UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Classificação Temática Multirrótulo de Proposições Legislativas Utilizando Técnicas de Mineração de Textos e Aprendizado de Máquina

Thales dos Santos Oliveira

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Thales dos Santos Oliveira

Classificação Temática Multirrótulo de Proposições Legislativas Utilizando Técnicas de Mineração de Textos e Aprendizado de Máquina

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Fábio Manoel França Lobato

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Oliveira, Thales dos Santos
Classificação Temática Multirrótulo de Proposições Legislativas Utilizando Técnicas de Mineração de Textos e Aprendizado de Máquina / Thales dos Santos Oliveira ; orientador Fábio Manoel França Lobato. – São Carlos, 2024.
96 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Lobato, Fábio Manoel França, orient. II. Título.

Thales dos Santos Oliveira

Thematic Multi-label Classification of Legislative Proposals Using Text Mining and Machine Learning Techniques

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Fábio Manoel França Lobato

Original version

São Carlos

2024

AGRADECIMENTOS

A Deus, causa primordial de todas as coisas, que sempre esteve esteve ao meu lado me guiando e dando forças.

À Larissa, pela compreensão nos momentos em que me recolhi para dedicar a este trabalho.

Aos meus pais, por pavimentarem meu caminho de estudos para que eu chegasse até aqui.

Ao meu orientador, Professor Fábio Lobato, pelo conhecimento compartilhado, disponibilidade e incentivo.

Aos professores do MBA Inteligência Artificial e Big Data, que fizeram deste curso uma jornada empolgante de novas descobertas.

RESUMO

Oliveira, T. d. S. **Classificação Temática Multirrótulo de Proposições Legislativas Utilizando Técnicas de Mineração de Textos e Aprendizado de Máquina**. 2024. 96p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Em instituições públicas, a correta gestão da informação permite que dados relevantes para a população sejam organizados e acessíveis. Quando a gestão da informação é realizada corretamente, o processo pode ajudar a reduzir a redundância e os erros administrativos, promovendo um uso mais racional dos recursos públicos e atendendo a normas legais, como a Lei de Acesso à Informação (LAI). O Centro de Documentação e Informação é a unidade responsável pela gestão da informação na Câmara dos Deputados. Em sua estrutura institucional, a Seção de Indexação de Matérias Legislativas tem o dever de indexar as proposições legislativas submetidas à casa, atribuindo a cada uma delas termos e informações de forma a permitir sua catalogação e buscas futuras. Parte deste processo envolve a classificação das proposições em um ou mais temas, presentes em uma lista com 32 opções. Trata-se de um trabalho manual e que requer extrema atenção dos Analistas Legislativos, profissionais que realizam esta tarefa. A indexação manual sabidamente está sujeita a desafios relacionados ao condicionamento humano como subjetividade, conhecimento prévio e neutralidade do profissional. Considerando os desafios mencionados e a quantidade de proposições submetidas anualmente, torna-se interessante a busca de soluções para classificação automática de proposições por temas. Esta pesquisa teve como objetivo projetar, implementar e avaliar um sistema de classificação multirrótulo automática de temas de proposições legislativas por meio de técnicas de Mineração de Textos (MT). Para isso, utilizou-se dados disponibilizados pelo Portal da Câmara dos Deputados contendo as proposições legislativas submetidas entre os anos de 2013 e 2023, em que cada uma das proposições foi manualmente classificada pela Seção de Indexação de Matérias Legislativas. Este trabalho avaliou 53 combinações de classificadores dos tipos *Problem Transformation* e *Algorithm Classification* para que se pudesse entender a viabilidade da aplicação e quais os classificadores mais indicados para o problema. Os resultados demonstraram que soluções utilizando técnicas de Inteligência Artificial são promissoras, com algumas limitações, e que novas linhas de pesquisa têm potencial de levar a maiores valores de *F1-score* e *Subset Accuracy*. O trabalho impacta positivamente o processo de indexação de proposições legislativas uma vez que apresenta técnicas de Mineração de Textos como exemplo de auxílio viável a servidores públicos que realizam a classificação temática das proposições.

Palavras-chave: Classificação multirrótulo; Proposição legislativa; Indexação de docu-

mentos; Mineração de Textos; Inteligência Artificial; Serviço Público.

ABSTRACT

Oliveira, T. d. S. **Thematic Multi-label Classification of Legislative Proposals Using Text Mining and Machine Learning Techniques**. 2024. 96p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

In public institutions, the correct management of information allows relevant data to the population to be organized and accessible. When information management is carried out correctly, the process can help reduce redundancy and administrative errors, promoting a more rational use of public resources and meeting legal standards, such as the Access to Information Law (LAI). The Documentation and Information Center is the unit responsible for managing information in the Chamber of Deputies. Within its institutional structure, the Legislative Matters Indexing Section is responsible for indexing the legislative proposals submitted to the House, assigning each of them additional terms and information in order to allow their cataloging and future searches. Part of this process involves classifying the proposals into one or more thematic area, present in a list with 32 options. This is a manual work and requires extreme attention from Legislative Analysts, the professionals who perform this task. Manual indexing is known to be subject to challenges related to the human condition, such as subjectivity, prior knowledge, and professional neutrality. Considering the aforementioned challenges and the number of proposals submitted annually, it becomes interesting to seek solutions for automatically classifying proposals by its thematic areas. This research aimed to design, implement and evaluate a system of automatic multi-label classification of topics of legislative propositions through Text Mining techniques. For this, data made available by the Portal of the Chamber of Deputies containing the legislative propositions submitted between the years 2013 and 2023 were used, in which each of the propositions was manually classified by the Legislative Matters Indexing Section. This work evaluated 53 combinations of Problem Transformation and Algorithm Classification classifiers in order to understand the feasibility of the application and which classifiers are most suitable for the problem. The results showed that solutions using Artificial Intelligence techniques are promising, with some limitations, and that new lines of research have the potential to lead to higher values of F1-score and Subset Accuracy. The work has a positive impact on the indexing process of legislative propositions since it presents Text Mining techniques as an example of viable aid to public servants who carry out the thematic classification of propositions.

Keywords: Multi-label classification; Legislative proposal; Document indexing; Text Mining; Artificial Intelligence; Public Service.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
AM	Aprendizado de Máquina
BERT	*Bidirectional Encoder Representations from Transformers*
CCJC	Comissão de Constituição e Justiça e de Cidadania
CEDI	Centro de Documentação e Informação da Câmara dos Deputados
CELEG	Coordenação de Organização da Informação Legislativa
CFT	Comissão de Finanças e Tributação
CGU	Controladoria Geral da União
CNN	Redes Neurais Convolucionais
COBEC	Coordenação de Preservação de Conteúdos Informacionais
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CSV	*Comma-separated Value*
DIKW	*Data-Information-Knowledge-Wisdom*
DST	*Data Science Trajectories*
ETL	*Extract, Transform, and Load*
FN	Falso Negativo
FP	Falso Positivo
IA	Inteligência Artificial
ISO	Organização Internacional para Padronização
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Databases*
LAI	Lei de Acesso à Informação
LLM	*Large Language Models*
MT	Mineração de Textos

NLG	*Natural Language Generation*
NLP	*Natural Language Processing*
NLU	*Natural Language Understanding*
ODS	Objetivos de Desenvolvimento Sustentável
PEC	Proposta de Emenda à Constituição
PL	Projeto de Lei
PLN	Processamento de Línguas Naturais
RNN	Redes Neurais Recorrentes
SEMMA	*Sample, Explore, Modify, Model, Assess*
SETAP	Seção de Gestão de Taxonomias e Políticas de Indexação
SGM	Secretaria-Geral da Mesa
SIDEX	Seção de Indexação de Matérias Legislativas
SILEG	Sistema de Informações Legislativas
STF	Supremo Tribunal Federal
TECAD	Tesauro da Câmara dos Deputados
TF-IDF	*Term Frequency - Inverse Document Frequency*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

A eficiência no serviço público é uma preocupação constante. Ações e decisões devem ser tomadas tendo como base o Princípio da Economicidade, garantindo o melhor uso dos recursos disponíveis com a máxima eficiência, eficácia e efetividade (FEDERAL, 2020, p. 64). Esforços em busca de processos mais eficientes frequentemente envolvem a implementação de novas tecnologias e sistemas de gestão de processos. Um problema específico e desafiador é a indexação manual de documentos, que está sujeita ao julgamento, subjetividade, conhecimento prévio e experiência dos profissionais responsáveis (NAVES, 2007). De acordo com Desordi e Bona (2020, p. 9), uma abordagem que vem sendo adotada por órgãos públicos envolve a implantação de sistemas baseados em Inteligência Artificial (IA) para apoiar o desenvolvimento das funções desempenhadas pelos servidores públicos. Técnicas de IA têm sido utilizadas em diversos campos como saúde e educação, auxiliando em tomadas de decisão e na automação de tarefas (BRASIL, 2023).

Como consequência, governos de diversos países, como o Brasil, têm buscado regulamentar seu uso dentro da estrutura pública (BRASIL, 2023). Neste sentido, o Congresso Nacional é o órgão constitucional que exerce, dentre outras, as funções de elaborar, debater, aperfeiçoar e aprovar as leis no âmbito federal. É constituído pelo Senado e pela Câmara dos Deputados, que foram estabelecidos na primeira Constituição do país, em 1824. O Senado Federal é composto por representantes das unidades federativas. Portanto, para garantir igualdade na formulação e aprovação de leis, cada Unidade Federativa possui três senadores. Os membros da Câmara dos Deputados, por outro lado, representam a população. Consequentemente, as unidades federativas possuem entre 8 e 70 representantes, conforme sua densidade populacional, totalizando 513 deputados por legislatura (BRASIL, 2024b).

Para que normas legais sejam instauradas no Brasil, é necessária a formulação de proposições legislativas, ou seja, propostas formais submetidas à apreciação do Senado, da Câmara ou do Congresso Nacional. As proposições legislativas podem incluir Propostas de Emenda à Constituição (PEC), Projetos de Lei (PL), Emendas, entre outras. Projetos de lei que têm origem e aprovação na Câmara dos Deputados são submetidos à revisão pelo Senado. Da mesma forma, projetos apresentados e aprovados pelos senadores devem passar pelo crivo dos deputados antes de serem encaminhadas para a sanção presidencial, virando, assim, uma lei (BRASIL, 2024c).

Milhares de proposições são feitas anualmente por membros do Congresso. Porém, a criação e submissão das propostas é apenas parte do processo que ocorre antes das votações. Uma das etapas intermediárias é a indexação das propostas. Nela, funcionários do Centro de Documentação e Informação da Câmara dos Deputados (CEDI), na função de pessoas

indexadoras, analisam as proposições e realizam sua indexação manualmente por meio de um sistema de gerenciamento interno chamado Sistema de Informações Legislativas (SILEG). O processo de indexação das proposições legislativas é importante pois assegura o cumprimento da Lei de Acesso à Informação, à qual o governo está sujeito, ao garantir a recuperação dos documentos pelos cidadãos por meio de pesquisas no site da Câmara dos Deputados (BRASIL, 2016).

Uma das fases do processo de indexação de proposições é a seleção das áreas temáticas, na qual a pessoa indexadora, após ler o documento de inteiro teor, identifica em quais temas a proposição se encaixa. O sistema SILEG disponibiliza uma lista de temas constituída por trinta e dois itens como Comunicações, Economia, Educação e Saúde e permite a escolha de múltiplos temas para uma mesma proposição (BRASIL, 2016).

Em processos de categorização manual, fatores humanos como subjetividade, nível de experiência, falta de atenção e até cansaço podem gerar resultados errados ou aquém do esperado (NAVES, 2007). Ao analisar o contexto legislativo do Brasil e os desafios da indexação manual de documentos, identificou-se uma potencial problemática. Considerando o paradigma *human-in-the-loop*, a Seção de Indexação de Matérias Legislativas (SIDEX), responsável pela indexação de proposições na Câmara dos Deputados, foi contactada para auxiliar na validação do problema e de hipóteses levantadas. Segundo o setor, cada servidor indexa, em média, 130 proposições por mês. Além disso, foram levantadas as seguintes hipóteses:

- Um sistema de recomendação de áreas temáticas para as proposições de lei na câmara dos deputados reduziria o tempo total de indexação dos servidores;
- Um sistema de recomendação de áreas temáticas para as proposições de lei na câmara dos deputados diminuiria possíveis erros cometidos pelos servidores anotadores, aumentando assim a confiabilidade no processo de indexação;
- Um sistema de recomendação de áreas temáticas para as proposições de lei na câmara dos deputados auxiliaria novos servidores que ainda não estão familiarizados com todas áreas temáticas disponíveis para seleção a serem mais assertivos na classificação.

Considerando este cenário, percebe-se a possibilidade de auxiliar os profissionais responsáveis pela indexação de proposições legislativas a realizar seu trabalho do modo mais assertivo possível.

A classificação multirrótulo é uma das tarefas do subcampo da Inteligência Artificial, denominado Aprendizado de Máquina (AM), na qual é possível atribuir uma ou mais categorias a uma única entrada (HERRERA *et al.*, 2016). Desta forma, visando a contínua assertividade no trabalho de classificação de temas de proposições legislativas e tendo como base as proposições previamente classificadas pelo CEDI, questiona-se:

Q1 *"É possível automatizar a classificação de novas proposições legislativas por tema utilizando técnicas de Mineração de Textos?"*

Q2 *"Quais métricas são viáveis de serem utilizadas para avaliar a solução proposta em termos de eficiência e acurácia?"*

Q3 *"Quais técnicas de Aprendizado de Máquina resultam em classificações mais eficazes no contexto do trabalho?"*

À luz do cenário apresentado e dos questionamentos supraditos, este trabalho tem como objetivo geral projetar, implementar e avaliar um sistema de categorização automática de temas de proposições legislativas por meio de técnicas de Mineração de Textos. No Capítulo 2 é apresentada a Fundamentação Teórica que abrange os conceitos utilizados para o desenvolvimento da pesquisa. Nele também é detalhado o processo de classificação de Proposições Legislativas realizada pelo CEDI. O Capítulo 3 contém a descrição de trabalhos relacionados, sendo um deles com objetivo similar ao nosso. No Capítulo 4 apresentamos a abordagem metodológica utilizada e as ferramentas escolhidas para os experimentos realizados. O Capítulo 5 possui a descrição dos resultados dos experimentos realizados. O trabalho é concluído no Capítulo 6, onde apresentamos os desafios encontrados, as contribuições para a área e possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A fim de proporcionar uma visão geral sobre a solução a ser apresentada, este capítulo inicia abordando a indexação manual e automática de documentos. Seguimos expondo os conceitos de Ciência de Dados; Inteligência Artificial e Aprendizado de Máquina; Categorias de Algoritmos de Classificação; além de debater técnicas de Processamento de Línguas Naturais. Por fim, nos aprofundamos no processo de indexação de propostas legislativas adotado pelo Centro de Documentação e Informação da Câmara dos Deputados.

2.1 Indexação Manual e Automática de Documentos

A Associação Brasileira de Normas Técnicas (ABNT) define indexação como sendo o "*ato de identificar e descrever o conteúdo de um documento com termos representativos dos seus assuntos e que constituem uma linguagem de indexação*" (ABNT, 1992). Em outras palavras, Anderson e Pérez-Carballo (2001, p. 233, tradução nossa) definem que indexação "*significa simplesmente apontar ou indicar o conteúdo, significado, propósito e características de mensagens, textos e documentos*".

Na indexação manual, os conceitos são identificados por meio de uma análise intelectual. A norma NBR 12.676 de agosto de 1992, elaborada pela ABNT, define que ela consiste de três estágios que, na prática, podem se sobrepor. Em cada um deles, é esperado o uso de instrumentos de indexação, como tesauros, códigos de classificação, cabeçalhos de assunto, etc. São eles:

1. **Exame do documento e estabelecimento de seu conteúdo:** Este estágio consiste na leitura e compreensão do documento. A norma indica a leitura de algumas partes do documento para facilitar sua compreensão, como título e subtítulo; resumo; sumário; introdução; ilustrações, diagramas, tabelas e seus títulos; palavras ou grupos de palavras em destaque e referências bibliográficas;

2. **Identificação dos conceitos presentes no assunto:** Aqui, a pessoa indexadora identifica sistematicamente os conceitos que são os essenciais na descrição do assunto. Ao escolher tais conceitos, a pessoa indexadora deve considerar as consultas que podem ser feitas no momento de uma busca e ter em mente a finalidade para a qual são usados os termos de indexação;

3. **Tradução desses conceitos nos termos de uma linguagem de indexação:** Por fim, a pessoa indexadora converte os conceitos identificados no estágio anterior em termos adotados pela instituição, que os representam de forma precisa e são mantidos em instrumentos de referência como, por exemplo um tesauro.

A Organização Internacional para Padronização (ISO), define tesauro como um

> vocabulário controlado e estruturado no qual conceitos são representados por termos, organizados de forma que as relações entre conceitos são explicitadas, e termos preferenciais são acompanhados por entradas que conduzem a sinônimos e quase-sinônimos (ISO, 2011, tradução nossa).

O processo de indexação como um todo é complexo e impõe dificuldades à pessoa indexadora por lhe demandar esforços cognitivos. Desta forma, é relevante que sejam abordados certos aspectos que podem impactar a qualidade da indexação de um documento.

De acordo com Pinheiro (1978, p. 109), a indexação envolve julgamento e isso pode gerar discrepâncias na escolha de termos escolhidos por indexadores para representação do conteúdo de um documento. Indexar, segundo a autora, é um processo altamente subjetivo, o que torna a consistência absoluta praticamente impossível. A consistência da indexação depende das condições de desempenho, da experiência dos indexadores e de seus instrumentos de auxílio como manuais e vocabulários controlados. Quando estes instrumentos são adotados, há o aumento considerável da consistência.

Naves (2007, p. 191-192; p. 202-203) afirma que os fatores mais relevantes relacionados à influência da pessoa indexadora no processo de análise de assunto são a subjetividade, o conhecimento prévio e sua formação e experiência. O trabalho conclui, entre outros pontos, que a compreensão do texto é primordial no processo de análise de assuntos e que há necessidade da pessoa indexadora buscar contribuições de áreas disciplinares diversas. No trabalho também é enfatizado que a deficiência na formação e a falta de especialização do profissional levam a dificuldades no processo de análise. A autora ainda explicita que maior experiência da pessoa indexadora somada com o domínio da técnica de indexação lhe proporciona mais segurança, maior vivência e maior capacidade de decisão diante de situações complexas.

A subjetividade é classificada por Naves (2007, p. 191) como a situação em que "*diferentes indivíduos criam diferentes figuras ou ideias de uma mesma informação externa, por causa de suas inclinações pessoais e afetivas, que certamente interferem no trabalho por eles desenvolvido*". Já Rubi (2017, p. 294) aponta que a análise de assunto é revestida de uma subjetividade característica uma vez que é realizada a partir da leitura do documento pela pessoa indexadora. Em outro trabalho, Rubi (2009, p. 83) afirma que comprovadamente, o processo de indexação é imerso em subjetividade em virtude se ser realizado por seres humanos que utilizam seu conhecimento prévio em diversas áreas para identificar e selecionar conceitos de um documento. Para diminuir a subjetividade, o processo de indexação deve seguir normas e possuir uma política de indexação bem definida a fim

de nortear com diretrizes e critérios o trabalho da pessoa indexadora, reunidos em um manual de indexação.

Como já pontuado, o conhecimento prévio é um fator de influência da pessoa indexadora no processo de indexação. Para compreender um texto, as pessoas utilizam todo seu conhecimento prévio armazenado na memória de longo prazo. O conhecimento prévio facilita a compreensão do texto, uma vez que oferece uma estrutura na qual o conteúdo do documento possa ser relacionado (NEVES; DIAS; PINHEIRO, 2006). Booth (2013, p. 34) também aponta a importância dos conhecimentos gerais do profissional. Segundo a autora, eles fornecem uma base sobre a qual as informações apresentadas em um documento podem ser sobrepostas, ajudando a pessoa indexadora na compreensão e interpretação. Booth afirma que para manter um bom padrão de conhecimentos gerais, o profissional precisa se manter informado sobre acontecimentos atuais por meio de qualquer meio a ele disponível, como Internet, televisão e jornais.

Booth (2013, p. 36) também aponta outra questão que pode impactar o resultado da indexação: a neutralidade da pessoa indexadora. A autora explica que ao iniciar a análise de um documento, toda pessoa indexadora traz consigo suas crenças, preconceitos, ideias, "fatos", conhecimentos gerais e "sabedoria convencional". Também aponta que grande parte deste pacote é útil para auxiliar na compreensão, interpretação e representação do conteúdo do documento. Em casos em que documentos tratam de forma crítica ou controversa de um assunto, pode haver um contraste com opiniões pessoais. Contudo, a autora pondera que não é necessário que uma pessoa indexadora seja totalmente a favor de tudo em um documento, mas a indexação deve representar o tom e também o conteúdo. O profissional pode ter discordância com certas partes do documento, no entanto, isso não deve refletir no índice. Embora o índice seja uma obra em si, elaborada pela pessoa indexadora para demonstrar o conhecimento amplo e especializado, bem como a expertise técnica do profissional, ele não deve revelar as convicções, posturas ou juízos pessoais do profissional.

A relação entre o nível de experiência de pessoas indexadoras e a sua habilidade técnica é mencionada por Neves, Dias e Pinheiro (2006, p. 143), que citam uma experiência realizada cujos resultados indicaram que profissionais mais experientes empregaram estratégias específicas que não foram adotadas por indexadores novatos. Além disso, foi observado que tanto a habilidade na utilização da linguagem de indexação quanto a familiaridade com o tema exercem influência significativa no trabalho dos indexadores.

Por fim, podemos citar a falta de tempo como outro fator dificultante na atividade de indexação. Neves, Dias e Pinheiro (2006, p. 142) ponderam que o cotidiano da pessoa indexadora concentra-se no ato da leitura, com a finalidade de permitir que a informação contida nos documentos sejam acessadas por usuários de sistemas de recuperação de informação. A leitura integral de documentos demanda um tempo que, por vezes, o

profissional não possui, sendo ele instruído a limitar-se a partes do documento, como o título e o resumo.

Tendo em vista os desafios impostos aos indexadores no processo de indexação manual, o aumento acelerado de informações geradas e a disponibilidade de recursos computacionais em constante avanço, justificam-se os esforços de desenvolvimento de técnicas e ferramentas de indexação automática.

Dentre as diversas definições de "indexação automática" presentes na literatura, Maron (1961, p. 404) determina que o termo denota do problema de decidir, de forma mecânica, a qual categoria (assunto ou área do conhecimento) um dado documento pertence, ou seja, decidir automaticamente sobre o que ele trata.

Segundo Anderson e Pérez-Carballo (2001, p. 232) humanos e máquinas usam diferentes abordagens na indexação. Humanos examinam documentos e textos considerando suas mensagens e características. Já os computadores comparam os símbolos que compõem os textos consultando dados contextuais, como um tesauro; aplicando indexação sintática ou de padrões para identificar unidades maiores de texto; ou ainda calculando atributos para documentos baseados em dados disponíveis.

A utilização da indexação automatizada tem sido tema de debates entre os estudiosos. Leiva (2010, p. 57-60) dá exemplos de justificativas defendidas por autores partidários a ela, como a de que indexação humana é subjetiva, lenta e cara; o aumento da quantidade de documentos eletrônicos favorece o desenvolvimento de pesquisas na área; a indexação automática baseada em Processamento de Línguas Naturais (PLN) oferece alternativas atraentes para indexação de documentos; e evita-se inconsistências causadas por um ou por diferentes profissionais na análise de um mesmo documento.

A indexação automática não é unanimidade na tentativa de atenuação dos desafios impostos pela indexação manual. de Keyser (2012, p. 51-57) reúne argumentos de diferentes autores em favor da indexação manual. Alguns deles são de que a indexação automática não fornece uma visão geral coerente dos termos do índice; não resolve os problemas de sinônimos ou variações; não leva em consideração o contexto; e não permite a busca por assuntos relacionados.

Devido a sua importância, a pesquisa sobre os procedimentos de indexação automática tem avançado rapidamente como uma das áreas mais promissoras da Inteligência Artificial, visando a criação automática de bancos de dados textuais, além da produção e atualização de dicionários (ROBREDO, 1991, p. 130). Guimarães *et al.* (2019, p. 117) completa este raciocínio ao afirmar que no âmbito da Organização do Conhecimento, a Inteligência Artificial desempenha um papel crucial ao fornecer assistência técnica na análise da informação, particularmente na classificação, descrição e indexação de documentos em ambientes digitais. Sua contribuição para a recuperação de informações resulta em uma

maior exaustividade e especificidade, gerando resultados mais relevantes para os usuários.

2.2 Ciência de Dados

Dados são uma abstração de entidades do mundo real (pessoa, objeto ou evento). Cada entidade é tipicamente descrita como um número de atributos. Por exemplo, alguns atributos de um livro são autor, título, gênero, editora, data de lançamento, número de páginas, quantidade de palavras, etc. Uma base de dados consiste nos dados relativos a uma coleção de entidades na qual cada entidade é descrita em termos de um conjunto de atributos. De maneira simplificada, podemos descrever uma base de dados com uma matriz de dados $n \times m$ na qual n é o número de entidades (linhas) e m é o número de atributos (colunas) (KELLEHER; TIERNEY, 2018).

Dados são comumente classificados em três tipos: estruturados, não estruturados e semi-estruturados. Gandomi e Haider (2015, p. 138) os define como:

- **Dados estruturados:** São os dados tabulares encontrados em planilhas ou em bancos de dados relacionais;
- **Dados não estruturados:** Não possuem a organização estrutural necessária para que máquinas possam utilizá-la para análises. São exemplos os arquivos de textos, imagens, áudios e vídeos;
- **Dados semi-estruturados:** Dados deste tipo não seguem padrões rígidos com relação à sua estrutura. Arquivos do tipo *Extensible Markup Language* (XML) e *JavaScript Object Notation* (JSON) possuem marcações definidas pelo usuário que os torna interpretáveis por máquinas.

2.2.1 Conhecimento Gerado por Dados

Ackoff (1989) afirma que dados não têm valor até que sejam processados em uma forma utilizável. Em seu trabalho, o autor descreve os tipos de conteúdo da mente humana por meio de uma hierarquia sendo que a sabedoria está no topo, passando por compreensão, conhecimento, informação e, na base, dados. Cada uma dessas categorias inclui a categoria anterior no sentido de que, por exemplo, não há sabedoria sem compreensão e não há compreensão sem conhecimento. Além disso, o autor demonstra a percepção de que, "em média, 40% do conteúdo das mentes humanas consiste de dados, 30% de informação, 20% de conhecimento e 10% de compreensão e, virtualmente, nada de sabedoria". Cada um dos tipos descritos por Ackoff podem ser sumarizados como:

- **Dados:** São produtos de observação, símbolos que representam propriedades de objetos, eventos e seus ambientes. Não têm valor até que sejam processados em uma forma utilizável.

- **Informação:** São extraídos dos dados por meio de análise.
- **Conhecimento:** É o que permite a transformação de informação em instruções. Pode ser obtido por meio de outra pessoa que o tenha por meio de instruções, ou extraindo-o por meio da experiência.
- **Compreensão/Inteligência:** A habilidade de aumentar a eficiência.
- **Sabedoria:** A habilidade de aumentar a eficácia, de atribuir significado e avaliar o valor das informações em um contexto mais amplo. Envolve uma compreensão mais profunda dos propósitos e implicações das informações, bem como a capacidade de aplicar percepções de forma ética e construtiva.

Bellinger, Castro e Mills (2004) discutem as classificações de Ackoff e expõem o entendimento de que a compreensão não seria um nível separado na hierarquia, mas sim um assistente na transição de um estágio para o outro. Neste caso, haveria uma compreensão de relações na transição entre dados e informação; uma compreensão de padrões entre informação e conhecimento; e uma compreensão de princípios entre conhecimento e sabedoria. Essa relação é ilustrada na Figura 1.

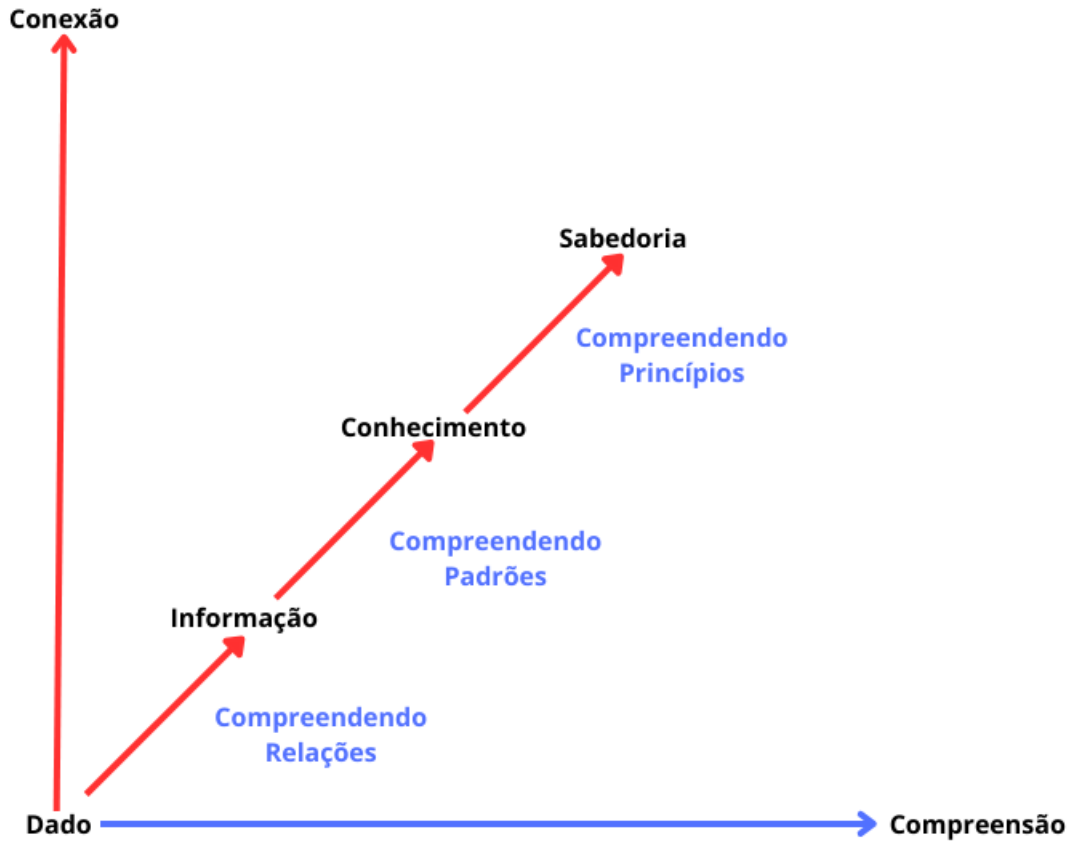

Figura 1 – Hierarquia DIKW.

Fonte: Adaptado de Bellinger, Castro e Mills (2004). Tradução nossa.

Em seu artigo, Rowley (2007) analisa o que se popularizou como a Hierarquia *Data-Information-Knowledge-Wisdom (DIKW)*, Hierarquia da Informação ou Pirâmide do Conhecimento buscando provocar o debate a respeito de conceitos fundamentais de gerenciamento de informações, sistemas de informação e gestão do conhecimento. Após uma detalhada revisão da literatura, a autora conclui que a hierarquia não é citada explicitamente em muitos livros mas está implícita em várias definições de dados, informação, conhecimento e sabedoria. Seu trabalho também indica que a sabedoria é um conceito negligenciado na literatura de gestão de conhecimento e sistemas de informação, sugerindo que haja maior debate sobre a natureza da sabedoria individual e organizacional.

2.2.2 *Big Data*

O termo *Big Data* tem sido utilizado com ampla frequência nos últimos anos. Gandomi e Haider (2015, p. 138) mencionam que muitas vezes a característica mais comum associada a este termo é a quantidade de dados em questão. Porém, há outras características relevantes a serem consideradas para que se possa afirmar que um projeto está lidando com *Big Data*.

As três primeiras características, ou dimensões, que foram identificadas como sendo desafiadoras em gestão de dados são: o Volume, a Variedade e a Velocidade (LANEY, 2001 apud GANDOMI; HAIDER, 2015). Posteriormente, outras dimensões foram percebidas como pertinentes, como a Veracidade, a Variabilidade e o Valor. Uma breve descrição sobre cada dimensão é feita por Gandomi e Haider (2015, p. 138-139) a seguir:

- **Volume:** É a quantidade escalar dos dados. Em termos de *Big Data*, o volume é relativo e varia de acordo com outros fatores, como o tempo e tipo dos dados. Com relação ao tempo, o que é considerado "grande" atualmente pode não ser no futuro por conta das capacidades crescentes de armazenamento. Já com relação aos tipos de dados, duas bases de dados do mesmo tamanho mas de tipos diferentes podem necessitar de tecnologias diferentes para sua gestão, por exemplo, dados tabulares e vídeos;

- **Variedade:** Refere-se à heterogeneidade em uma base de dados, ou seja, a capacidade de processamento de dados estruturados, semi-estruturados e não estruturados;

- **Velocidade:** Proporção na qual os dados são gerados e a velocidade com que devem ser analisados e utilizados;

- **Veracidade:** Representa a falta de confiabilidade inerente a algumas fontes de dados. A necessidade de lidar com dados imprecisos e incertos é outra faceta de *Big Data*. Este problema é mitigado por meio de ferramentas para gestão de dados incertos;

- **Variabilidade (e Complexidade):** A Variabilidade trata da variação nas taxas de fluxos de dados já que, frequentemente, a Velocidade não é consistente e apresenta picos e vales. A Complexidade refere-se ao fato dos dados utilizados em *Big Data* serem obtidos de fontes diferentes, adicionando a necessidade de conectar, combinar, limpar e transformar estes dados;

- **Valor:** Indica o quanto os dados são úteis para um determinado fim. *Big Data* são comumente caracterizados por seu relativo "baixo valor de densidade", que significa que dados coletados na sua forma original geralmente possuem um valor baixo em relação ao seu volume. Entretanto, um alto valor pode ser obtido ao se analisar grandes volumes desses dados.

Tendo em vista as dimensões apresentadas, o armazenamento dos dados constitui um desafio. Sendo assim, é importante definirmos sucintamente o que é uma *data warehouse*. De acordo com Aalst (2016, p. 127) trata-se de um repositório lógico único de dados transacionais e operacionais de uma organização que extrai dados de outros sistemas operacionais. Sua finalidade é unificar as informações de forma que possam ser utilizadas para outras finalidades.

Segundo Vassiliadis (2009, p. 1), o trabalho necessário para que *data warehouses* forneçam aos usuários acesso a informações integradas e gerenciáveis é desafiador e, na prática, gera alguns problemas. O autor enumera a necessidade de transformar o dado a ser adicionado na base em um formato que possa ser consultado e recuperado pelo usuário; a importância de limpar os dados para que os mesmos fiquem livres de ruídos e inconsistências, além de completos e confiáveis; e a exigência de se atualizar os dados constantemente para que o usuário utilize sempre sua versão mais recente.

Extract, Transform, and Load (ETL), ou Extrair, Transformar e Carregar (tradução nossa), descreve os processos e ferramentas usadas para auxiliar o mapeamento, junção e movimento de dados entre bases de dados distintas (KELLEHER; TIERNEY, 2018, p. 74). Aalst (2016, p. 127) explica que neste processo são realizadas as seguintes tarefas:

1. Extração de dados de fontes externas;

2. Transformação dos dados para que se adéquem às necessidades operacionais;

3. Carregamento dos dados nos sistemas-alvo, como um *data warehouse* ou um banco de dados relacional.

Vassiliadis e Simitsis (2009) ressalta que apesar dos processos de ETL serem a espinha dorsal da arquitetura de um *data warehouse*, eles têm outras aplicações, como os *mashups*. Trata-se de aplicações que integram dados obtidos dinamicamente por meio de

invocações a serviços web provenientes de mais de uma fonte. De acordo com o autor, a filosofia por trás desta operação é o ETL e cita o *Google Maps* como exemplo de *mashup*.

2.2.3 Processo de Ciência de Dados

Ao longo dos anos acadêmicos e profissionais têm se esforçado para estabelecer padrões na área de Mineração de Dados[1]. Acadêmicos focam na criação de *frameworks* para Mineração de Dados visando o estabelecimento de uma linguagem padrão. Já profissionais buscam definir processos e metodologias a fim de guiar a implementação de aplicações de Mineração de Dados (AZEVEDO; SANTOS, 2008).

Cielen, Meysman e Ali (2016) explicam que seguir uma abordagem estruturada para Ciência de Dados ajuda a aumentar as chances de sucesso em um projeto a baixos custos. Além disso, os autores enumeram seis fases como sendo tipicamente usadas em projetos de Ciência de Dados, ressaltando que tal abordagem pode não ser viável para todos os projetos. São elas:

1. **Definir o objetivo da pesquisa:** Garantir que todos os envolvidos entendam o "o que", o "como" e o "por que" do projeto;
2. **Obtenção dos dados:** Encontrar dados adequados às necessidades do projeto e obter acesso a eles;
3. **Preparação dos dados:** Transformar os dados em "estado bruto" em dados que possam ser utilizados por modelos, realizando a detecção e correção de diferentes tipos de erros, além da combinação de dados provenientes de diferentes fontes;
4. **Exploração dos dados:** Entender os dados em profundidade buscando por padrões, correlações e desvios por meio de técnicas de visualização e descrição;
5. **Modelagem dos dados:** Tentar obter os *insights* ou fazer as predições definidas para o projeto;
6. **Apresentação dos resultados:** Demonstrar os resultados juntamente com a análise do que foi feito.

Cielen, Meysman e Ali (2016) esclarecem, contudo, que em um projeto de Ciência de Dados o progresso não se dá de forma linear da primeira à última fase e sim de forma iterativa entre elas.

[1] Nesta seção o termo Mineração de Dados é utilizado algumas vezes de acordo com a referência analisada. Trata-se de um termo que, segundo Martínez-Plumed *et al.* (2021), tem sido adotado com menor frequência. Eles afirmam que "Ciência de Dados é um termo muito mais comumente usado que Mineração de Dados no contexto de descoberta de conhecimento" (tradução nossa).

Segundo Azevedo e Santos (2008), tem havido grande crescimento e consolidação no campo de Mineração de Dados e alguns dos esforços têm sido na busca para se estabelecer padrões na área. Estas buscas resultaram na criação de diversas metodologias com padrões que definem um conjunto de passos com o objetivo de guiar implementações de aplicações de Mineração de Dados. As mais relevantes são o *Knowledge Discovery in Databases* (KDD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), o *Sample, Explore, Modify, Model, and Assess* (SEMMA) (SAS Institute, 2017) e o *Cross-Industry Standard Process for Data Mining* (CRISP-DM) (WIRTH; HIPP, 2000).

Martínez-Plumed *et al.* (2021) apresentam um modelo por eles desenvolvido, o *Data Science Trajectories* (DST), que pode ser considerado como uma evolução dos ciclos dispostos no CRISP-DM, KDD e SEMMA. Ele expande o CRISP-DM ao incluir atividades exploratórias, como exploração de objetivos, exploração das fontes de dados e exploração de valor de dados. Os autores pontuam que quando projetos de Ciência de Dados se tornam mais exploratórios, o projeto pode seguir caminhos variados, exigindo modelos mais flexíveis. No artigo é sugerido um modelo baseado em trajetórias e como ele pode ser usado para categorizar projetos de Ciência de Dados (direcionados por objetivos, exploratórios ou de gerenciamento de dados).

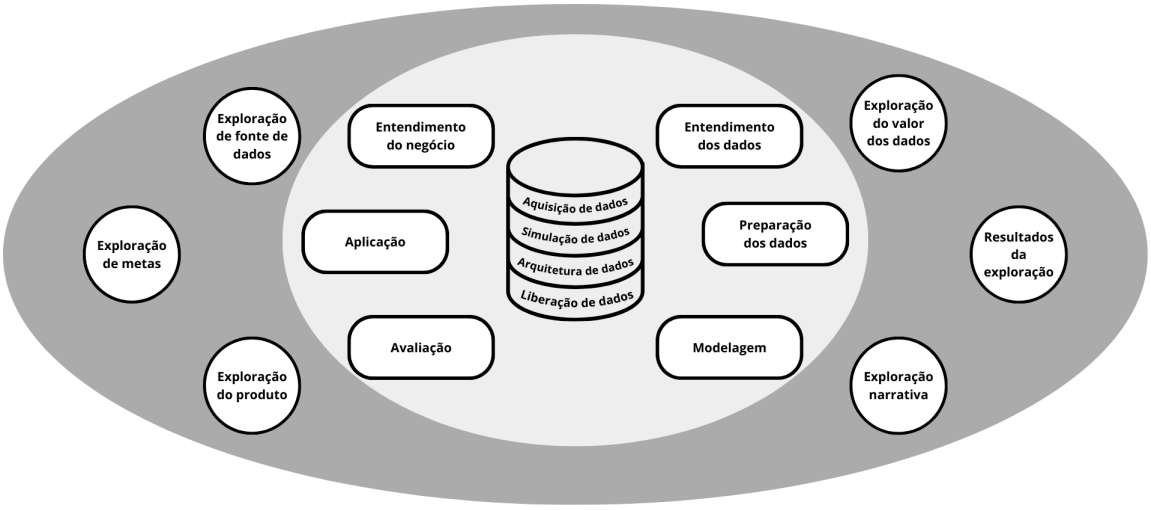

Figura 2 – Exemplo de trajetória em um projeto de Ciência de Dados.

Fonte: Adaptado de Martínez-Plumed *et al.* (2021). Tradução nossa.

A Figura 2 apresenta um diagrama geral que inclui as possíveis atividades que podem fazer parte de um projeto de Ciência de Dados. O mapa possui um formato elíptico com um círculo no centro onde, na parte externa ao círculo estão dispostas as atividades exploratórias; na parte interna do círculo estão dispostas as atividades orientadas a objetivos do CRISP-DM; e no centro as atividades de gerenciamento de dados. Por meio dele é possível traçar diversas trajetórias diferentes a partir das especificidades de cada projeto.

Em síntese, DST auxilia os cientistas de dados a planejar e organizar suas atividades de maneira mais eficaz, permitindo a inclusão ou exclusão de etapas conforme necessário. Sua abordagem flexível ajuda no gerenciamento da variedade de caminhos que um projeto de Ciência de Dados pode tomar.

2.3 Inteligência Artificial e Aprendizado de Máquina

Para Boden (2016, p. 1-2), Inteligência Artificial tenta fazer com que computadores executem as coisas que a mente pode fazer. Algumas delas, como raciocínio, são descritas como "inteligentes", ao contrário de outras, como a visão. Porém, todas elas envolvem habilidades psicológicas que nos permite concluir um certo objetivo. Como exemplos, podemos citar a percepção, a associação, a predição, o planejamento e o controle motor. Segundo a autora, IA tem dois objetivos: o tecnológico, de usar computadores para que eles realizem tarefas úteis; e o científico, de usar conceitos de IA e modelos para ajudar a responder questões sobre humanos e outros seres vivos.

Faceli *et al.* (2021) explicam que desde os anos 1970, houve uma expansão do uso da IA para a solução de problemas reais. O constante aumento da complexidade dos problemas a serem solucionados computacionalmente e o aumento do volume de dados gerados por diversos setores, motivou o desenvolvimento de ferramentas computacionais sofisticadas e autônomas para aquisição de conhecimento. Nesse contexto, aparece o Aprendizado de Máquina, uma subárea da Inteligência Artificial. Goodfellow, Bengio e Courville (2016) definem Aprendizado de Máquina como a habilidade de sistemas de IA adquirirem seu próprio conhecimento por meio da extração de padrões de dados brutos.

De acordo com Cozman e Neri (2021), a área de Aprendizado de Máquina tem sido predominantemente impulsionada por métodos que identificam padrões em grandes bases de dados. Isso tem levado a uma ênfase significativa no aprendizado de máquina estatístico, no qual os dados desempenham um papel central. Além das técnicas estatísticas, abordagens inspiradas em conceitos biológicos também desempenham um papel crucial, destacando-se as redes neurais. Os autores destacam que estas são funções compostas por camadas de neurônios artificiais que conseguem extrair padrões de alta complexidade a partir de dados, o que viabiliza tarefas de difícil automação.

O artigo da Brasil (2023) mostra que soluções que utilizam técnicas de AM têm sido amplamente utilizados em tarefas como o do diagnóstico de doenças, escolhas de melhores rotas e tradução de textos.

Tais tarefas podem ser divididas tecnicamente em Preditivas e Descritivas. De acordo com o Google (ca. 2020), a análise preditiva é uma forma de análise de dados que visa responder à pergunta: "O que pode acontecer depois?". Trata-se do processo de usar dados para prever resultados futuros. Neste processo, são usadas técnicas de análise de

dados, AM, IA e modelos estatísticos para prever padrões que tenham capacidade de prever comportamentos futuros. Já a análise descritiva tenta responder "O que aconteceu?". Faceli *et al.* (2021, p. 38-39) explica que ao invés de predizer um valor, padrões são extraídos de um conjunto de dados. Esses algoritmos não usam o conhecimento "supervisores externos", utilizando, assim, o paradigma de aprendizado não supervisionado. Tarefas descritivas também são utilizadas para associar valores de um subconjunto de atributos preditivos a valores de outro subconjunto.

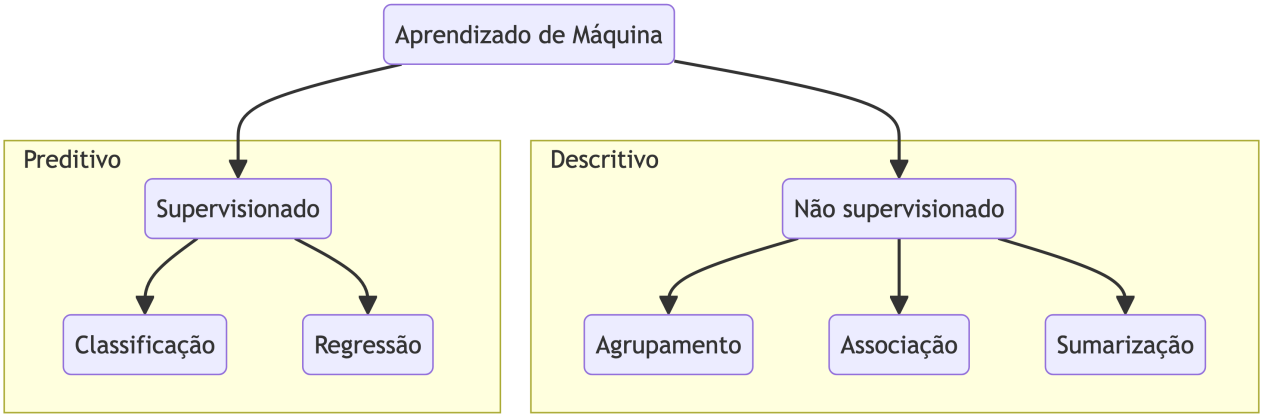

Figura 3 – Hierarquia de aprendizado.

Fonte: Adaptado de Faceli *et al.* (2021, p. 40).

Na Figura 3 são apresentadas, de forma hierárquica, as categorias de algoritmos de aprendizado supervisionado (tarefas preditivas) e não supervisionado (tarefas descritivas). A primeira categoria é dividida em Classificação e Regressão. Já a segunda, em Agrupamento, Associação e Sumarização.

Considerando a solução a ser proposta neste trabalho, direcionamos o foco para tarefas preditivas, mais especificamente em tarefas de classificação de textos.

2.3.1 Categorias de Algoritmos de Classificação

O livro de Herrera *et al.* (2016, p. 11-15) explica os algoritmos de classificação de maneira didática. De acordo com os autores, os algoritmos de classificação visam criar, a partir de dados previamente rotulados, um modelo capaz de predizer o rótulo (ou classe) para outros dados nunca antes vistos. Para melhor entendimento, os autores pontuam que pode-se pensar nos dados como sendo uma tabela que representa um conjunto de atributos. Esses atributos são divididos em dois subconjuntos (ou colunas): o primeiro contém os atributos de entrada, ou seja, as variáveis que atuarão como preditores; o segundo contém os atributos de saída, a classe ou rótulo atribuído a cada instância. A correlação entre os

atributos de entrada e de saída é analisada pelo modelo que, uma vez treinado, pode ser usado para processar novas amostras de dados obtendo uma previsão de classe.

Herrera *et al.* (2016, p. 11-15) também listam cinco categorias diferentes de classificação. Elas não se aplicam apenas à classificação de textos mas também a outros tipos de dados como imagens, vídeos e áudios. Estas categorias são resumidas na Tabela 1 considerando a quantidade de saídas e seus tipos.

- **Classificação Binária:** Possuem apenas um atributo de saída e podem assumir dois valores diferentes: verdadeiro ou falso. Exemplo de aplicação: Diagnóstico de doenças (Doente ou Não doente);

- **Classificação Multiclasse:** Também possuem apenas um atributo de saída, porém, podem assumir qualquer valor dentro de um conjunto finito e discreto no contexto de cada aplicação à qual se dedica o algoritmo. Exemplo de aplicação: Tipos de macarrão (*Fettuccine* ou *Penne* ou *Spagetti*);

- **Classificação Multirrótulo:** Ao contrário das anteriores, cada uma das instâncias de dados tem associado um vetor de saídas, em vez de apenas um valor. O tamanho do vetor é baseado na quantidade de rótulos diferentes no conjunto de dados. Cada elemento do vetor é um valor binário que representa se o rótulo é relevante ou não para a amostra. Exemplo de aplicação: Um filme pode ser comédia e suspense, outro pode ser terror e um terceiro pode ser ação, ficção científica e drama;

- **Classificação Multidimensional:** Assim como Classificadores Multirrótulo, possuem um vetor de saída associado a cada instância, ao invés de apenas um valor. Contudo, cada item deste vetor pode assumir qualquer valor de um conjunto predefinido, não se limitando a ser binário. Exemplo de aplicação: Categorização e sub-categorização de músicas (*Rock* [*Pop Rock* e/ou *Indie Rock* e/ou *Ska*] e/ou *Funk* [*Groovie* e/ou *Soul Music* e/ou *Funk Rock*]);

- **Aprendizado de Múltiplas Instâncias:** Aprende um rótulo de classe comum para um conjunto de vetores de recursos de entrada. É um problema muito diferente dos demais, pois cada instância de dados lógicos é definida não apenas por um vetor de características de entrada, mas por uma coleção de instâncias físicas, cada uma com um conjunto de atributos de entrada.

Como pode ser visto na Tabela 1, o número de saída e o tipo de saída determina a categoria de classificação. Por exemplo, na classificação binária cada instância tem apenas uma saída (primeira coluna) e esta é binária (segunda coluna). Já na multirrótulo, há a possibilidade de n saídas por instância, sendo elas também binárias.

Número de saídas	Tipo de saída	Categoria de classificação
1 por instância	Binária	Binária
1 por instância	Multivalorada	Multiclasse
n por instância	Binária	Multirrótulo
n por instância	Multivalorada	Multidimensional
1 por n instâncias	Binária/Multivalorada	Múltiplas Instâncias

Tabela 1 – Categorias de classificação de acordo com a saída a ser prevista.

Fonte: Adaptado de Herrera *et al.* (2016).

2.3.2 Métricas de Desempenho de Classificadores

A avaliação de modelos de Aprendizado de Máquina é uma etapa crucial no desenvolvimento de sistemas preditivos, especialmente para classificadores. A escolha das métricas de desempenho adequadas é fundamental para garantir que o modelo atenda às necessidades do problema em questão.

As subseções a seguir possuem as definições de Matriz de Confusão, Acurácia, Precisão, Revocação e *F1-score*, apresentadas por Cabral (2021, p. 42-43); além da definição de *Subset Accuracy* por Herrera *et al.* (2016); e as definições de Macro e Micromedia por Joachims (2002) e Tarekegn, Giacobini e Michalak (2021).

2.3.2.1 Matriz de Confusão

Trata-se de uma tabela que auxilia na visualização do desempenho de algoritmos que possuem o objetivo de prever as classes de uma variável, como mostrado na Tabela 2.

- **Verdadeiro Positivo:** quantidade de dados corretamente previstos como pertencentes à classe;
- **Verdadeiro Negativo:** quantidade de dados corretamente rejeitados como pertencentes à classe;
- **Falso Positivo:** quantidade de dados incorretamente previstos como pertencentes à classe;
- **Falso Negativo:** quantidade de dados incorretamente rejeitados como pertencentes à classe.

A Tabela 2 representa a Matriz de Confusão, que é composta de duas linhas e duas colunas contendo o número de Falsos Positivos (FP), Falsos Negativos (FN), Verdadeiros

	Previsto Positivo	**Previsto Negativo**
Real Positivo	VP	FN
Real Negativo	FP	VN

Tabela 2 – Matriz de Confusão.

Positivos (VP) e Verdadeiros Negativos (VN). Ela resume os resultados das previsões do modelo em comparação com os resultados reais.

Como será visto adiante, a Matriz de Confusão é uma ferramenta que auxilia no cálculo das métricas de desempenho como Acurácia, Precisão, Revocação e *F1-score*.

2.3.2.2 Acurácia

A acurácia refere-se à proporção de previsões corretas em relação ao total de resultados previstos.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.1}$$

2.3.2.3 Precisão

A precisão avalia a proporção de verdadeiros positivos em relação ao total de elementos recuperados corretamente.

$$Precisão = \frac{VP}{VP + FP} \tag{2.2}$$

2.3.2.4 Revocação

A revocação mede falsos negativos contra verdadeiros positivos, ou seja, é o oposto da precisão.

$$Revocação = \frac{VP}{VP + FN} \tag{2.3}$$

2.3.2.5 *F1-score*

O *F1-score* é a média harmônica de precisão e revocação.

$$F1\text{-}score = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} = \frac{2 \times \text{VP}}{2 \times \text{VP} + \text{FP} + \text{FN}} \tag{2.4}$$

No contexto de Classificação Multirrótulo, outra métrica se torna relevante, a Acurácia de Subconjuntos, ou *Subset Accuracy*.

2.3.2.6 Subset Accuracy

De acordo com Herrera *et al.* (2016, p. 57), esta é a possivelmente a métrica mais estrita. Nela, todos os rótulos (preditos e reais) são comparados avaliando-se a igualdade total. Quanto maior a quantidade de rótulos, menor a chance do classificador produzir de forma exata a saída correta.

$$SubsetAccuracy = \frac{1}{n}\sum_{i=1}^{n}[\![\mathbf{Y}_i = \mathbf{Z}_i]\!] \tag{2.5}$$

onde:

- n é o número total de amostras,
- $\mathbf{Y}_i$ é o vetor de rótulos verdadeiros da i-ésima amostra,
- $\mathbf{Z}_i$ é o vetor de rótulos previstos da i-ésima amostra,

2.3.2.7 Macro e Micromedia

Ainda no contexto de Classificação Multirrótulo, Joachims (2002, p. 30-31) explica que frequentemente é útil calcular o desempenho médio de um algoritmo de aprendizagem em múltiplos conjuntos de treinamento. No caso particular da Classificação Multirrótulo, geralmente há interesse de identificar quão bem todos os rótulos podem ser previstos, não apenas um único. Para conseguir isso, é possível se utilizar da macromedia (*macro-averaging*) e micromedia (*micro-averaging*). Por meio delas é possível que os resultados de m tarefas binárias possam ter sua média calculada para se obter um único valor de performance.

Joachims (2002, p. 31) define que a macromedia corresponde à maneira padrão de calcular uma média aritmética. A medida de desempenho (precisão, revocação, etc.) é calculada separadamente para cada um dos m experimentos. A média é calculada como a média aritmética da medida de desempenho em todos os experimentos. Como exemplo, a métrica *F1-score* seria calculada da seguinte forma:

$$F1^{macro} = \frac{1}{m}\sum_{i=1}^{m} F1_i \tag{2.6}$$

Com relação à micromedia, Joachims (2002, p. 31) explica que ela não calcula a média da medida de desempenho resultante, mas, em vez disso, faz a média das matrizes de confusão. Para cada célula da matriz, a média aritmética é computada, levando a uma matriz de confusão média. Com base nessa tabela, a medida de desempenho é computada. Novamente, como exemplo usamos a métrica *F1-score*:

$$F1^{micro} = \frac{2 \times \mathrm{VP}^{avg}}{2 \times \mathrm{VP}^{avg} + \mathrm{FP}^{avg} + \mathrm{FN}^{avg}} \tag{2.7}$$

Tarekegn, Giacobini e Michalak (2021, p. 7) enfatizam que quanto maior os valores da Acurácia, Precisão, Revocação e *F1-score*, maior a performance do algoritmo de aprendizagem.

Joachims (2002, p. 10 e 12) afirma que a maior parte das tarefas de classificação de texto recaem sobre a Classificação Multirrótulo. Ele também explica que na classificação de textos a entrada de dados consiste de linguagem natural, ou seja, línguas faladas por humanos. Um dos problemas fundamentais quando se lida com linguagens naturais é que o contexto tem uma influência importante no sentido de um texto. A palavra "banco" é usada como exemplo pelo autor, já que pode significar uma instituição financeira, um móvel ou ao acúmulo de areia às margens de um rio. Tendo isso em vista, é importante analisarmos as técnicas atuais de Processamento de Linguagens Naturais.

2.4 Processamento de Línguas Naturais

O Processamento de Línguas Naturais é, de acordo com Caseli e Nunes (2024, p. 8), "*um campo de pesquisa que tem como objetivo investigar e propor métodos e sistemas de processamento computacional da linguagem humana*". O termo tem origem no inglês, *Natural Language Processing* (NLP), e também é comumente descrito como Processamento de Linguagens Naturais. Neste trabalho, demos preferência ao termo "língua natural" como uma tradução mais precisa da expressão "*natural language*", uma vez que "Natural", neste caso, se refere às línguas faladas por humanos, ao contrário das demais linguagens (de programação, matemáticas, etc.).

O PLN é dividido em duas áreas: *Natural Language Understanding* (NLU) e *Natural Language Generation* (NLG). A NLU é relativa ao processamento que busca analisar e interpretar a língua. Ela é utilizada, por exemplo, em uma interação com *chatbots*, onde uma entrada de texto é processada para que sistema possa decidir o que deve fazer: fornecer uma resposta ou executar uma ação. No caso da NLG, o objetivo é a geração de linguagem natural, como ocorre com o ChatGPT (CASELI; NUNES, 2024, p. 10-11).

Do ponto de vista computacional, Caseli e Nunes (2024, p. 76-85) também apontam as aplicações de PLN podem ser desenvolvidas seguindo-se as seguintes etapas:

- **Pré-processamento:** Nessa etapa, algumas das tarefas comuns são: segmentação do texto em sentenças (sentenciação), separação de palavras (tokenização), tokenização em sub-palavras (vetorização de *subtokens*), normalização de palavras (lematização e radicalização), entre outras;

- **Processamento de conteúdo dos textos:** Aqui, ocorre a etiquetagem morfossintática das palavras em relação às suas classes gramaticais (tarefa de PoS tagging);

- **Análise morfológica:** Por fim, realiza-se a anotação automática de atributos morfológicos (tarefa de anotação de *feats* ou *features* morfológicas).

Jurafsky e Martin (2024, p. 4-5) ajuda na definição dos processos mais comuns que compõem a etapa de pré-processamento. Segundo eles, a normalização de textos significa convertê-los a uma forma padronizada. Uma das tarefas iniciais ao se um texto é separar ou, *tokenizar*, palavras, frases ou até mesmo caracteres, dependendo da granularidade desejada. Cada uma dessas partes é chamada de *token*. Outra parte da normalização é a *lematização*, a tarefa de determinar que duas palavras possuem a mesma raiz, ou seja, a mesma origem, apesar de serem diferentes. Os autores usam como exemplo as palavras "aprender", "aprendendo" e "aprendizado". Uma simplificação da *lematização* é a *stemização*, na qual os sufixos são removidos das palavras. De acordo com os autores, a normalização também inclui a segmentação de sentenças, ou sentenciação, na qual o texto é dividido em frases utilizando-se limitadores como os pontos final, de exclamação e de interrogação.

Na etiquetagem morfossintática, também conhecida como *Part-of-speech tagging* ou *PoS tagging*, as etiquetas gramaticais são associadas a cada palavra de um texto com base em sua classe gramatical e características morfológicas. Atuam na identificação das funções sintáticas e morfológicas das palavras dentro de uma sentença. Trata-se de algo essencial para aplicações como a geração de resumos automatizados. (CASELI; NUNES, 2024, p. 82).

Concluindo o processo, Caseli e Nunes (2024, p. 84) também explicam que a anotação de atributos morfológicos consiste em identificar informações sobre as características gramaticais e morfológicas das palavras em um texto, como número, gênero, modo, tempo verbal, pessoa, entre outras. Seu objetivo é capturar e estruturar essas informações gramaticais de forma organizada, permitindo que algoritmos de PLN e AM possam processar adequadamente a estrutura e as relações linguísticas presentes em textos.

2.5 Indexação de Proposições Legislativas no Congresso Nacional

Proposição Legislativa é toda matéria sujeita a deliberação na Câmara dos Deputados, como Projetos de Lei e Propostas de Emenda à Constituição (BRASIL, 2024a). Proposições estão sujeitas ao processo legislativo, que compreende a elaboração, análise e votação de diversos tipos de propostas e cada um deles segue uma tramitação diferente. Como exemplo, podemos tomar a tramitação de Projetos de Lei.

Segundo (BRASIL, 2024c), o processo de tramitação de um PL se inicia com a sua publicação. Estes projetos podem ser propostos por deputados ou senadores; por qualquer comissão da Câmara, do Senado ou do Congresso Nacional; pelo presidente da República; pelo Supremo Tribunal Federal (STF); pelos tribunais superiores; pelo procurador-geral da República; e pelos cidadãos por meio de iniciativas populares. Após a publicação, ocorre a análise do conteúdo do projeto por Comissões Permanentes ou Comissões Especiais. Na fase de análise de admissibilidade, as propostas que criam gastos ou tratam de finanças públicas devem ser aprovadas pela Comissão de Finanças e Tributação (CFT) e, por fim, todas as propostas são avaliadas pela Comissão de Constituição e Justiça e de Cidadania (CCJC) segundo sua consonância com a Constituição. A votação no Plenário é o próximo passo, caso alguma das comissões citadas anteriormente não aprove a proposição. Em seguida, após aprovação no Plenário a proposição segue para o Senado, onde será analisada e votada, ou para sanção ou veto do Presidente da República. Na Figura 4, o processo legislativo aqui descrito é detalhado por meio de um infográfico no qual são apresentadas os atores, comissões, casas legislativas envolvidas até a sanção pela Presidência da República.

2.5.1 A Lei de Acesso à Informação

Em 2011, foi sancionada a Lei nº 12.527 (Lei de Acesso à Informação), que, de acordo com seu art. 3º, tem como diretriz assegurar o direito de acesso à informação observando-se a publicidade como regra e o sigilo como exceção; a divulgação de informações de interesse público, independentemente de solicitações; o desenvolvimento da cultura de transparência na administração pública e do seu controle social (BRASIL, 2011).

Para definir as diretrizes para o cumprimento desta lei, as casas parlamentares, Câmara dos Deputados e Senado Federal, decretaram, respectivamente, o Ato da Mesa nº 45/2012 (BRASIL, 2012b) e o Ato da Comissão Diretora nº 9/2012 (BRASIL, 2012a). Além disso, tendo em vista a necessidade de adoção de procedimentos específicos para atender a LAI em seu art. 9º, inciso I, alínea b, que assegura o acesso a informações públicas com o fim de informar sobre a tramitação de documentos, a Câmara dos Deputados promulgou o Ato da Mesa nº 80/2013 (BRASIL, 2013b). Ele delibera sobre a Política de Indexação de Conteúdos Informacionais da Câmara dos Deputados e define o Tesauro da Câmara dos Deputados (TECAD) como o documento a ser seguido como instrumento de linguagem documentária no processo de indexação.

O Ato da Mesa nº 80/2013 também estabelece o Centro de Documentação e Informação da Câmara dos Deputados como o órgão responsável por coordenar e supervisionar a implantação da Política de Indexação de Conteúdos Informacionais da Câmara dos Deputados.

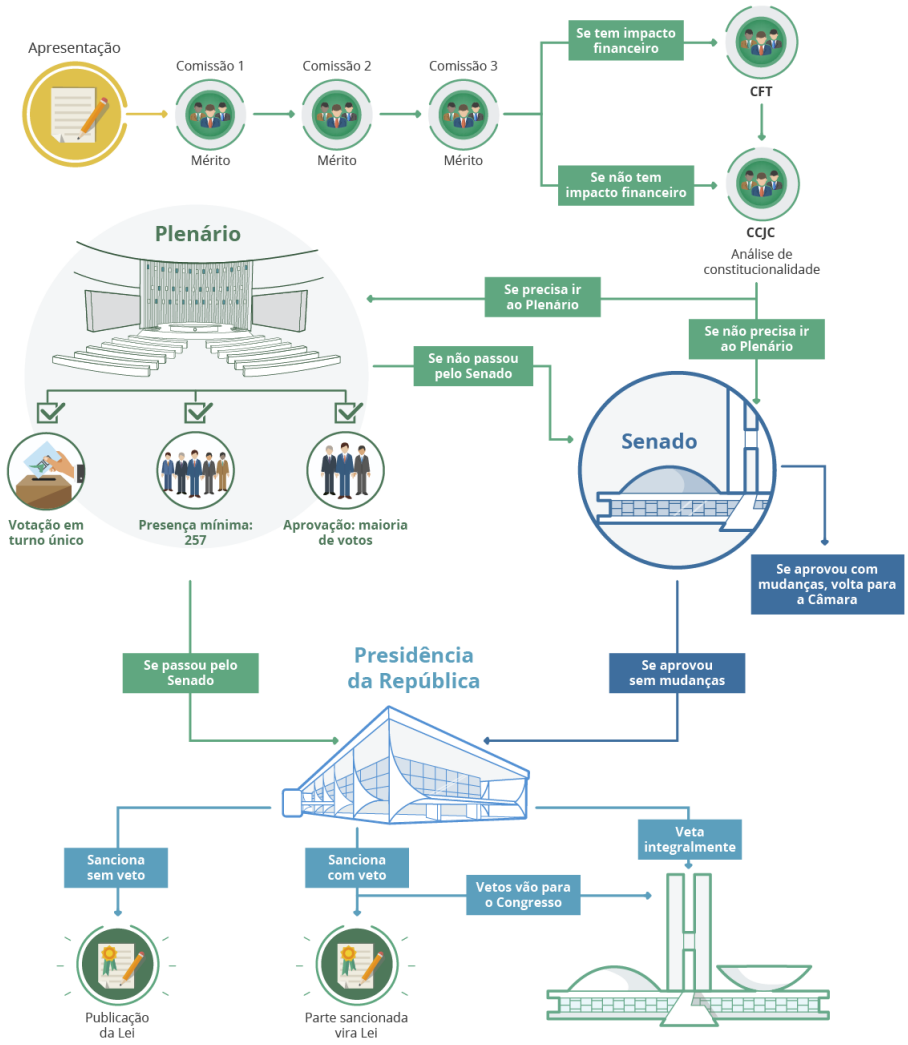

Figura 4 – Processo Legislativo: Projeto de Lei.

Fonte: Brasil (2024c).

2.5.2 O Centro de Documentação e Informação da Câmara dos Deputados

O CEDI foi criado por meio da Resolução da Câmara dos Deputados nº 20/1971 com a finalidade de, entre outras, coordenar, controlar e dirigir as atividades de informação; registrar a tramitação das proposições; arquivar e avaliar documentos; e editar publicações (BRASIL, 1971). A logomarca do CEDI é apresentada na Figura 5.

Dentre as atividades realizadas pelo Centro, ressaltam-se a curadoria da Biblioteca da Câmara dos Deputados, que conta com acervo de milhares de obras raras e uma centena de periódicos; a manutenção do Arquivo da Câmara dos Deputados, que gera documentos históricos e facilita seu acesso para a população; e a responsabilidade de conservação e restauração do acervo cultural da Câmara dos Deputados (BRASIL, ca. 2020). Uma demonstração prática da sua importância para a preservação do patrimônio histórico

Figura 5 – Logomarca CEDI.

Fonte: Portal da Câmara dos Deputados (2024).

nacional é o trabalho de recuperação e restauração de itens avariados durante a invasão do Congresso Nacional, em 8 de janeiro de 2023 (BBC NEWS BRASIL, 2023). O livro "Restaurando a Democracia: A preservação da memória da Câmara para futuras gerações" (FILHO *et al.*, 2024) revela os trabalhos realizados pelo CEDI, por meio da Coordenação de Preservação de Conteúdos Informacionais (COBEC) para reintegração de obras aos acervos e seus espaços de exibição.

Em 2013, a Câmara dos Deputados, por meio do Ato da Mesa nº 125, Anexo I, definiu a estrutura administrativa do CEDI (BRASIL, 2013a, p. 32-33). Dentre as Coordenações pertencentes ao CEDI, a que possui maior relevância para dados gerados no processo legislativo é a Coordenação de Organização da Informação Legislativa (CELEG). A estrutura da CELEG é exibida na Figura 6, na qual são mostrados todos os Serviços e Seções sob o seu controle. Em seu Anexo III, este ato também estabelece que a indexação de proposições legislativas e a descrição de seu conteúdo é de responsabilidade da Seção de Indexação de Matérias Legislativas (BRASIL, 2013a, p. 63).

Atualmente, a SIDEX é composta por quatro funcionárias: três servidoras públicas concursadas com formação superior em Biblioteconomia e uma estagiária. Internamente essas profissionais possuem o título de "Analistas Legislativos – Documentação e Informação Legislativa". Em média, um servidor indexa cento e trinta projetos por mês. Por meio do trabalho realizado por essa seção, cidadãos que utilizam o site da Câmara dos Deputados conseguem encontrar proposições submetidas pelos seus representantes (SIDEX, 2024).

Existem dezenas de tipos de proposições no processo legislativo. A SIDEX é

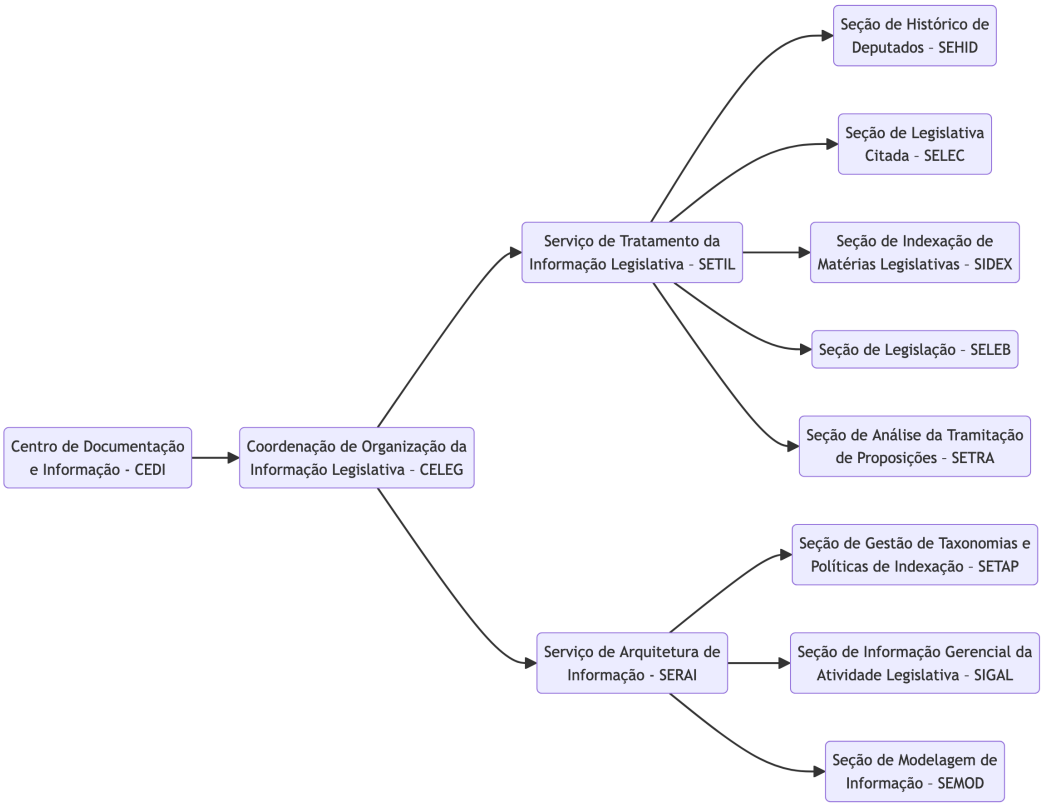

Figura 6 – Estrutura da CELEG.

Fonte: Elaborada pelo autor.

responsável pela análise dos seguintes (BRASIL, 2016):

- Avisos
- Consulta
- Emenda do Senado
- Medida Provisória
- Projeto de Decreto Legislativo
- Proposta de Emenda à Constituição
- Proposta de Fiscalização e Controle
- Projeto de Lei
- Projeto de Lei Complementar
- Projeto de Lei de Conversão
- Projeto de Resolução
- Projeto de Resolução do Congresso Nacional
- Requerimento de Instituição de CPI
- Representação
- Mensagens
- Indicação
- Requerimento de Informações
- Solicitação de Informação ao TCU

O sistema interno utilizado no processo de indexação de proposições é o Sistema

de Informações Legislativas. Trata-se de um sistema usado por várias seções da CELEG e foi desenvolvido com o intuito de automatizar todas as etapas do processo legislativo da Câmara dos Deputados. Ele é responsável pelo cadastro e acompanhamento do ciclo de vida de proposições até seu arquivamento ou transformação em lei. O SILEG também gerencia a agenda das comissões, com a publicação da pauta e do resultado das reuniões e votações (BRASIL, 2024d). Na Figura 7 é apresentada uma captura de tela do SILEG destacando pastas de trabalho utilizadas pela SIDEX no processo de indexação.

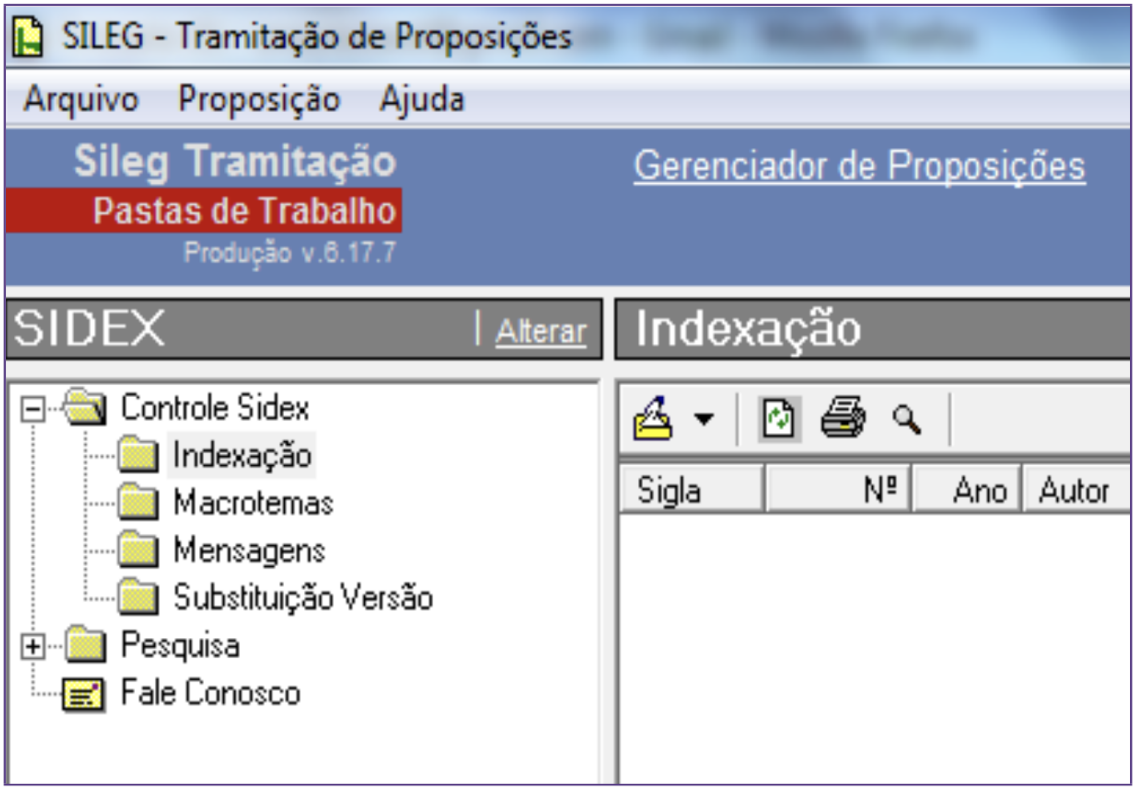

Figura 7 – Tela do SILEG.

Fonte: Brasil (2016).

Para que as proposições estejam disponíveis para a SIDEX, é necessário que haja o registro das mesmas no sistema SILEG. Esta etapa preliminar pode ser resumida da seguinte forma:

1. As proposições, após serem criadas, são registradas em um aplicativo chamado Autenticador. Ele tem a função de gerar um código de barras a partir do conteúdo de um documento eletrônico, para ser impresso em papel. Este código é alterado sempre que o documento eletrônico, por algum motivo, é modificado;

2. A proposição é impressa e protocolada, e sua versão eletrônica é inserida na base do SILEG;

3. Uma vez no sistema, o documento eletrônico torna-se público e disponível na Internet. Nesse momento, a proposição é numerada e enviada para a caixa de trabalho de setores da Casa que lidam com a mesma, como a SIDEX.

O processo de tratamento da informação de proposições legislativas pode ser dividido em seis etapas, como demonstra o Manual de Indexação de Proposição Legislativa (BRASIL, 2016). São elas a pesquisa de proposição, conferência da ementa, indexação, elaboração da explicação de ementa, elaboração de notas e seleção das áreas temáticas. Tais etapas podem ou não ser aplicadas a depender do tipo de proposição a ser analisada, com exceção da conferência de ementa, que acontece para todos eles.

De forma sucinta, o processo pode ser descrito da seguinte forma:

1. **Pesquisa de Proposições:** O servidor faz a leitura do documento de inteiro teor da proposição e verifica se ele já existe na base do SILEG. Esta pesquisa é importante para padronizar os termos a serem usados na indexação;

2. **Conferência da Ementa:** A ementa é preenchida no SILEG pelo Gabinete do Deputado ou pela Secretaria-Geral da Mesa (SGM) que, segundo Brasil (2020, p. 86), é o "principal órgão de assessoramento da Mesa Diretora, também responsável pelo recebimento e encaminhamento das proposições e pelo acompanhamento dos trabalhos legislativos". A SIDEX compara a ementa registrada no sistema com a ementa do inteiro teor da proposição e, caso necessário, atualiza o sistema com o texto correto;

3. **Indexação:** Na etapa de indexação, a SIDEX analisa os assuntos tratados no inteiro teor da proposição e seleciona os termos que representam esses assuntos. Os termos a serem utilizados estão presentes no Tesauro da Câmara dos Deputados (TECAD). Quando nenhum dos termos presentes no TECAD representam o assunto da proposição, é realizado um pedido de inclusão de um novo termo;

4. **Elaboração da Explicação da Ementa:** A explicação da ementa é necessária em dois casos distintos: ementas que trazem o número da norma a ser alterada, mas não as informações sobre a finalidade do projeto; e ementas que trazem o objetivo do projeto, mas não o número da lei que está sendo modificada;

5. **Elaboração de Notas:** Em alguns casos, é necessária a adição de informações adicionais, chamadas notas, nos seguintes casos:

    - **Ementa que apresenta erro de digitação:** A ementa precisa ser transcrita com os erros corrigidos;

- **Registro de Termo Livre:** Necessário nos casos em que termos não estejam presentes na ementa mas são indispensáveis para representar o assunto do projeto, além de não haver sinônimos ou outro termo no TECAD para representar a informação;
- **Apelidos dados pela imprensa ao projeto de lei:** Acréscimo de nomes não-oficiais dados a algumas proposições, como PEC da Bengala;

6. **Seleção das Áreas Temáticas:** Temas da Proposição, Macrotemas ou Áreas Temáticas são os assuntos nos quais as proposições em análise são classificadas. Os temas auxiliam o sistema de busca e facilitam a recuperação da informação. Alguns exemplos de Áreas Temáticas disponíveis no SILEG são: Comunicações, Economia, Educação e Saúde. Atualmente, a responsável pelo estudo dos temas é a Seção de Gestão de Taxonomias e Políticas de Indexação (SETAP).

Após o término do tratamento da informação, os dados são salvos no SILEG e a proposição impressa é entregue à Chefia da Seção para que seja realizada a revisão da indexação.

3 TRABALHOS RELACIONADOS

Na literatura, diversos trabalhos buscam automatizar processos de classificação de textos utilizando técnicas de Aprendizagem de Máquina e Processamento de Línguas Naturais. O objetivo desta seção é descrever alguns destes trabalhos apesentando as diversas estratégias por eles abordadas.

Andrade (2015) objetivou criar um modelo para classificação automática de denúncias feitas pelo portal da Controladoria Geral da União (CGU) a fim de demonstrar sua viabilidade em comparação à triagem manual feita por funcionários do órgão. Para isso, inicialmente a autora utilizou os algoritmos *Random Forest*, *Decision Tree*, *Naive Bayes* e SVM, sem obter resultados satisfatórios. Dentre 64 categorias possíveis, o melhor resultado obtido foi o SVM, que alcançou uma precisão de 59%. Utilizou-se, então outra abordagem de classificador baseado em árvore de Huffman à qual se deu o nome de CAH+MDL. Após diversos ajustes, esta nova abordagem levou a uma precisão de 84%. A autora calculou a taxa de acertos da triagem manual por meio da avaliação da quantidade de reencaminhamentos de denúncias triadas para uma nova área, que era de 10%, levando a uma assertividade de 90%. Com uma diferença de 6%, os processos manual e automático tiveram assertividade próxima, o que demonstrou a viabilidade da solução.

Alfiani, Imamah e Yuhana (2021) realizaram um estudo sobre classificação de materiais de aprendizagem utilizando técnicas de Classificação Multirrótulo. Materiais de aprendizagem são recursos científicos das áreas de química, física e biologia para estudantes do ensino médio. Os autores explicam que na Aprendizagem Adaptativa pode-se ajustar os materiais de aprendizagem com base nas habilidades individuais dos alunos. Tais materiais podem ser categorizados em tópicos e sub-tópicos, sendo possível que cada material pertença a múltiplos tópicos e sub-tópicos simultaneamente. O trabalho utilizou 448 materiais de aprendizagem que foram manualmente classificados por professores utilizando-se 10 tópicos e 41 sub-tópicos. Os tópicos e sub-tópicos (rótulos) foram extraídos do Regulamento do Ministério da Educação e Cultura da República da Indonésia. Experimentos foram realizados utilizando combinações entre métodos de transformação de problemas (*Binary Relevance*, *Label Powerset* e *Classifier Chain*) e algoritmos de classificação de rótulo único (*Naive Bayes*, *Random Forest* e SVM). Os resultados mostraram que a melhor combinação foi a de Binary Relevance com SVM. Esta combinação obteve as maiores Acurácia e *F-measure*, além do menor valor de *Hamming Loss*.

Morales-Hernández, Jagüey e Becerra-Alonso (2022) compararam a performance entre diferentes modelos de classificação multirrótulo para artigos de pesquisa alinhados aos Objetivos de Desenvolvimento Sustentável (ODS), que são uma coleção de 17 metas globais estabelecidas pela Assembleia Geral das Nações Unidas em 2015 (ONU, 2024). Eles

fazem parte de uma resolução chamada “Agenda 2030”, que fornece um plano de ação para a paz, prosperidade, e proteção do planeta. A motivação do trabalho foi obter um modelo que ajudasse instituições e pesquisadores a alinhar produtos científicos a estes objetivos. Foram feitas combinações entre métodos de transformação de problemas (*One-Versus-Rest*, *Binary Relevance*, *Label Powerset* e *Classifier Chain*) e algoritmos de classificação de rótulo único (*Naive Bayes*, *Logistic Regression*, *Random Forest* e SVM). Comparou-se, então, a performance destas combinações utilizando *datasets* balanceados e desbalanceados. Os resultados mostraram que *Label Powerset* com SVM foi a combinação que gerou a maior acurácia 91% e que o *Label Powerset* foi o melhor método de transformação independentemente do estado do *dataset* (balanceado ou desbalanceado).

Vale (2022) explorou técnicas de Aprendizado de Máquina e Mineração de Textos com o objetivo de criar um classificador automático que permita definir os temas de uma proposição legislativa baseando-se em documentos classificados manualmente pela Câmara dos Deputados. Os dados utilizados foram obtidos do Portal da Câmara dos Deputados e foi usado o texto de inteiro teor das proposições para treinar os modelos. Os textos de inteiro teor foram pre-processados e utilizou-se a combinação do método de transformação *MultiOutput* e de 14 algoritmos de classificação diferentes, entre eles *Linear SVC*, *Naive Bayes*, *Random Forest* e *Logistic Regression*. Os valores de acurácia obtidos para estas combinações não passaram de 44,36% com o método *MultiOutput* e *Linear SVC* com seleção de atributos. Diante de tais resultados, o autor utilizou uma implementação própria da abordagem *Multiple Single-label*, que consiste em transformar problemas de classificação multirrótulo em múltiplos problemas de classificação binária. Com esta abordagem, obteve-se uma acurácia máxima de 97,38% com o *Linear SVC* com seleção de atributos.

O trabalho apresentado por Vale (2022) compartilha o mesmo objetivo do nosso, porém diverge em alguns aspectos metodológicos, como mostrado na próxima seção. Por exemplo, o período definido pelo autor para os anos de submissão das proposições utilizadas foi de quatro anos (2018 a 2021). Nosso trabalho utiliza dados de onze anos, entre 2013 a 2023 visando uma melhor representação das classes; maior precisão dos modelos utilizados na avaliação; e uma diminuição de possíveis impactos causados por eventuais *outliers*.

Além disso, a metodologia de pré-processamento da base de dados difere entre os trabalhos em alguns aspectos. Um dos passos consiste em baixar um arquivo externo em formato *Portable Document Format* (PDF), *Hypertext Markup Language* (HTML) ou *Microsoft Word Binary File Format* (DOC) para cada proposição existente na base de dados e, em seguida, extrair o texto nele contido. Trata-se do Inteiro Teor da proposição, ou seja, a íntegra do texto que foi submetido pelo seu autor. Vale (2022) executa este passo em um estágio inicial do processo, antes da devida limpeza dos dados, o que pode ter levado a um aumento desnecessário no tempo de execução tendo em vista que houveram

proposições posteriormente removidas da base. Em contraste, nosso trabalho executa este passo ao fim do processo de limpeza, garantindo que apenas arquivos relevantes fossem baixados e tratados. Contudo, Vale adiciona em sua metodologia um passo para extração de textos de arquivos HTML, o que não é feito em nosso trabalho. A vasta maioria dos documentos de Inteiro Teor são disponibilizados em formato PDF, portanto optamos em excluir documentos em HTML e DOC, este último também foi descartado por Vale.

Outra distinção importante é o fato do nosso trabalho ter eliminado da base de dados os tipos de proposições que não são indexadas pelo CEDI, utilizando para isso a lista presente na seção 2.5.2. Com essa decisão garantimos que todos os dados utilizados na classificação foram submetidos ao processo de indexação, também detalhado na seção 2.5.2. Vale (2022) utilizou o critério de verificar se a proposição era ou não classificada por tema, ou seja, a proposição era descartada caso não houvesse pelo menos um tema associado a ela.

Também é importante ressaltar a distinção entre as métricas de avaliação utilizadas, uma vez que utilizamos o *F1-score* como métrica principal, ao contrário de Vale (2022) que preferiu medir a acurácia dos classificadores.

Em resumo, os trabalhos relacionados apresentados fornecem uma base valiosa para o desenvolvimento de novos projetos de classificação multirrótulo de textos, além de mostrar sua viabilidade técnica. Na próxima seção será apresentada a metodologia deste trabalho. Ela foi construída utilizando técnicas e práticas identificadas nos trabalhos aqui expostos, com o objetivo de apresentar uma abordagem eficaz para o problema.

4 MÉTODO

Este capítulo descreve a metodologia e os procedimentos utilizados na condução deste trabalho. Para responder às questões de pesquisa de maneira rigorosa e confiável, optou-se pela abordagem metodológica *Data Science Trajectories*, por ser mais flexível com relação ao caminho de exploração dos dados e às atividades realizadas para se atingir os objetivos, como explicado por Martínez-Plumed *et al.* (2021). Além disso, o trabalho busca contribuir com trabalhos futuros por meio da exposição das técnicas empregadas no desenvolvimento da solução.

Todos os recursos computacionais e bibliotecas utilizadas nesse trabalho são descritos no Anexo B e a trajetória desenhada para o trabalho é mostrada na Figura 8.

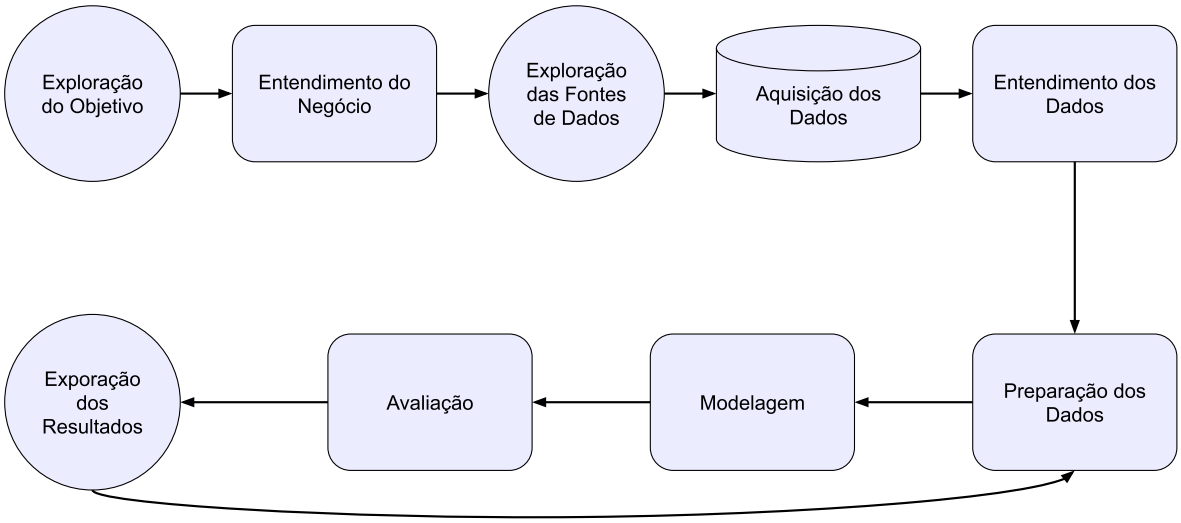

Figura 8 – Trajetória utilizada no trabalho.

Fonte: Elaborada pelo autor.

4.1 Exploração dos Objetivos

Considerando os desafios inerentes ao processo de indexação manual de proposições legislativas realizado pelo Centro de Documentação e Informação da Câmara dos Deputados, descritos no Capítulo 2, este trabalho objetiva:

1. **Propor uma solução para classificação automatizada das proposições por suas áreas temáticas utilizando técnicas de Mineração de Textos.** Apesar da classificação temática de proposições ser apenas parte do processo de indexação

realizado pelo CEDI, trata-se de uma etapa importante e que impacta a capacidade de pesquisa das proposições por cidadãos e funcionários da Câmara dos Deputados.

4.2 Entendimento do Negócio

O entendimento do processo empregado pela SIDEX na indexação de proposições legislativas se deu, em maior parte, pela leitura do Manual de Indexação de Proposição Legislativa (BRASIL, 2016). Trata-se de um documento institucional destinado a padronizar e orientar o processo de indexação de documentos legislativos. Ele que detalha o passo a passo que deve ser seguido pelos Analistas Legislativos que trabalham na SIDEX ao indexar as proposições. Seu objetivo é garantir a consistência e a precisão na recuperação de informações, facilitando o acesso e a pesquisa de documentos legislativos por meio de um sistema organizado de indexação. Além disso, a sessão foi contactada por *e-mail* para questionamentos relacionados à documentação utilizada como auxílio no processo de indexação; aos funcionários em atuação na seção; ao sistema utilizado no processo de indexação (SILEG); e ao processo de indexação em si. Os questionamentos feitos e as respostas da SIDEX encontram-se no Apêndice A.

Os trabalhos expostos no Capítulo 2 complementam o Manual de Indexação de Proposição Legislativa apresentando alguns desafios enfrentados em processos de indexação. Dentre outros, foi esclarecido que a indexação envolve julgamento da pessoa indexadora, e isso pode gerar discrepâncias na escolha de termos escolhidos (PINHEIRO, 1978); que os fatores mais relevantes relacionados à influência do profissional no processo de análise de assunto são a subjetividade, o conhecimento prévio e sua formação e experiência (NAVES, 2007); e que a análise de assunto é revestida de uma subjetividade característica uma vez que é realizada a partir da leitura do documento pela pessoa indexadora (RUBI, 2017).

O conhecimento adquirido nesta fase contribuiu para a tomada de decisões técnicas, como a filtragem de dados, explicitada adiante.

4.3 Exploração das Fontes de Dados

Nesta etapa, buscou-se entender de onde e como os dados seriam extraídos. Todas as Proposições Legislativas submetidas no Congresso Nacional são disponibilizadas no portal da Câmara dos Deputados. No período em que esta fase foi executada, haviam três formas para obtenção dos dados abertos: uma *Application Programming Interface* (API), uma página de Pesquisa Simplificada e arquivos separados por ano.

Após análise das opções, constatou-se que a API exportava apenas arquivos em XML e JSON. Além disso, não possuía informação sobre a temática, sendo necessária a consulta de outra API que checa a temática individualmente por id da proposição. Já a Pesquisa Simplificada, apesar de simples de usar e de exportar os resultados em diversos

Figura 9 – Proposições por ano de apresentação.

Fonte: Portal da Câmara dos Deputados (2024).

formatos, incluindo CSV, não disponibiliza todos os tipos de proposições classificados pela SIDEX e nem possui informação sobre temática.

Uma vez que as formas de obtenção dos dados mostradas anteriormente não atendiam às necessidades do trabalho, definiu-se o uso dos arquivos separados por ano. Estes são disponibilizados em cinco formatos diferentes, incluindo o CSV. Todos os dados necessários para este trabalho estão disponibilizados em três conjuntos de arquivos: "*Proposições por ano de apresentação*", "*Classificação temática das proposições*" e "*Autores das Proposições por ano de apresentação*".

- **Proposições:** Contém dados sobre proposições submetidas à Câmara dos Deputados, como identificador universal (URI), ano, tipo, ementa, URL para documento de Inteiro Teor, etc;
- **Classificação Temática das Proposições:** Cada registro corresponde a um tema na qual as proposições foram classificadas pelo CEDI. Uma proposição pode ser representada em várias linhas do arquivo, caso tenha sido classificada como tendo mais de um tema;
- **Autores das Proposições:** Possui dados relativos aos autores das proposições. Uma proposição pode ser representada em várias linhas do arquivo, caso possua mais

Figura 10 – Classificação temática das proposições.

Fonte: Portal da Câmara dos Deputados (2024).

Figura 11 – Autores das Proposições por ano de apresentação.

Fonte: Portal da Câmara dos Deputados (2024).

de um autor.

Os registros de proposições são datados a partir de 1934 da seguinte forma:

- **1934-1945:** Algumas das proposições que tramitaram na Câmara;
- **1946-2000:** Proposições de tipos que poderiam se tornar (ou se tornaram) leis e normas jurídicas;
- **2001-atualmente:** Todas as proposições tramitadas na Câmara.

4.4 *Extração*: Aquisição dos Dados

Foi definido o período entre os anos de 2013 e 2023 como amostra para este trabalho. Os arquivos foram baixados em formato CSV para cada ano do período, totalizando 33 arquivos. Para isso, foi utilizado um *web crawler* implementado em Python que baixou cada um dos arquivos e os organizou em pastas distintas, organizadas por tipo de documento (Proposições, Classificação Temática das Proposições e Autores das Proposições).

Nesta etapa, analisou-se um arquivo de cada conjunto. Os atributos de cada um deles são listados na Tabela 7 (Proposições por ano de apresentação), Tabela 8 (Autores das Proposições por ano de apresentação) e Tabela 9 (Classificação temática das Proposições), todos presentes no Anexo C. O objetivo foi entender com um pouco mais de detalhamento os dados disponibilizados pela Câmara dos Deputados e tomar decisões que seriam implementadas na próxima etapa, a de pré-processamento dos dados.

A grande maioria dos dados disponíveis não foi considerada necessária para a finalidade do trabalho, permitindo uma redução drástica do número de atributos, principalmente no conjunto Proposições. As informações relevantes para a classificação automática são o Inteiro Teor das proposições, os temas associados a elas e seus proponentes com os respectivos partidos e Unidades Federativas. Além disso, não julgou-se necessário, nesta etapa, realizar qualquer tipo de análise mais detalhada para entendimento da correlação dos atributos.

Notou-se que o conjunto Proposições não contém o texto de Inteiro Teor das proposições, apenas suas ementas. Como estes são textos muito curtos e que podem deixar de fora contextualizações relevantes, entendeu-se que utilizá-la poderia impactar o resultado final do trabalho. Contudo, a coluna “urlInteiroTeor” contém um *link* para este documento. Os formatos de cada documento disponibilizado podem variar entre PDF, HTML e DOC.

4.5 *Transformação*: Preparação dos Dados

Após o entendimento dos dados obtidos, iniciou-se seu pre-processamento. Para o tratamento dos dados utilizou-se o Pandas, uma biblioteca para análise e manipulação de dados com recursos para sua exploração, limpeza e processamento. Cada um dos conjuntos de arquivos foi unido para que formassem três *dataframes* com os dados de todos os anos, ou seja, as proposições (*proposals*), as classificações temáticas (*proposals_themes*) e os autores (*proposals_authors*). Após a união de cada conjunto, foi realizada a seleção de atributos para cada *dataframe* mantendo apenas as colunas relevantes para o trabalho.

Não se realizaram mais tratamentos no *dataframe proposals_authors* a partir deste momento. Seus dados foram salvos em um arquivo CSV para uso futuro ("proposicoes_-autores.csv").

Os dataframes *proposals* e *proposals_themes* foram unidos em um só: *proposals*. Nele, foi realizada uma filtragem dos dados removendo as proposições cujo tipo não é classificado pelo CEDI, garantindo que os dados restantes foram submetidos ao processo de indexação de proposições legislativas.

Tendo em vista o processo de classificação multirrótulo em que os algoritmos disponíveis trabalham com números e não com textos, os valores categóricos dos temas foram transformados em *one-hot encoding*. Ao final deste processo, o *dataframe* contava com 74.109 proposições.

Como explicado na sessão anterior, o Inteiro Teor de cada proposição está disponível por meio de *links*. Os arquivos PDF foram baixados localmente por meio de um *script* e seus dos textos foram extraídos utilizando-se o PyMuPDF, uma biblioteca para extração de dados, análise, conversão e manipulação de arquivos PDF. Ao fim do processo de extração dos textos, observou-se falha em 1.356 arquivos, ou seja, 1,83% do total. Estes foram removidos do *dataframe*, resultando na quantidade definitiva de proposições da base de dados: 72.753 proposições.

Esta etapa foi realizada em duas fases: Análise dos Temas das Proposições e Análise dos Textos das Proposições. Na primeira, tentamos extrair algumas informações por meio da análise dos temas das proposições. Na segunda, utilizamos técnicas de PLN para compreender melhor o conteúdo dos textos de Inteiro Teor extraídos dos PDF's a fins de ajustar as técnicas que serão utilizadas de forma definitiva no pré-processamento dos textos.

4.5.1 Análise dos Temas das Proposições

Com a base de dados contendo as informações relevantes para o processo de classificação, utilizamos a bibliotecas Seaborn e Matplotlib para gerar os diagramas utilizados.

Inicialmente, desejou-se entender qual era a quantidade de proposições classificadas em cada um dos temas. Observou-se que os temas com mais proposições foram “Saúde”, “Direitos Humanos e Minorias” e “Administração Pública”. Já os três com menos proposições foram “Ciências Exatas e da Terra”, “Ciências Sociais e Humanas” e “Direito Constitucional”. Este dado é relevante pois demonstra o desbalanceamento dos dados, já que a quantidade de proposições com tema mais frequente é 3.288 vezes maior que as com tema menos frequente. A lista completa pode ser vista na Tabela 3.

Temas	**Quantidade de Proposições**
Saúde	13.153
Direitos Humanos e Minorias	10.869
Administração Pública	10.745
Finanças Públicas e Orçamento	9.144
Educação	7.174
Trabalho e Emprego	5.674
Defesa e Segurança	5.170
Viação, Transporte e Mobilidade	4.417
Direito Penal e Processual Penal	4.237
Comunicações	4.135
Economia	4.013
Meio Ambiente e Desenvolvimento Sustentável	3.538
Cidades e Desenvolvimento Urbano	3.290
Indústria, Comércio e Serviços	3.220
Agricultura, Pecuária, Pesca e Extrativismo	3.204
Energia, Recursos Hídricos e Minerais	3.142
Previdência e Assistência Social	2.864
Direito e Defesa do Consumidor	2.445
Homenagens e Datas Comemorativas	2.432
Direito Civil e Processual Civil	2.152
Esporte e Lazer	1.904
Arte, Cultura e Religião	1.459

Continua na próxima página

Temas	Quantidade de Proposições
Relações Internacionais e Comércio Exterior	1.453
Política, Partidos e Eleições	1.234
Ciência, Tecnologia e Inovação	1.229
Processo Legislativo e Atuação Parlamentar	1.136
Estrutura Fundiária	766
Direito e Justiça	602
Turismo	538
Direito Constitucional	237
Ciências Sociais e Humanas	6
Ciências Exatas e da Terra	4

Tabela 3 – Quantidades de proposições por tema.
Fonte: Elaborada pelo autor.

Também percebemos a importância de entender quais as maiores e menores correlações entre os temas das proposições. Estes valores foram calculados e, em seguida, foi gerado o *heatmap* exibido na Figura 12. As cinco maiores e menores correlações são detalhadas na Tabela 4.

Os resultados mostram que mesmo a correlação mais alta entre os temas (Ciência, Tecnologia e Inovação e Comunicações) pode ser considerada desprezível (MUKAKA, 2012).

4.5.2 Análise dos Textos das Proposições

O foco nesta fase foi entender o conteúdo dos textos que foram extraídos dos seus respectivos arquivos em PDF. Como se trata de um processo automático, temos pouco controle sobre como essa extração é de fato realizada, requerendo, assim, uma exploração manual.

Para facilitar o processo de análise, decidiu-se realizar o pré-processamento inicial do texto de Inteiro Teor de 30% das proposições. Utilizou-se para isso a biblioteca spaCy, desenvolvida para processamento de Línguas Naturais, com o modelo de linguagem para português, além da biblioteca NLTK, que provê um conjunto de bibliotecas de processamento de texto para classificação, tokenização, lematização, entre outras.

Os passos executados no processo foram:

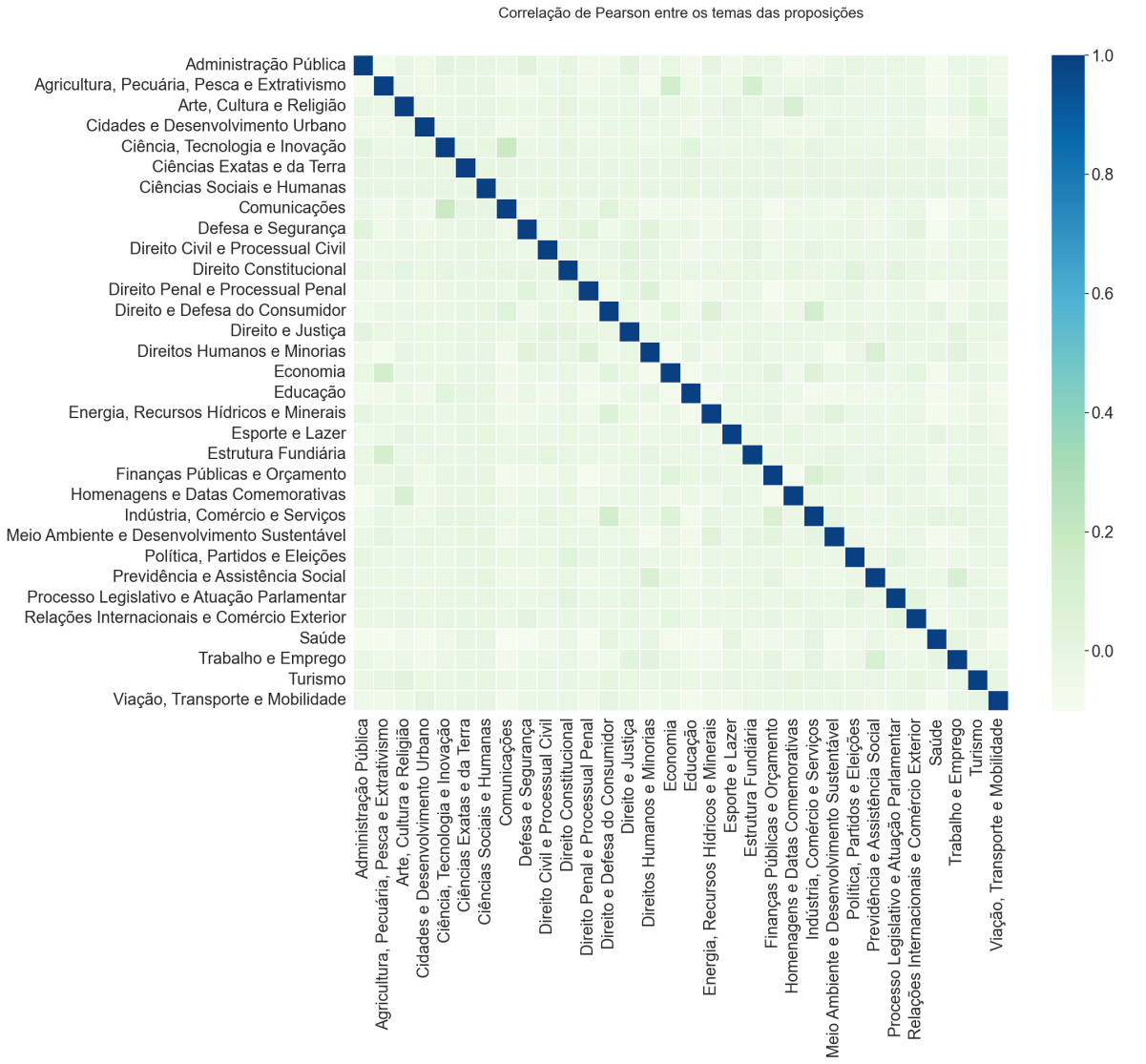

Figura 12 – Correlação de Pearson entre os temas das proposições.

Fonte: Elaborada pelo autor.

1. Transformação dos caracteres alfabéticos para minúsculos;
2. Tokenização;
3. Remoção de *tokens* indesejados: pontuações, espaços, aspas, colchetes, *emails* e URL's;
4. Remoção de *stopwords*, também chamadas de palavras irrelevantes.

Para auxiliar a análise, também foi gerada uma lista com a frequência de cada *token* gerado a partir da amostra.

Ao comparar aleatoriamente textos extraídos com seus respectivos arquivos em PDF, percebeu-se que:

Maiores Correlações		
Tema 1	**Tema 2**	**Valor**
Ciência, Tecnologia e Inovação	Comunicações	0,179270
Direito e Defesa do Consumidor	Indústria, Comércio e Serviços	0,137131
Agricultura, Pecuária, Pesca e Extrativismo	Economia	0,129193
Agricultura, Pecuária, Pesca e Extrativismo	Estrutura Fundiária	0,115031
Arte, Cultura e Religião	Homenagens e Datas Comemorativas	0,100491
Menores Correlações		
Tema 1	**Tema 2**	**Valor**
Comunicações	Saúde	-0,100666
Direito Penal e Processual Penal	Saúde	-0,088913
Defesa e Segurança	Saúde	-0,084753
Direito Penal e Processual Penal	Finanças Públicas e Orçamento	-0,084724
Saúde	Viação, Transporte e Mobilidade	-0,081295

Tabela 4 – Maiores e menores correlações entre os temas das proposições.

Fonte: Elaborada pelo autor.

- Quando o processo de extração é realizado em tabelas, podem ser gerados *tokens* que agregam pouco sentido ao texto na ordem em que são colocados;

- Os textos originais em PDF possuem muitos códigos alfanuméricos de uso interno da Câmara dos Deputados. Estes podem ser exaustivamente repetidos em cabeçalhos e/ou rodapés, o que gera muitos *tokens* que não são representativos para o conteúdo da proposição;

- Além dos códigos alfanuméricos, textos semelhantes a “Assinado eletronicamente por: <nome_próprio>” podem aparecer em várias páginas de um mesmo PDF, o que também não é representativo para o conteúdo da proposição;

- *Tokens* distintos foram criados a partir de uma mesma palavra, possivelmente devido a espaçamentos ou quebras de linha nos documentos originais. Por exemplo: “quí” e “micas”; “urbaní” e “stico”;

- *Tokens* foram formados a partir de palavras existentes com outros caracteres. Por exemplo: “federação:3”; “instituicao&q” e “enem)1”;

- Resquícios de URL's e *emails* não foram deletadas pelo spaCy, provavelmente devido a quebras de linha ou espaçamento nos documentos originais;
- Há falta de padronização em alguns *tokens* gerados, por exemplo representando valores monetários: "r$" e "50.000,00" (separados) e "r$25.600,00" (unidos).

Considerando essas observações e a viabilidade de possíveis ajustes que fossem precisos e que garantissem a o texto não fosse alterado em sua essência, decidiu-se que apenas mais dois passos deveriam ser adicionados ao pré-processamento dos textos:

5. Remoção de códigos alfanuméricos que possuam formatos dos exemplos: "705f12e311", "cd233908409700" e "d5848237-5758-5dfe-70d4-29a92e7b3e96";
6. Lematização.

4.6 *Carregamento*: Dados Salvos em *Dataset* Final

Na última fase do processo ETL, os textos de Inteiro Teor das proposições foram unidos ao *dataframe proposals*. Os dados foram e salvos em um arquivo CSV, garantindo que que estejam estejam organizados e prontos para serem utilizados em etapas posteriores da pesquisa. A descrição dos dados pode ser vista na Figura 13.

4.7 Preparação dos Dados: Pré-processamento e Vetorização dos Textos

Foi executado o pré-processamento dos textos de 100% das proposições por meio dos passos enumerados na sessão anterior. Em seguida, foi feita a vetorização dos dados utilizando-se os cálculos de *Term Frequency - Inverse Document Frequency* (TF-IDF).

Ao fim dos cálculos de TF-IDF de cada proposição, os 10 termos mais relevantes, juntamente com seu valor, foram salvos no *dataset* na coluna "PalavrasMaisRelevantes".

4.8 Modelagem

Nesta etapa foi realizado o processo de classificação automática das proposições. Definiu-se que seriam testados diversos algoritmos para tentar obter os melhores resultados levando em conta a assertividade do modelo e seu tempo de execução.

Considerando que este trabalho se propõe a resolver um problema de classificação multirrótulo, foram considerados "*Problem Transformation Methods*", ou seja, métodos que transformam um problema de classificação multirrótulo em um ou mais problemas de classificação binária; e "*Algorithm Adaptation Methods*" que adaptam algoritmos de classificação que originalmente não suportam problemas multirrótulo para lidar diretamente com múltiplos rótulos. Cada um dos métodos juntamente com os algoritmos de classificação utilizados, quando aplicados, são descritos na Tabela 5.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 72753 entries, 0 to 72752
Data columns (total 37 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Id                                            72753 non-null  int64
 1   Tipo                                          72753 non-null  object
 2   Numero                                        72753 non-null  int64
 3   Ano                                           72753 non-null  int64
 4   InteiroTeor                                   72753 non-null  object
 5   Administração Pública                         72753 non-null  int64
 6   Agricultura, Pecuária, Pesca e Extrativismo   72753 non-null  int64
 7   Arte, Cultura e Religião                      72753 non-null  int64
 8   Cidades e Desenvolvimento Urbano              72753 non-null  int64
 9   Ciência, Tecnologia e Inovação                72753 non-null  int64
 10  Ciências Exatas e da Terra                    72753 non-null  int64
 11  Ciências Sociais e Humanas                    72753 non-null  int64
 12  Comunicações                                  72753 non-null  int64
 13  Defesa e Segurança                            72753 non-null  int64
 14  Direito Civil e Processual Civil              72753 non-null  int64
 15  Direito Constitucional                        72753 non-null  int64
 16  Direito Penal e Processual Penal              72753 non-null  int64
 17  Direito e Defesa do Consumidor                72753 non-null  int64
 18  Direito e Justiça                             72753 non-null  int64
 19  Direitos Humanos e Minorias                   72753 non-null  int64
 20  Economia                                      72753 non-null  int64
 21  Educação                                      72753 non-null  int64
 22  Energia, Recursos Hídricos e Minerais         72753 non-null  int64
 23  Esporte e Lazer                               72753 non-null  int64
 24  Estrutura Fundiária                           72753 non-null  int64
 25  Finanças Públicas e Orçamento                 72753 non-null  int64
 26  Homenagens e Datas Comemorativas              72753 non-null  int64
 27  Indústria, Comércio e Serviços                72753 non-null  int64
 28  Meio Ambiente e Desenvolvimento Sustentável   72753 non-null  int64
 29  Política, Partidos e Eleições                 72753 non-null  int64
 30  Previdência e Assistência Social              72753 non-null  int64
 31  Processo Legislativo e Atuação Parlamentar    72753 non-null  int64
 32  Relações Internacionais e Comércio Exterior   72753 non-null  int64
 33  Saúde                                         72753 non-null  int64
 34  Trabalho e Emprego                            72753 non-null  int64
 35  Turismo                                       72753 non-null  int64
 36  Viação, Transporte e Mobilidade               72753 non-null  int64
dtypes: int64(35), object(2)
memory usage: 20.5+ MB
```

Figura 13 – *Dataset* após limpeza de dados.

Fonte: Elaborada pelo autor.

Tendo em mente os recursos computacionais reduzidos que foram utilizados na implementação do trabalho (detalhados no Anexo B), a avaliação dos métodos/algoritmos de classificação seguiu os seguintes passos:

- **Executar todos os classificadores:**
  - Execução das 53 combinações de classificadores indicados na Tabela 5;
  - Utilização de 30% do *dataset*;
  - Sem configuração de hiperparâmetros;

Tipo do Método	Método	Algoritmo de Classificação
Problem Transformation	*Binary Relevance*	*BernoulliNB*
		ExtraTreesClassifier
	Classifier Chain	*KNeighborsClassifier*
		LinearSVC
	Label Powerset	*LogisticRegression*
		MultinomialNB
	One Versus Rest	*PassiveAggressiveClassifier*
		Perceptron
	Multi Output	*RandomForestClassifier*
		SGDClassifier
Algorithm Adaptation	*BRkNNaClassifier*	N/A
	BRkNNbClassifier	
	MLkNN	

Tabela 5 – Métodos de classificação utilizados.

Fonte: Elaborada pelo autor.

  - Definição de 30 minutos como tempo máximo de execução para cada classificador;
  - A assertividade foi calculada utilizando o valor da micromedia de *F1-score*. Segundo Hafeez *et al.* (2023), ela é considerada uma das métricas mais relevantes e utilizadas na literatura em casos de aprendizados desbalanceados[1];
  - Seleção dos 5 melhores classificadores para utilização nos próximos passos.

- **Realizar ajuste de hiperparâmetros:**
  - Utilização do *GridSearchCV* para cálculo dos melhores parâmetros a serem usados pelos classificadores obtidos no passo anterior;
  - Utilização de 1% do *dataset*.

- **Executar os melhores classificadores:**
  - Execução dos 5 melhores classificadores com e sem ajustes de hiperparâmetros;
  - Utilização de 100% do *dataset*;
  - Comparação dos valores obtidos.

No capítulo seguinte discutiremos os resultados obtidos pela metodologia descrita.

[1] No artigo, o autor usa o termo *F-score* ao invés de *F1-score*.

5 AVALIAÇÃO EXPERIMENTAL

Este capítulo apresenta os resultados obtidos a partir dos experimentos realizados com o objetivo de classificar textos de proposições legislativas por temas. O estudo envolveu a aplicação de 53 modelos de classificação multirrótulo em um conjunto de dados composto por 72.753 textos contendo o inteiro teor das proposições e 32 rótulos no formato *one-hot-encoding*.

A avaliação dos modelos foi realizada utilizando a métrica de desempenho *F1-score* (micro). Para fins de análise, foi também calculada a *Subset Accuracy*. Após a execução de todos os classificadores utilizando 30% dos dados, os cinco melhores resultados obtidos são detalhados na Tabela 6.

Método	**Algoritmo de Classificação**	***F1-score***	**Subset Accuracy**	**Tempo de Treinamento (Segundos)**
Classifier Chain	*LinearSVC*	0,751	0,531	53,630
One Vs Rest	*LinearSVC*	0,746	0,518	3,950
Binary Relevance	*LinearSVC*	0,746	0,518	33,740
Classifier Chain	*SGDClassifier*	0,735	0,520	79,130
MultiOutput Classifier	*SGDClassifier*	0,726	0,506	2,310

Tabela 6 – Cinco melhores classificadores após execução com 30% dos dados e sem ajuste de hiperparâmetros.

Fonte: Elaborada pelo autor.

É relevante frisar que na classificação geral os métodos *One Vs Rest (LinearSVC)* e *MultiOutput Classifier (LinearSVC)* obtiveram os mesmos valores para as métricas utilizadas, a única diferença foi o tempo de execução: 3,95 segundos no primeiro e 4,02 no segundo, sendo este o critério de desempate adotado.

Em uma tentativa de melhorar os valores obtidos antes da execução da etapa de ajuste de hiperparâmetros, foram realizadas alterações na etapa de pré-processamento do texto de Inteiro Teor. Aumentou-se a lista de *stopwords*; abreviações foram trocadas pela palavra original; siglas de estados foram trocados por seu nome; numerais, datas e horas foram removidos; e o processo de lematização foi substituído pela *stemização*.

Tais alterações não surtiram efeito relevante nos valores das métricas e, dado o aumento no trabalho de implementação e aumento no tempo de execução do pré-processamento do texto, esta mudança foi considerada ineficaz. A comparação dos valores

obtidos para a métrica *F1-score* é feita na Figura 14.

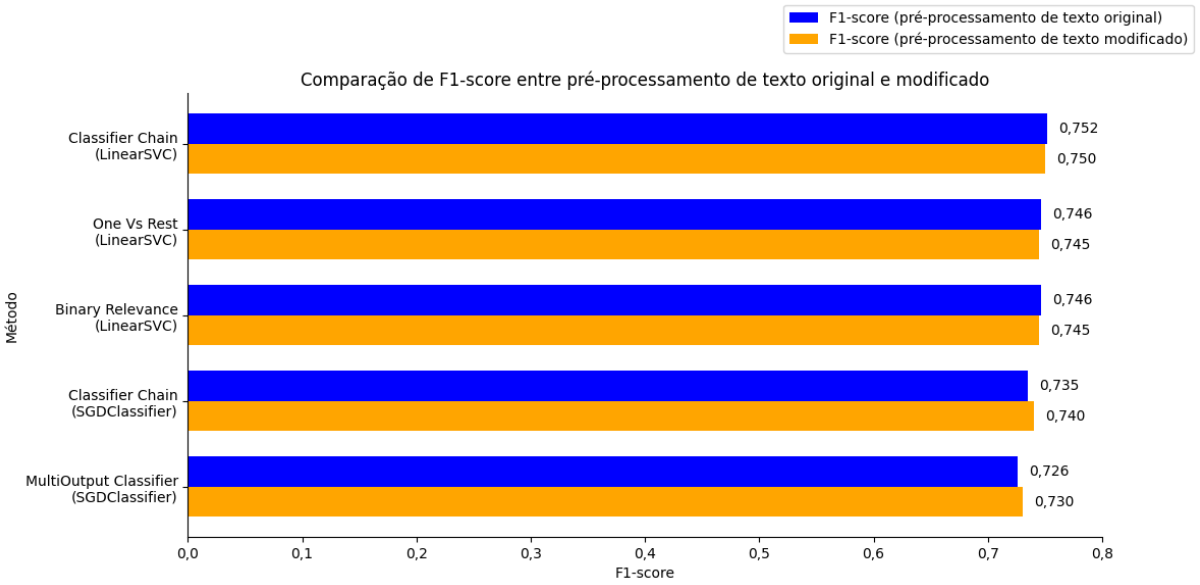

Figura 14 – *F1-score* dos cinco melhores classificadores após execução com 30% dos dados (com e sem alterações no pré-processamento do texto).

Fonte: Elaborada pelo autor.

O ajuste dos hiperparâmetros dos melhores classificadores foi realizado utilizando-se 1% dos dados devido limitações computacionais. Os resultados obtidos são apresentados a seguir, onde os parâmetros em negrito representam aqueles cujos valores definidos como melhores são os valores padrão utilizados pelos métodos do Scikit-learn:

- ***ClassifierChain (LinearSVC)***
  - Melhores Parâmetros: 'C': 10.0, **'loss': 'squared_hinge'**, **'max_iter': 1000**, 'penalty': 'l1'
  - Melhor *F1-score*: 0,571
- ***OneVsRestClassifier (LinearSVC)***
  - Melhores Parâmetros: 'C': 10.0, 'loss': 'hinge', **'max_iter': 1000**, **'penalty': 'l2'**
  - Melhor *F1-score*: 0,493
- ***BinaryRelevance (LinearSVC)***
  - Melhores Parâmetros: 'C': 10.0, **'loss': 'squared_hinge'**, **'max_iter': 1000**, 'penalty': 'l1'

  - Melhor *F1-score*: 0,553

- ***ClassifierChain (SGDClassifier)***
  - Melhores Parâmetros: ‘alpha’: 0.001, ‘loss’: ‘modified_huber’, **‘max_iter’: 1000**, ‘penalty’: ‘l1’
  - Melhor *F1-score*: 0,585

- ***MultiOutputClassifier (SGDClassifier)***
  - Melhores Parâmetros: ‘alpha’: 0.001, ‘loss’: ‘modified_huber’, **‘max_iter’: 1000**, ‘penalty’: ‘l1’
  - Melhor *F1-score*: 0,579

Após execução dos melhores classificadores utilizando-se 100% dos dados, concluímos que os classificadores performaram em média 7,36% melhor quando não houve ajuste dos seus hiperparâmetros. Este pode ser um indicador de que a execução do *GridSearchCV* com apenas 1% da base foi insuficiente para identificar os melhores parâmetros a serem utilizados. Na Figura 15 é feita a comparação das duas execuções.

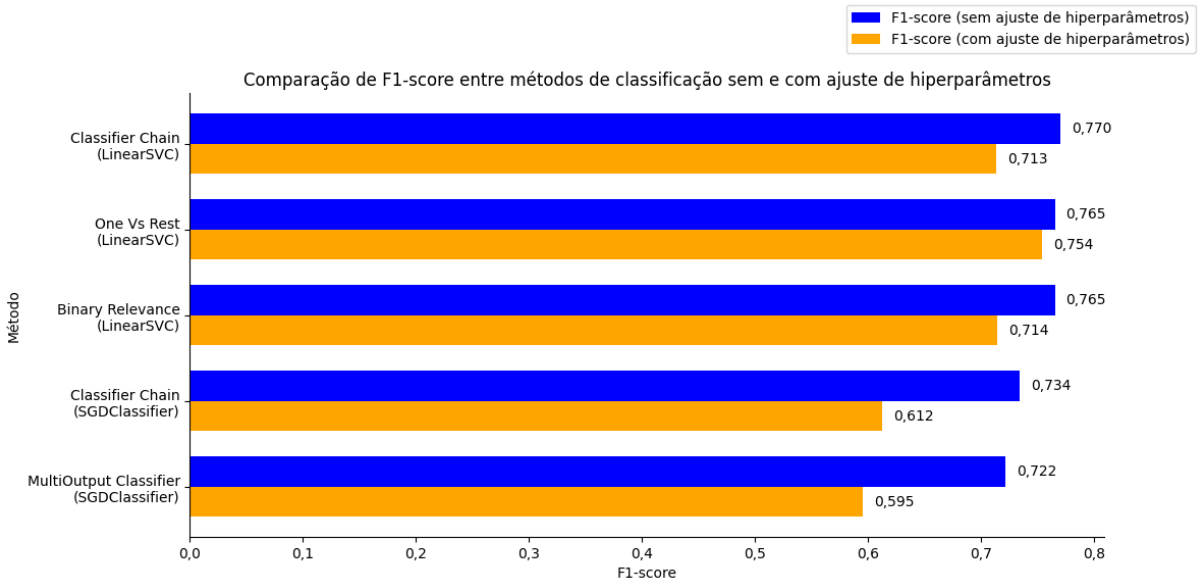

Figura 15 – *F1-score* dos cinco melhores classificadores após execução com 100% dos dados (com e sem ajuste de hiperparâmetros).

Fonte: Elaborada pelo autor.

Considerando os critérios utilizados, a melhor combinação observada foi a do método *Classifier Chain* utilizando o classificador *LinearSVC*, que obteve o *F1-score* de 77,03%. Apesar de ser um resultado que não possibilita a substituição da indexação manual,

pode-se considerar a utilização das técnicas apresentadas de forma integrada ao processo de indexação das proposições a fim de auxiliar os indexadores por meio de sugestões de temas a serem utilizados em um dado documento.

Para que haja um nível alto de confiabilidade com relação à classificação de todos os temas possíveis em uma proposição, é necessário que haja um alto valor de *Subset Accuracy*. Os resultados mostraram que os classificadores com melhores valores de *F1-score* também apresentaram os melhores valores de *Subset Accuracy*. Esta métrica também sofreu variações irrelevantes após as modificações feitas no processo de pré-processamento do texto e no ajuste de hiperparâmetros, como pode ser visto na Figura 16.

O método *Classifier Chain* utilizando o classificador *LinearSVC* também apresentou o melhor resultado de *Subset Accuracy*, com o valor de 54,97%. Tal valor não permite que esta solução seja considerada suficiente para a classificação temática precisa quando se deseja uma correspondência exata de todos os temas de uma proposição.

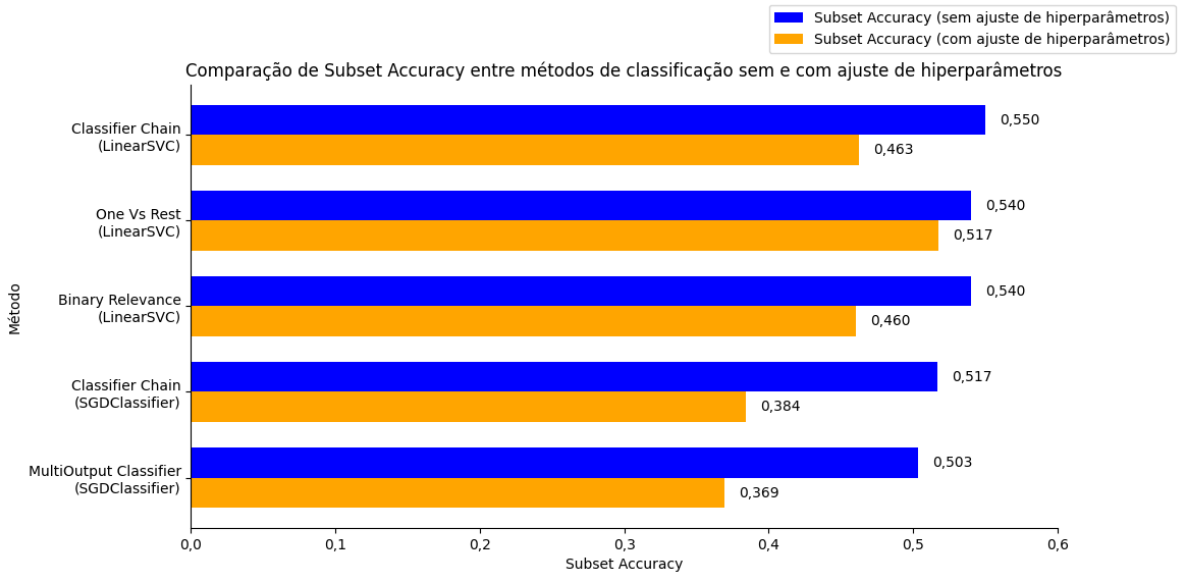

Figura 16 – *Subset Accuracy* dos cinco melhores classificadores após execução com 100% dos dados (com e sem ajuste de hiperparâmetros).

Fonte: Elaborada pelo autor.

Os resultados obtidos demonstram o potencial das abordagens propostas para a classificação multirrótulo de textos legislativos, com a aplicação de algoritmos de ML e técnicas de PLN. A utilização do TF-IDF em conjunto com os métodos de classificação multirrótulo elencados alcançaram resultados promissores de *F1-score*, refletindo a capacidade de tais métodos em lidar com a complexidade e a multiplicidade de rótulos presentes nos dados legislativos, apesar dos baixos valores de *Subset Accuracy*. Esses resultados abrem caminho para futuras pesquisas e desenvolvimento de soluções mais robustas e eficientes na classificação de textos no contexto legislativo.

A devolutiva para a SIDEX será realizada por meio da entrega da cópia digital deste trabalho e de materiais complementares, como o código-fonte da implementação, caso solicitado. Havendo interesse por parte da seção, uma reunião pode ser marcada para apresentação mais detalhada dos resultados. Além disso, será oferecido suporte contínuo para auxiliar nas discussões do progresso de uma possível implementação no SILEG e sugerir ajustes conforme o avanço das ações propostas.

6 CONCLUSÕES

A eficiência no serviço público é um objetivo contínuo e dinâmico que requer uma abordagem heterogênea. Este trabalho investigou o desafio da indexação manual de documentos, que é muitas vezes sujeita a julgamento subjetivo e pode ser inconsistente. Exploramos, em particular, a implementação de técnicas de Aprendizado de Máquina na automação da classificação de proposições legislativas, o que se mostrou uma abordagem promissora para auxiliar Analistas Legislativos da Câmara dos Deputados, permitindo que eles se concentrem em atividades de maior valor agregado. Esta abordagem não somente buscou melhorar a assertividade da classificação temática, mas também aumentar a satisfação dos funcionários, ao reduzir o número de tarefas tediosas e repetitivas.

Os resultados obtidos mostraram que o desbalanceamento da base de dados utilizada, ou seja, a diferença na quantidade proposições previamente classificadas em um ou outro tema, pode ter gerado um baixo valor de *Subset Accuracy* (54,97%), apesar do valor maior para a métrica *F1-score* (77,03%). Ambos valores foram observados para o método *Classifier Chain* utilizando o classificador *LinearSVC*, que obteve o melhor desempenho entre os classificadores testados. As limitações computacionais também se apresentaram como um desafio pelo tempo de execução, principalmente nas etapas de comparação entre classificadores e de ajustes de hiperparâmetros.

As contribuições desta pesquisa são apenas o começo de um caminho que se mostra promissor, onde a inovação tecnológica se alinha com a melhoria contínua dos processos e serviços públicos.

6.1 Desafios Encontrados

Durante o desenvolvimento deste trabalho, enfrentamos diversos desafios que impactaram a pesquisa e a implementação da solução proposta. Primeiramente, a necessidade da utilização de um *web crawler* para a realização dos downloads de arquivos PDF contendo o Inteiro Teor das proposições, além da extração dos textos destes documentos. Apesar de ser a única solução possível no momento, este processo adiciona informações que se repetem a cada página e são pouco ou nada relevantes ao texto de Inteiro Teor. Estas são provenientes principalmente de cabeçalhos, rodapés e assinaturas eletrônicas. Além disso, *tokens* podem ser erroneamente criados devido a separação de palavras por sílabas no texto. A falta de padronização em abreviações é outro fator de atenção, pois gera *tokens* diferentes que representam a mesma palavra. Apesar do resultado da extração de textos dos arquivos PDF conter tais problemas, estes causaram pouco impacto no resultado final, já que testes feitos com 30% dos dados mostraram que os classificadores tiveram resultados semelhantes comparando textos com e sem ajustes extras na fase de preprocessamento.

Além disso, a limitação de recursos computacionais foi um obstáculo importante, já que a implementação de sistemas de IA requer infraestrutura robusta. O maior, mas não único, impacto percebido foi na execução do *GridSearchCV* para ajuste dos hiperparâmetros. Não foi possível executa-lo utilizando a quantidade desejável de dados (30%). Utilizando apenas 1% dos dados, o *GridSearchCV* apresentou configurações para os classificadores que, ao serem utilizadas com 100% dos dados, geraram resultados em média 7,36% piores do que quando não se ajustou os hiperparâmetros.

6.2 Contribuições

Este trabalho contribuiu significativamente para uma melhor compreensão do processo de indexação de proposições legislativas realizadas pela SIDEX. Contactamos a seção via *email* e vários questionamentos foram respondidos, o que ajudou a complementar o Manual de Indexação de Proposição Legislativa (BRASIL, 2016) disponibilizado no site da Câmara dos Deputados. Tais questionamentos são disponibilizados no Anexo A. Os objetivos do estudo se pautaram parcialmente pelas respostas fornecidas pela seção e podem ser de grande valia para a mesma em caso de pesquisas futuras para implementações semelhantes dentro do sistema interno por ela utilizado, o SILEG.

A metodologia adotada também contribui para trabalhos futuros pois identificou alguns caminhos que podem se mostrar mais promissores para a melhoria das métricas adotadas, como:

- **Identificação de classificadores promissores:** O trabalho identificou o *LinearSVC* e o *SGDClassifier* como os melhores classificadores dentre os testados para o problema;
- **Necessidade de melhor balanceamento da base de dados:** Na fase de aquisição de dados, ao invés de se obter os dados por faixa de anos, pode-se filtrar toda a base de dados disponível para que se obtenha o máximo de proposições possíveis para os temas identificados como menos frequentes, como “Ciências Exatas e da Terra” e “Ciências Sociais e Humanas”, por exemplo;
- **Ajustes de hiperparâmetros:** Utilizando-se mais recursos computacionais, a execução do *GridSearchCV* com pelo menos 30% dos dados pode identificar melhores parâmetros para os classificadores.

6.3 Trabalhos Futuros

Os resultados obtidos neste estudo abrem várias possibilidades para trabalhos futuros. Primeiramente, é essencial continuar a investigar e desenvolver algoritmos de IA que possam ser mais facilmente integrados nos sistemas existentes, como o SILEG. A pesquisa

pode se aprofundar na criação de modelos adaptativos que aprendam continuamente com novos dados, melhorando suas previsões e eficiência ao longo do tempo.

Outra direção promissora envolve a aplicação de redes neurais, em particular as redes neurais transformadoras, como o *Bidirectional Encoder Representations from Transformers* (BERT), pode melhorar a análise e categorização das proposições legislativas. Essas redes têm a capacidade de entender o contexto de palavras em textos longos, o que pode aprimorar significativamente a precisão na classificação e indexação dos documentos. Além disso, a implementação de redes neurais convolucionais (CNN's) e recorrentes (RNN's) para a análise de padrões temporais e espaciais em dados pode proporcionar entendimentos valiosos para a tomada de decisões e otimização de processos.

Para aprimorar ainda mais a eficiência nos processos de classificação multirrótulo, sugerimos a aplicação de técnicas de *undersampling*. Essas técnicas podem ajudar a equilibrar o conjunto de dados, especialmente em cenários onde há uma grande disparidade entre as classes, o que acontece em nossa base de dados.

Explorar o potencial dessas tecnologias avançadas permitirá não apenas a automação de tarefas repetitivas, mas também a identificação de novas oportunidades de melhoria nos fluxos de trabalho, contribuindo para um serviço público mais eficiente e orientado para resultados.

REFERÊNCIAS

AALST, W. van der. **Process Mining: Data Science in Action**. 2ª. ed. New York, NY, USA: Springer Publishing Company, Incorporated, 2016. ISBN 978-3-662-49851-4.

ACKOFF, R. L. From data to wisdom. **Journal of Applied Systems Analysis**, v. 16, p. 3–9, 1989.

ALFIANI, F. S.; IMAMAH; YUHANA, U. L. Categorization of learning materials using multilabel classification. *In*: **2021 International Conference on Electrical and Information Technology (IEIT)**. [*S.l.: s.n.*], 2021. p. 167–171. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9587387>. Acesso em: 30 jun. 2024.

ANDERSON, J. D.; PéREZ-CARBALLO, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval. part i: Research, and the nature of human indexing. **Information Processing & Management**, v. 37, n. 2, p. 231–254, 2001. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457300000261>. Acesso em: 30 jun. 2024.

ANDRADE, P. H. M. A. d. **Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos**: um estudo da automatização da triagem de denúncias na CGU. 2015. 54 f. Dissertação (Dissertação (Mestrado Profissional em Computação Aplicada)) — Universidade de Brasília, Brasília, 2015. Disponível em: <http://repositorio2.unb.br/jspui/handle/10482/21004>. Acesso em: 30 jun. 2024.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). **NBR 12676**: Métodos para análise de documentos - determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro, 1992. 4 p. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/7880407/mod_resource/content/2/NormaBrasilenaIndizacionIsidoroGilLeiva.pdf>.

AZEVEDO, A.; SANTOS, M. F. Kdd, semma and crisp-dm: a parallel overview. *In*: **IADIS European Conf. Data Mining**. [*S.l.: s.n.*], 2008. p. 182–185. Disponível em: <https://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf>. Acesso em: 26 jun. 2024.

BBC NEWS BRASIL. **Documentário BBC | 8 de Janeiro: o dia que abalou o Brasil**. Youtube, 2023. Disponível em: <https://www.youtube.com/watch?v=MxciQQRUMNk>. Acesso em: 24 fev. 2024.

BELLINGER, G.; CASTRO, D.; MILLS, A. **Data, information, knowledge, and wisdom**. 2004. Disponível em: <https://www.systems-thinking.org/dikw/dikw.htm>. Acesso em: 14 abr. 2024.

BODEN, M. A. **AI: Its Nature and Future**. Oxford, UK: Oxford University Press UK, 2016. ISBN 978-0-19-877798-4.

BOOTH, P. **Indexing: The Manual of Good Practice**. De Gruyter, 2013. ISBN 9783110948592. Disponível em: <https://books.google.com.br/books?id=IgsgAAAAQBAJ>.

BRASIL. **Resolução da Câmara dos Deputados nº 20, de 30 de novembro de 1971**. Brasília, DF, 1971. 1 p. Dispõe sobre a organização administrativa da Câmara dos Deputados e determina outras providências. Disponível em: <https://www2.camara.leg.br/legin/fed/rescad/1970-1979/resolucaodacamaradosdeputados-20-30-novembro-1971-321275-norma-pl.html>. Acesso em: 18 fev. 2024.

BRASIL. **Lei nº 12.527, de 18 de novembro de 2011**. Brasília, DF, 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Disponível em: <https://www2.camara.leg.br/legin/fed/lei/2011/lei-12527-18-novembro-2011-611802-norma-pl.html>. Acesso em: 17 fev. 2024.

BRASIL. **Ato da Comissão Diretora nº 9, de 16 de maio de 2012**. Brasília, DF, 2012. Regulamenta, no âmbito do Senado Federal, a Lei nº 12.527, de 18 de novembro de 2011, que dispõe sobre o acesso aos dados, informações e documentos de interesse da sociedade e do Estado. Disponível em: <https://www12.senado.leg.br/transparencia/leg/pdf/normas/ATC92012_compilado.pdf/view>. Acesso em: 18 fev. 2024.

BRASIL. **Ato da Mesa nº 45, de 16 de julho de 2012**. Brasília, DF, 2012. 92 p. Dispõe sobre a aplicação, no âmbito da Câmara dos Deputados, da Lei de Acesso à Informação - Lei nº 12.527, de 18 de novembro de 2011, e dá outras providências. Disponível em: <https://www2.camara.leg.br/legin/int/atomes/2012/atodamesa-45-16-julho-2012-773823-norma-cd-mesa.html>. Acesso em: 18 fev. 2024.

BRASIL. **Ato da Mesa nº 125, de 19 de dezembro de 2013**. Brasília, DF, 2013. 29 p. Dispõe sobre a estrutura administrativa do Centro de Documentação e Informação da Câmara dos Deputados e dá outras providências. Disponível em: <https://www2.camara.leg.br/legin/int/atomes/2013/atodamesa-125-19-dezembro-2013-777735-norma-cd-mesa.html>. Acesso em: 18 fev. 2024.

BRASIL. **Ato da Mesa nº 80, de 31 de janeiro de 2013**. Brasília, DF, 2013. 92 p. Dispõe sobre a Política de Indexação de Conteúdos Informacionais, o Tesauro da Câmara dos Deputados e dá outras providências. Disponível em: <https://www2.camara.leg.br/legin/int/atomes/2013/atodamesa-80-31-janeiro-2013-775250-norma-cd-mesa.html>. Acesso em: 18 fev. 2024.

BRASIL. **Manual de Indexação de Proposição Legislativa**. Brasília, DF, 2016. Disponível em: <https://bd.camara.leg.br/bd/bitstream/handle/bdcamara/29179/manual_indexacao_legislativa.pdf>. Acesso em: 18 fev. 2024.

BRASIL. **Glossário de Termos Legislativos**. 2ª. ed. Brasília, DF: Grupo de Trabalho Permanente de Integração da Câmara dos Deputados com o Senado Federal, Subgrupo Glossário Legislativo, 2020. ISBN 978-65-5676-016-2.

BRASIL. **Uso de inteligência artificial pelo poder público será sujeito a regulamentação**. Senado Federal: Agência Senado, 2023. Disponível em: <https://www12.senado.leg.br/noticias/materias/2023/05/12/uso-de-inteligencia-artificial-pelo-poder-publico-sera-sujeito-a-regulamentacao>. Acesso em: 25 fev. 2024.

BRASIL. **Acesso à Informação**: Processo legislativo. Câmara dos Deputados, 2024. Disponível em: <https://www2.camara.leg.br/transparencia/acesso-a-informacao/copy_of_perguntas-frequentes/lei-de-acesso-a-informacao>. Acesso em: 15 fev. 2024.

BRASIL. **Atribuições**. Congresso Nacional, 2024. Disponível em: <https://www.congressonacional.leg.br/institucional/atribuicoes>. Acesso em: 25 fev. 2024.

BRASIL. **Entenda o Processo Legislativo**. Câmara dos Deputados, 2024. Disponível em: <https://www.camara.leg.br/entenda-o-processo-legislativo/>. Acesso em: 17 fev. 2024.

BRASIL. **Soluções de Atividades Parlamentares**: Sileg. Câmara dos Deputados, 2024. Disponível em: <https://www.camara.leg.br/internet/servicos-tic/atv-parlamentar-ficha-sileg.htm>. Acesso em: 25 fev. 2024.

BRASIL. **Visita Técnica Especial**. Congresso Nacional, ca. 2020. Disponível em: <https://www2.congressonacional.leg.br/visite/arquivos/folder-visita-informacao-e-documentacao>. Acesso em: 24 fev. 2024.

BRASIL, B. N. **3 áreas em que a inteligência artificial já está melhorando nossas vidas**. 2023. Disponível em: <https://www.bbc.com/portuguese/articles/crgl4mx5nvno>. Acesso em: 15 fev. 2024.

CABRAL, M. O. **Detecção de postagens com informações falsas sobre a pandemia do Covid-19 na rede social Instagram**. 2021. 108 p. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de São Carlos, São Carlos, 2021. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/15074>. Acesso em: 03 ago. 2024.

CASELI, H. d. M.; NUNES, M. d. G. V. (ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2ª. ed. BPLN, 2024. ISBN 978-65-00-95750-1. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/>.

CIELEN, D.; MEYSMAN, A.; ALI, M. **Introducing Data Science: Big Data, Machine Learning, and more, using Python tools**. Manning Publications Co., 2016. Disponível em: <https://livebook.manning.com/book/introducing-data-science/chapter-2/>. Acesso em: 26 jun. 2024.

COZMAN, F. G.; NERI, H. O que, afinal, é inteligência artificial? *In*: COZMAN, F. G.; PLONSKI, G. A.; NERI, H. (ed.). **Inteligência artificial: avanços e tendências**. Universidade de São Paulo: Instituto de Estudos Avançados, 2021. p. 21–29. ISBN 978-65-87773-13-1. Disponível em: <https://www.livrosabertos.abcd.usp.br/portaldelivrosUSP/catalog/book/650>. Acesso em: 19 mar. 2024.

de Keyser, P. Automatic indexing versus manual indexing. *In*: de Keyser, P. (ed.). **Indexing**. Chandos Publishing, 2012, (Chandos Information Professional Series). p. 39–63. ISBN 978-1-84334-292-2. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9781843342922500027>.

DESORDI, D.; BONA, C. D. A inteligência artificial e a eficiência na administração pública. **Revista de Direito**, v. 12, n. 02, p. 01–22, 2020. Disponível em: <https://periodicos.ufv.br/revistadir/article/view/9112>. Acesso em: 2 jun. 2024.

FACELI, K. *et al.* **Inteligência Artificial**: Uma abordagem de aprendizado de máquina. 2ª, versão kindle. ed. Rio de Janeiro, RJ: LTC, 2021. ISBN 978-85-216-3749-3.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. *In*: _____. **Advances in Knowledge Discovery and Data Mining**. Usa: American Association for Artificial Intelligence, 1996. p. 1–34. ISBN 0262560976. Disponível em: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>. Acesso em: 26 jun. 2024.

FEDERAL, G. de Trabalho Permanente de Integração da Câmara dos Deputados com o S. (ed.). **Glossário de termos orçamentários**. 1ª. ed. Brasília, DF: Grupo de Trabalho Permanente de Integração da Câmara dos Deputados com o Senado Federal, 2020. ISBN 978-65-5676-062-9. Disponível em: <https://bd.camara.leg.br/bd/bitstream/handle/bdcamara/40193/glossario_termos_orcamentarios.pdf>.

FILHO, J. R. R. C. *et al.* **Restaurando a Democracia**: A preservação da memória da câmara para futuras gerações. 1ª. ed. Brasília, DF: Edições Câmara, 2024. ISBN 978-85-402-0972-5.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, v. 35, n. 2, p. 137–144, 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0268401214001066>. Acesso em: 12 abr. 2024.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA, USA; London, UK: The MIT Press, 2016. ISBN 978-0-262-33737-3.

GOOGLE. **O que é análise preditiva?** ca. 2020. Disponível em: <https://cloud.google.com/learn/what-is-predictive-analytics?hl=pt-br>. Acesso em: 17 mar. 2024.

GUIMARÃES, J. A. C. *et al.* Ethical challenges in archival knowledge organization: the description of personal data for long-term preservation. *In*: HAYNES, D.; VERNAU, J. (ed.). **The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization**. Ergon, 2019. p. 113–126. ISBN 978-3-95650-550-8. Disponível em: <https://www.researchgate.net/publication/334833409_Ethical_Challenges_in_Archival_Knowledge_Organization_the_description_of_personal_data_for_long-term_preservation>. Acesso em: 19 mar. 2024.

HAFEEZ, A. *et al.* Addressing imbalance problem for multi label classification of scholarly articles. **IEEE Access**, v. 11, p. 74500–74516, 2023. Disponível em: <https://ieeexplore.ieee.org/abstract/document/10177756>. Acesso em: 1 ago. 2024.

HERRERA, F. *et al.* **Multilabel Classification: Problem Analysis, Metrics and Techniques**. 1ª. ed. Cham, Switzerland: Springer Publishing Company, Incorporated, 2016. ISBN 978-3-319-41111-8.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 25964-1**: Information and documentation — thesauri and interoperability with other vocabularies — part 1: Thesauri for information retrieval. Geneva, Switzerland, 2011. Disponível em: <https://www.iso.org/obp/ui/en/#iso:std:53657:en>.

JOACHIMS, T. **Learning to Classify Text Using Support Vector Machines**: Methods, theory and algorithms. New York, NY, USA: Springer Science+Business Media New York, 2002. ISBN 978-1-4615-0907-3.

JURAFSKY, D.; MARTIN, J. H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Não publicado. 2024.

KELLEHER, J. D.; TIERNEY, B. **Data Science**. Cambridge, MA, USA: MIT Press, 2018. (MIT Press Essential Knowledge Series). ISBN 978-0-262-53543-4.

LANEY, D. **3D Data Management: Controlling Data Volume, Velocity, and Variety**. 2001.

LEIVA, I. G. **La automatización de la indización: propuesta teórica-metodológica. Aplicación en el área de biblioteconomía y documentación**. 2010. Disponível em: <https://digitum.um.es/digitum/bitstream/10201/13286/1/GilLeiva.pdf>.

MARON, M. E. Automatic indexing: An experimental inquiry. **Journal of the ACM**, Association for Computing Machinery, New York, NY, USA, v. 8, n. 3, p. 404–417, jul. 1961. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/321075.321084>. Acesso em: 15 abr. 2024.

MARTíNEZ-PLUMED, F. *et al.* CRISP-DM twenty years later: From data mining processes to data science trajectories. **IEEE Transactions on Knowledge and Data Engineering**, v. 33, n. 8, p. 3048–3061, 2021. Disponível em: <https://ieeexplore.ieee.org/document/8943998>. Acesso em: 15 abr. 2024.

MORALES-HERNáNDEZ, R. C.; JAGüEY, J. G.; BECERRA-ALONSO, D. A comparison of multi-label text classification models in research articles labeled with sustainable development goals. **IEEE Access**, v. 10, p. 123534–123548, 2022. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9954368>. Acesso em: 30 jun. 2024.

MUKAKA, M. M. A guide to appropriate use of correlation coefficient in medical research. **Malawi Medical Journal**, v. 24, n. 3, p. 69–71, 2012. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/>. Acesso em: 2 jun. 2024.

NAVES, M. M. L. Estudo de fatores interferentes no processo de análise de assunto. **Perspectivas em Ciência da Informação**, v. 6, n. 2, nov. 2007. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/23378>. Acesso em: 9 mar. 2024.

NEVES, D. A. d. B.; DIAS, E. W.; PINHEIRO, M. V. Uso de estratégias metacognitivas na leitura do indexador. **Ciência da Informação**, v. 35, n. 3, p. 141–152, 2006. Disponível em: <http://www.scielo.br/pdf/ci/v35n3/v35n3a14.pdf>. Acesso em: 2 mar. 2024.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **Programa das Nações Unidas para o Desenvolvimento**. 2024. Disponível em: <https://www.undp.org/pt/brazil/objetivos-de-desenvolvimento-sustentavel>. Acesso em: 24 fev. 2024.

PINHEIRO, L. V. R. Medidas de consistência da indexação; interconsistência. **Ciência da Informação**, v. 7, n. 2, dez. 1978. Disponível em: <https://revista.ibict.br/ciinf/article/view/116>. Acesso em: 6 mar. 2024.

ROBREDO, J. Indexação automática de textos: uma abordagem otimizada e simples. **Ciência da Informação**, v. 20, n. 2, ago. 1991. Disponível em: <https://revista.ibict.br/ciinf/article/view/348>. Acesso em: 6 mar. 2024.

ROWLEY, J. The wisdom hierarchy: representations of the DIKW hierarchy. **Journal of Information Science**, v. 33, n. 2, p. 163–180, 2007. Disponível em: <https://journals.sagepub.com/doi/epdf/10.1177/0165551506070706>. Acesso em: 6 mar. 2024.

RUBI, M. P. Os princípios da política de indexação na análise de assunto para catalogação: especificidade, exaustividade, revocação e precisão na perspectiva dos catalogadores e usuários. *In*: FUJITA, M. S. L. (ed.). **A indexação de livros: a percepção de catalogadores e usuários de bibliotecas universitárias. Um estudo de observação do contexto sociocognitivo com protocolos verbais**. São Paulo: Editora UNESP; São Paulo: Cultura Acadêmica, 2009. p. 81–94. Disponível em: <https://static.scielo.org/scielobooks/wcvbc/pdf/boccato-9788579830150.pdf>. Acesso em: 19 mar. 2024.

RUBI, M. P. Diretrizes teórico-metodológicas sobre leitura documentária para indexação. *In*: FUJITA, M. S. L.; NEVES, D. A. d. B.; DAL'EVEDOVE, P. R. (ed.). **Leitura documentária: estudos avançados para a indexação**. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2017. p. 291–308. ISBN 978-85-7983-917-7. Disponível em: <https://www.marilia.unesp.br/Home/Publicacoes/leitura-documetnaria---ebook.pdf>. Acesso em: 9 mar. 2024.

SAS Institute. **Introduction to SEMMA**. [*S.l.*], 2017. 26 jun. 2024. Disponível em: <https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm>.

SIDEX. **Monografia MBA USP: Questionamentos sobre o processo de indexação de proposições**. 2024. [mensagem pessoal].Mensagem recebida por <sidex.cedi@camara.leg.br> em 29 fev. 2024.

TAREKEGN, A. N.; GIACOBINI, M.; MICHALAK, K. A review of methods for imbalanced multi-label classification. **Pattern Recognition**, v. 118, 2021. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320321001527>. Acesso em: 1 ago. 2024.

VALE, D. C. d. **Classificação Temática de Propostas Legislativas Apresentadas à Câmara dos Deputados por Meio da Aplicação de Modelos de Machine Learning**. 2022. Trabalho de Conclusão de Curso, Belo Horizonte, Minas Gerais, 2022.

VASSILIADIS, P. A survey of extract-transform-load technology. **International Journal of Data Warehousing and Mining**, v. 5, p. 1–27, jul. 2009. Disponível em: <https://www.researchgate.net/publication/220613761_A_Survey_of_Extract-Transform-Load_Technology>. Acesso em: 25 jun. 2024.

VASSILIADIS, P.; SIMITSIS, A. Extraction, transformation, and loading. *In*: _____. **Encyclopedia of Database Systems**. Boston, MA: Springer US, 2009. p. 1095–1101. Disponível em: <https://doi.org/10.1007/978-0-387-39940-9_158>. Acesso em: 25 jun. 2024.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. *In*: **Practical application of knowledge discovery and data mining, PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON THE PRACTICAL APPLICATION OF KNOWLEDGE DISCOVERY AND DATA MINING**. Practical Application Co., 2000. p. 29–40. Disponível em: <https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf>. Acesso em: 26 jun. 2024.

ANEXOS

ANEXO A – QUESTIONAMENTOS FEITOS À SIDEX POR *E-MAIL*

Quanto à documentação:

- A versão mais recente encontrada para o Manual de Indexação de Proposições Legislativas foi publicada em 2016 (ISBN 978-85-402-0514-7). Existe uma versão mais recente?

  Resposta: *A versão em uso ainda é essa.*

  – Caso exista, qual o link para o documento?

    Nota: *Não houve resposta para esta pergunta.*

  – Caso não exista, poderia confirmar se houve alguma alteração relevante no processo de indexação?

    Resposta: *Em 2018 houve um estudo sobre as áreas temáticas que passaram a ser agrupadas em 32 temas.*

Quanto aos funcionários da SIDEX:

- Quantas pessoas estão alocados na Seção de Indexação de Matérias Legislativas (SIDEX)?

  Resposta: *Quatro pessoas.*

- Quantos funcionários são servidores concursados, estagiários e com cargos especiais?

  Resposta: *Três concursados e uma estagiária.*

- Quantos funcionários são responsáveis pela indexação de proposições?

  Resposta: *Três.*

- Apenas servidores concursados realizam a indexação de proposições?

  Resposta: *Não. A estagiária em Biblioteconomia realiza, sob supervisão.*

- Qual a formação (graduação) dos funcionários que realizam a indexação de proposições? Biblioteconomia?

  Resposta: *Sim. Biblioteconomia.*

- Os funcionários responsáveis pela classificação de proposições recebem algum título específico? Por exemplo: “Agente de Indexação” ou “Analista de Indexação”.

  Resposta: *Não. São Analistas Legislativos – Documentação e Informação Legislativa.*

Quanto ao sistema SILEG:

- Existe alguma funcionalidade do Sistema de Informações Legislativas (SILEG) que se utiliza de técnicas de Inteligência Artificial?
  - Se sim, o que ela faz? Houve a criação de um protótipo para classificação das proposições em áreas temáticas.

    Resposta: *Essa é uma ferramenta que ainda está em desenvolvimento.*
  - Há a especificação de quais técnicas são aplicadas?

    Resposta: *Não há.*
- Quantas áreas temáticas estão disponíveis para seleção no SILEG? O Manual de Indexação indica 38, porém, a API de Dados Abertos da Câmara indica 32.

  Resposta: *Em 2018 foi realizado um estudo e atualmente são 32 áreas temáticas.*

Quanto ao processo de indexação:

- Há algum tipo de "revisão por pares" depois que a classificação é realizada?

  Resposta: *Não.*
- Quantas proposições, em média, um servidor indexa por semana/mês?

  Resposta: *Em média, um servidor indexa 130 (cento e trinta) projetos em um mês.*
- Caso o SILEG sugerisse as áreas temáticas a serem definidas para as proposições,
  - haveria uma redução perceptível no tempo total de indexação?

    Resposta: *Sim.*
  - ajudaria a reduzir possíveis erros cometidos pelo servidor?

    Resposta: *Sim.*
  - aumentaria a confiabilidade da indexação?

    Resposta: *Sim.*
  - auxiliaria novos servidores que ainda não estão familiarizados com todas áreas temáticas disponíveis para seleção?

    Resposta: *Sim.*
- Tendo em vista as atividades envolvidas no processo de indexação, haveria valor na existência de um *dashboard* que possibilitasse a filtragem de proposições por áreas temáticas e/ou termos do Tesauro da Câmara (Tecad)?

  Resposta: *Sim.*

ANEXO B – RECURSOS COMPUTACIONAIS E FERRAMENTAS UTILIZADAS

B.1 Recursos Computacionais

Máquina e Configuração:

Para a execução dos experimentos e desenvolvimento do projeto, foi utilizada uma máquina com a seguinte configuração:

- **Processador:** Apple M2 Pro
- **Memória RAM:** 16 GB
- **Armazenamento:** 500 GB
- **Sistema Operacional:** macOS Sonoma 14.6

B.2 Linguagem de Programação

Python (v. 3.9.6)

Python foi escolhido devido às suas capacidades robustas de manipulação de dados, desenvolvimento rápido e a grande quantidade de recursos disponíveis para tarefas de Aprendizado de Máquina.

B.3 Bibliotecas e Ferramentas

Visual Studio Code (v. 1.9.2)

IDE de código aberto que possui extensões oferecidas pela própria desenvolvedora (Microsoft) que permitem o desenvolvimento em Python.

Matplotlib (v. 3.9.0) e Seaborn (v. 0.13.2)

Utilizadas para a visualização de dados. A visualização é uma parte importante para a análise exploratória e para a apresentação dos resultados de maneira compreensível.

NLTK (v. 3.8.1) e SpaCy (v. 3.7.4)

Para Processamento de Línguas Naturais. NLTK é utilizado para tarefas básicas de PLN, enquanto SpaCy é utilizado para tarefas avançadas e de alto desempenho.

NumPy (v. 1.26.4)

Usado para suporte à arrays de grandes dimensões e matrizes, além de possuir funções matemáticas de alto desempenho.

Pandas (v. 2.2.2)

Utilizado para a manipulação e análise de dados. O pandas é essencial para a limpeza dos dados, operações de transformação e análise exploratória.

Scikit-Learn (v. 1.2.2)

Biblioteca principal utilizada para a construção e avaliação dos modelos de Aprendizado de Máquina. Inclui diversas ferramentas para pré-processamento de dados, seleção de modelos, validação cruzada e métricas de desempenho.

Scikit-Multilearn (v. 0.2.0)

É uma extensão do Scikit-Learn focada em problemas de classificação multirrótulo.

SciPy (v. 1.13.0)

Fornece uma grande coleção de algoritmos numéricos e funções para matemática, ciência e engenharia.

ANEXO C – BASES DE DADOS EXTRAÍDAS

C.1 Proposições por Ano de Apresentação

Atributos	Tipo	Descrição
id	int64	Identificador único da proposição
uri	object	URL para informações detalhadas sobre a proposição
siglaTipo	object	Sigla que representa o tipo da proposição (ex. PL, PEC, MPV)
numero	int64	Número da proposição no ano de apresentação
ano	int64	Ano em que a proposição foi apresentada
codTipo	int64	Código numérico do tipo de proposição
descricaoTipo	object	Descrição textual do tipo de proposição
ementa	object	Descrição breve do conteúdo da proposição
ementaDetalhada	object	Descrição detalhada do conteúdo da proposição
keywords	object	Palavras-chave associadas à proposição para facilitar a busca e categorização
dataApresentacao	object	Data em que a proposição foi apresentada
uriOrgaoNumerador	object	URL do órgão responsável pela numeração
uriPropAnterior	float64	URL da proposição anterior, quando aplicável (para proposições que sejam continuidade de outras)
uriPropPrincipal	object	URL proposição principal, quando aplicável (para proposições que estejam agrupadas)

Continua na próxima página

Atributos	Tipo	Descrição
uriPropPosterior	object	URL da proposição posterior, quando aplicável (para proposições que sejam continuidade de outras)
urlInteiroTeor	object	URL para o texto completo da proposição
urnFinal	float64	URL final da proposição (parece estar em desuso)
ultimoStatus_dataHora	object	Data e hora do último status registrado da proposição
ultimoStatus_sequencia	int64	Sequência do último status registrado, indicando a ordem dos eventos na tramitação
ultimoStatus_uriRelator	object	URL do relator da proposição no último status registrado
ultimoStatus_idOrgao	float64	Identificador único do órgão responsável pelo último status registrado
ultimoStatus_siglaOrgao	object	Sigla do órgão responsável pelo último status registrado (ex. CCJC)
ultimoStatus_uriOrgao	object	URL do órgão responsável pelo último status registrado
ultimoStatus_regime	object	Regime de tramitação no último status registrado (ex. Urgência, Prioridade, Ordinário)
ultimoStatus_descricaoTramitacao	object	Descrição da tramitação no último status registrado
ultimoStatus_idTipoTramitacao	int64	Identificador único do tipo de tramitação no último status registrado
ultimoStatus_descricaoSituacao	object	Descrição da situação no último status registrado
ultimoStatus_idSituacao	float64	Identificador único da situação no último status registrado

Continua na próxima página

Atributos	Tipo	Descrição
ultimoStatus_despacho	object	Despacho do Presidente da Câmara dos Deputados com a distribuição da proposição para as comissões no último status registrado
ultimoStatus_apreciacao	object	Tipo de apreciação requerida para a proposição (ex. conclusiva, terminativa)
ultimoStatus_url	object	URL para acessar informações detalhadas sobre o último status da proposição

Tabela 7 – Atributos dos arquivos contendo as proposições por ano de apresentação.
Fonte: Elaborada pelo autor.

C.2 Autores das Proposições por Ano de Apresentação

Atributos	Tipo	Descrição
idProposicao	int64	Identificador único da proposição
uriProposicao	object	URL para informações detalhadas sobre a proposição
idDeputadoAutor	float64	Identificador único do deputado autor
uriAutor	object	URL para mais informações sobre o autor da proposição
codTipoAutor	int64	Código do tipo de autor
tipoAutor	object	Tipo de autor da proposição (ex. Deputado, Comissão, Mesa Diretora)
nomeAutor	object	Nome completo do deputado ou entidade que é autor da proposição
siglaPartidoAutor	object	Sigla do partido político ao qual o deputado autor pertence no momento da proposição
uriPartidoAutor	object	URL para mais informações sobre o partido do autor
siglaUFAutor	object	Sigla da unidade federativa do autor

Continua na próxima página

Atributos	Tipo	Descrição
ordemAssinatura	int64	Número que indica a ordem em que os autores assinaram a proposição
proponente	int64	Número que indica se o autor listado é o proponente principal da proposição

Tabela 8 – Atributos dos arquivos contendo os autores das proposições por ano de apresentação.
Fonte: Elaborada pelo autor.

C.3 Classificação Temática das Proposições

Atributos	Tipo	Descrição
uriProposicao	object	URL para informações detalhadas sobre a proposição
siglaTipo	object	Abreviação que indica o tipo de proposição (ex. PL, PEC, PDC)
numero	int64	Número da proposição no ano de apresentação
ano	int64	Ano em que a proposição foi apresentada
codTema	int64	Código numérico que representa o tema ao qual a proposição pertence
tema	object	Nome do tema da proposição (ex. Educação, Saúde, Segurança
relevancia	int64	Importância ou prioridade do tema da proposição

Tabela 9 – Atributos dos arquivos contendo a classificação temática das proposições.
Fonte: Elaborada pelo autor.