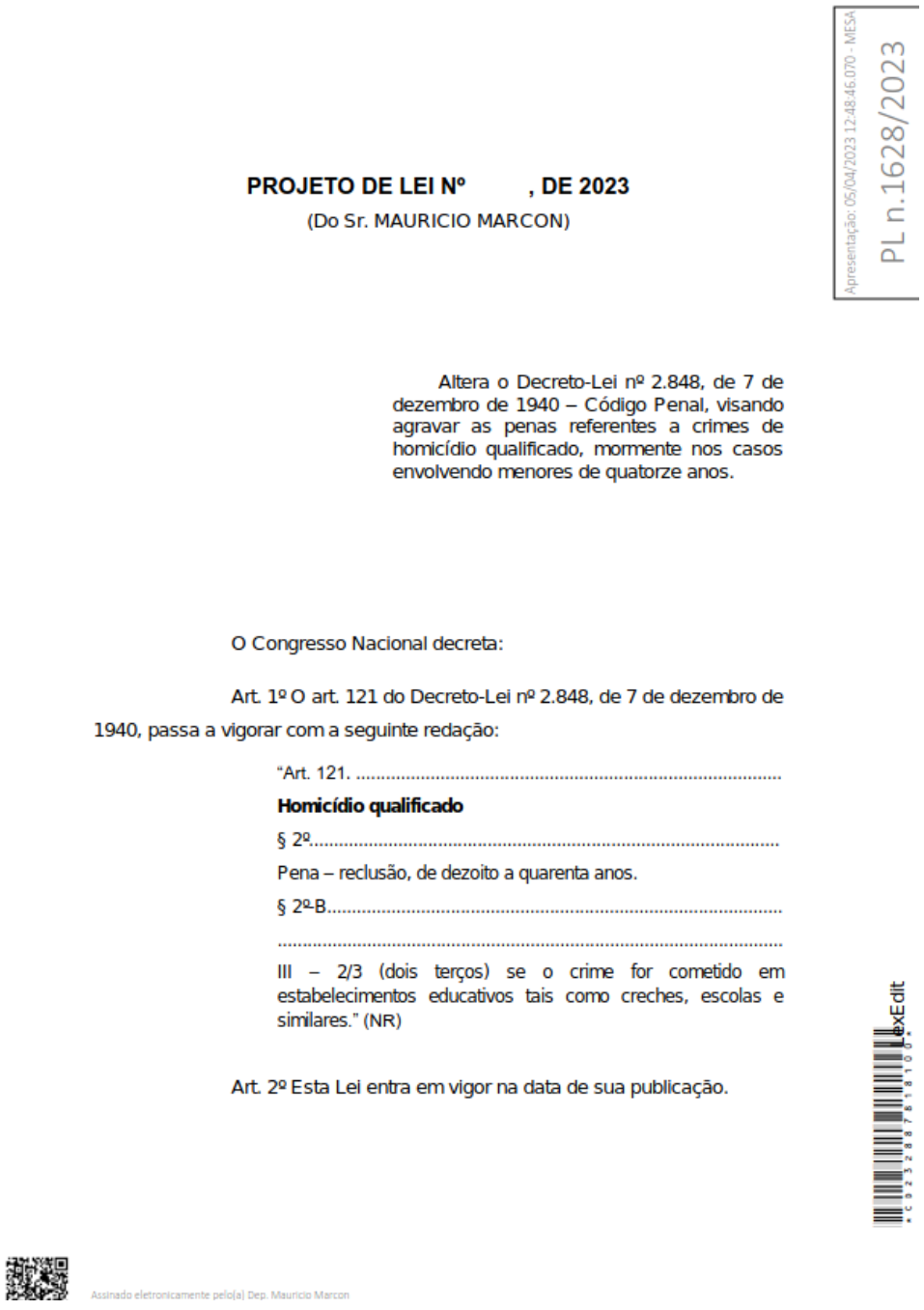
Com os resultados das etapas anteriores a coluna *clusters* da tabela *proposições* foi preenchida com o título dado para o *cluster* daquela respectiva proposição. Essa base de dados foi exportada para um arquivo de definição, o qual pode ser importado facilmente por qualquer banco de dados PostgreSQL com a versão compatível. Além disso, todo o código das etapas anteriores também

[9] <https://gemini.google.com/>

está disponível em um repositório aberto do GitHub presente no anexo, junto com os modelos produzidos na etapa anterior e um arquivo CSV contendo todas as proposições analisadas e o texto presente na primeira página de suas ementas.

Seguem exemplos de textos de algumas proposições e suas respectivas classificações:

Figura 3.19: *Proposição classificada como: Segurança Pública e Proteção de Vulneráveis.*



Apresentação: 05/04/2023 12:48:46.070 - MESA
PL n.1628/2023

PROJETO DE LEI Nº , DE 2023

(Do Sr. MAURICIO MARCON)

Altera o Decreto-Lei nº 2.848, de 7 de dezembro de 1940 – Código Penal, visando agravar as penas referentes a crimes de homicídio qualificado, mormente nos casos envolvendo menores de quatorze anos.

O Congresso Nacional decreta:

Art. 1º O art. 121 do Decreto-Lei nº 2.848, de 7 de dezembro de 1940, passa a vigorar com a seguinte redação:

"Art. 121. .....

Homicídio qualificado

§ 2º.....

Pena – reclusão, de dezoito a quarenta anos.

§ 2º-B.....

.....

III – 2/3 (dois terços) se o crime for cometido em estabelecimentos educativos tais como creches, escolas e similares." (NR)

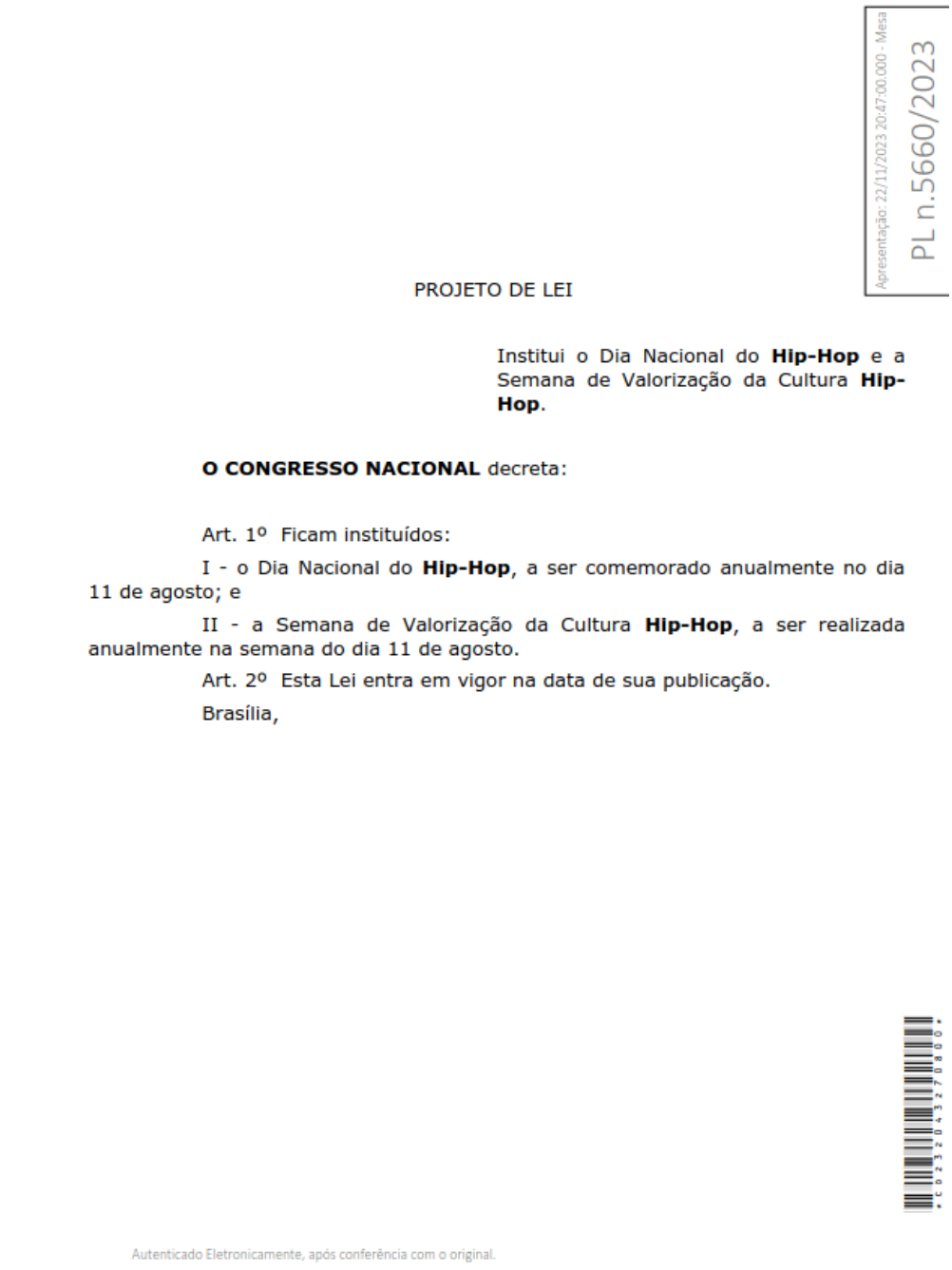
Art. 2º Esta Lei entra em vigor na data de sua publicação.

LexEdit

Assinado eletronicamente pelo(a) Dep. Mauricio Marcon

Fonte: Elaborado pelo autor.

Figura 3.20: *Proposição classificada como: Trabalho,Educação e Cultura.*



Apresentação: 22/11/2023 20:47:00.000 - Mesa
PL n.5660/2023

PROJETO DE LEI

Institui o Dia Nacional do **Hip-Hop** e a Semana de Valorização da Cultura **Hip-Hop**.

O CONGRESSO NACIONAL decreta:

Art. 1º Ficam instituídos:

I - o Dia Nacional do **Hip-Hop**, a ser comemorado anualmente no dia 11 de agosto; e

II - a Semana de Valorização da Cultura **Hip-Hop**, a ser realizada anualmente na semana do dia 11 de agosto.

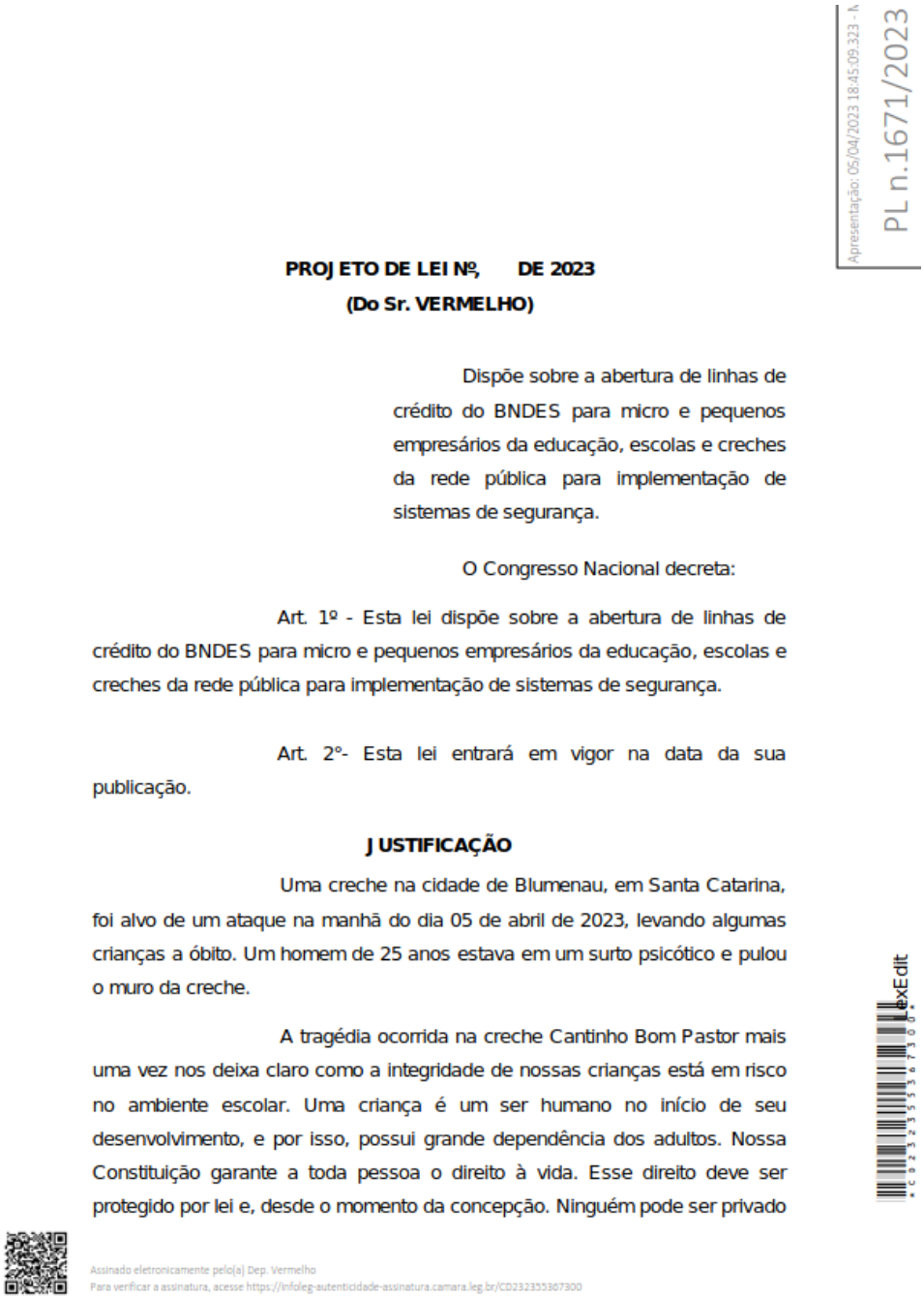
Art. 2º Esta Lei entra em vigor na data de sua publicação.

Brasília,

Autenticado Eletronicamente, após conferência com o original.

Fonte: Elaborado pelo autor.

Figura 3.21: *Proposição classificada como: Trabalho,Educação e Cultura.*



Apresentação: 05/04/2023 18:45:09.323 - N
PL n.1671/2023

PROJETO DE LEI Nº, DE 2023
(Do Sr. VERMELHO)

Dispõe sobre a abertura de linhas de crédito do BNDES para micro e pequenos empresários da educação, escolas e creches da rede pública para implementação de sistemas de segurança.

O Congresso Nacional decreta:

Art. 1º - Esta lei dispõe sobre a abertura de linhas de crédito do BNDES para micro e pequenos empresários da educação, escolas e creches da rede pública para implementação de sistemas de segurança.

Art. 2°- Esta lei entrará em vigor na data da sua publicação.

JUSTIFICAÇÃO

Uma creche na cidade de Blumenau, em Santa Catarina, foi alvo de um ataque na manhã do dia 05 de abril de 2023, levando algumas crianças a óbito. Um homem de 25 anos estava em um surto psicótico e pulou o muro da creche.

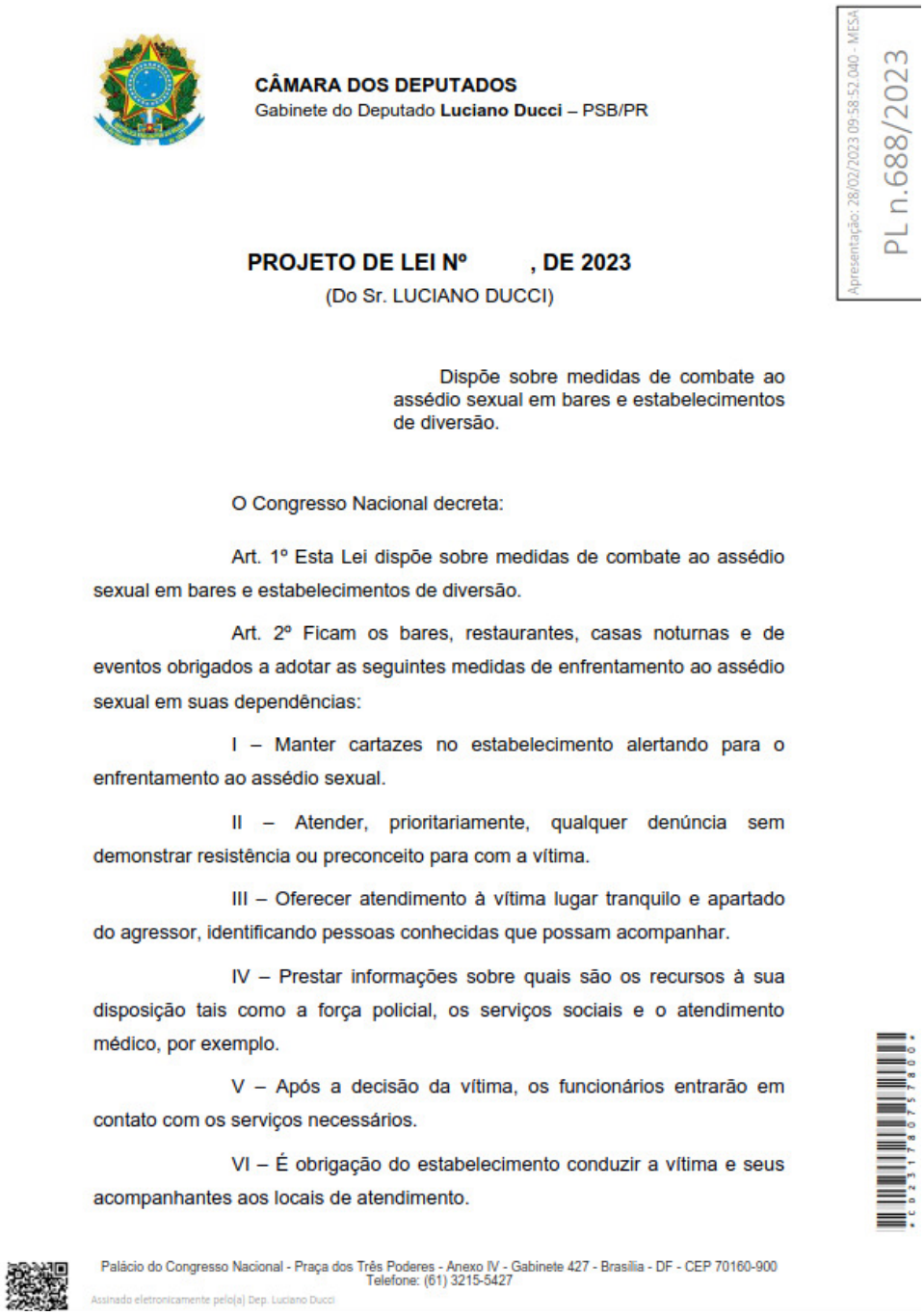
A tragédia ocorrida na creche Cantinho Bom Pastor mais uma vez nos deixa claro como a integridade de nossas crianças está em risco no ambiente escolar. Uma criança é um ser humano no início de seu desenvolvimento, e por isso, possui grande dependência dos adultos. Nossa Constituição garante a toda pessoa o direito à vida. Esse direito deve ser protegido por lei e, desde o momento da concepção. Ninguém pode ser privado

LexEdit

Assinado eletronicamente pelo(a) Dep. Vermelho
Para verificar a assinatura, acesse https://infoleg-autenticidade-assinatura.camara.leg.br/CD232355367300

Fonte: Elaborado pelo autor.

Figura 3.22: *Proposição classificada como: Saúde,Educação,Cultura e Proteção de Vulneráveis.*



CÂMARA DOS DEPUTADOS
Gabinete do Deputado **Luciano Ducci** – PSB/PR

Apresentação: 28/02/2023 09:58:52.040 - MESA
PL n.688/2023

PROJETO DE LEI Nº , DE 2023
(Do Sr. LUCIANO DUCCI)

Dispõe sobre medidas de combate ao assédio sexual em bares e estabelecimentos de diversão.

O Congresso Nacional decreta:

Art. 1º Esta Lei dispõe sobre medidas de combate ao assédio sexual em bares e estabelecimentos de diversão.

Art. 2º Ficam os bares, restaurantes, casas noturnas e de eventos obrigados a adotar as seguintes medidas de enfrentamento ao assédio sexual em suas dependências:

I – Manter cartazes no estabelecimento alertando para o enfrentamento ao assédio sexual.

II – Atender, prioritariamente, qualquer denúncia sem demonstrar resistência ou preconceito para com a vítima.

III – Oferecer atendimento à vítima lugar tranquilo e apartado do agressor, identificando pessoas conhecidas que possam acompanhar.

IV – Prestar informações sobre quais são os recursos à sua disposição tais como a força policial, os serviços sociais e o atendimento médico, por exemplo.

V – Após a decisão da vítima, os funcionários entrarão em contato com os serviços necessários.

VI – É obrigação do estabelecimento conduzir a vítima e seus acompanhantes aos locais de atendimento.

Palácio do Congresso Nacional - Praça dos Três Poderes - Anexo IV - Gabinete 427 - Brasília - DF - CEP 70160-900
Telefone: (61) 3215-5427

Assinado eletronicamente pelo(a) Dep. Luciano Ducci

Fonte: Elaborado pelo autor.

Capítulo 4

Considerações finais e trabalhos futuros

Neste capítulo serão apresentadas as conclusões obtidas ao longo do desenvolvimento deste trabalho, destacando os principais resultados alcançados e as lições aprendidas. Também serão discutidas as perspectivas de trabalhos futuros, explorando as possibilidades de expansão do que foi realizado neste trabalho, bem como novas áreas de pesquisa que podem ser exploradas a partir dos resultados obtidos.

4.1 Considerações finais

O presente trabalho explorou a utilização de técnicas de PLN (Processamento de Linguagem Natural) e aprendizado de máquina para analisar e categorizar proposições legislativas da Câmara dos Deputados do Brasil[1]. O desenvolvimento de um modelo de agrupamento (clustering) permitiu a classificação automática dessas proposições em tópicos relevantes, facilitando a compreensão e o acompanhamento das atividades parlamentares pela sociedade. Através da aplicação de técnicas de visão computacional e Reconhecimento ótico de caracteres, foi possível extrair informações dos arquivos PDF das proposições, superando as limitações dos campos de metadados disponibilizados pela Câmara.

É importante ressaltar que os métodos de reconhecimento ótico de caracteres utilizados neste trabalho podem apresentar falhas (NAGY G.; NARTKER, 2000), e com isso podem ocorrer incoerências nas classificações das proposições utilizando arquivos PDF. Para solucionar este problema seria necessário obter o texto das proposições na íntegra diretamente da Câmara dos Deputados, removendo a necessidade de analisar os caracteres graficamente. Por esses motivos é necessário que na utilização dos agrupamentos feitos neste trabalho esteja explicito que podem haver incoerências no resultado do modelo.

A utilização de modelos de linguagem de grande escala (LLMs) para a nomeação dos clusters pode demonstrar o potencial dessas ferramentas na interpretação e categorização automatizada de grandes volumes de dados textuais. Sobre estas limitações é preciso ressaltar a falta de coesão no campo *keywords* disponibilizado pela API da Câmara dos Deputados. A escassez de proposi-

[1] <http://camara.leg.br>

ções com esse campo devidamente preenchido, somada à inconsistência na atribuição de palavras-chave, evidenciou a necessidade de buscar fontes alternativas de informação para a categorização dos temas. Essa limitação ressalta a importância de aprimorar a gestão e a padronização dos metadados das proposições, garantindo maior confiabilidade e utilidade para pesquisas e análises futuras.

A disponibilização de uma base de dados relacional estruturada e de fácil acesso, contendo informações sobre deputados, partidos e proposições, representa uma contribuição significativa para a pesquisa e o desenvolvimento de novas aplicações que promovam a transparência e a participação cidadã na democracia brasileira. O presente trabalho reforça a importância da utilização de tecnologias da informação e comunicação (TICs) para aprimorar o acesso à informação e fortalecer a democracia no Brasil. Esta base de dados (e todo o processo de aquisição e processamento dos dados disponibilizados pela Câmara dos Deputados) será disponibilizada na plataforma de código aberto GitHub[2].

4.2 Trabalhos futuros

O presente trabalho abre caminho para diversas possibilidades de pesquisa e desenvolvimento futuro. Algumas das principais áreas de expansão incluem:

- **Aprimoramento do modelo de agrupamento**: Investigar a aplicação de diferentes algoritmos de clustering e técnicas de pré-processamento de texto para aprimorar a precisão e a qualidade dos agrupamentos gerados.
- **Análise temporal**: Expandir a base de dados para incluir proposições de outros anos e legislaturas, permitindo a análise da evolução dos temas e das atividades parlamentares ao longo do tempo.
- **Incorporação de mais dados**: Integrar outros dados disponibilizados pela Câmara dos Deputados, como discursos parlamentares, votações e notícias, para enriquecer a análise e obter uma visão mais completa das atividades legislativas.
- **Desenvolvimento de aplicações**: Criar aplicações e visualizações interativas que utilizem a base de dados gerada para facilitar o acesso e a compreensão das informações sobre as proposições e os deputados, promovendo a participação cidadã e o acompanhamento das atividades parlamentares.
- **Análise de sentimento**: Aplicar técnicas de análise de sentimento para avaliar a polarização e o tom dos discursos e proposições, identificando tendências e padrões no debate político.

Em resumo, o presente trabalho estabelece uma base para futuras pesquisas e desenvolvimentos na área de análise de dados legislativos, com o potencial de contribuir para o fortalecimento

[2] <https://github.com/arthur-mts/tcc>

da democracia e da transparência no Brasil. A combinação de técnicas de processamento de linguagem natural, aprendizado de máquina, visão computacional e inteligência artificial oferece um caminho para a criação de ferramentas e soluções que promovam a participação cidadã e o acompanhamento das atividades parlamentares.

Referências Bibliográficas

BALLARD, D. H. B. C. M. *Computer Vision*. : Department of Computer Science University of Rochester, New York, 1982. 31

BATISTA C.; VIANA, F. Tecnologias da informação e comunicação (tics) e democracia. *Instituto de Ciência Política da Universidade de Brasília*, 2006. 1

BERRY Michael J. A.; LINOFF, G. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. : Wiley Computer Publishing, 1997. 5

CASTELLS, M. *A sociedade em rede. São Paulo: Paz e Terra*. 1999. <https://periodicos.ufpb.br/ojs/index.php/ies/article/view/337>. Acesso em: 23 ago. 2024. 1

ENAP. *Ciência de dados em políticas públicas: uma experiência de formação*. : Escola Nacional de Administração Pública (Enap), 2022. Disponível em <http://repositorio.enap.gov.br/handle/1/7472>. Acesso em 28 de agosto de 2024. 1

FERNANDES, M. S. Tenho dito: uma aplicação para análise de discursos parlamentares utilizando técnicas de processamento de linguagem natura. *Trabalho de Conclusão de Curso – Universidade de Brasília*, 2017. ix, 11, 12

HAN J.; KAMBER, M. P. J. *Data Mining: Concepts and Techniques*. : Morgan Kaufmann Publishers, 2012. 17, 18

IETF. *RFC 9110. Hypertext Transfer Protocol Version 2 (HTTP/2).* 2022. Disponível em <https://httpwg.org/specs/rfc9110.html>. Acesso em: 8 ago. 2024. 6

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 1972. 7

JURAFSKY D.; MARTIN, J. H. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. : Stanford University, 2024. Disponível em <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>. Acesso em 28 de agosto de 2024. 9

LUHN, H. P. Keyword-in-context index for technical literature. *American Documentation 11 (4): 288–295.*, 1960. 7

MARODIN G.; FRANK, A. G. T. G. L. N. T. Lean product development and lean manufacturing: Testing moderation effects. *International Journal of Production Economics, v. 203, n. June*, 2018. 11

MAX, S. *Democracia eletrônica para quem? : quem são, o que querem e como os cidadãos avaliam o portal da Câmara dos Deputados*. Dissertação (Mestrado) — Universidade de Brasília, Instituto de Ciência Política, 2012. ix, 1, 13, 14, 15

MOLINARI, A. M. *Aprendendo com discursos : uma análise em alta dimensão da ideologia e polarização política na Câmara dos Deputados Federais do Brasil*. Dissertação (Mestrado) — Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas (FACE), 2020. Disponível em <http://repositorio2.unb.br/jspui/handle/10482/39876>. Acesso em 28 de agosto de 2024. ix, 12

MULHOLLAND C.; DE OLIVEIRA, S. R. *Uma Nova Cara Para a Política? Considerações sobre Deepfakes e Democracia.* 2021. <https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5773>. Acesso em: 23 ago. 2024. 1

MULLER A.; GUIDO, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. : O'Reilly Media, 2016. 8

NAGY G.; NARTKER, T. A. V. R. Optical character recognition: An illustrated guide to the frontier. *SPIE Vol. 3967, 58-69*, 2000. 7, 45

PETERSON, A. S. A. Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, 2018. ix, 15

PRINCE, S. J. *Computer Vision: Models, Learning, and Inference*. : Cambridge University Press, 2012. Disponívem em <https://udlbook.github.io/cvbook/>. Acesso em 24 de setembro de 2024. 6

RAJARAMAN A.; ULLMAN, J. *Mining of Massive Datasets*. : Cambridge University Press, 2012. 33

RARFORD, A. e. a. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, 2019. Disponível em <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>. Acesso em 28 de agosto de 2024. 10

RIBEIRO R. A.; FARINA, R. M. F. F. Diversificando na prÁtica a utilizaÇÃo dos mÉtodos Ágeis: Scrum e kanban – desenvolvimento de uma locadora de veÍculos. *Revista Científica Semana Acadêmica*, 2023. 11

ROVER, A. J. *A democracia digital possível.* 2006. <https://periodicos.ufsc.br/index.php/sequencia/article/view/15202>. Acesso em: 23 ago. 2024. 2

TAN P. N.; STEINBACH, M. K. V. *Introduction to Data Mining*. : Pearson Education Limited, 2014. Disponível em <https://www.ceom.ou.edu/media/docs/upload/Pang-Ning_Tan_Michael_Steinbach_Vipin_Kumar_-_Introduction_to_Data_Mining-Pe_NRDK4fi.pdf>. Acesso em 29 de agosto de 2024. 6, 9

VASWANI, A. e. a. Attention is all you need. *31st Conference on Neural Information Processing System*, 2017. 10

ZUCCO C.; POWER, T. Fragmentation without cleavages? endogenous fractionalization in the brazilian party system. *Forthcoming in Comparative Politics*, 2019. Disponível em <http://dx.doi.org/10.2139/ssrn.3466149>. Acesso em 28 de agosto de 2024. 13

Apêndices

Anexo A

Fundamentação teorica

A.1 Contagem de proposições legislativas por tipo no ano de 2023

Anexo disponível em <https://drive.google.com/file/d/1TebkzHzDEqEBU6GKKKQY8Jd78SsMNQmU/view?usp=drive_link>

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande - Código INEP: 25137409
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

versão final do TCC

Assunto:	versão final do TCC
Assinado por:	Arthur Soares
Tipo do Documento:	Anexo
Situação:	Finalizado
Nível de Acesso:	Ostensivo (Público)
Tipo do Conferência:	Cópia Simples

Documento assinado eletronicamente por:

- **Arthur Mauricio Thomaz Soares**, **ALUNO (201911250022) DE BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO - CAMPINA GRANDE**, em 04/10/2024 19:17:22.

Este documento foi armazenado no SUAP em 04/10/2024. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1268425
Código de Autenticação: 7e9a220cb0